Information
Sciences
Institute

OKN-AKBC
May 22nd,Amherst, USA

OKG-SOFT: AN OPEN KNOWLEDGE GRAPH WITH MACHINE READABLE SCIENTIFIC SOFTWARE METADATA

Daniel Garijo, Maximiliano Osorio, Deborah Khider,
Varun Ratnakar and Yolanda Gil

University of Southern California,
Information Sciences Institute

@dgarijov

Science is changing: Open Science

Open data

figshare

DRYAD

The Dataverse Network™ Project

Impact and credit

Open access

creative commons

open source initiative

Open publications

PLOS

Policy Guidelines FOR THE DEVELOPMENT AND PROMOTION OF OPEN ACCESS

wikigenes collaborative publishing

Open source software

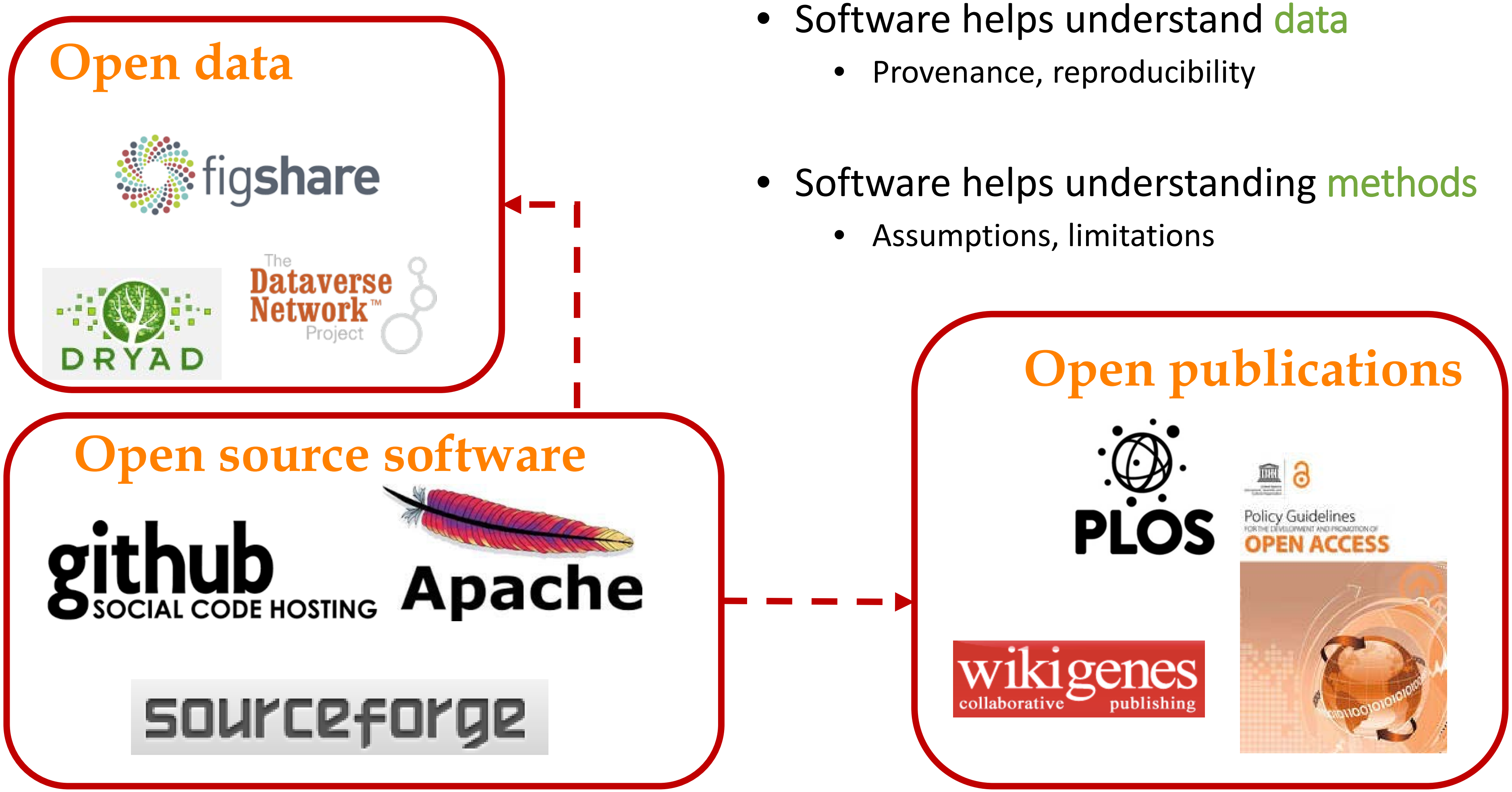
github SOCIAL CODE HOSTING

Apache

sourceforge

The importance of Scientific Software

- Software helps understand data
  - Provenance, reproducibility
- Software helps understanding methods
  - Assumptions, limitations



Open data

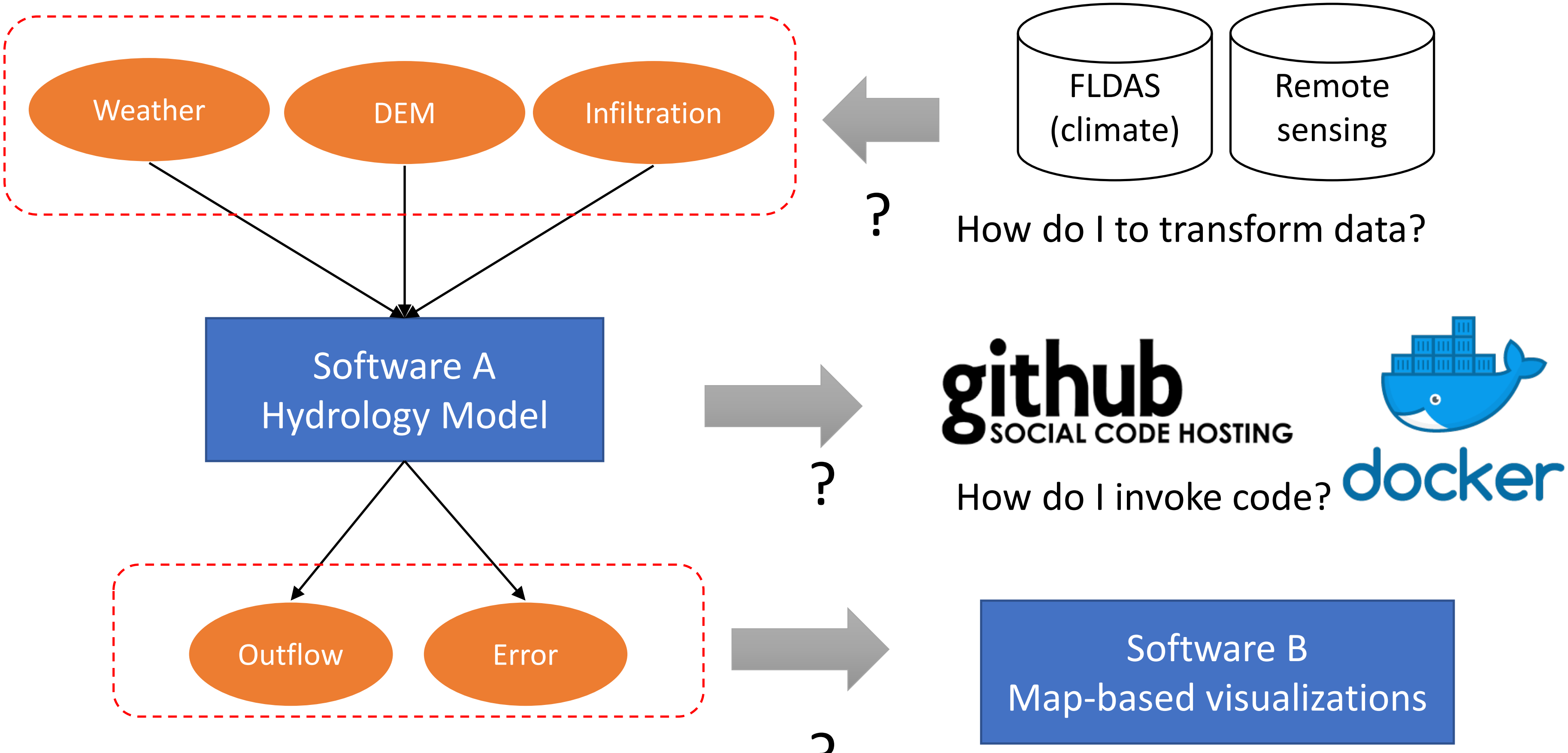
Open source software

Open publications

figshare
DRYAD
The Dataverse Network™ Project
github SOCIAL CODE HOSTING
Apache
sourceforge
PLOS
Policy Guidelines FOR THE DEVELOPMENT AND PROMOTION OF OPEN ACCESS
wikigenes collaborative publishing

Why is it difficult to reuse Scientific Software?

Let's imagine we want to reuse existing work:



How do I to transform data?

How do I invoke code?

How do I interpret results?

Outline

1. Requirements help scientific software reusability

2. Our current approach for representing scientific software metadata

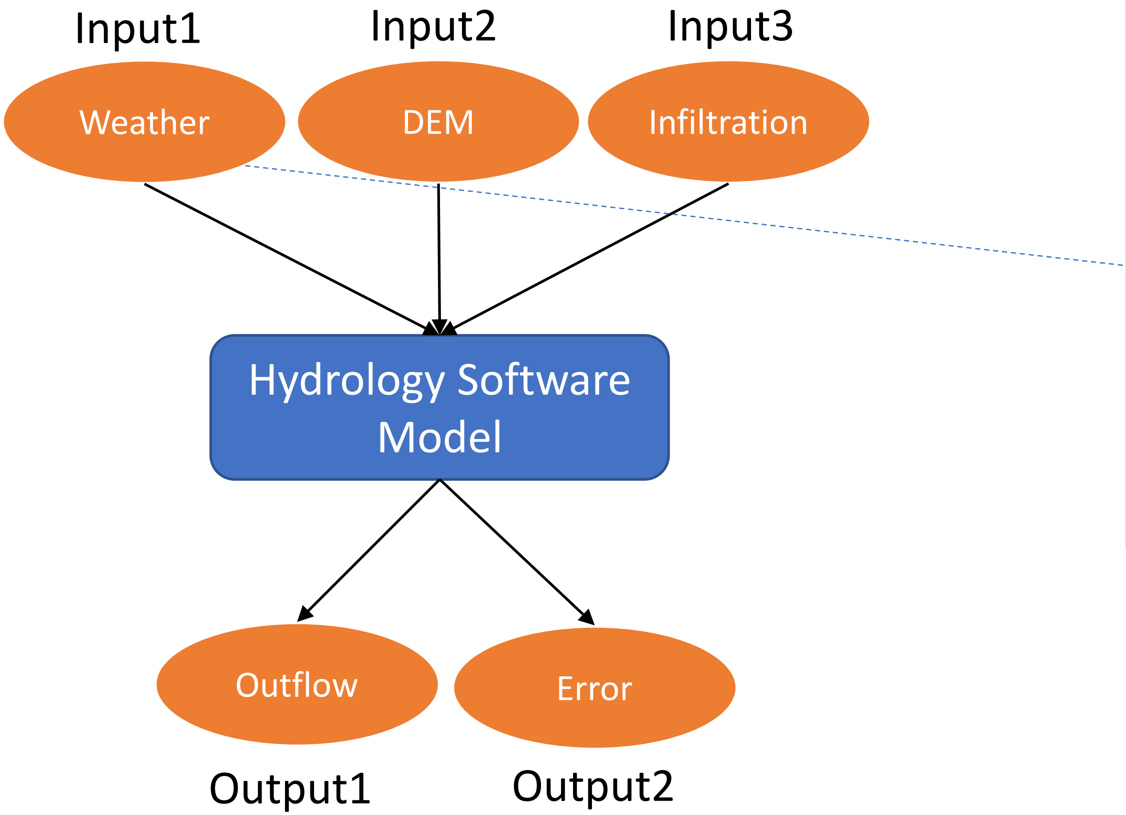
3. A framework to query, explore, exploit and publish software metadata

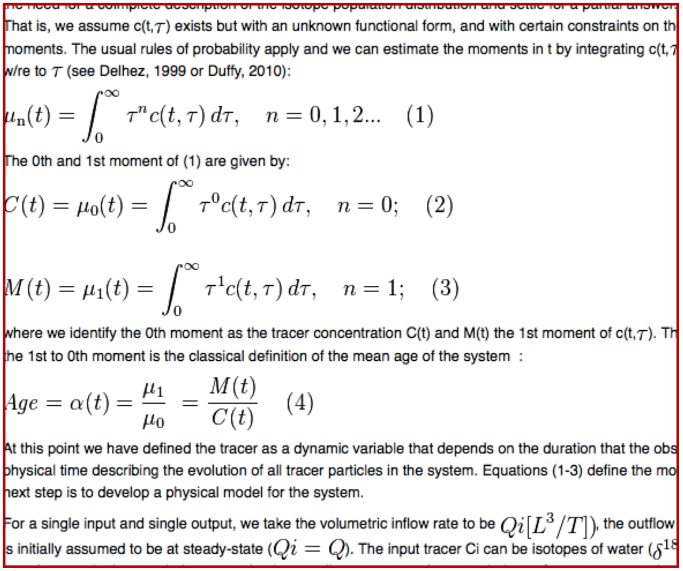
Outline

1. **Requirements help scientific software reusability**
2. Our current approach for representing scientific software metadata
3. A framework to query, explore, exploit and publish software metadata

Requirements for Software Reusability

1. Exposing software inputs, outputs and their corresponding variables




he need for a complete description of the isotope population distribution and settle for a partial answer.
That is, we assume c(t,τ) exists but with an unknown functional form, and with certain constraints on th
moments. The usual rules of probability apply and we can estimate the moments in t by integrating c(t,τ
w/re to τ (see Delhez, 1999 or Duffy, 2010):

$$\mu_n(t) = \int_0^\infty \tau^n c(t,\tau)\, d\tau, \quad n = 0,1,2... \quad (1)$$

The 0th and 1st moment of (1) are given by:

$$C(t) = \mu_0(t) = \int_0^\infty \tau^0 c(t,\tau)\, d\tau, \quad n = 0; \quad (2)$$

$$M(t) = \mu_1(t) = \int_0^\infty \tau^1 c(t,\tau)\, d\tau, \quad n = 1; \quad (3)$$

where we identify the 0th moment as the tracer concentration C(t) and M(t) the 1st moment of c(t,τ). Th
he 1st to 0th moment is the classical definition of the mean age of the system :

$$Age = \alpha(t) = \frac{\mu_1}{\mu_0} = \frac{M(t)}{C(t)} \quad (4)$$

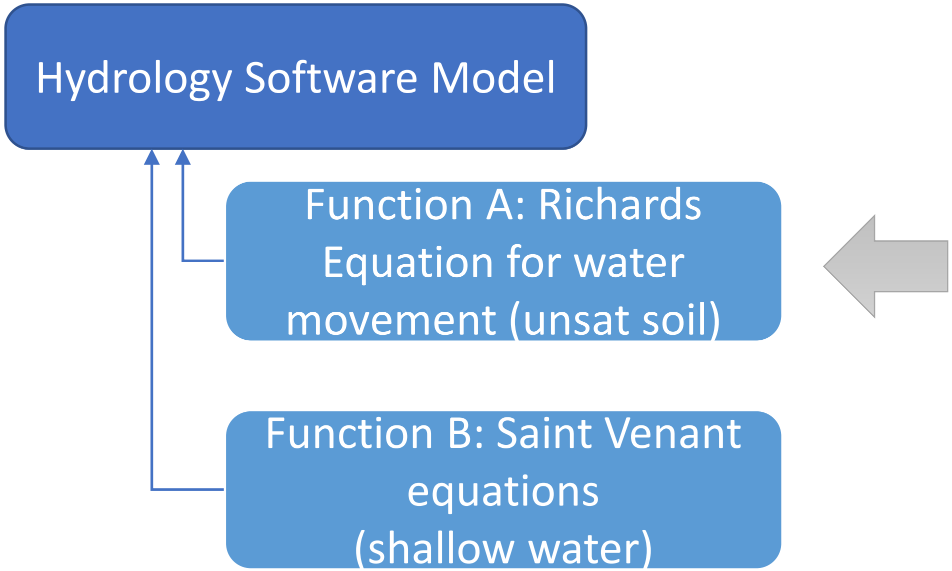
At this point we have defined the tracer as a dynamic variable that depends on the duration that the obs
physical time describing the evolution of all tracer particles in the system. Equations (1-3) define the mo
next step is to develop a physical model for the system.

For a single input and single output, we take the volumetric inflow rate to be $Qi[L^3/T]$), the outflow
s initially assumed to be at steady-state ($Qi = Q$). The input tracer Ci can be isotopes of water (δ^{18}

- Land surface temperature (degC)
- Precipitation rate (mm/h)
- Land surface wind speed (m/day)
- Net radiation (MJ/(day m^2)

Requirements for Software Reusability

1. Exposing software inputs, outputs and their corresponding variables
2. Capturing the functions of the software being used



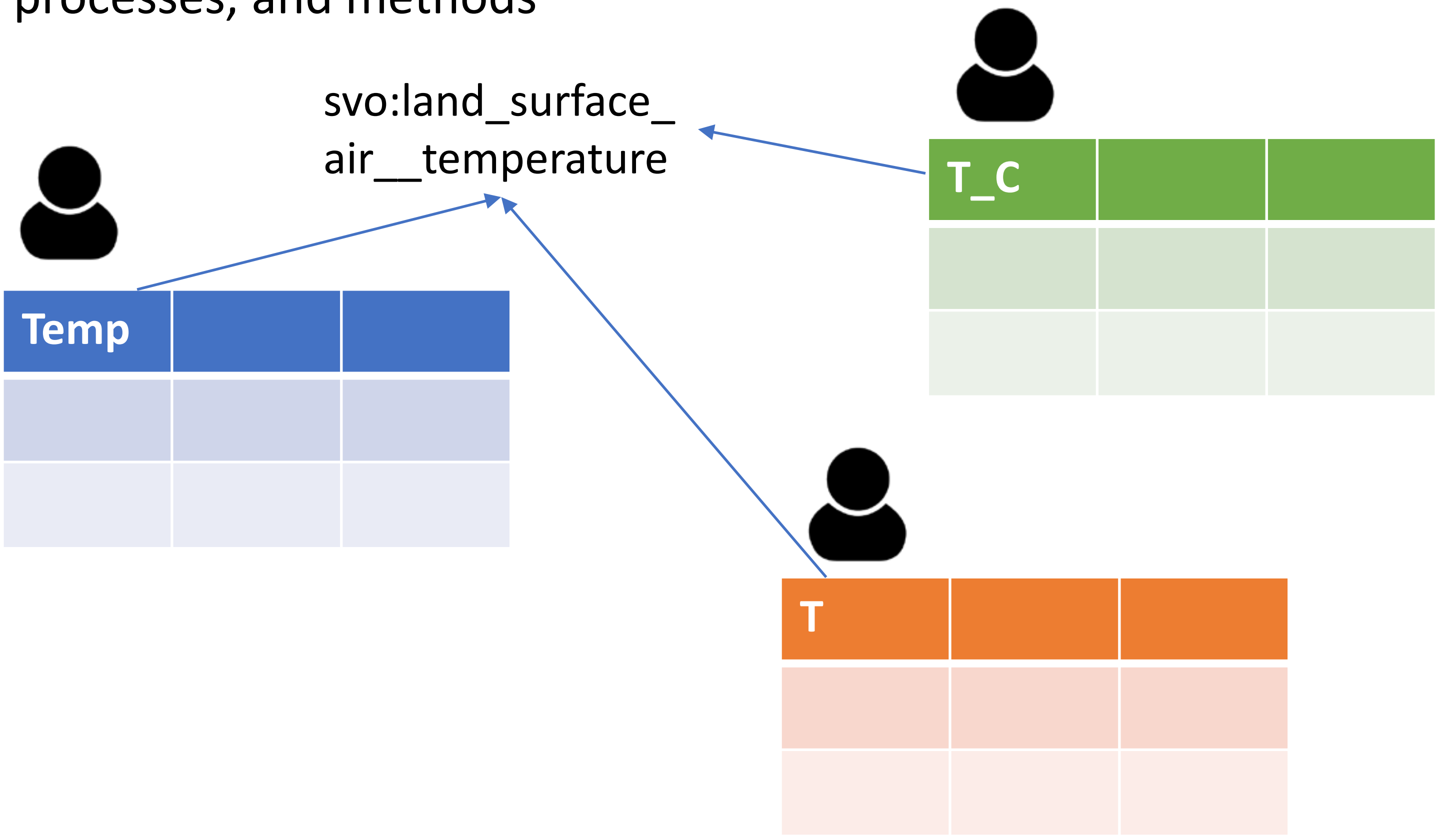
Hydrology Software Model

Function A: Richards Equation for water movement (unsat soil)

Function B: Saint Venant equations (shallow water)

Requirements for Software Reusability

1. Exposing software inputs, outputs and their corresponding variables
2. Capturing the functions of software being used
3. Using principled ontologies with structured names for model variables, processes, and methods



svo:land_surface_ air__temperature

Temp

T_C

T

Requirements for Software Reusability

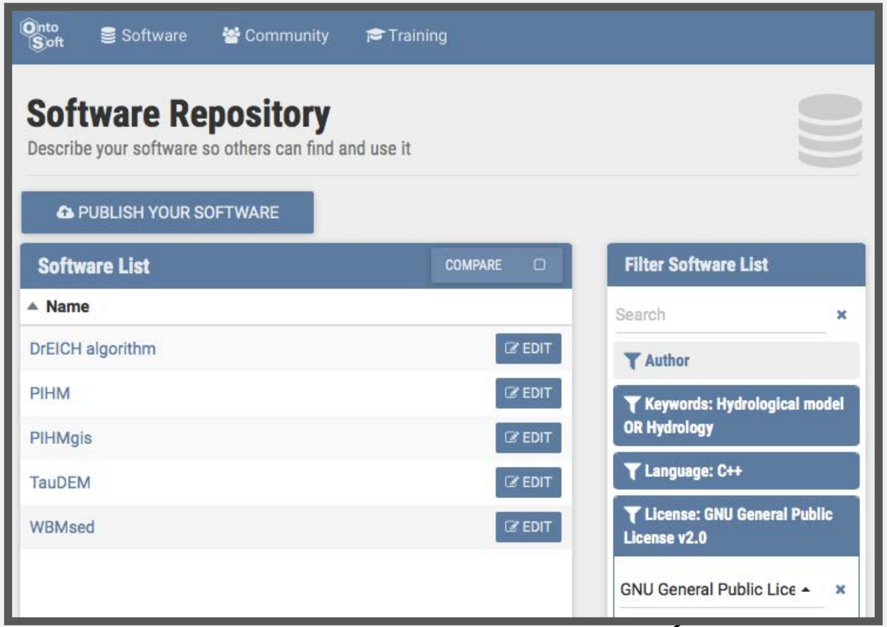
1. Exposing software inputs, outputs and their corresponding variables
2. Capturing the functions of software being used
3. Using principled ontologies with structured names for model variables, processes, and methods
4. Capture the semantic structure of software invocations

Dependencies?
Sample runs?
Invocation command?
Is data supposed to be in the same folder?
Default arguments/Configuration files?
Volumes?
Do I have to log in in the image

Outline

1. Requirements help scientific software reusability
2. **Our current approach for representing scientific software metadata**
3. A framework to query, explore, exploit and publish software metadata

Prior Work: OntoSoft Software Metadata Registry



Software Repository

Describe your software so others can find and use it

PUBLISH YOUR SOFTWARE

Software List COMPARE □

▲ Name

Name	
DrEICH algorithm	EDIT
PIHM	EDIT
PIHMgis	EDIT
TauDEM	EDIT
WBMsed	EDIT

Filter Software List

Search ×

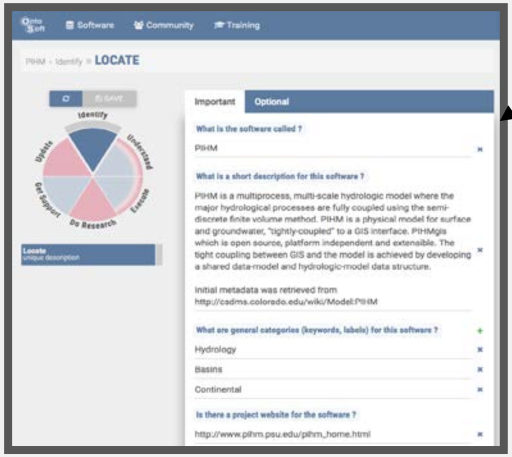
Author

Keywords: Hydrological model OR Hydrology

Language: C++

License: GNU General Public License v2.0

GNU General Public Lice ▲ ×



PIHM › Identify » LOCATE

Locate
unique description

Important | Optional

What is the software called ?

PIHM ×

What is a short description for this software ?

PIHM is a multiprocess, multi-scale hydrologic model where the major hydrological processes are fully coupled using the semi-discrete finite volume method. PIHM is a physical model for surface and groundwater, "tightly-coupled" to a GIS interface. PIHMgis which is open source, platform independent and extensible. The tight coupling between GIS and the model is achieved by developing a shared data-model and hydrologic-model data structure.

Initial metadata was retrieved from http://csdms.colorado.edu/wiki/Model:PIHM ×

What are general categories (keywords, labels) for this software ? +

Hydrology ×

Basins ×

Continental ×

Is there a project website for the software ?

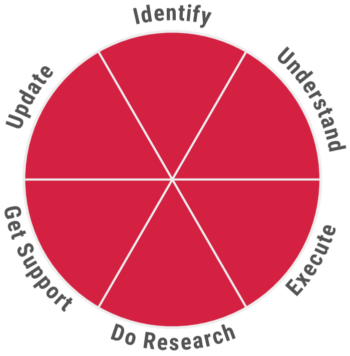
http://www.pihm.psu.edu/pihm_home.html ×

Finding Software

OntoSoft

Model and Software Metadata Registry

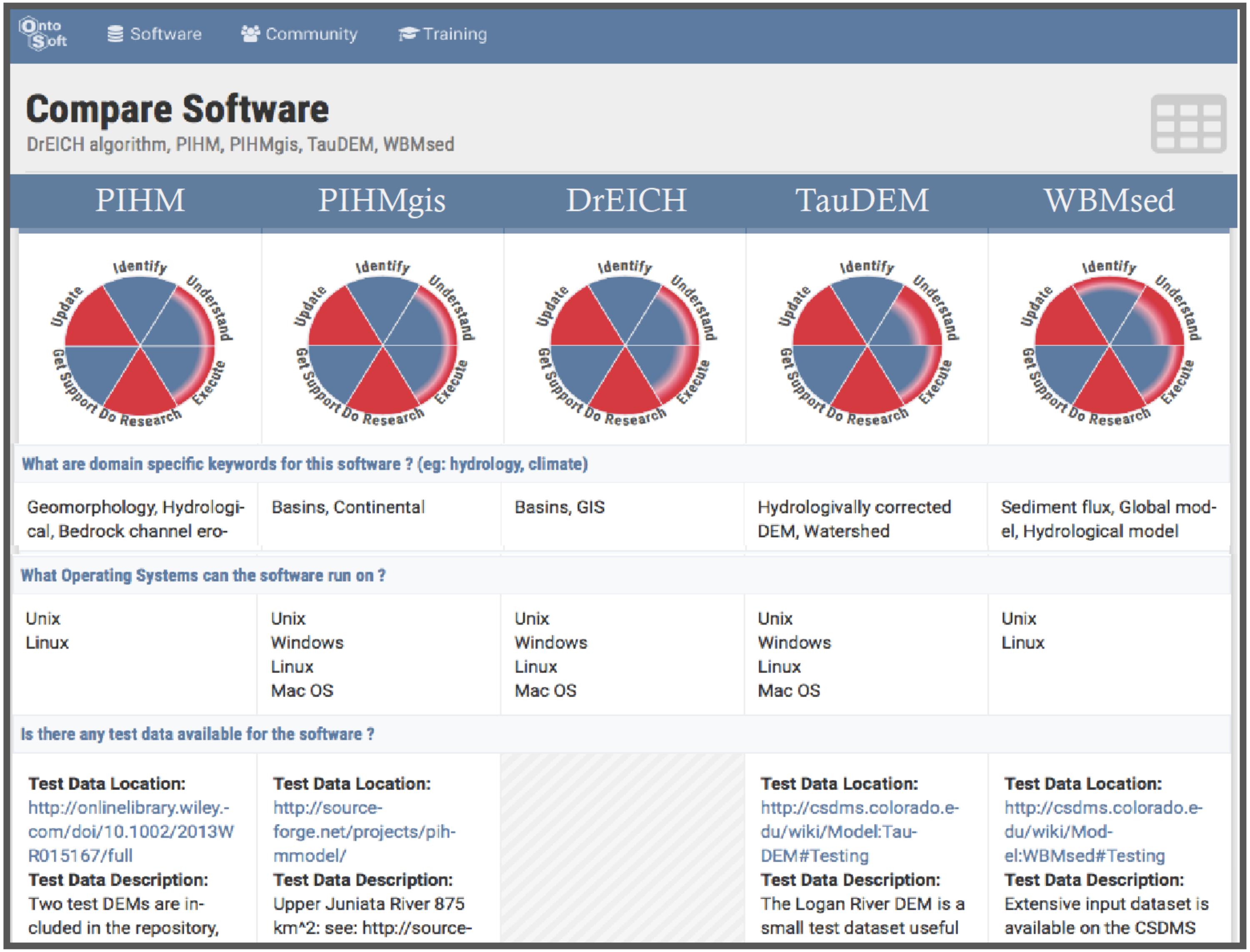
- Complements code repositories to make them understandable
- Software metadata designed for scientists
- Metadata is curated by decentralized communities of users
- Training scientists on best practices



http://ontosoft.org

[Gil et al 2015]: OntoSoft: Capturing Scientific Software Metadata Eighth ACM International Conference on Knowledge Capture, Palisades, NY, 2015

Prior Work: OntoSoft Software Metadata Registry

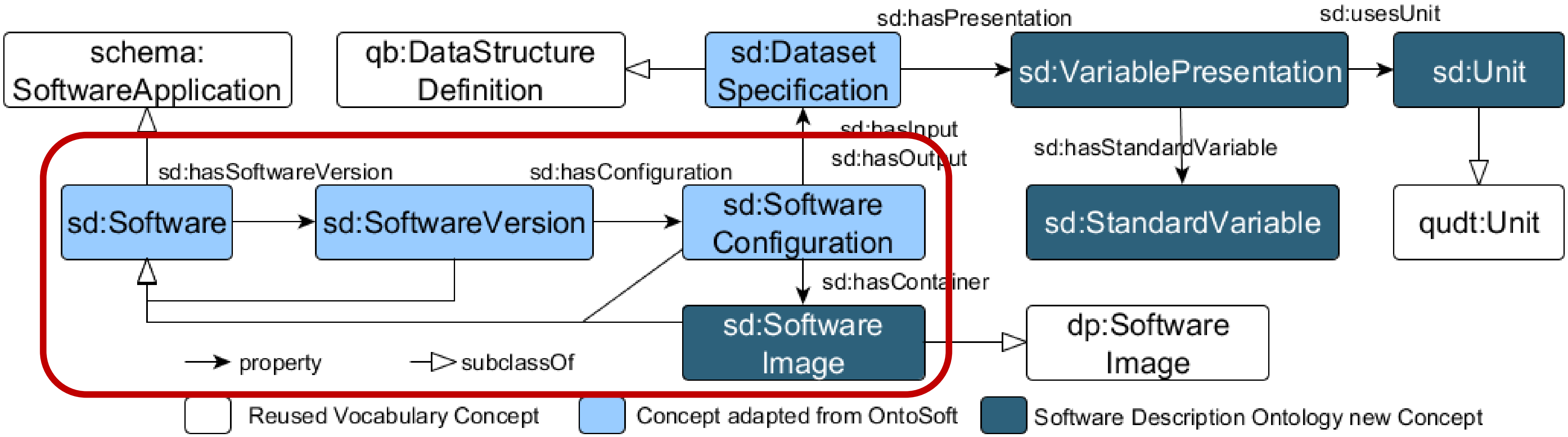


Compare Software

DrEICH algorithm, PIHM, PIHMgis, TauDEM, WBMsed

PIHM	PIHMgis	DrEICH	TauDEM	WBMsed
What are domain specific keywords for this software ? (eg: hydrology, climate)				
Geomorphology, Hydrological, Bedrock channel ero-	Basins, Continental	Basins, GIS	Hydrologivally corrected DEM, Watershed	Sediment flux, Global model, Hydrological model
What Operating Systems can the software run on ?				
Unix Linux	Unix Windows Linux Mac OS	Unix Windows Linux Mac OS	Unix Windows Linux Mac OS	Unix Linux
Is there any test data available for the software ?				
Test Data Location: http://onlinelibrary.wiley.com/doi/10.1002/2013WR015167/full **Test Data Description:** Two test DEMs are included in the repository,	**Test Data Location:** http://sourceforge.net/projects/pihmmodel/ **Test Data Description:** Upper Juniata River 875 km^2: see: http://source-		**Test Data Location:** http://csdms.colorado.edu/wiki/Model:TauDEM#Testing **Test Data Description:** The Logan River DEM is a small test dataset useful	**Test Data Location:** http://csdms.colorado.edu/wiki/Model:WBMsed#Testing **Test Data Description:** Extensive input dataset is available on the CSDMS

Evolving OntoSoft: Software Description Ontology

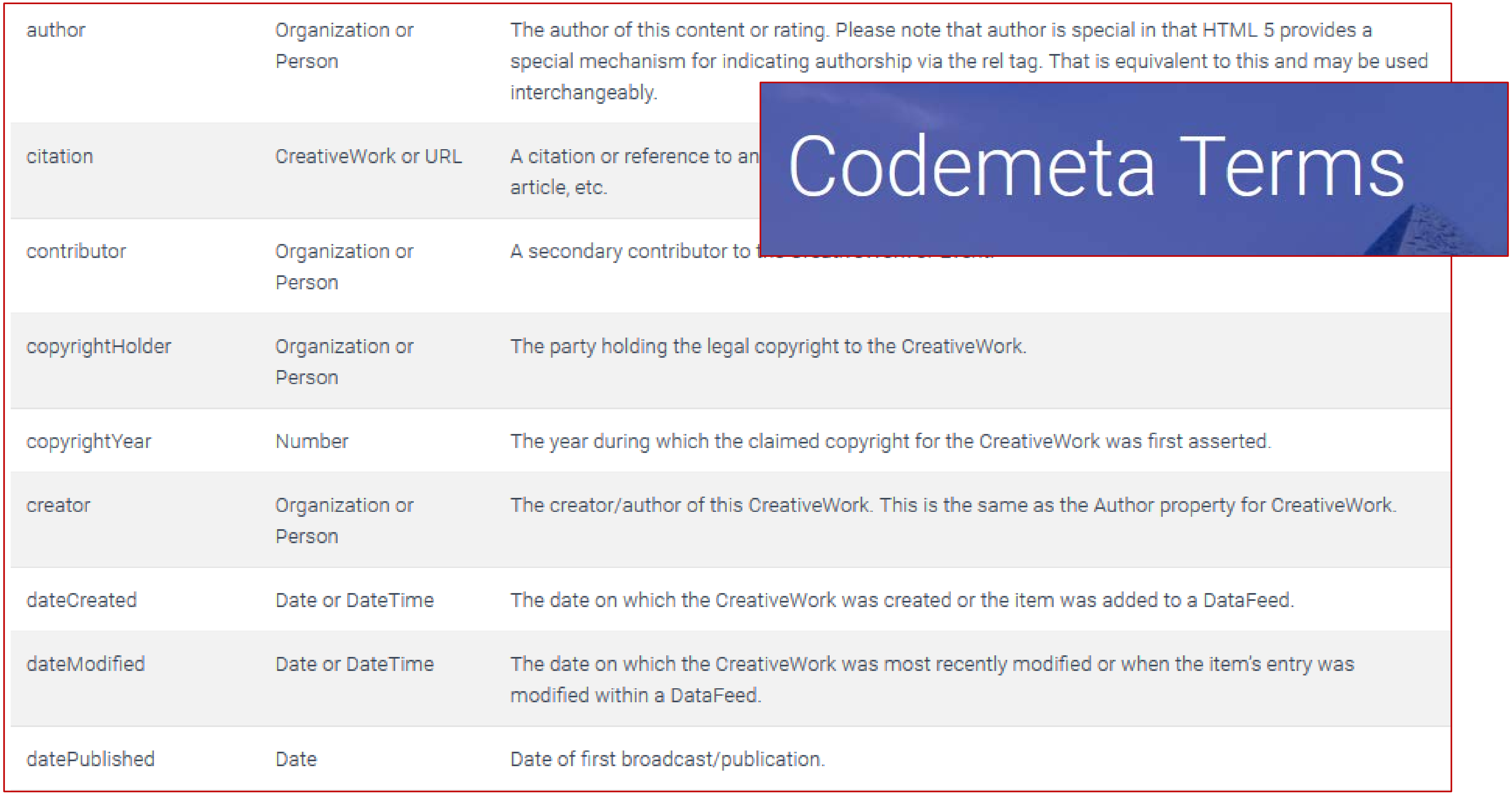


Extensions:

- Schema.org (software metadata)
- W3C Data Cubes (Contents of inputs and outputs)
- NASA QUDT (Units)
- DockerPedia (Software images)
- Scientific Variables Ontology (Standard Variables)

https://w3id.org/okn/o/sd#

Evolving OntoSoft: Extending schema.org and Codemeta

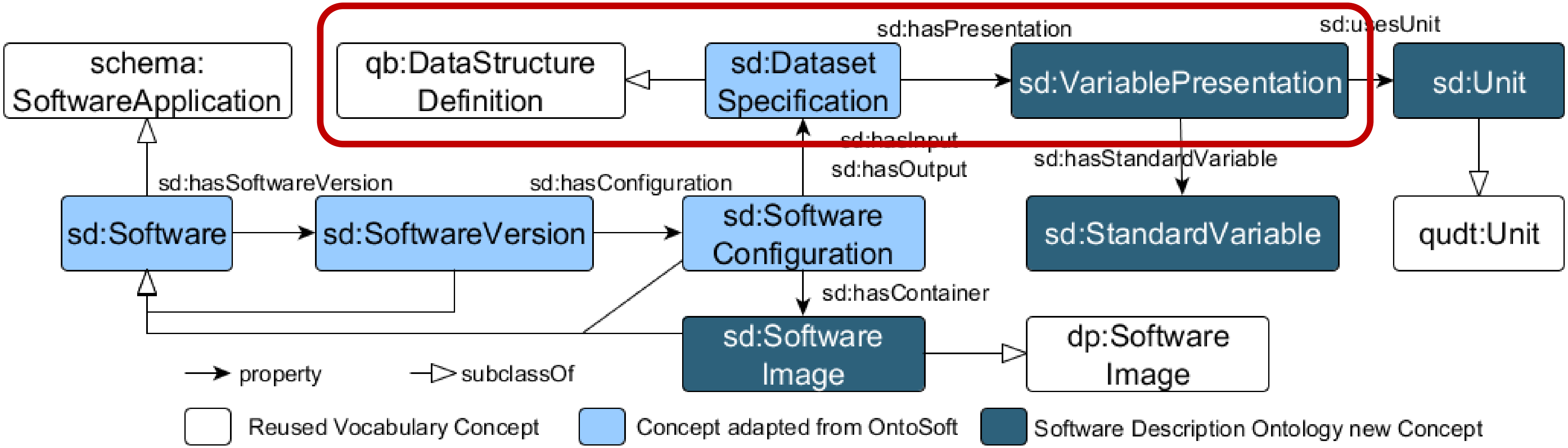


Codemeta Terms

author	Organization or Person	The author of this content or rating. Please note that author is special in that HTML 5 provides a special mechanism for indicating authorship via the rel tag. That is equivalent to this and may be used interchangeably.
citation	CreativeWork or URL	A citation or reference to an article, etc.
contributor	Organization or Person	A secondary contributor to the CreativeWork or Event.
copyrightHolder	Organization or Person	The party holding the legal copyright to the CreativeWork.
copyrightYear	Number	The year during which the claimed copyright for the CreativeWork was first asserted.
creator	Organization or Person	The creator/author of this CreativeWork. This is the same as the Author property for CreativeWork.
dateCreated	Date or DateTime	The date on which the CreativeWork was created or the item was added to a DataFeed.
dateModified	Date or DateTime	The date on which the CreativeWork was most recently modified or when the item's entry was modified within a DataFeed.
datePublished	Date	Date of first broadcast/publication.

https://w3id.org/okn/o/sd#

Evolving OntoSoft: Software Description Ontology

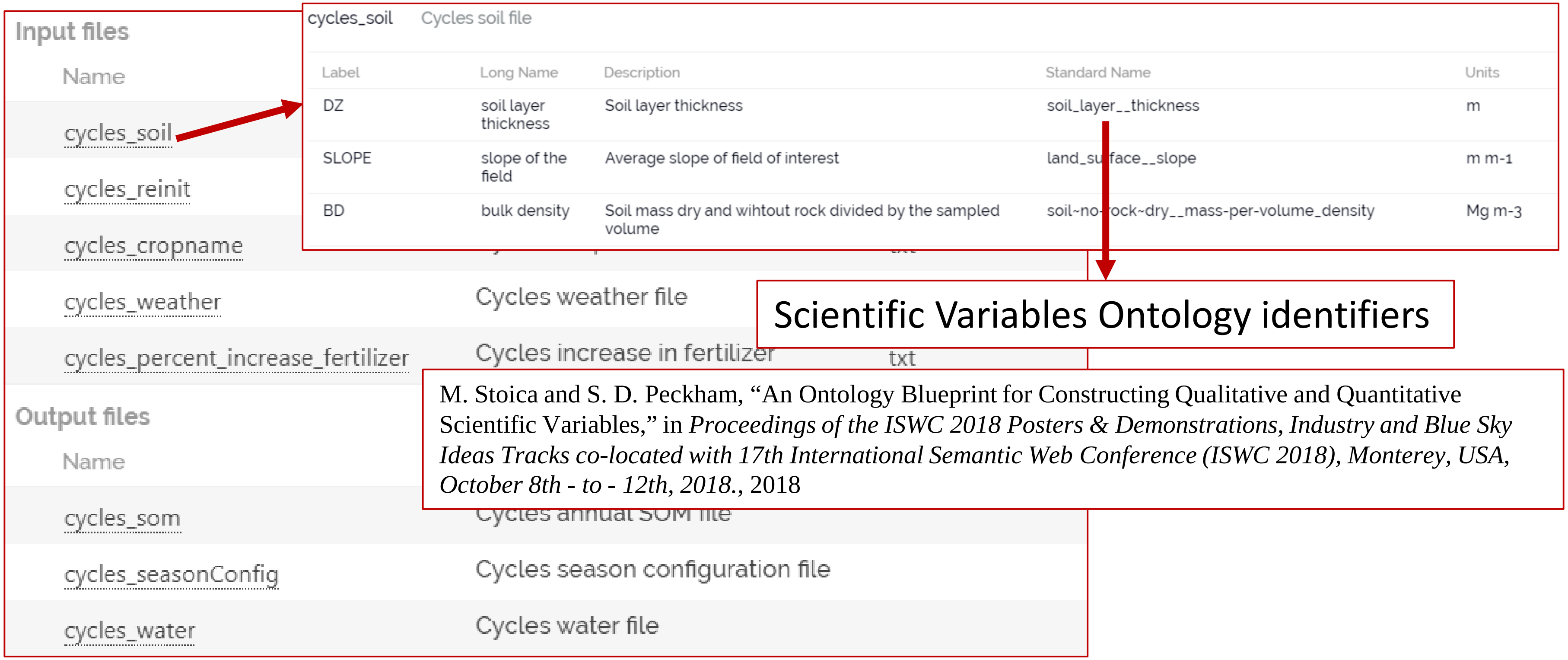


Extensions:

- Schema.org (software metadata)
- W3C Data Cubes (Contents of inputs and outputs)
- NASA QUDT (Units)
- DockerPedia (Software images)
- Scientific Variables Ontology (Standard Variables)

https://w3id.org/okn/o/sd#

Describing Input/Output files, parameters and variables



Input files

Name		
cycles_soil		
cycles_reinit		
cycles_cropname		txt
cycles_weather	Cycles weather file	
cycles_percent_increase_fertilizer	Cycles increase in fertilizer	txt

cycles_soil Cycles soil file

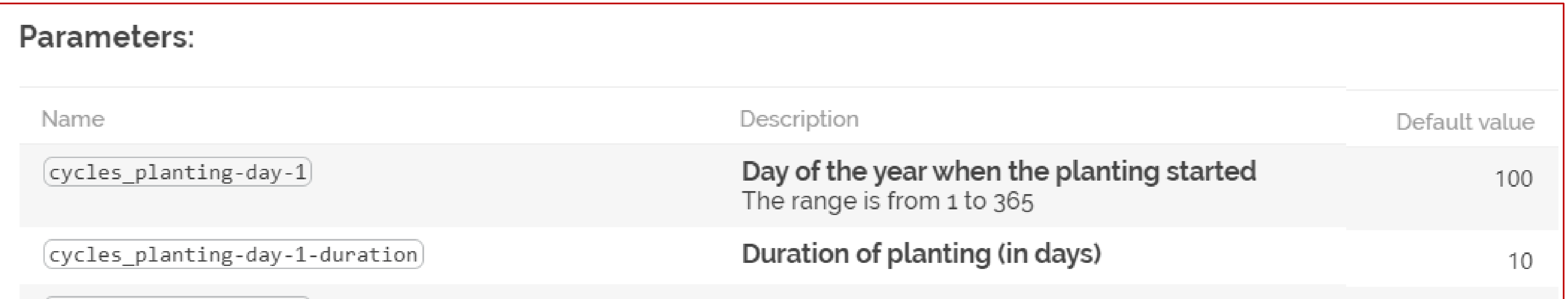
Label	Long Name	Description	Standard Name	Units
DZ	soil layer thickness	Soil layer thickness	soil_layer__thickness	m
SLOPE	slope of the field	Average slope of field of interest	land_surface__slope	m m-1
BD	bulk density	Soil mass dry and wihtout rock divided by the sampled volume	soil~no-rock~dry__mass-per-volume_density	Mg m-3

Scientific Variables Ontology identifiers

M. Stoica and S. D. Peckham, "An Ontology Blueprint for Constructing Qualitative and Quantitative Scientific Variables," in *Proceedings of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks co-located with 17th International Semantic Web Conference (ISWC 2018), Monterey, USA, October 8th - to - 12th, 2018.*, 2018

Output files

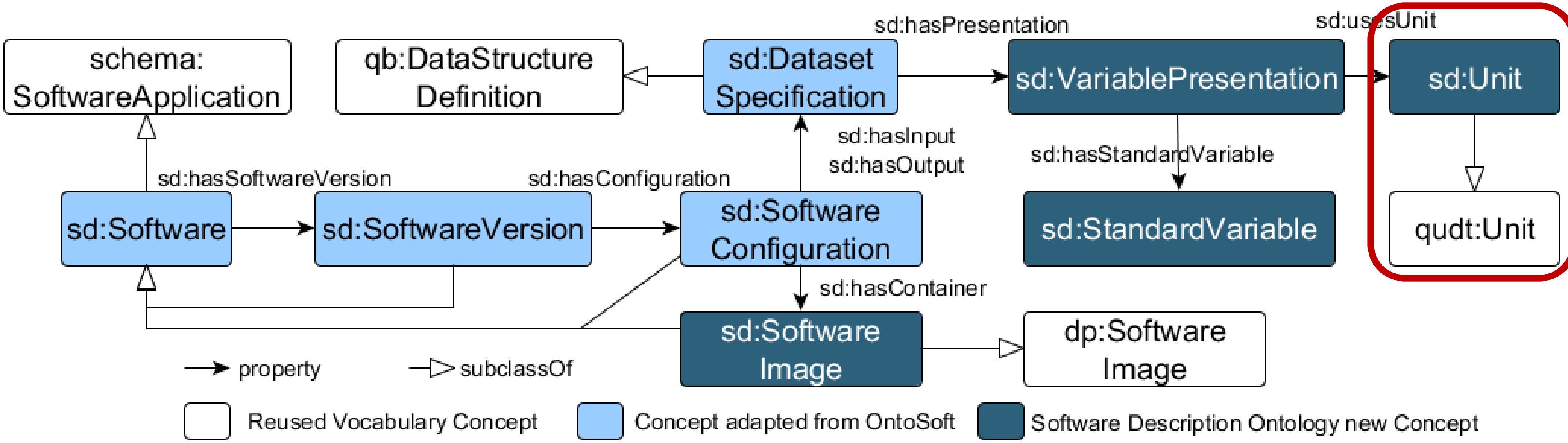
Name	
cycles_som	Cycles annual SOM file
cycles_seasonConfig	Cycles season configuration file
cycles_water	Cycles water file



Parameters:

Name	Description	Default value
cycles_planting-day-1	**Day of the year when the planting started** The range is from 1 to 365	100
cycles_planting-day-1-duration	**Duration of planting (in days)**	10

Evolving OntoSoft: Software Description Ontology



Extensions:

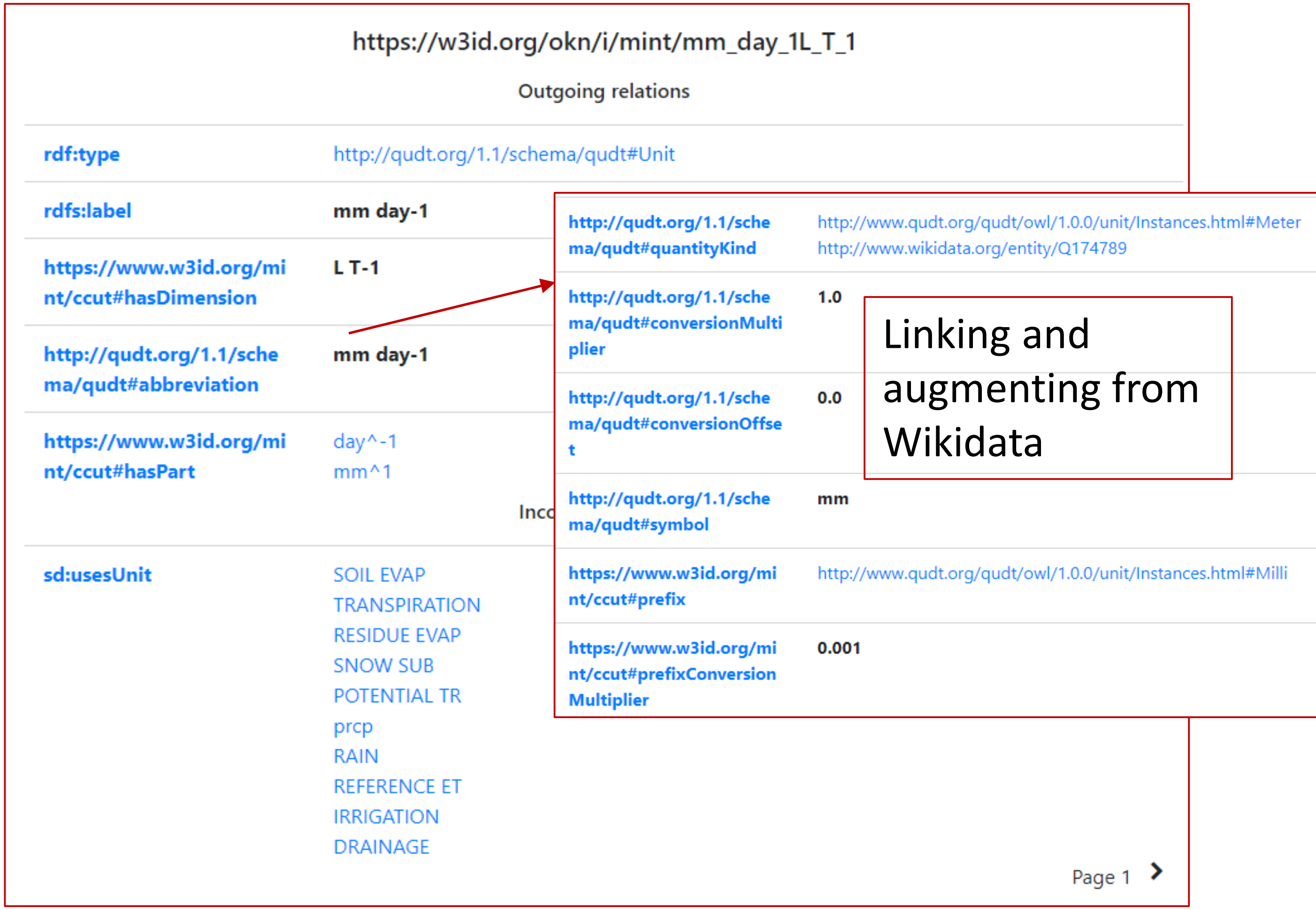
- Schema.org (software metadata)
- W3C Data Cubes (Contents of inputs and outputs)
- NASA QUDT (Units)
- DockerPedia (Software images)
- Scientific Variables Ontology (Standard Variables)

https://w3id.org/okn/o/sd#

Machine readable representation of units

“RAIN”
mm/day

CCUT



https://w3id.org/okn/i/mint/mm_day_1L_T_1

Outgoing relations

rdf:type	http://qudt.org/1.1/schema/qudt#Unit
rdfs:label	mm day-1
https://www.w3id.org/mint/ccut#hasDimension	L T-1
http://qudt.org/1.1/schema/qudt#abbreviation	mm day-1
https://www.w3id.org/mint/ccut#hasPart	day^-1 mm^1

Inco

sd:usesUnit	SOIL EVAP TRANSPIRATION RESIDUE EVAP SNOW SUB POTENTIAL TR prcp RAIN REFERENCE ET IRRIGATION DRAINAGE

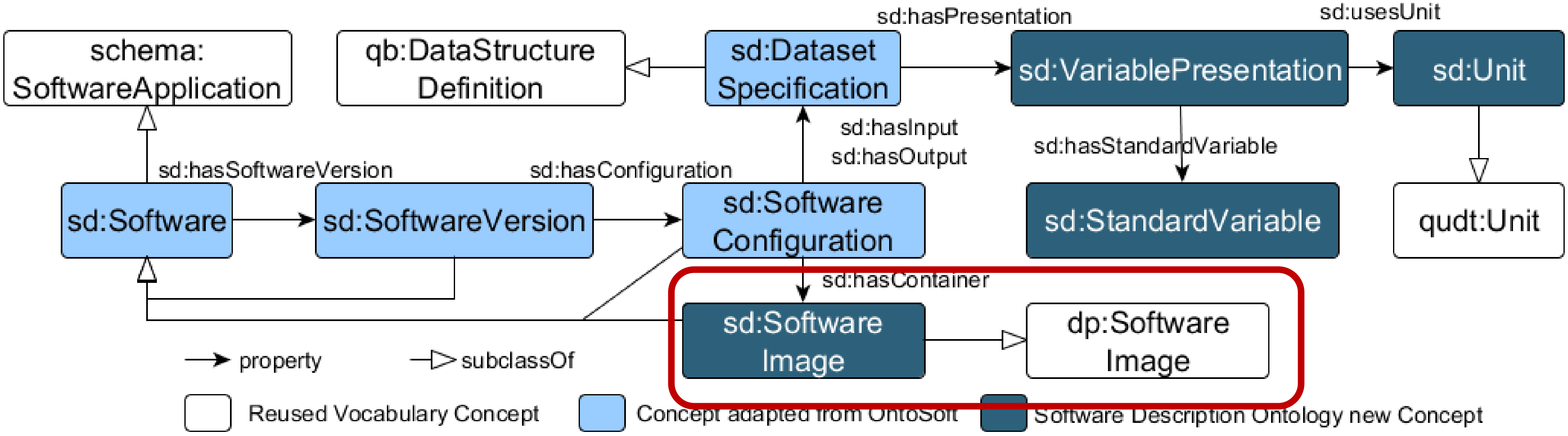
Page 1 ›

http://qudt.org/1.1/schema/qudt#quantityKind	http://www.qudt.org/qudt/owl/1.0.0/unit/Instances.html#Meter http://www.wikidata.org/entity/Q174789
http://qudt.org/1.1/schema/qudt#conversionMultiplier	1.0
http://qudt.org/1.1/schema/qudt#conversionOffset	0.0
http://qudt.org/1.1/schema/qudt#symbol	mm
https://www.w3id.org/mint/ccut#prefix	http://www.qudt.org/qudt/owl/1.0.0/unit/Instances.html#Milli
https://www.w3id.org/mint/ccut#prefixConversionMultiplier	0.001

Linking and augmenting from Wikidata

- B. Shbita, A. Rajendran, J. Pujara, and C. Knoblock, Parsing, Representing and Transforming Units of Measure, in Modeling the World's Systems, 2019.

Evolving OntoSoft: Describing Containers



Extensions:

- Schema.org (software metadata)
- W3C Data Cubes (Contents of inputs and outputs)
- NASA QUDT (Units)
- DockerPedia (Software images)
- Scientific Variables Ontology (Standard Variables)

https://w3id.org/okn/o/sd#

Evolving OntoSoft: Describing Containers

Image tag

“mintproject/ pihm2cycles”

Docker pedia

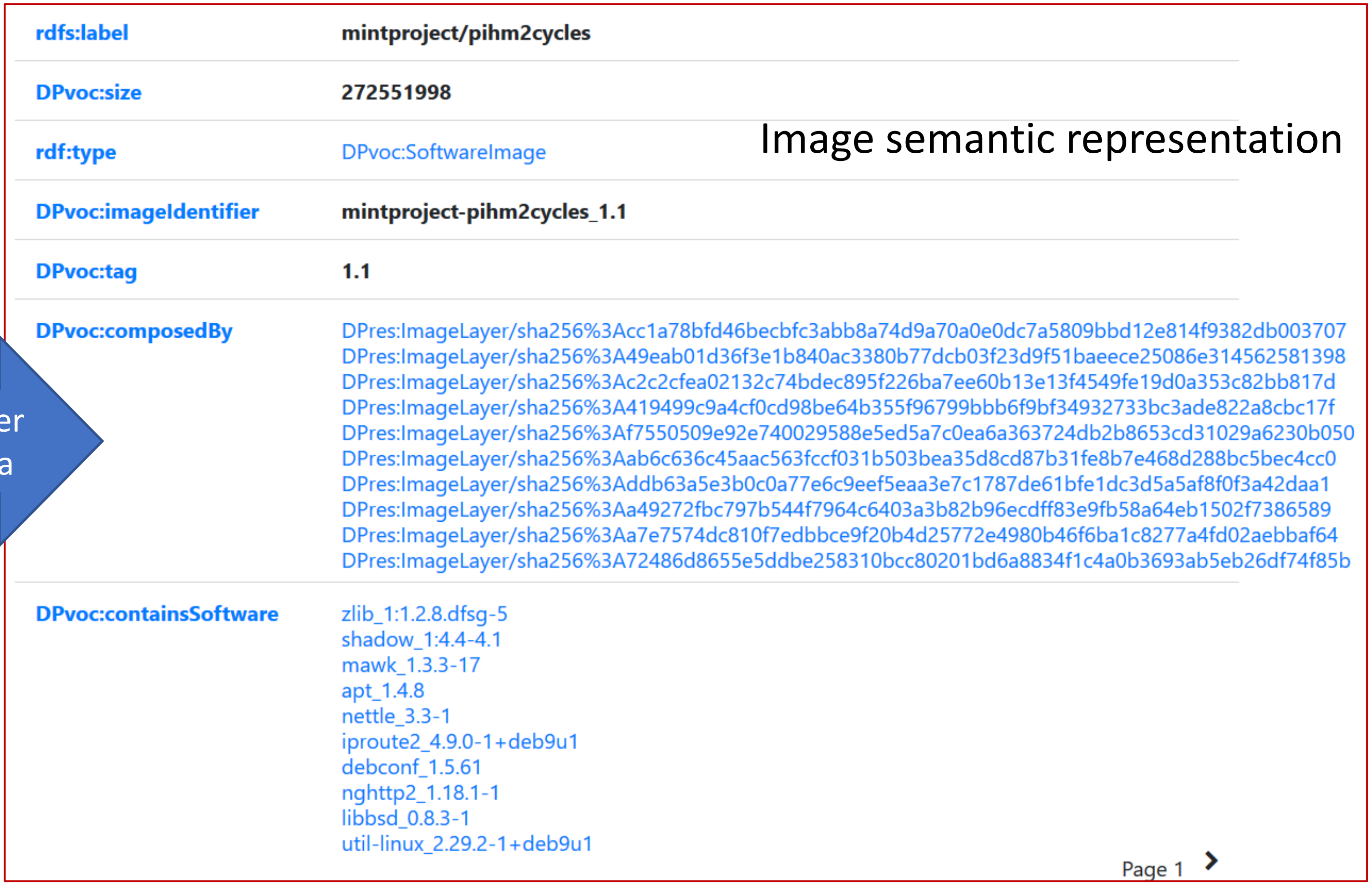


Image semantic representation

rdfs:label	mintproject/pihm2cycles
DPvoc:size	272551998
rdf:type	DPvoc:SoftwareImage
DPvoc:imageIdentifier	mintproject-pihm2cycles_1.1
DPvoc:tag	1.1
DPvoc:composedBy	DPres:ImageLayer/sha256%3Acc1a78bfd46becbfc3abb8a74d9a70a0e0dc7a5809bbd12e814f9382db003707 DPres:ImageLayer/sha256%3A49eab01d36f3e1b840ac3380b77dcb03f23d9f51baeece25086e314562581398 DPres:ImageLayer/sha256%3Ac2c2cfea02132c74bdec895f226ba7ee60b13e13f4549fe19d0a353c82bb817d DPres:ImageLayer/sha256%3A419499c9a4cf0cd98be64b355f96799bbb6f9bf34932733bc3ade822a8cbc17f DPres:ImageLayer/sha256%3Af7550509e92e740029588e5ed5a7c0ea6a363724db2b8653cd31029a6230b050 DPres:ImageLayer/sha256%3Aab6c636c45aac563fccf031b503bea35d8cd87b31fe8b7e468d288bc5bec4cc0 DPres:ImageLayer/sha256%3Addb63a5e3b0c0a77e6c9eef5eaa3e7c1787de61bfe1dc3d5a5af8f0f3a42daa1 DPres:ImageLayer/sha256%3Aa49272fbc797b544f7964c6403a3b82b96ecdff83e9fb58a64eb1502f7386589 DPres:ImageLayer/sha256%3Aa7e7574dc810f7edbbce9f20b4d25772e4980b46f6ba1c8277a4fd02aebbaf64 DPres:ImageLayer/sha256%3A72486d8655e5ddbe258310bcc80201bd6a8834f1c4a0b3693ab5eb26df74f85b
DPvoc:containsSoftware	zlib_1:1.2.8.dfsg-5 shadow_1:4.4-4.1 mawk_1.3.3-17 apt_1.4.8 nettle_3.3-1 iproute2_4.9.0-1+deb9u1 debconf_1.5.61 nghttp2_1.18.1-1 libbsd_0.8.3-1 util-linux_2.29.2-1+deb9u1

https://dockerpedia.inf.utfsm.cl/

- M. Osorio, H. Vargas, and C. Buil Aranda, “DockerPedia: a Knowledge Graph of Docker Images,” in Proceedings of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks co-located with 17th International Semantic Web Conference (ISWC 2018), Monterrey, 2018.

Outline

1. Requirements help scientific software reusability

2. Our current approach for representing scientific software metadata

3. **A framework to query, explore, exploit and publish software metadata**

OKG-SOFT: Framework

Software Model Catalog contains:

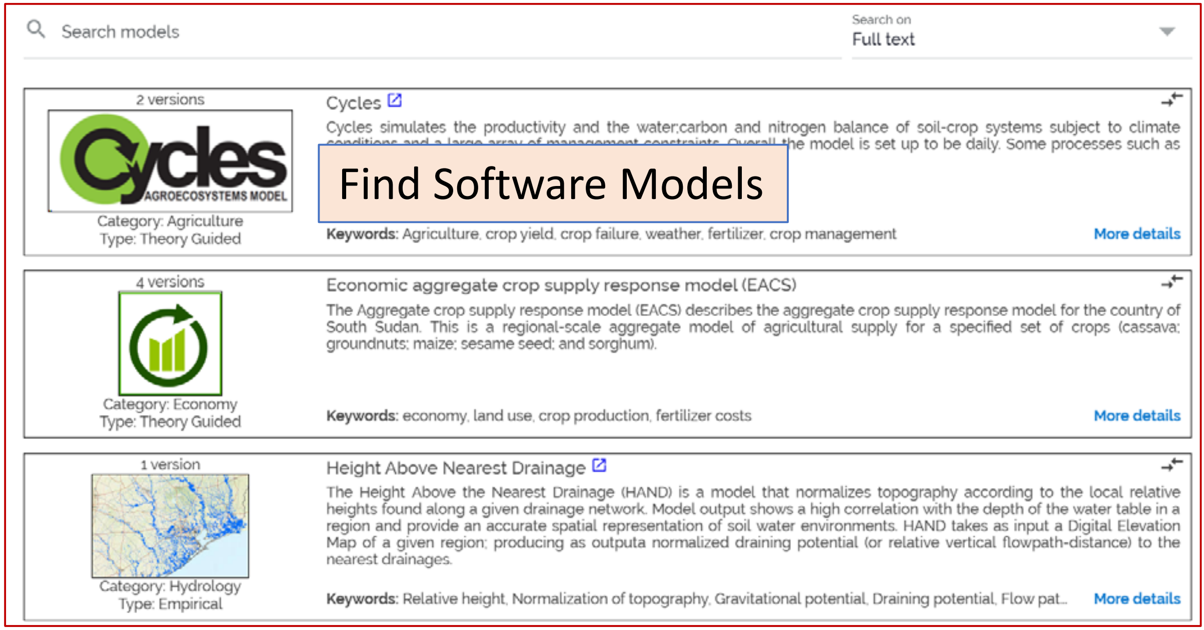
- Models from hydrology, agriculture and economy, their versions and model configurations.
  - More than 200 variables mapped to SVO.
  - All models are executable through scientific workflows
  - Most contents are added manually (expert users) **collaboratively**

- Automated unit transformations
- Automated software image description
- Semi-automated Wikidata linking

APIs:

- SPARQL endpoint
- REST APIs (GET/POST) https://query.mint.isi.edu/api/mintproject/MINT-ModelCatalogQueries#/
- Python clients

Exploitation: Exploring Scientific Software Model Metadata



Search models

Search on
Full text

2 versions

Category: Agriculture
Type: Theory Guided

Cycles

Cycles simulates the productivity and the water;carbon and nitrogen balance of soil-crop systems subject to climate conditions and a large array of management constraints. Overall the model is set up to be daily. Some processes such as

Keywords: Agriculture, crop yield, crop failure, weather, fertilizer, crop management

More details

4 versions

Category: Economy
Type: Theory Guided

Economic aggregate crop supply response model (EACS)

The Aggregate crop supply response model (EACS) describes the aggregate crop supply response model for the country of South Sudan. This is a regional-scale aggregate model of agricultural supply for a specified set of crops (cassava; groundnuts; maize; sesame seed; and sorghum).

Keywords: economy, land use, crop production, fertilizer costs

More details

1 version

Category: Hydrology
Type: Empirical

Height Above Nearest Drainage

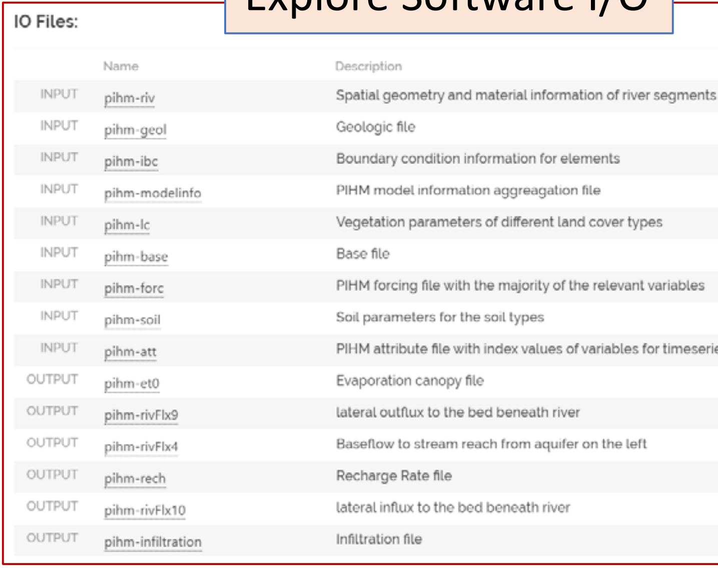
The Height Above the Nearest Drainage (HAND) is a model that normalizes topography according to the local relative heights found along a given drainage network. Model output shows a high correlation with the depth of the water table in a region and provide an accurate spatial representation of soil water environments. HAND takes as input a Digital Elevation Map of a given region; producing as outputa normalized draining potential (or relative vertical flowpath-distance) to the nearest drainages.

Keywords: Relative height, Normalization of topography, Gravitational potential, Draining potential, Flow pat...

More details

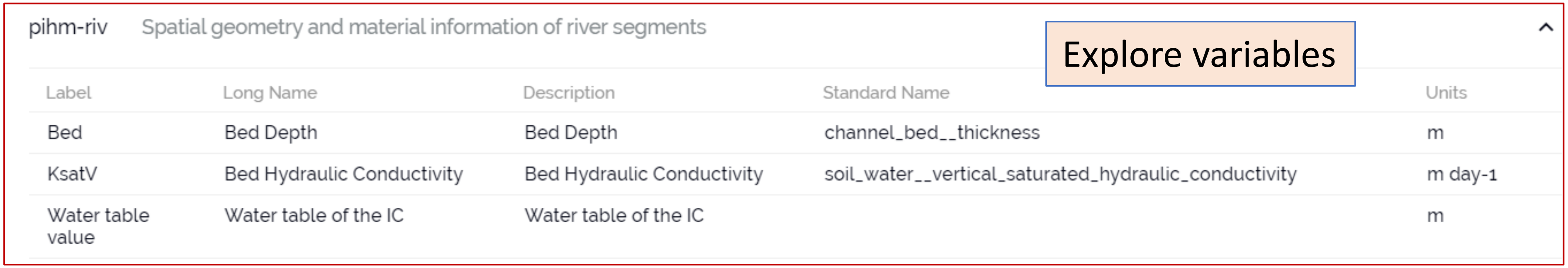
Find Software Models

Explore Software I/O



IO Files:

	Name	Description
INPUT	pihm-riv	Spatial geometry and material information of river segments
INPUT	pihm-geol	Geologic file
INPUT	pihm-ibc	Boundary condition information for elements
INPUT	pihm-modelinfo	PIHM model information aggreagation file
INPUT	pihm-lc	Vegetation parameters of different land cover types
INPUT	pihm-base	Base file
INPUT	pihm-forc	PIHM forcing file with the majority of the relevant variables
INPUT	pihm-soil	Soil parameters for the soil types
INPUT	pihm-att	PIHM attribute file with index values of variables for timeserie
OUTPUT	pihm-et0	Evaporation canopy file
OUTPUT	pihm-rivFlx9	lateral outflux to the bed beneath river
OUTPUT	pihm-rivFlx4	Baseflow to stream reach from aquifer on the left
OUTPUT	pihm-rech	Recharge Rate file
OUTPUT	pihm-rivFlx10	lateral influx to the bed beneath river
OUTPUT	pihm-infiltration	Infiltration file



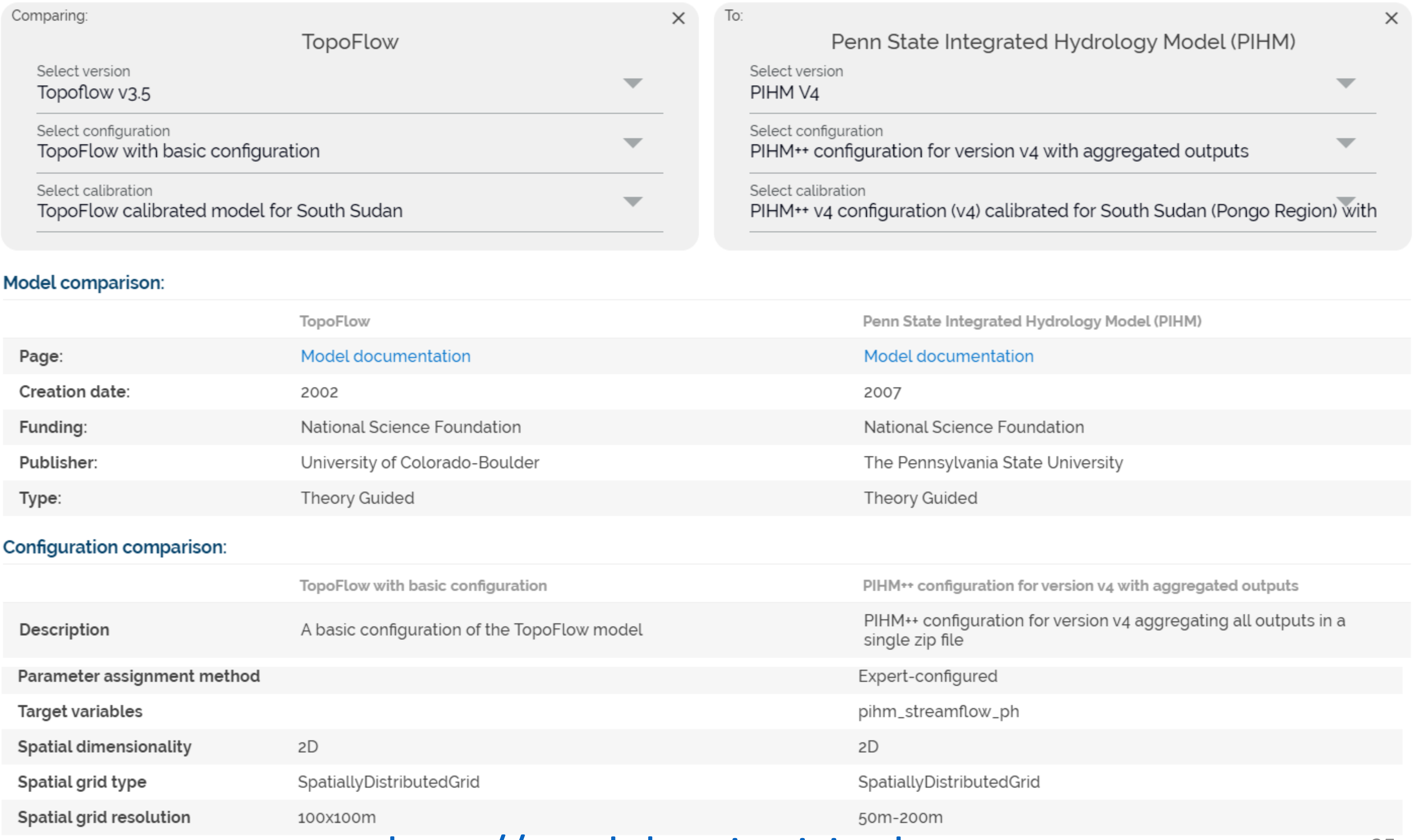
pihm-riv Spatial geometry and material information of river segments

Label	Long Name	Description	Standard Name	Units
Bed	Bed Depth	Bed Depth	channel_bed__thickness	m
KsatV	Bed Hydraulic Conductivity	Bed Hydraulic Conductivity	soil_water__vertical_saturated_hydraulic_conductivity	m day-1
Water table value	Water table of the IC	Water table of the IC		m

Explore variables

http://models.mint.isi.edu

Exploitation: Comparing Scientific Software Models



Comparing:

TopoFlow

Select version
Topoflow v3.5

Select configuration
TopoFlow with basic configuration

Select calibration
TopoFlow calibrated model for South Sudan

To:

Penn State Integrated Hydrology Model (PIHM)

Select version
PIHM V4

Select configuration
PIHM++ configuration for version v4 with aggregated outputs

Select calibration
PIHM++ v4 configuration (v4) calibrated for South Sudan (Pongo Region) with

Model comparison:

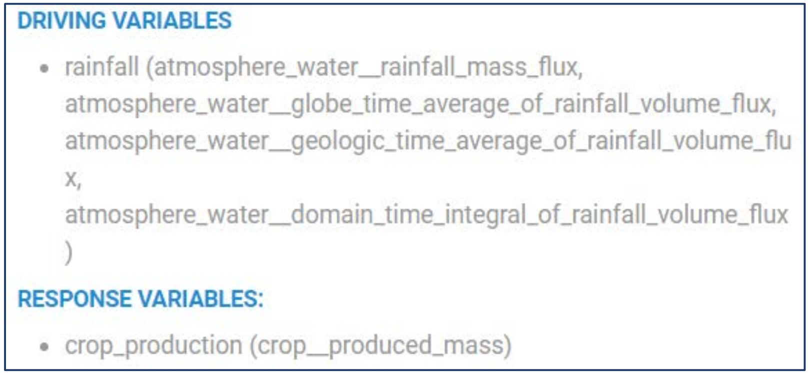
	TopoFlow	Penn State Integrated Hydrology Model (PIHM)
Page:	Model documentation	Model documentation
Creation date:	2002	2007
Funding:	National Science Foundation	National Science Foundation
Publisher:	University of Colorado-Boulder	The Pennsylvania State University
Type:	Theory Guided	Theory Guided

Configuration comparison:

	TopoFlow with basic configuration	PIHM++ configuration for version v4 with aggregated outputs
Description	A basic configuration of the TopoFlow model	PIHM++ configuration for version v4 aggregating all outputs in a single zip file
Parameter assignment method		Expert-configured
Target variables		pihm_streamflow_ph
Spatial dimensionality	2D	2D
Spatial grid type	SpatiallyDistributedGrid	SpatiallyDistributedGrid
Spatial grid resolution	100x100m	50m-200m

http://models.mint.isi.edu

Exploitation: Towards Automated Software Composition

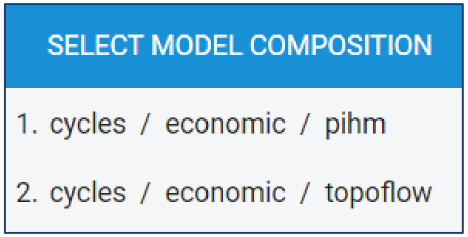


DRIVING VARIABLES

- rainfall (atmosphere_water__rainfall_mass_flux, atmosphere_water__globe_time_average_of_rainfall_volume_flux, atmosphere_water__geologic_time_average_of_rainfall_volume_flux, atmosphere_water__domain_time_integral_of_rainfall_volume_flux)

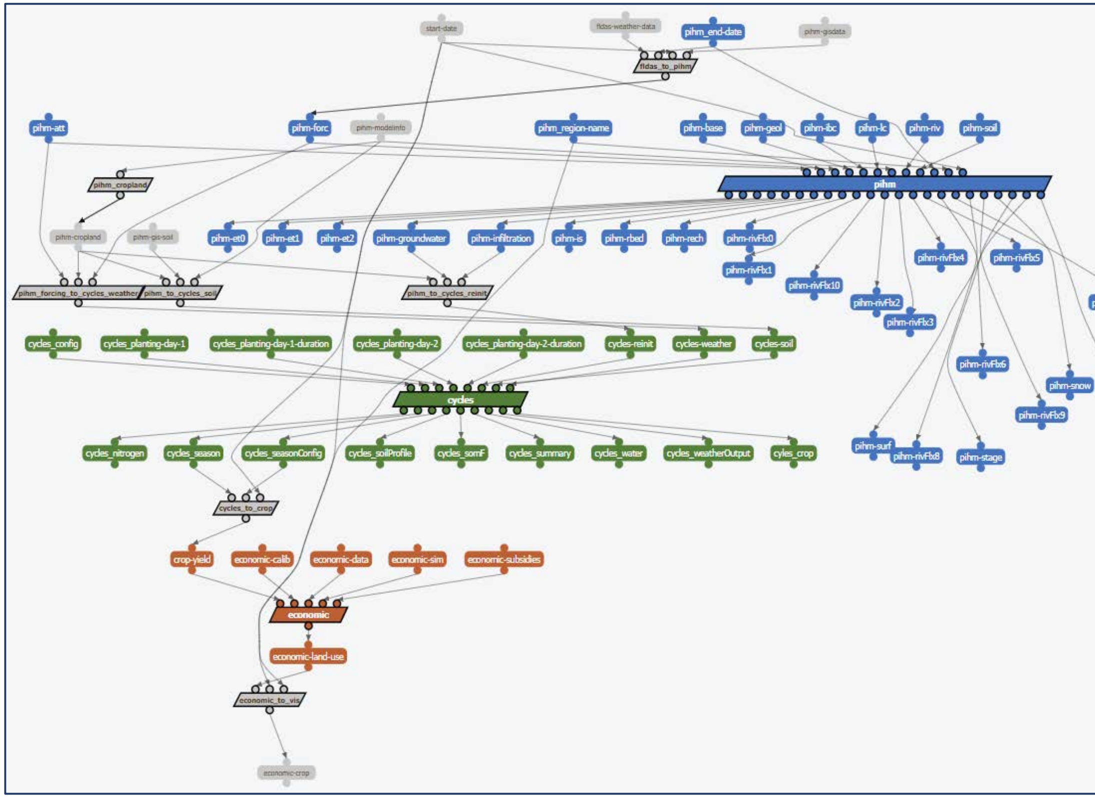
RESPONSE VARIABLES:

- crop_production (crop__produced_mass)



SELECT MODEL COMPOSITION

1. cycles / economic / pihm
2. cycles / economic / topoflow



Summary

Scientific software reusability is crucial to understand

- Existing data
- Published methods

1. Requirements for scientific software reusability include
   - Expose inputs, outputs, variables and software invocation details!

2. Our approach for capturing and structuring scientific software

3. A framework to query, explore, exploit and publish software metadata

Help us making your software more reusable

Contact me: dgarijo@isi.edu

We would like to thank Scott Peckham, Maria Stoica, Chris Duffy, Lele Shu, Kelly Cobourn, Zeya Zhang Suzanne Pierce, Armen Kemanian, Rajiv Mayani, Jay Puajara, Basel Shbita, Dhruv Pattel, Rohit Mayura, Amrish Goel and Anuj Doiphode