Computation Institute

SOLE: Linking Research Papers with Science Objects

Tanu Malik, Quan Pham, Ian Foster

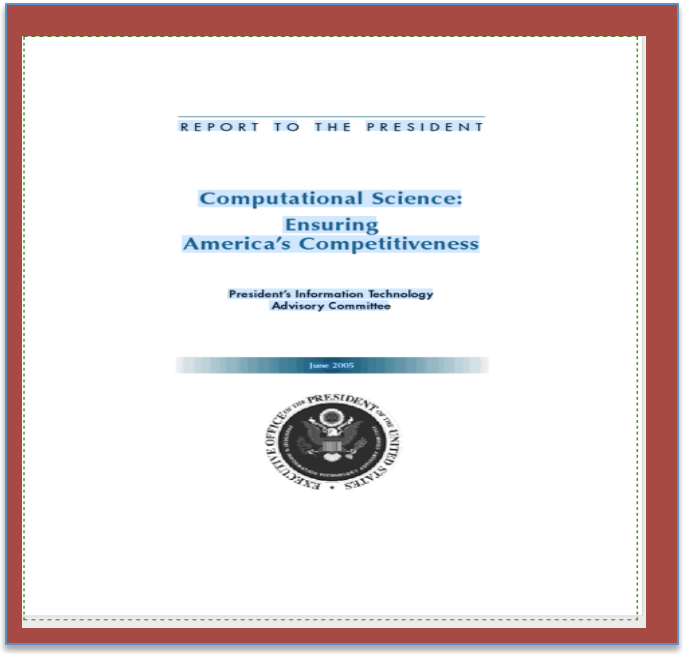
Computation Institute, Dept. of Computer Science

University of Chicago

Microsoft eScience, 2012
Chicago

The boon of computational science

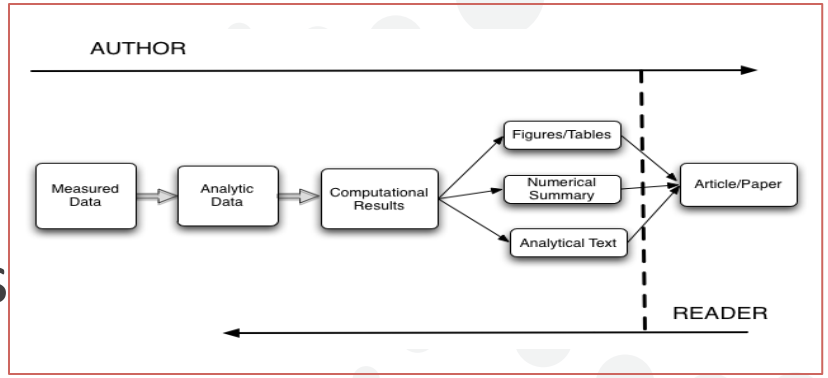
Computational science provides a unique window through which researchers can investigate problems that are otherwise impractical or impossible to address, ranging from scientific investigations of the biochemical processes of the human brain and the fundamental forces of physics shaping the universe, 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 on Computational Science: Ensuring America's Competitiveness, June 2005

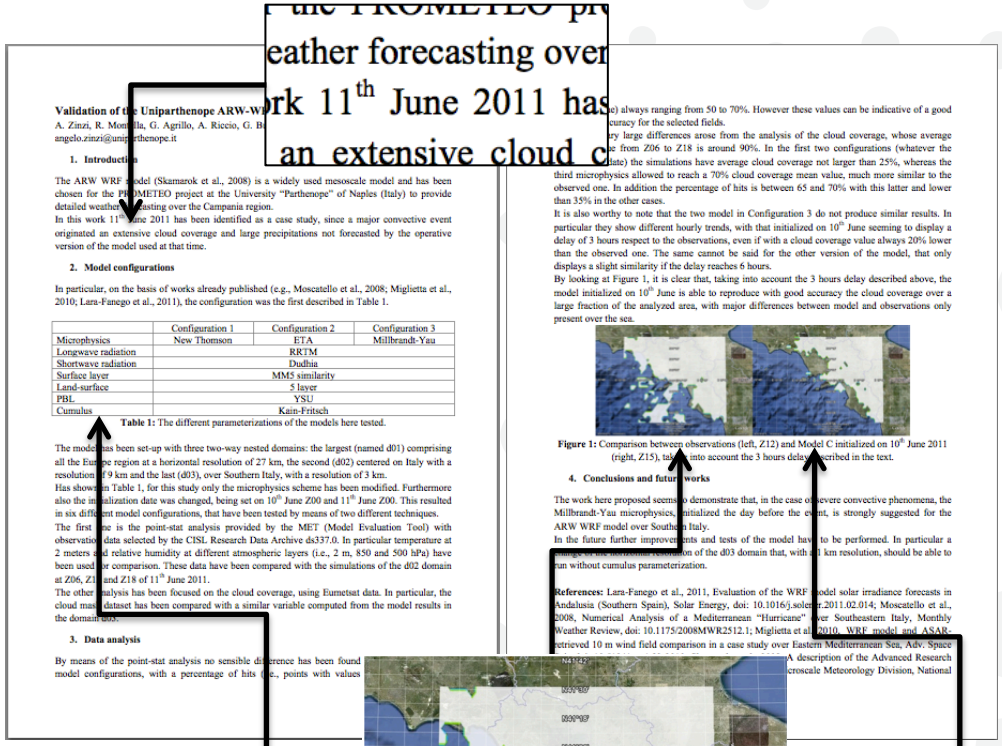
The scourge of computational science

- 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



	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.

Example 2: Policy Paper

special conditions or circumstances that differentiate them from more recent statutes. Iowa, the earliest RPS (1983), had requirements so small (∼ 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 (∼ $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

1 2 3 4 5

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).

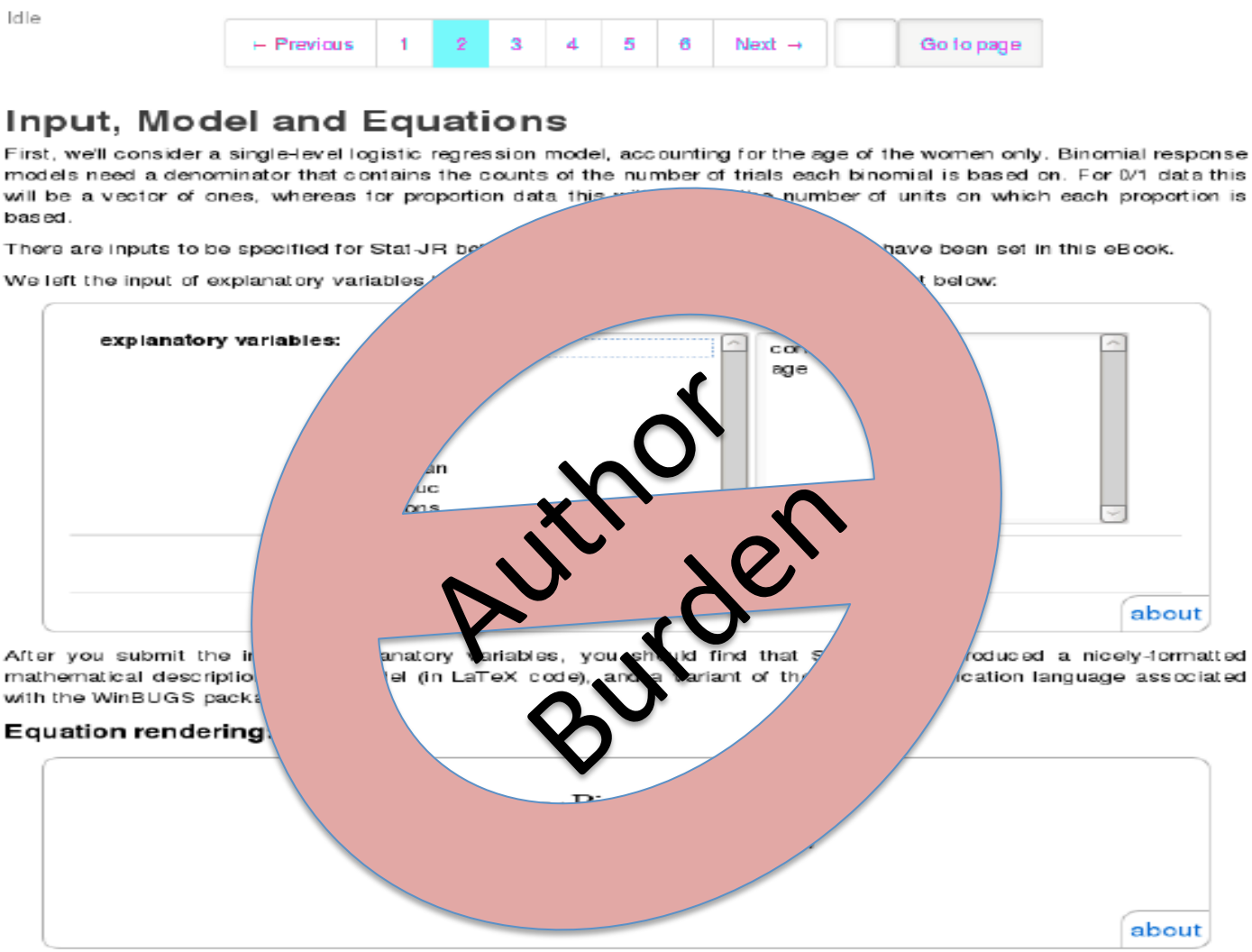
Current state-of-the-art

- 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*

Readability desired by readers and reviewers

- 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.

Well-Connected



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.

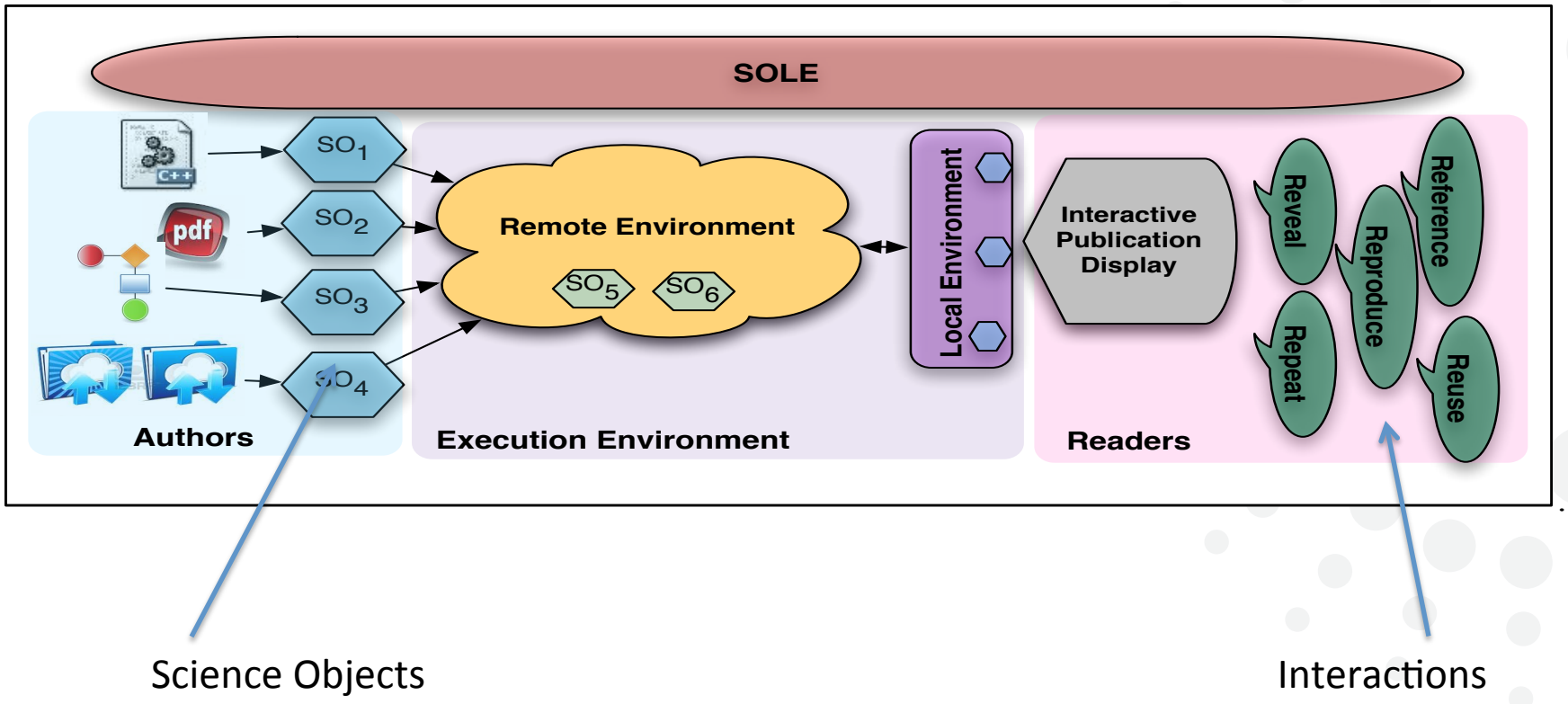
Relieving Author Burden and Improving Readability

- 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

SOLE: Science Object and Linking

- 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



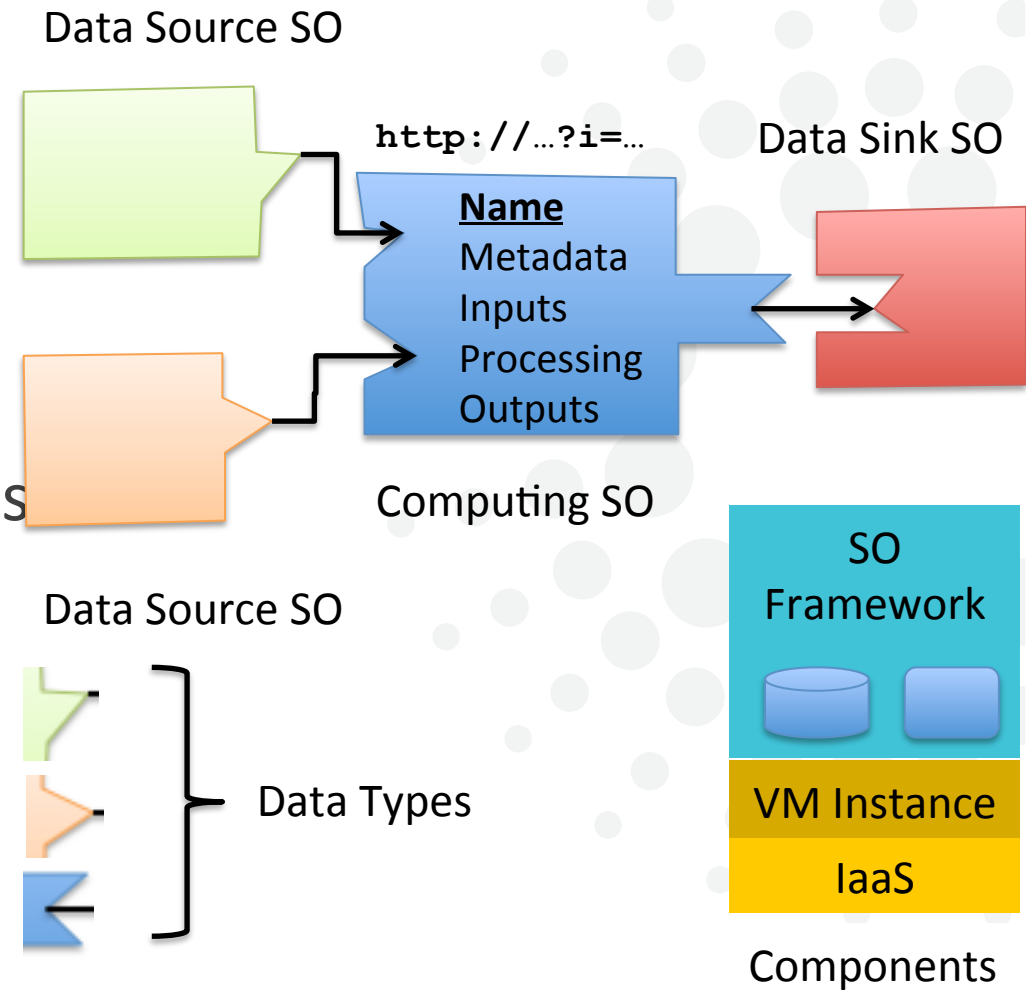
Interactions

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@
- Close tag format: ###t@
- Tag naming: tag1|tag2|…|tagn (fluidinfo like)
- Element attributes: Key/Value

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

source( "~/thesis/code/peel.R")
source( "~/thesis/code/maps.Rʃ")
###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

                              # studyArea used to work out RMSE
                              # calcs and tables
##studyArea <- "thumb”
peelBands <- peelClasses +1
###t@
```

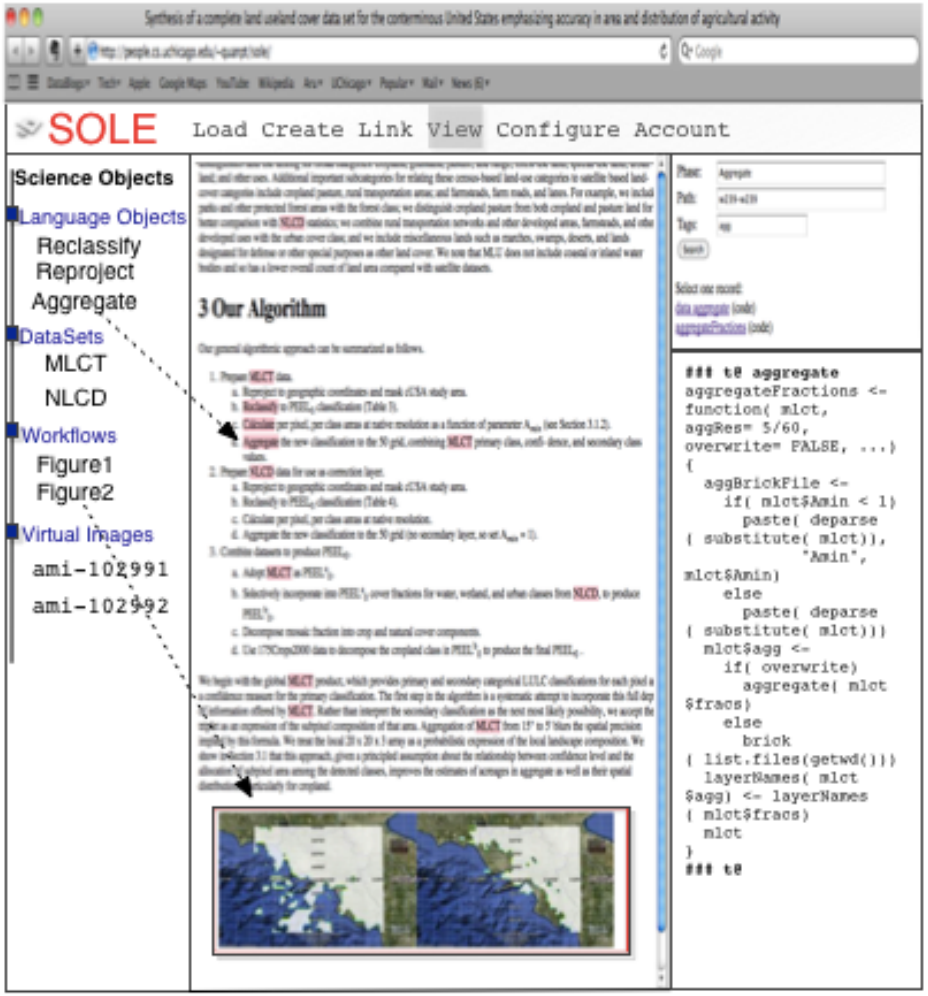
Source Code Association

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

1. Authors identify science objects with human-readable tags

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

2. Run SOLE; Create SO



Step 3: Search SO with tag name and link

3. Search SO with tag name and link

Annotation Association

#t@tbl reclassification of MLCT

	MLCT/IGBP	$PEEL_0$
0	water	water
1	evergreen needleleaf forest	forest
2	deciduous needleleaf forest	
3	evergreen broadleaf forest	
4	deciduous broadleaf forest	
5	mixed forests	
6	closed shrublands	shrub
7	open shrublands	
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	barren
16	barren or sparsely vegetated	

Workflow Association

- Workflow input
- Software unit tag
- Inputs and Outputs

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

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

Workflow Association

SOLE

Create Load Tag Account

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

Master of ArtsM.A. Geography & Environmental Studies

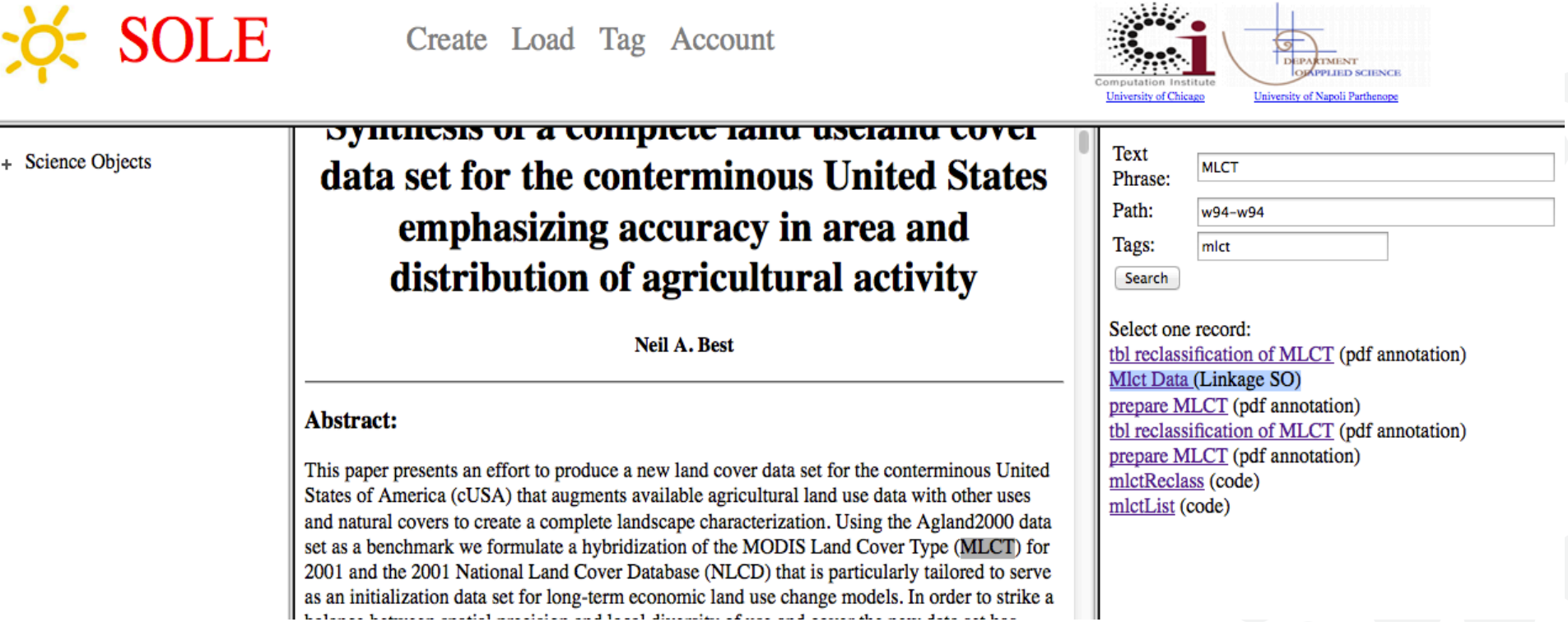
August 2011

Synthesis of a complete land us
data set for the conterminous U
emphasizing accuracy in a
distribution of agricultural

Neil A. Best

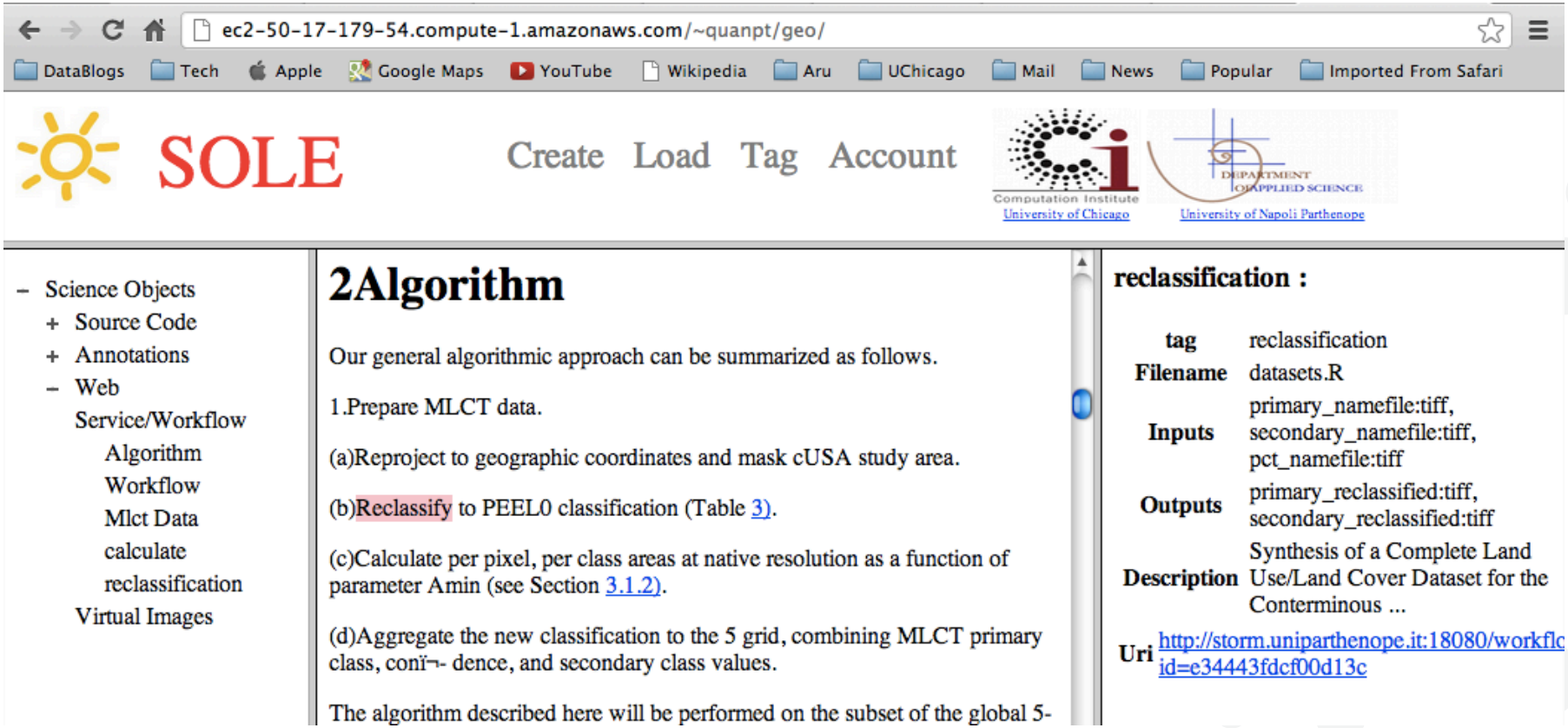
- The defined SOs for the paper

Attach the SO to a word in the paper



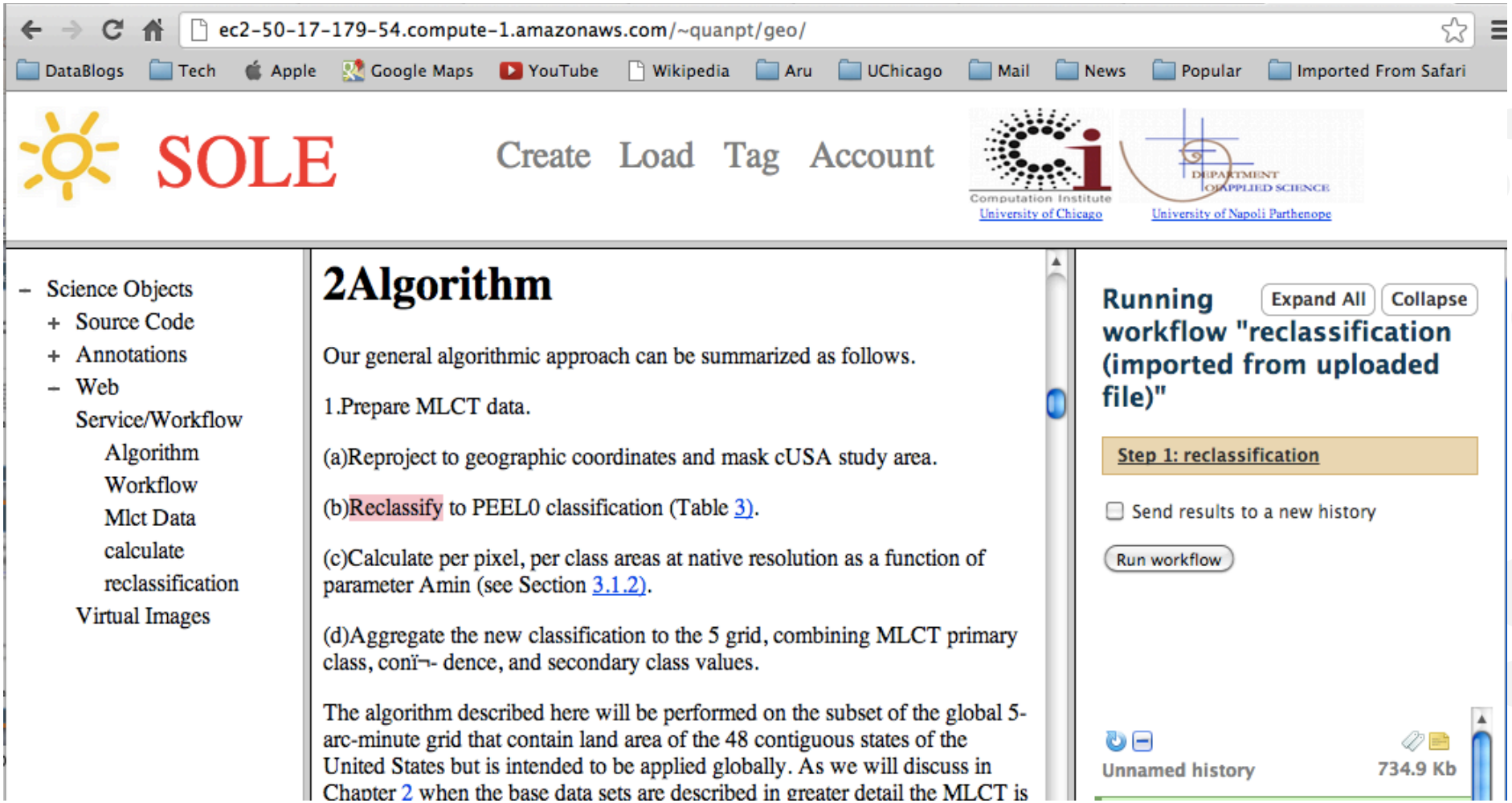
- Attach the SO to a word in the paper

Workflow Asociation

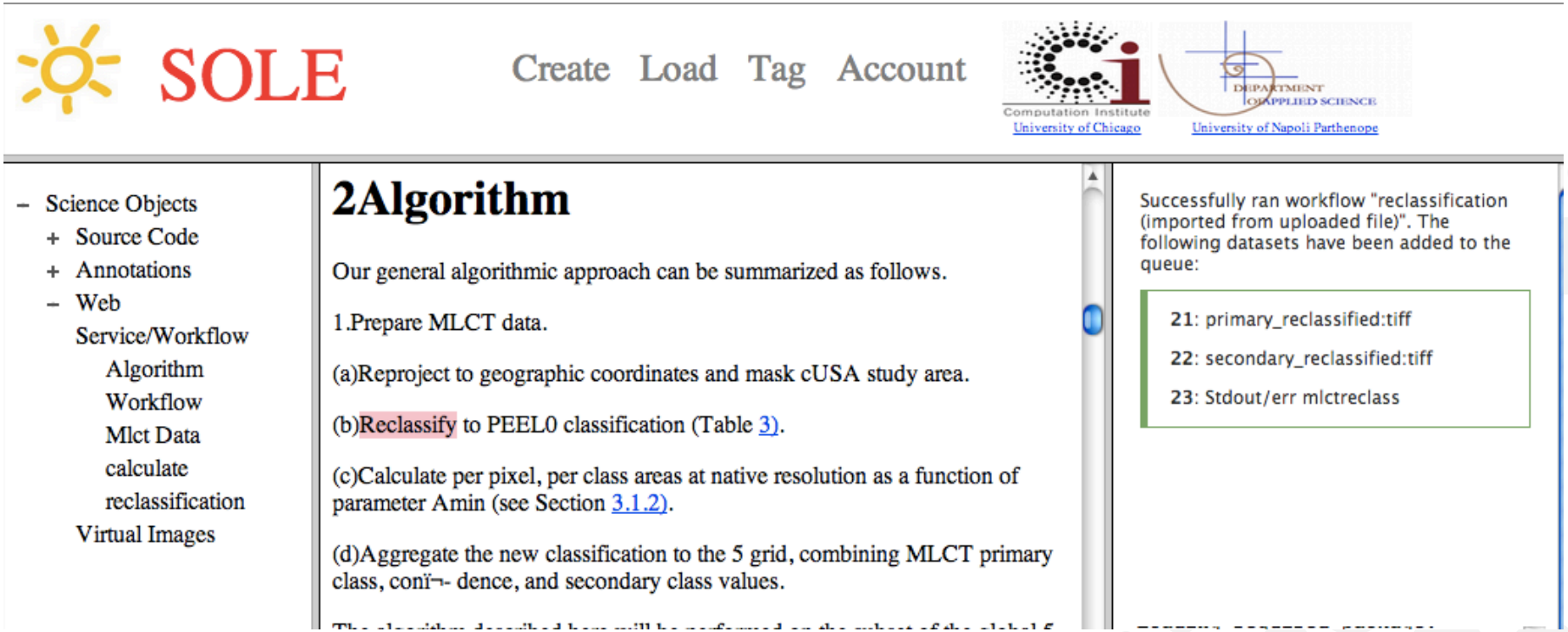


- Attach the SO to a word in the paper

Workflow Run



Workflow Run



Digital Repository Association

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

Science Object

- 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

Science Object (Cntd)

- 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

Demo time

- 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