

# SOLE: Linking Research Papers with Science Objects

Tanu Malik, Quan Pham, Ian Foster

*Computation Institute, Dept. of Computer Science  
University of Chicago*

Microsoft eScience, 2012  
Chicago

# The boon of computational science



Computational science provides a unique window through which researchers can investigate problems that are otherwise impractical or impossible to address, ranging from scientific investigations of the biochemical processes of the human brain and the fundamental forces of physics shaping the universe, to analysis of the spread of infectious disease or airborne toxic agents in a terrorist attack, to supporting advanced industrial methods with significant economic benefits, such as rapidly designing more efficient airplane wings computationally rather than through expensive and time-consuming wind tunnel experiments.

REPORT TO THE PRESIDENT

**Computational Science:  
Ensuring  
America's Competitiveness**

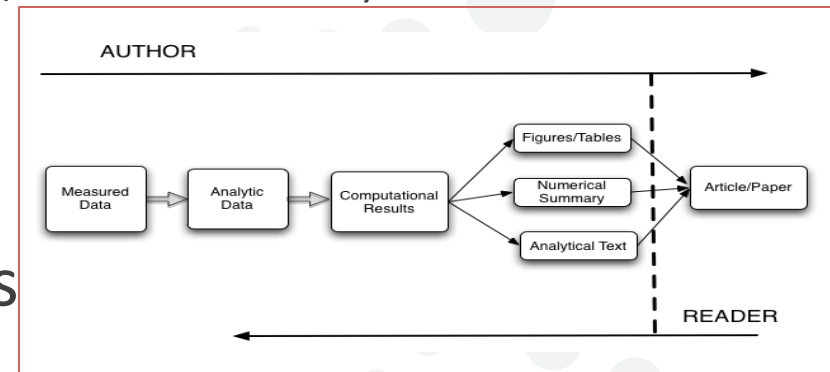
President's Information Technology  
Advisory Committee

June 2005



*Report to the President on  
Computational Science:  
Ensuring America's Competitiveness, June 2005*

- Inputs to computational science are not linked with its outputs.
  - *Inputs*: Large quantities of data, complex data manipulation and/or numerical simulation use of large and often distributed software stacks, etc. (software, data, computation)
  - *Outputs*: Research papers (text-based, non-interactive)
- Authors and Readers approach computational science from opposite directions
- Post hoc association of computational science inputs to outputs.



# Example 1: Computational Paper



- The paper is about weather forecast model validation:
  - 2 input datasets
  - Different model setup
  - Post-processed output as overlay images
- Science Object deployment plan:
  - Datasets are from GFS
  - A dataset could be related to its initialization date
  - Table 1 describes the experiments
  - The model has to run on a HPC resource
  - Figure 1 describes the results

**Validation of the Uniparthenope ARW-WRF model for weather forecasting over the area around the city of Naples on 11<sup>th</sup> June 2011 has shown an extensive cloud cover**

Validation of the Uniparthenope ARW-WRF model for weather forecasting over the area around the city of Naples on 11<sup>th</sup> June 2011 has shown an extensive cloud cover

always ranging from 50 to 70%. However these values can be indicative of a good accuracy for the selected fields.

large differences arose from the analysis of the cloud coverage, whose average value from Z06 to Z18 is around 90%. In the first two configurations (whatever the date) the simulations have average cloud coverage not larger than 25%, whereas the third microphysics allowed to reach a 70% cloud coverage mean value, much more similar to the observed one. In addition the percentage of hits is between 65 and 70% with this latter and lower than 35% in the other cases.

It is also worthy to note that the two model in Configuration 3 do not produce similar results. In particular they show different hourly trends, with that initialized on 10<sup>th</sup> June seeming to display a delay of 3 hours respect to the observations, even if with a cloud coverage value always 20% lower than the observed one. The same cannot be said for the other version of the model, that only displays a slight similarity if the delay reaches 6 hours.

By looking at Figure 1, it is clear that, taking into account the 3 hours delay described above, the model initialized on 10<sup>th</sup> June is able to reproduce with good accuracy the cloud coverage over a large fraction of the analyzed area, with major differences between model and observations only present over the sea.

	Configuration 1	Configuration 2	Configuration 3
Microphysics	New Thomson	ETA	Millbrandt-Yau
Longwave radiation		RRTM	
Shortwave radiation		Dudhia	
Surface layer		MM5 similarity	
Land-surface		5 layer	
PBL		YSU	
Cumulus		Kain-Fritsch	

**Table 1:** The different parameterizations of the models here tested.

The model has been set-up with three two-way nested domains: the largest (named d01) comprising all the Europe region at a horizontal resolution of 27 km, the second (d02) centered on Italy with a resolution of 9 km and the last (d03), over Southern Italy, with a resolution of 3 km. Has shown in Table 1, for this study only the microphysics scheme has been modified. Furthermore also the initialization date was changed, being set on 10<sup>th</sup> June Z00 and 11<sup>th</sup> June Z00. This resulted in six different model configurations, that have been tested by means of two different techniques. The first one is the point-stat analysis provided by the MET (Model Evaluation Tool) with observational data selected by the CISEL Research Data Archive ds337.0. In particular temperature at 2 meters and relative humidity at different atmospheric layers (i.e., 2 m, 850 and 500 hPa) have been used for comparison. These data have been compared with the simulations of the d02 domain at Z06, Z18 and Z18 of 11<sup>th</sup> June 2011.

The other analysis has been focused on the cloud coverage, using Eumetsat data. In particular, the cloud mass fraction has been compared with a similar variable computed from the model results in the domain d03.

**4. Conclusions and future works**

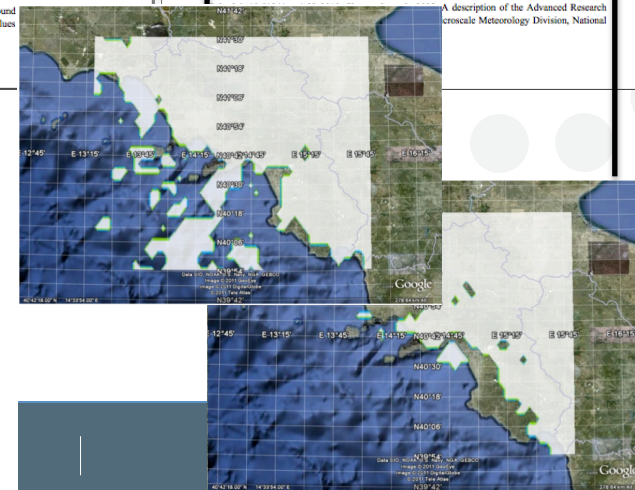
The work here proposed seems to demonstrate that, in the case of severe convective phenomena, the Millbrandt-Yau microphysics initialized the day before the event, is strongly suggested for the ARW WRF model over Southern Italy.

In the future further improvements and tests of the model have to be performed. In particular a further refinement of the d03 domain that, with a 1 km resolution, should be able to run without cumulus parameterization.

**References:** Lara-Fanego et al., 2011, Evaluation of the WRF model solar irradiance forecasts in Andalusia (Southern Spain), Solar Energy, doi: 10.1016/j.solener.2011.02.014; Moscatello et al., 2008, Numerical Analysis of a Mediterranean "Hurricane" over Southeastern Italy, Monthly Weather Review, doi: 10.1175/2008MWR2512.1; Miglietta et al., 2010, WRF model and ASAR-derived 10 m wind field comparison in a case study over Eastern Mediterranean Sea, Adv. Space Res., doi: 10.1016/j.asr.2010.03.014; A description of the Advanced Research crosscale Meteorology Division, National



Figure 1: Comparison between observations (left, Z12) and Model C initialized on 10<sup>th</sup> June 2011 (right, Z15), taking into account the 3 hours delay described in the text.



**Table 1:** The different parameterizations of the models here tested.

# Example 2: Policy Paper



special conditions or circumstances that differentiate them from more recent statutes. Iowa, the earliest RPS (1983), had requirements so small ( $\sim 1\%$  of state electricity sales; [4, 12]) that they could not be expected to affect the renewables industry significantly. Maine (2000) allowed existing facilities to contribute to the renewable portfolio and the RPS was initially met by existing hydropower rather than new construction [6]. Texas (1999) possesses such anomalously strong wind resources that development of windpower in the state could be driven largely by the federal Production Tax Credit and Investment Tax Credits with the state RPS playing a much less significant role. (The Production Tax Credit and Investment Tax Credits, henceforth “PTC”, reimburse qualifying renewable generators for up to 30% of the installed cost. See Appendix A.1 for further discussion.) Current Texas REC prices remain so low ( $\sim \$1$  per MWh; [13]) that the state RPS is not a significant subsidy for windpower in Texas, and current construction implies that Texas wind capacity will reach its 10 GW target almost fifteen years ahead of RPS-mandated requirements [14]. Prediction of the expected evolution of renewables implementation under the 2006-2011 statutes therefore requires new analysis.

State	Existing	Under Construction	Rank (Existing)
Texas	9,727	350	1
Iowa	3,670	0	2
California	2,739	443	3
Oregon	2,095	201	4
Washington	1,964	735	5
Illinois	1,848	587	6
Minnesota	1,818	677	7
New York	1,274	95	8
Colorado	1,248	552	9
Indiana	1,238	99	10

# Example 3: Data Mining Paper



- **Dataset description:** The NCI-60 data set contains anticancer screening results for more than 40,000 compounds. **It is publicly available** in the PubChem BioAssay database(38) as 73 bioassays with the name of NCI human tumor cell line growth inhibition assay under the DTP/NCI data source. In this work, only the **top 60 bioassays** (referred hereafter as NCI-60) with the largest number of tested compounds were selected (Supporting Information, Table S1). Relevant bioactivity data were downloaded at the PubChem FTP site (<ftp://ftp.ncbi.nlm.nih.gov/pubchem/Bioassay>, accessed on December 9, 2010). A total of 5083 compounds were found commonly tested in all of the 60 bioassays. Additional data set characteristics are summarized in Supporting Information.
- **Dataset description:** The Burnham Center for Chemical Genomics (BCCG) has launched a screening campaign for aqueous solubility against the NIH Molecular Libraries Small Molecule Repository (MLSMR), which contains more than 350000 compounds. The resultant bioassay (PubChem AID: 1996) was deposited publicly in the PubChem BioAssay database.(31) **As of June 18, 2010, this bioassay stored experimental solubility data for 47567 compounds.** The solubility data can be downloaded from the PubChem FTP site (<ftp://ftp.ncbi.nlm.nih.gov/pubchem/Bioassay/>). All compounds were measured using a standard protocol under the same conditions. (32) We consider that data set compiled from a single source, e.g., those used in this work, is more advantageous for statistical studies than those compiled from various sources (Supporting Information, Table S1).

- Online digital repositories
  - PubMed has over 21.78 million abstract records
  - Open Centr
- Author
- Author
  - Share
  - Make software packages available
  - Create and share virtual images
- *Loosely-connected: Disconnected with the claims and findings in the paper*





- Tightly-integrated
  - Link a concept in paper to its implementation in source code;
  - Link a dataset description to its metadata and digital object identifiers (DOIs);
  - Link a figure in the paper to its derivation and workflow, and
  - Link data values referenced from another paper sources to the exact location in that other paper's PDF source.





Idle

← Previous 1 2 3 4 5 6 Next → Go to page

## Input, Model and Equations

First, we'll consider a single-level logistic regression model, accounting for the age of the women only. Binomial response models need a denominator that contains the counts of the number of trials each binomial is based on. For 0/1 data this will be a vector of ones, whereas for proportion data this will be a number of units on which each proportion is based.

There are inputs to be specified for Stat-JR before the model can be run. These have been set in this eBook.

We left the input of explanatory variables blank. The following screenshot shows what you see below:

explanatory variables:

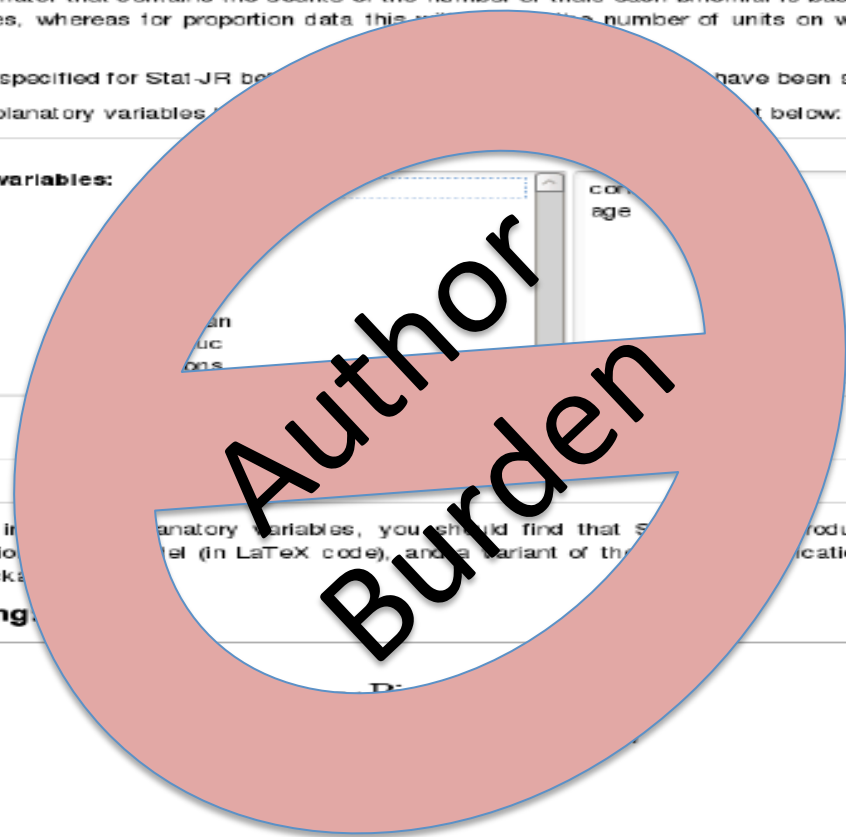
con  
age

about

After you submit the input, you should find that Stat-JR has produced a nicely-formatted mathematical description of the model (in LaTeX code), and a variant of the model in the TeX markup language associated with the WinBUGS package.

### Equation rendering.

about



Author  
Burden



- Transform
  - Each science object into a form amenable to linkage with a paper
  - Associate classes and functions in source code with URLs,
  - Record datasets in registries with locations and access methods
  - Cast data analysis pipelines as workflows with appropriate wrappers and web services that specify inputs and functional forms,
  - Associate with software on a adequately provisioned virtual image.
- Manage
  - The manner in which linkages are represented in papers.
  - Unwieldy URL usage, especially when an object is referenced multiple times.
- Present
  - Clicking on a science object link should lead to adequate presentation to the user.

# Relieving Author Burden and Improving Readability



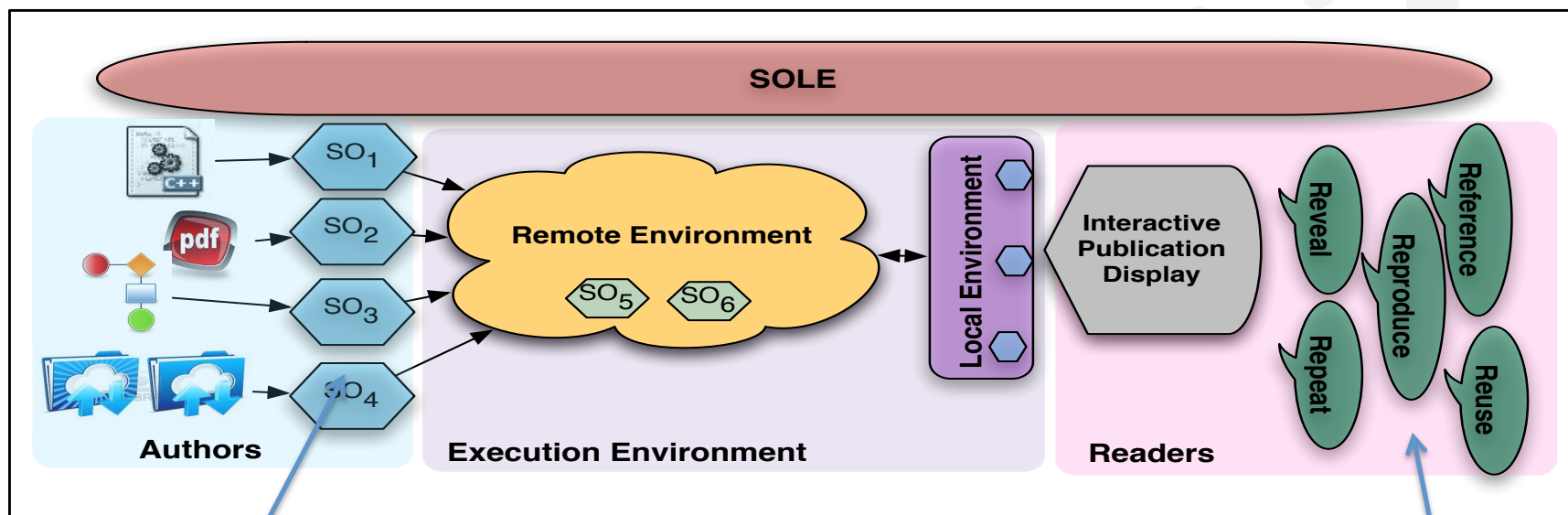
- **Automation:**
  - Simplify associations of data, source code, executions, complex flows, and provenance to the research paper.
  - An adequate representation of the associated
  - Curate the associations in a bibliography-like specification,
  - Provide multiple views of datasets and analyses.
  - In effect provide *publication-as-a-service*.
- **Interactive Performance:**
  - Must work just as effectively with remote operations, scheduling of resources and provide real-time access to distant instruments and data resources.
- **Usability:**
  - A usable interface to explore and analyze data in publications.
  - Maintain the document metaphor that has governed readership for centuries.
- **Sharing model:**
  - A supporting framework that provides effective means for data, model sharing and collaboration for diverse and distant groups of researchers and students to coordinate their work.
- **Attribution Model:**
  - The supporting framework must include usage statistics



- <http://www.ci.uchicago.edu/SOLE>
- Introduces the *publication-as-a-service*, by providing tools to authors to associate *science objects* with publications.
- Publication infrastructure for hosting interactive scientific papers
- Improves:
  - Transparency
  - Reproducibility
  - Repeatability



# SOLE Framework



Science Objects

Interactions

Purpose	Types of Interactions
Reuse	Reuse and share data, method and processes or any constituent part of it.
Repeat	Execute the processes in the same execution order as the original publications.
Reproduce	Repeat but with a different data, method, hardware, etc.
Reference	Be able to reference data, methods, and processes at various granularities.
Reveal	Be able to audit, review, and validate results.

Thanks to David DeRoure for the taxonomy.

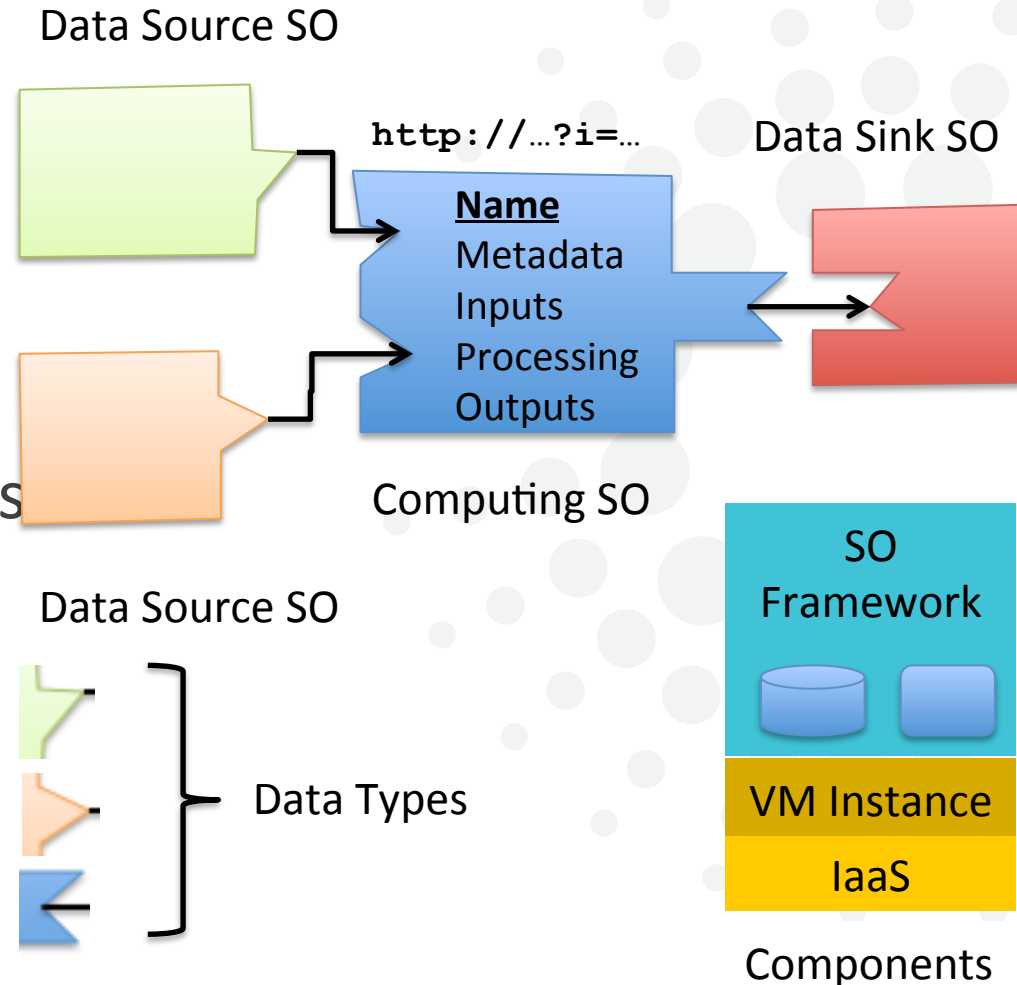
In: Bechhofer, S., De Roure, D., Gamble, M., Goble, C., Buchan, I., *Research objects: Towards exchange and reuse of digital knowledge.*

The Future of the Web for Collaborative Science, 2010.

# Science Objects



- Science Objects are created by authors
  - Simple tags in source code
  - SDKs
  - Interactive web portals



- Open tag  
format: `####t@`

```
####t@init|setup|import|working directory
options( prompt= " ", continue= " ", width= 60)
options(error= function(){
  ## recover()
  options( prompt= "> ", continue= "+ ", width= 80)
})
```
- Close tag  
format:  
`####t@`

```
source( "~/thesis/code/peel.R")
source( "~/thesis/code/maps.Rf")
####t@dataset directory
  texWd <- path.expand( "~/thesis/analysis")
  rasterWd <- path.expand( "~/thesis/data/analysis")
  dataPath <- path.expand( "~/thesis/data")
  setwd( rasterWd)
####t@

  overwriteRasters <- TRUE
  overwriteFigures <- TRUE
```
- Tag naming:  
tag1|tag2|...|  
tagn  
(fluidinfo like)

```

# studyArea used to work out RMSE
# calcs and tables
##studyArea <- "thumb"
peelBands <- peelClasses +1
####t@
```
- Element  
attributes:  
Key/Value



# Source Code Association



```
### to aggregate
aggregateFractions <- function
(mlot, aggRes= 5/60, overwrite=
FALSE, ...) {
  aggBrickFile <-
    if( mlot$Amin < 1)
      paste( deparse( substitute
(mlot)),
            "Amin", mlot$Amin)
    else
      paste( deparse( substitute
(mlot)))
  mlot$agg <-
    if( overwrite)
      aggregate( mlot$fraas)
    else
      brick( list.files(getwd())
            layerNames( mlot$agg) <-
            layerNames( mlot$fraas)
            mlot
          )
}
### to
```

```
>> ./sole.sh ./test.R
```

2. Run SOLE;  
Create SO

The screenshot shows the SOLE web interface. On the left, a sidebar lists categories: Science Objects, Language Objects (Reclassify, Reproject, Aggregate), DataSets (MLCT, NLCD), Workflows (Figure1, Figure2), and Virtual Images (ami-102991, ami-102992). The main content area displays '3 Our Algorithm' with a detailed description of the MLCT and NLCD data processing pipeline. A map of the United States is shown at the bottom of the algorithm section. On the right, there are search filters and a preview of the R code associated with the selected object.

Step 3: Search SO with tag name and link

3. Search SO with tag name and link

1. Authors identify science objects with human-readable tags

# Annotation Association



energy and low intensity of economic activity.

## #t@tbl reclassification of MLCT

	MLCT/IGBP	PEEL <sub>0</sub>
0	water	water
1	evergreen needleleaf forest	
2	deciduous needleleaf forest	
3	evergreen broadleaf forest	forest
4	deciduous broadleaf forest	
5	mixed forests	
6	closed shrublands	
7	open shrublands	shrub
8	woody savannas	
9	savannas	open
10	grasslands	
11	permanent wetlands	wetland
12	croplands	crop
13	urban	urban
14	cropland / natural vegetation mosaics	mosaic
15	permanent snow and ice	
16	barren or sparsely vegetated	barren



- Workflow input

```
###i:tiff@reclassification
primary_namefile <- "thumb_2001_lct1.tif"
secondary_namefile <- "thumb_2001_lct1_sec.tif"
pct_namefile <- "thumb_2001_lct1_pct.tif"
###i@
```

- Software unit tag

```
###sw@reclassification
thumb <- mlctList( primary_namefile,
                  secondary_namefile,
                  pct_namefile)
```

- Inputs and Outputs

```
###sw@
###i:tiff@calculate
###o:tiff@reclassification
primary_reclassified <- filename(thumb$pri)
secondary_reclassified <- filename(thumb$sec)
###o@
###i@

###sw@calculate
thumb <- mlctList( primary_reclassified,secondary_reclassified)
###sw@
```

```
###i:float@calculate
```

```
Amin <- 0.5
```



## SOLE

Create Load Tag Account

- Science Objects
  - + Source Code
  - + Annotations
- Web Service/Workflow
  - Algorithm Workflow
  - Mlct Data
  - calculate
  - reclassification
  - Virtual Images

Master of Arts M.A. Geography & Environmental Studies

August 2011

**Synthesis of a complete land use  
data set for the conterminous U.S.  
emphasizing accuracy in a  
distribution of agricultural**

Neil A. Best

- The defined SOs for the paper

# Attach the SO to a word in the paper



Create Load Tag Account



+ Science Objects

## Synthesis of a complete land use/land cover data set for the conterminous United States emphasizing accuracy in area and distribution of agricultural activity

Neil A. Best

### Abstract:

This paper presents an effort to produce a new land cover data set for the conterminous United States of America (cUSA) that augments available agricultural land use data with other uses and natural covers to create a complete landscape characterization. Using the Agland2000 data set as a benchmark we formulate a hybridization of the MODIS Land Cover Type (MLCT) for 2001 and the 2001 National Land Cover Database (NLCD) that is particularly tailored to serve as an initialization data set for long-term economic land use change models. In order to strike a balance between spatial precision and local diversity of uses and cover types, we detect the

Text   
Phrase:   
Path:   
Tags:

Select one record:

[tbl reclassification of MLCT](#) (pdf annotation)

[Mlct Data \(Linkage SO\)](#)

[prepare MLCT](#) (pdf annotation)

[tbl reclassification of MLCT](#) (pdf annotation)

[prepare MLCT](#) (pdf annotation)

[mlctReclass](#) (code)




[mlctList](#) (code)

- Attach the SO to a word in the paper



← → ↻ 🏠 ec2-50-17-179-54.compute-1.amazonaws.com/~quant/geo/ ☆ ☰

📁 DataBlogs 📁 Tech 🍏 Apple 🗺️ Google Maps 📺 YouTube 📄 Wikipedia 📁 Aru 📁 UChicago 📧 Mail 📰 News 📁 Popular 📁 Imported From Safari

 **SOLE**      Create Load Tag Account       Computation Institute  
University of Chicago       DEPARTMENT  
OF APPLIED SCIENCE  
University of Napoli Parthenope

---

- Science Objects
  - + Source Code
  - + Annotations
  - Web
    - Service/Workflow
      - Algorithm
      - Workflow
      - MLct Data
      - calculate reclassification
      - Virtual Images

## 2Algorithm

Our general algorithmic approach can be summarized as follows.

- 1.Prepare MLCT data.
  - (a)Reproject to geographic coordinates and mask cUSA study area.
  - (b)Reclassify to PEEL0 classification (Table 3).
  - (c)Calculate per pixel, per class areas at native resolution as a function of parameter Amin (see Section 3.1.2).
  - (d)Aggregate the new classification to the 5 grid, combining MLCT primary class, confidence, and secondary class values.

The algorithm described here will be performed on the subset of the global 5-

### reclassification :

<b>tag</b>	reclassification
<b>Filename</b>	datasets.R
<b>Inputs</b>	primary_namefile:tiff, secondary_namefile:tiff, pct_namefile:tiff
<b>Outputs</b>	primary_reclassified:tiff, secondary_reclassified:tiff
<b>Description</b>	Synthesis of a Complete Land Use/Land Cover Dataset for the Conterminous ...
<b>Uri</b>	<a href="http://storm.uniparthenope.it:18080/workflow?id=c34443fdcf00d13c">http://storm.uniparthenope.it:18080/workflow?id=c34443fdcf00d13c</a>

- Attach the SO to a word in the paper



← → ↻ 🏠 ec2-50-17-179-54.compute-1.amazonaws.com/~quanpt/geo/ ☆ ☰

DataBlogs Tech Apple Google Maps YouTube Wikipedia Aru UChicago Mail News Popular Imported From Safari

**SOLE** Create Load Tag Account

– Science Objects  
+ Source Code  
+ Annotations  
– Web  
Service/Workflow  
  Algorithm  
  Workflow  
  Mlct Data  
  calculate  
  reclassification  
Virtual Images

## 2Algorithm

Our general algorithmic approach can be summarized as follows.

- 1.Prepare MLCT data.
  - (a)Reproject to geographic coordinates and mask cUSA study area.
  - (b)Reclassify to PEEL0 classification (Table 3).
  - (c)Calculate per pixel, per class areas at native resolution as a function of parameter  $A_{min}$  (see Section 3.1.2).
  - (d)Aggregate the new classification to the 5 grid, combining MLCT primary class, confidence, and secondary class values.

The algorithm described here will be performed on the subset of the global 5-arc-minute grid that contain land area of the 48 contiguous states of the United States but is intended to be applied globally. As we will discuss in Chapter 2 when the base data sets are described in greater detail the MLCT is

**Running workflow "reclassification (imported from uploaded file)"** Expand All Collapse

**Step 1: reclassification**

Send results to a new history

Run workflow

Unnamed history 734.9 Kb



Create Load Tag Account



- Science Objects
  - + Source Code
  - + Annotations
- Web
  - Service/Workflow
    - Algorithm
    - Workflow
    - Mlct Data
    - calculate reclassification
    - Virtual Images

## 2Algorithm

Our general algorithmic approach can be summarized as follows.

1. Prepare MLCT data.

(a) Reproject to geographic coordinates and mask cUSA study area.

(b) **Reclassify** to PEEL0 classification (Table 3).

(c) Calculate per pixel, per class areas at native resolution as a function of parameter  $A_{min}$  (see Section 3.1.2).

(d) Aggregate the new classification to the 5 grid, combining MLCT primary class, confidence, and secondary class values.

Successfully ran workflow "reclassification (imported from uploaded file)". The following datasets have been added to the queue:

- 21: primary\_reclassified:tiff
- 22: secondary\_reclassified:tiff
- 23: Stdout/err mlctreclass



- Capture the traversal of links on the website.
- Present them as science objects
  - With search terms
  - Full lineage
  - Important URLs in the step





- Could be seen as an abstract component responding to events
- A fully described SO has 4 working contexts
  - **Acting**  
Performing data operations  
(datasets, instruments, program execution, virtual machine instancing)
  - **Visualizing**  
Rendering the object on a live media (web site)
  - **Printing**  
Rendering the object on a immutable media (printed paper)
  - **Interacting**  
The SO has to be managed using a web GUI
- Each context is enforced in a different environment
- Different contexts connect using distributed computing techniques



- The Acting context is executed in a virtualized environment
- Events:
  - **onInit**: fired when the SO is created
  - **onStageIn**:  
fired when data have to be moved in the SO local scratch area
  - **onRun**: fired when the SO has to do something as a program runs
  - **onStageOut**: fired when data have to be pushed back
  - **onFinalize**: fired before the SO has to be destroyed
- The VM could be instanced under the reader credentials

- <http://www.ci.uchicago.edu/SOLE>
- Two examples of research papers from the Center for Robust Decision making on Climate and Energy Policy (RDCEP)
  1. The author must associate the text and embedded figures with science objects that include datasets, algorithmic descriptions, computational analysis workflows, and workflow executions
  2. The author must associate descriptions in the paper with a set of data values, each of which is embedded in another research paper.
  3. PubChem papers with data descriptions and the association of search objects

# Conclusions



- The next generation of scientists will interact with living, reproducible, enhance-enabled publications.  
{ avoid to re-invent the wheel, speed-up science with deep social cooperation }
- High Performance Cloud Computing will be behind the scene  
{ HPCC means elastic scalability }
- SOLE is an application based on this vision thing.



Scientific Object Linking and Embedding

# Acknowledgements



- Neil Best, *RDCEP*
- Alison Brizius, *RDCEP*
- Don Middleton, *NCAR*
  
- Thanks to funding support from RDCEP



THE UNIVERSITY OF  
**CHICAGO**