

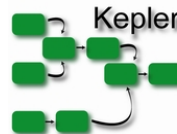
# Kepler/G-Pack: A Kepler Package Using the Google Cloud for Interactive Scientific Workflows (a.k.a. Koogole-Kuration Package)

*Gongjing Cao, Lei Dou, Quinn Hart, Bertram Ludaescher*  
*UC Davis*



**9th Biennial Ptolemy  
Miniconference**

**Berkeley, CA  
February 16, 2011**





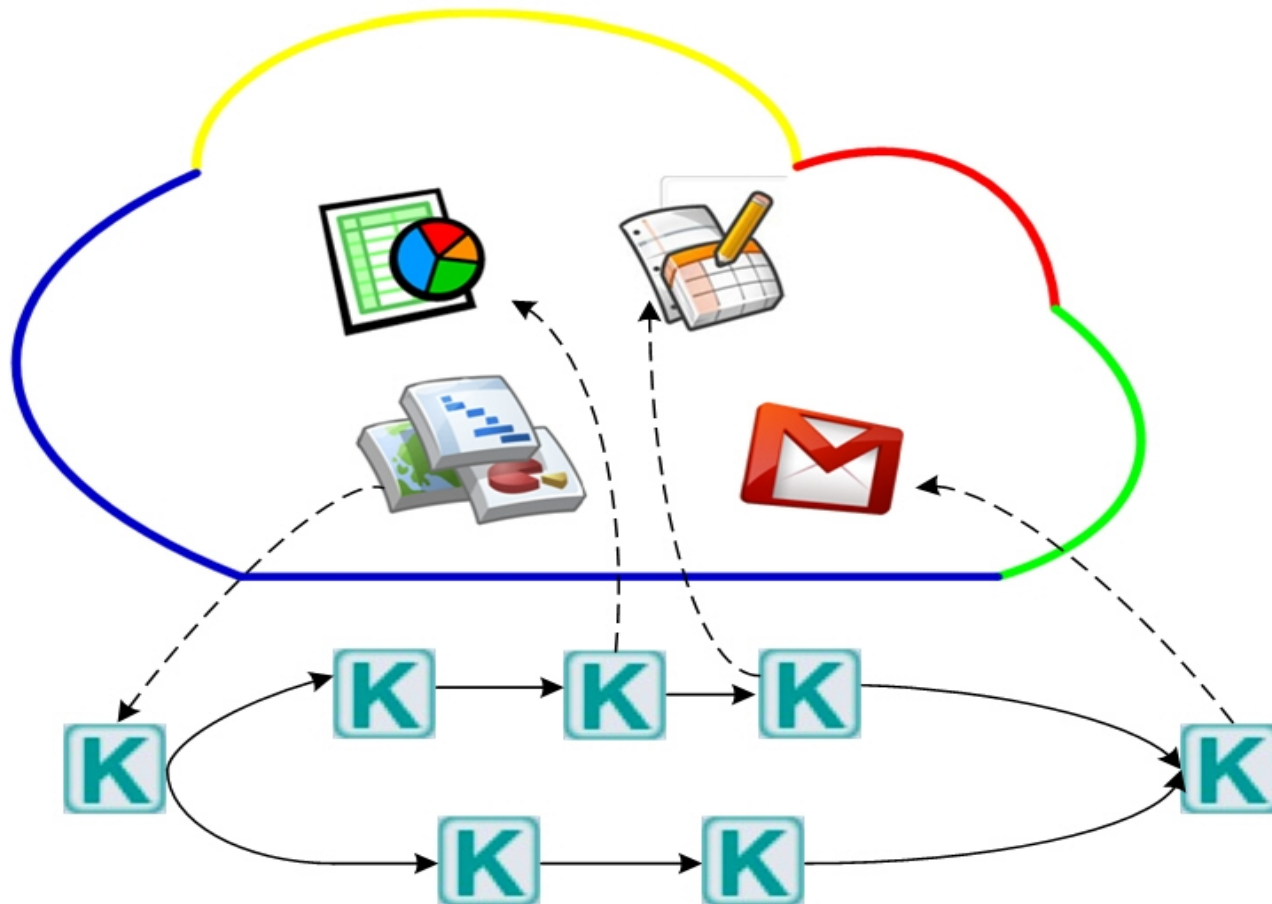
# Interactive Scientific Workflows

- Requirements for human interaction in scientific applications
  - dynamic branching based on scientists' runtime decision
  - semi-automatic data curation
- Category of human interaction

	<b>Synchronous</b>	<b>Asynchronous</b>
time cost	short time, instantly	unknown
people involved in interaction	workflow executor	people other than workflow executor
workflow blocking	yes	not necessarily
implementation approaches	graphical window web page/browser	dedicated server mail, polling, callback



# Google Cloud Computing in Kepler



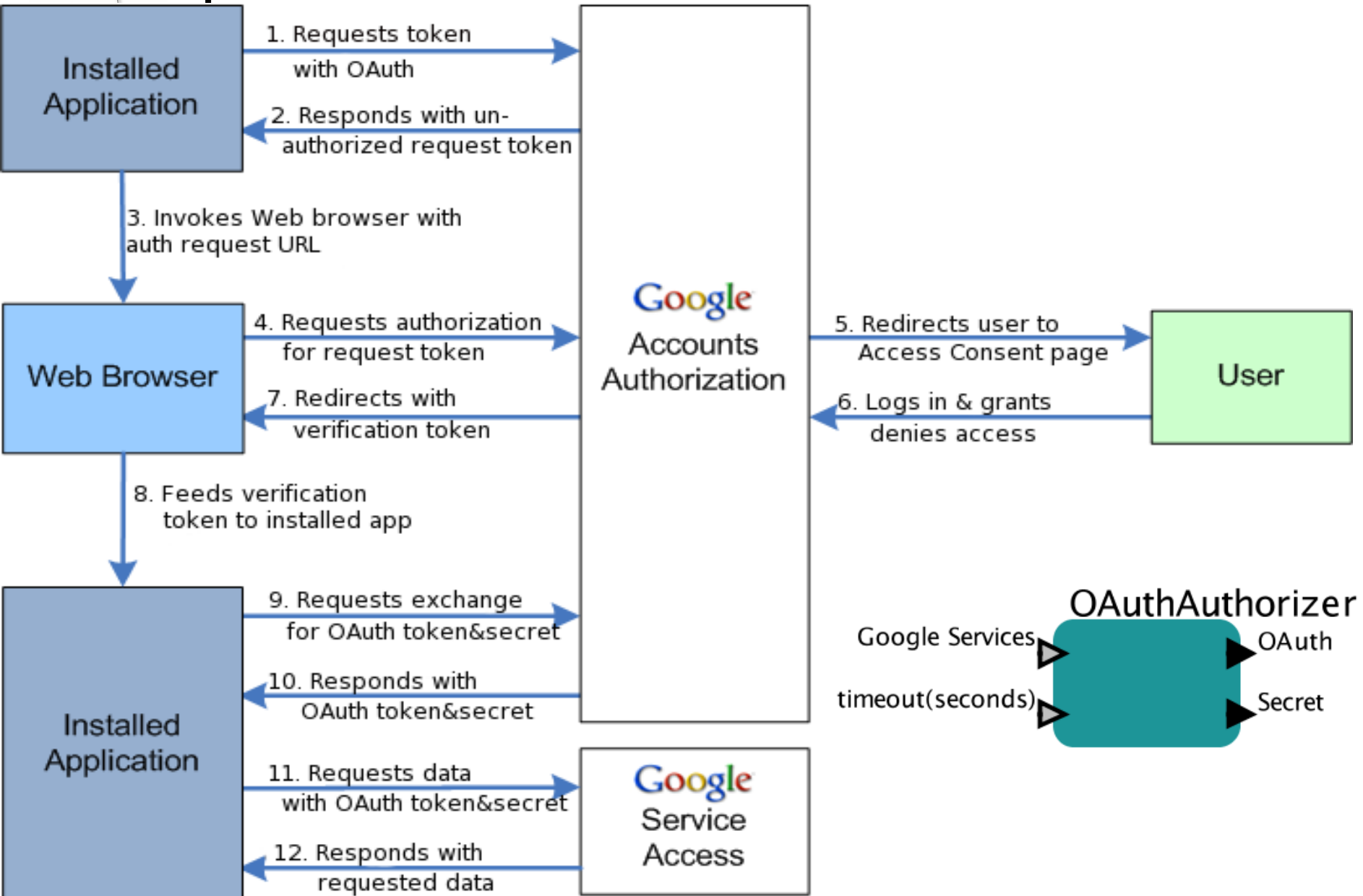


# Actors in Kepler/G-Pack

- Authorization
  - 1<sup>st</sup> step to acquire access to Google services
- Spreadsheet Operations
  - Various manipulations on Google Spreadsheet, like copy, share, import, export, query, audit.
- Data Analysis
  - Various operations especially for data curation purpose, like duplicates identification and fuse, data boundary inspection.
- Data Access
  - Google visualization datasource actor allows SQL-like access to Google cloud data
- Mail Service
  - MailSender Actor supports sending email through SMTP with UserName/Password or OAuth token/secret.

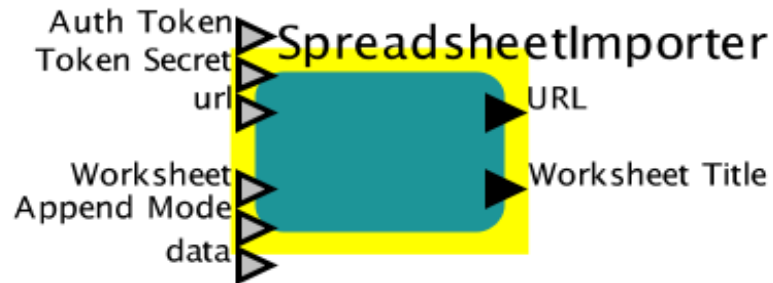


# OAuthAuthorizer Actor





# Spreadsheet Operation Actors



Actors	Functions
Importer	import data to a spreadsheet
Exporter	export data from a spreadsheet
Copy	copy a spreadsheet from a template
Share	share the spreadsheet with another user
Query	query data from the spreadsheet
Auditor PollingQuery	allow human interaction during the execution of the workflow



# Data Access VisualizationDataSource Actor

- Get the response from a servlet which is implemented with Google visualization datasource.

e.g. servlet:

<http://comet.cs.ucdavis.edu:8080/CimisVis/sql>

Table: daily

query:select \* where d\_date>date'2005-01-09' and d\_date<date'2005-01-20'

```

google.visualization.Query.setResponse({version:'0.6',status:'ok',sig:'1069066492',table:{cols:[{id:'d_max_at',label:'d_max_at',type:'number',pattern:''},{id:'d_min_at',label:'d_min_at',type:'number'},
{id:'d_dewp',label:'d_dewp',type:'number',pattern:''},{id:'d_date',label:'d_date',type:'date',
pattern:''},{id:'st_name',label:'st_name',type:'string',pattern:''},{id:'st_lat',label:'st_lat',ty
on',type:'number',pattern:''}],rows:[{c:[{v:52.0},{v:17.9},{v:8.8000002},{v:218.8},{v:10.2},{v:new
USDA},{v:36.335999},{v:285.0}],c:[{v:71.0},{v:13.3},{v:6.0999999},{v:280.39999},{v:5.9000001},{
USDA},{v:36.335999},{v:285.0}],c:[{v:54.0},{v:10.4},{v:2.3},{v:211.2},{v:4.9000001},{v:new Date
USDA},{v:36.335999},{v:285.0}],c:[{v:55.0},{v:8.3000002},{v:0.30000001},{v:126.9},{v:3.5999999}
USDA},{v:36.335999},{v:285.0}],c:[{v:20.0},{v:6.4000001},{v:4.8000002},{v:128.89999},{v:3.5},{v
USDA},{v:36.335999},{v:285.0}],c:[{v:38.0},{v:7.0},{v:4.0},{v:134.5},{v:3.5},{v:new Date(2005,0
USDA},{v:36.335999},{v:285.0}],c:[{v:21.0},{v:6.0999999},{v:4.0999999},{v:141.60001},{v:4.09999
USDA},{v:36.335999},{v:285.0}],c:[{v:25.0},{v:7.5999999},{v:4.8000002},{v:138.60001},{v:4.90000
USDA},{v:36.335999},{v:285.0}],c:[{v:29.0},{v:8.1000004},{v:6.0999999},{v:150.0},{v:5.6999998},
USDA},{v:36.335999},{v:285.0}],c:[{v:33.0},{v:7.3000002},{v:4.9000001},{v:145.10001},{v:5.30000
USDA},{v:36.335999},{v:285.0}]}}]);

```





# Data Access VisualizationDataSource Actor

- Actor parses JSON string to Kepler token
- array of RecordToken

The screenshot shows the Kepler IDE interface. On the left, a 'Search Components' panel is visible with a search bar and a list of ontologies. The main workspace displays a diagram with two actors: 'VisualizationDataSource' (highlighted in yellow) and 'Display' (a blue icon with 'T'). An arrow points from 'VisualizationDataSource' to 'Display'. Below the workspace, the 'Edit parameters for VisualizationDataSource' panel is open, showing the following configuration:

- URL: `"http://comet.cs.ucdavis.edu:8080/CimisVis/sql?table=daily&tq=select%20*%20where%20d_date%20<= '%202005-12-31' %20and%20d_date%20>= '%202005-01-01'"`
- firingsPerIteration: 1

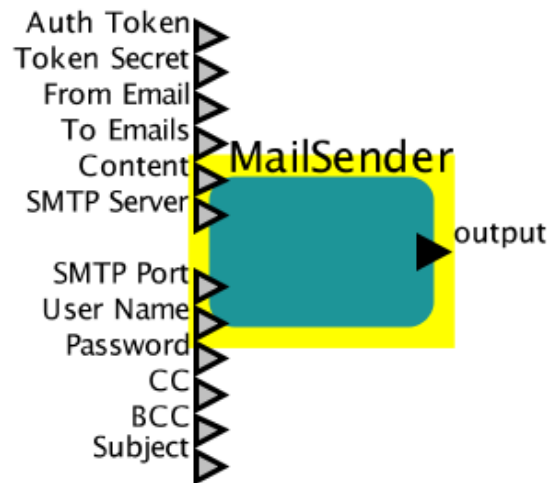
At the bottom of the panel, there are buttons for 'Help', 'Preferences', 'Restore Defaults', 'Remove', 'Add', and 'Commit'.





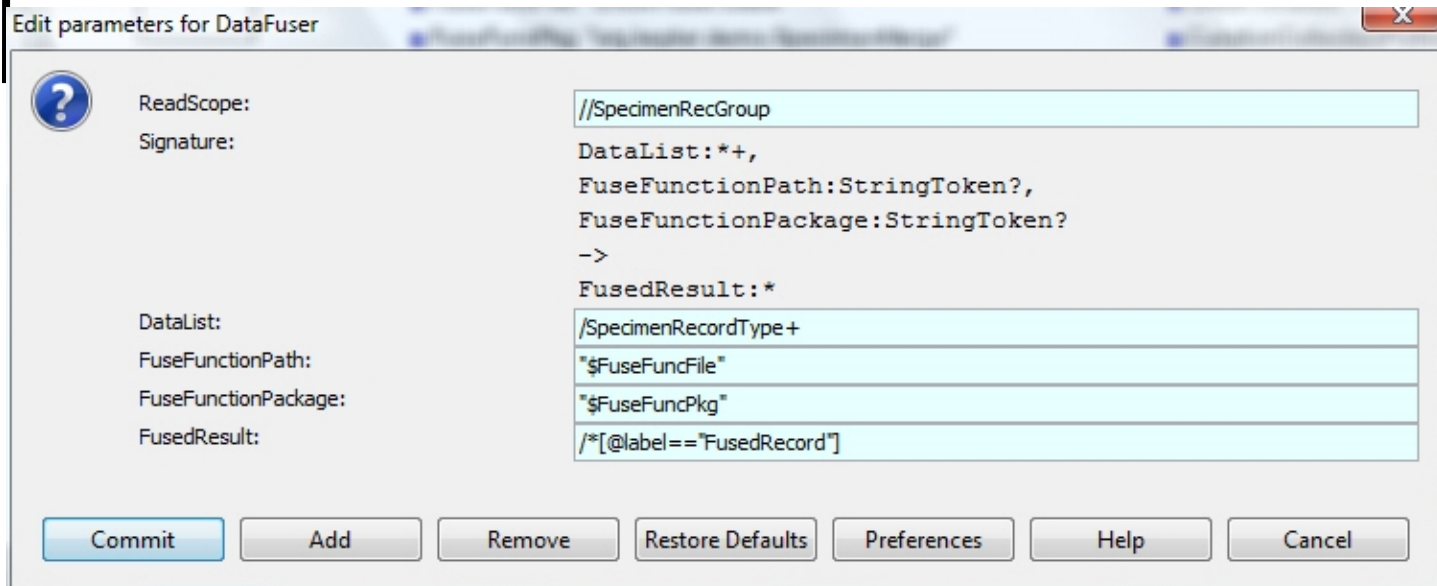
# MailSender Actor

- Gmail and other mail server supporting SMTP protocol
  - It supports sending email through SMTP with UserName/Password or OAuth token/secret.





# Data Analysis Actors (COMAD actors)



Actors	Functions
Clustering	Do clustering on a list of RecordToken by applying specified function for specified field.
DataFuser	Fuse a list of RecordToken with specified function
ConditionTester	Test whether specified condition is satisfied or not.

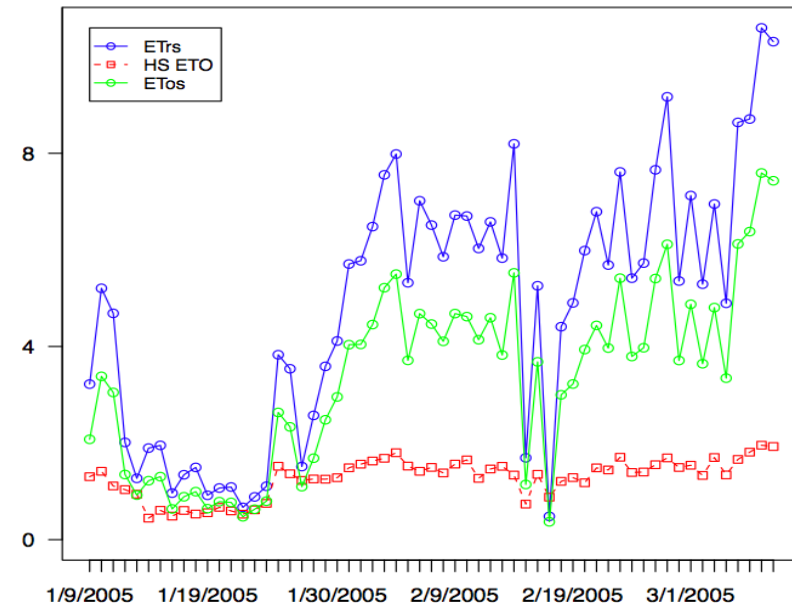
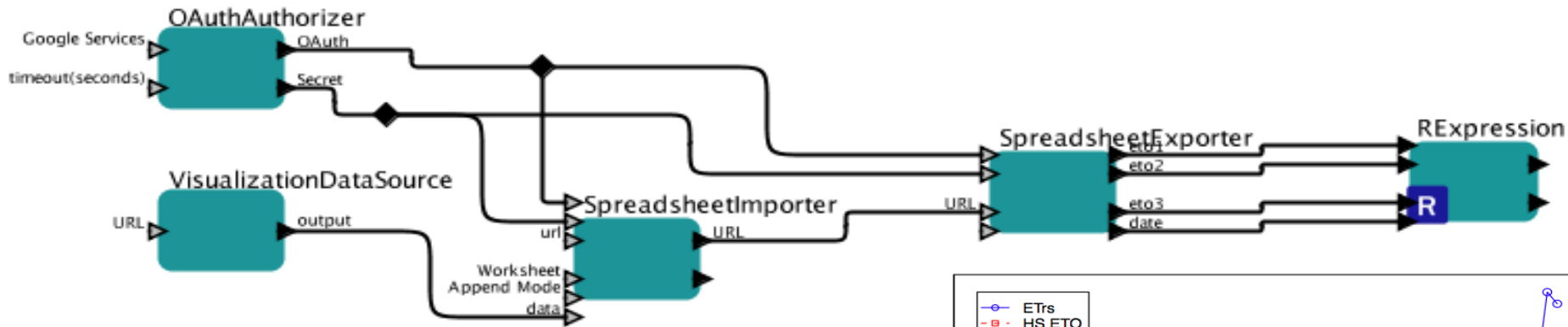


# Evapotranspiration Workflow



## Evapotranspiration Workflow - Koogle demo1

Spreadsheet is used as data storage and calculation tool during this demo.  
ETOs are calculated from CIMIS weather station data according to different models





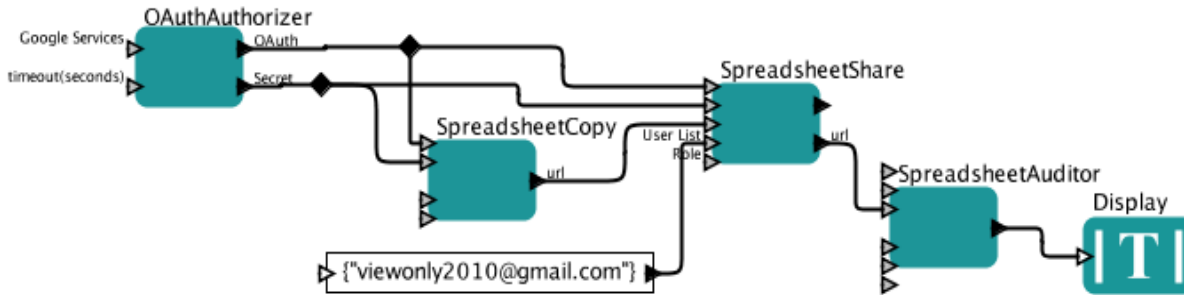
# Biofuel Refinery Workflow

SDF Director



biofuel refinery workflow

This workflow allows users to copy a spreadsheet from a templete, then use the existing data or edit the spreadsheet to help them make dicision in building a biofuel refinery.



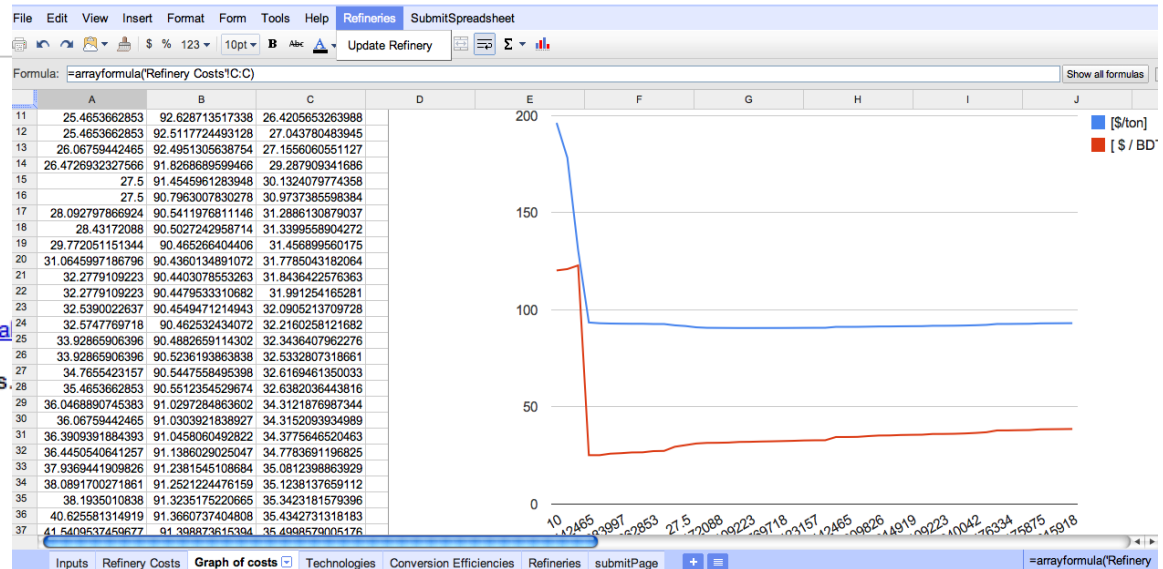
refinery-copy2 Inbox | X

★ from [kepler.ucd@gmail.com](mailto:kepler.ucd@gmail.com)  
 to [viewonly2010@gmail.com](mailto:viewonly2010@gmail.com)  
 date Fri, Feb 11, 2011 at 9:06 AM  
 subject refinery-copy2  
 mailed-by doclist.bounces.google.com

I've shared a document with you:

refinery-copy2  
<https://spreadsheets.google.com/ccc?key=0AgT1sEGRa>

It's not an attachment -- it's stored online at Google Docs





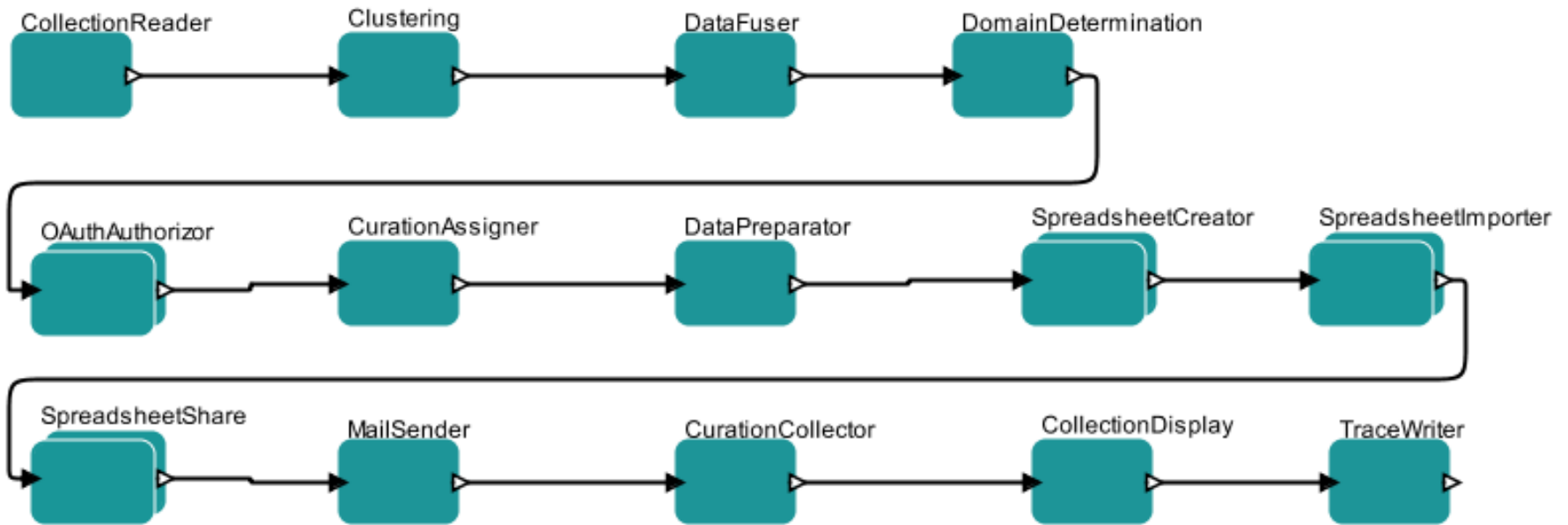
# SpecimenRecordMerge Workflow

ComadDirector



This workflow demonstrates how specimen data with duplicates are fused and curated semi-automatically.

1. Firstly the specimen records are imported from a file in csv format.
2. Secondly the duplicated sets of specimen records are identified through specific clustering function and for each duplicate set a fused record is generated.
3. Thirdly the original specimen records and the fused record are imported into a google spreadsheet for human curation. Meanwhile the corresponding curators are informed of such curation request.
4. Finally, the human curation result is collected once it's finished by the curator.





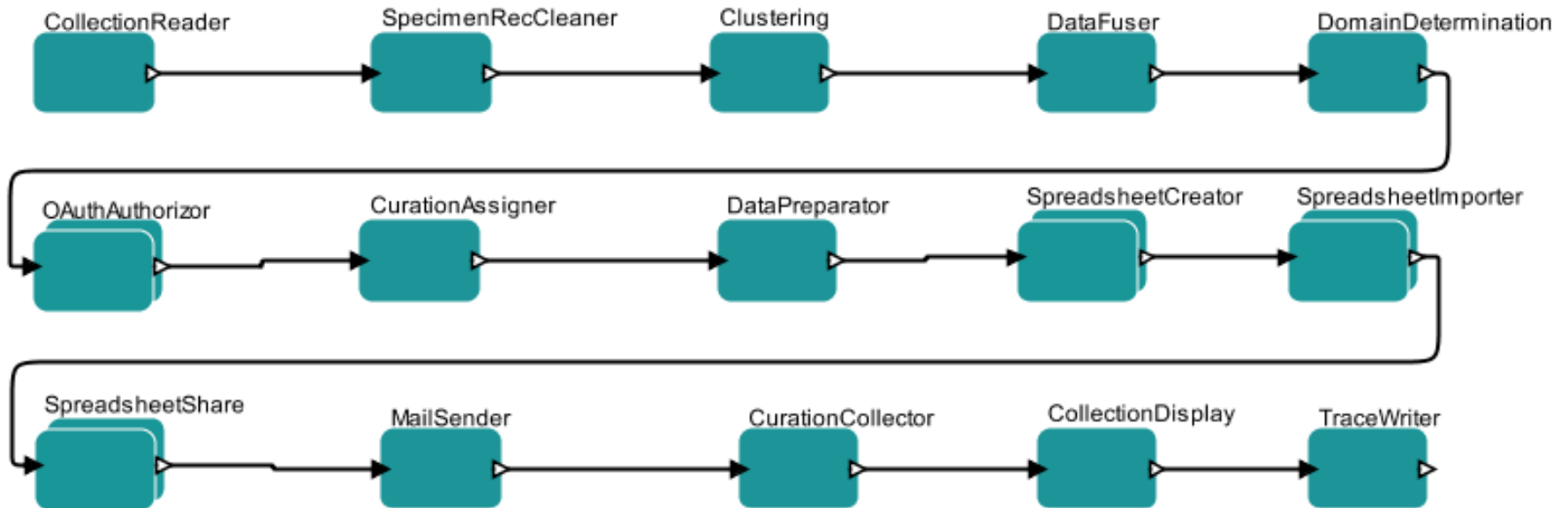
# AdvancedSpecimenRecordMerge Workflow

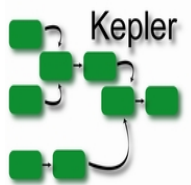
ComadDirector



This workflow makes the following improvement of SpecimenRecordMerge workflow. It's demonstrated how easy it is to reconfigure COMAD workflow to adapt to new functionalities.

1. Insert SpecimenRecCleaner actor to clean source data by removing "bad" specimen records with unclear collector of "et al."
2. Reconfigure Clustering actor to cluster specimen records against collector field with fuzzy match method. Therefore the specimen records collected at the same time and the same location but by different collector of "E. L. Morris" and "E. Morris" could be identified as duplicates.





# Thanks & Question