



USC Information
Sciences
Institute



FROM SOFTWARE METADATA REGISTRIES TO KNOWLEDGE GRAPHS: ONTOSOFT AND OKG-SOFT

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

University of Southern California,
Information Sciences Institute

@dgarijov

The importance of Scientific Software

Open data



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

Open source software

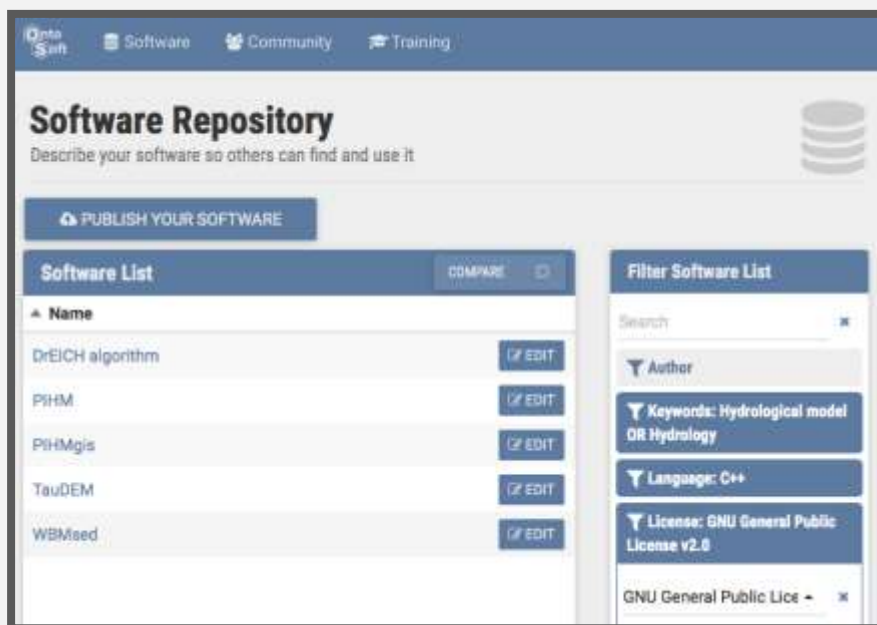


Open publications



Software registries help search, access and understand Scientific Software.

Prior Work: OntoSoft Software Metadata Registry



Finding Software



OntoSoft

Distributed 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>

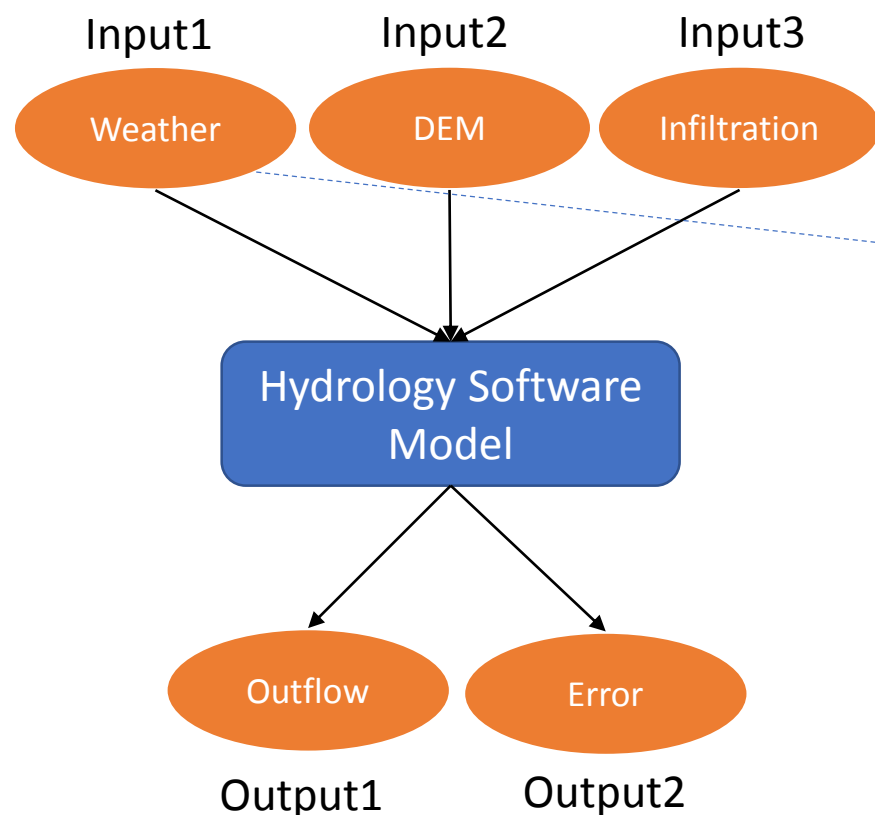
[Gil et al 2015]: OntoSoft: Capturing Scientific Software Metadata Eighth ACM International Conference on Knowledge Capture, Palisades, NY, 2015

Prior Work: OntoSoft Software Metadata Registry

Compare Software				
DrEICH algorithm, PIHM, PIHMgis, TauDEM, WBMsed				
PIHM	PIHMgis	DrEICH	TauDEM	WBMsed
Identify date Unders	Identify date Unders	Identify date Unders	Identify date Unders	Identify date Unders
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/pihm-model/ Test Data Description: Upper Juniata River 875 km ² : see: http://sourceforge.net/projects/pihm-model/		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

Requirements for Software Reusability

1. Exposing software inputs, outputs and their corresponding variables



That is, we assume $c(t, \tau)$ exists but with an unknown functional form, and with certain constraints on the moments. The usual rules of probability apply and we can estimate the moments in t by integrating $c(t, \tau)$ over τ (see Delhez, 1999 or Duffy, 2010):

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

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 $c(t, \tau)$. The 1st to 0th moment is the classical definition of the mean age of the system:

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

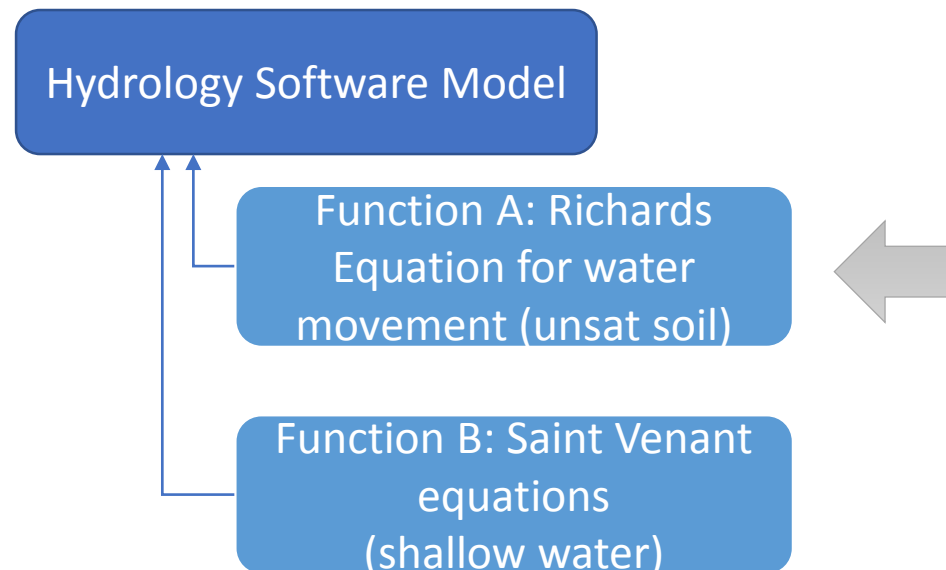
At this point we have defined the tracer as a dynamic variable that depends on the duration that the observed physical time describing the evolution of all tracer particles in the system. Equations (1-3) define the next step is to develop a physical model for the system.

For a single input and single output, we take the volumetric inflow rate to be $Q_i [L^3/T]$, the outflow is initially assumed to be at steady-state ($Q_i = Q_o$). The input tracer C_i 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 m²))

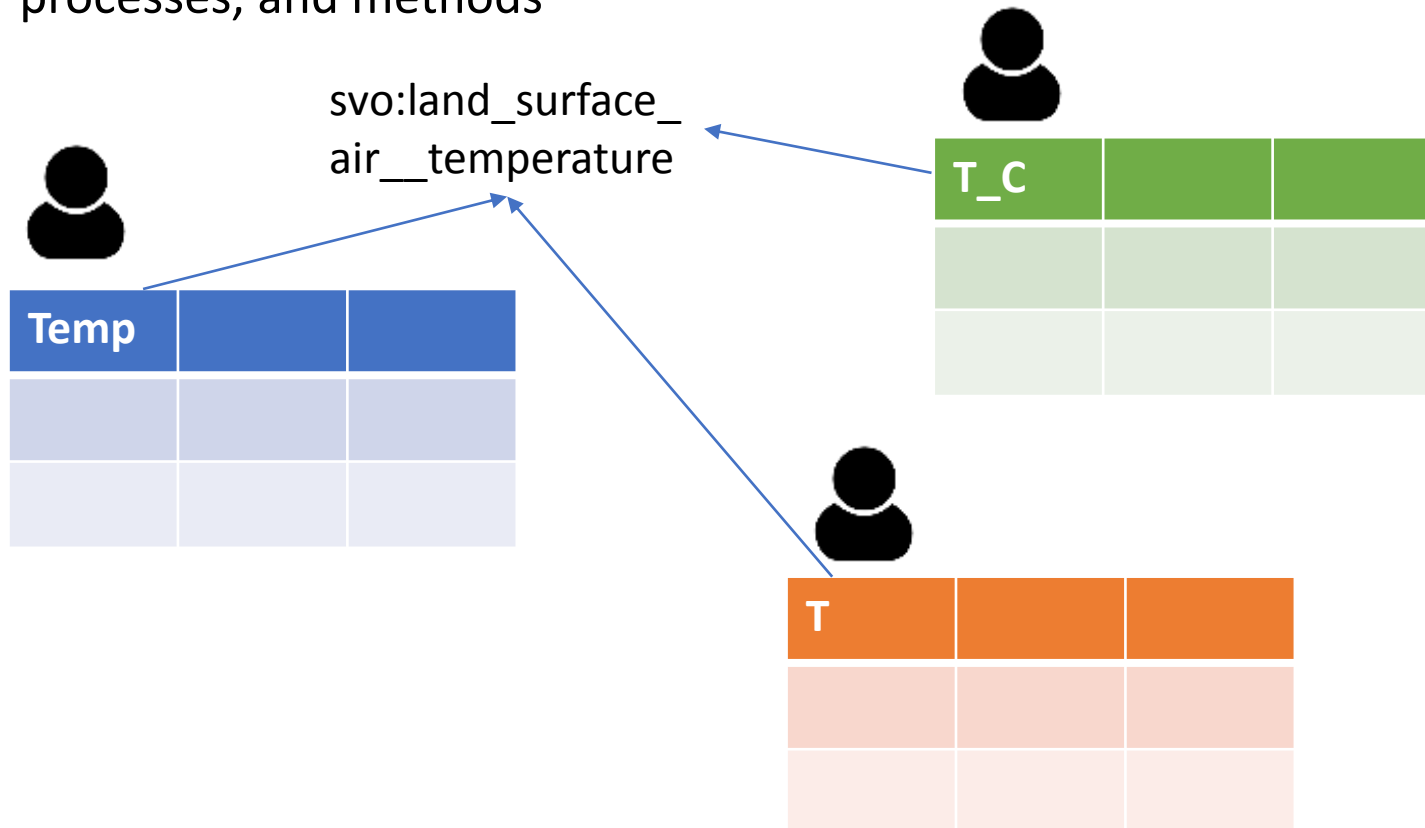
Requirements for Software Reusability

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



Requirements for Software Reusability

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



Requirements for Software Reusability

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

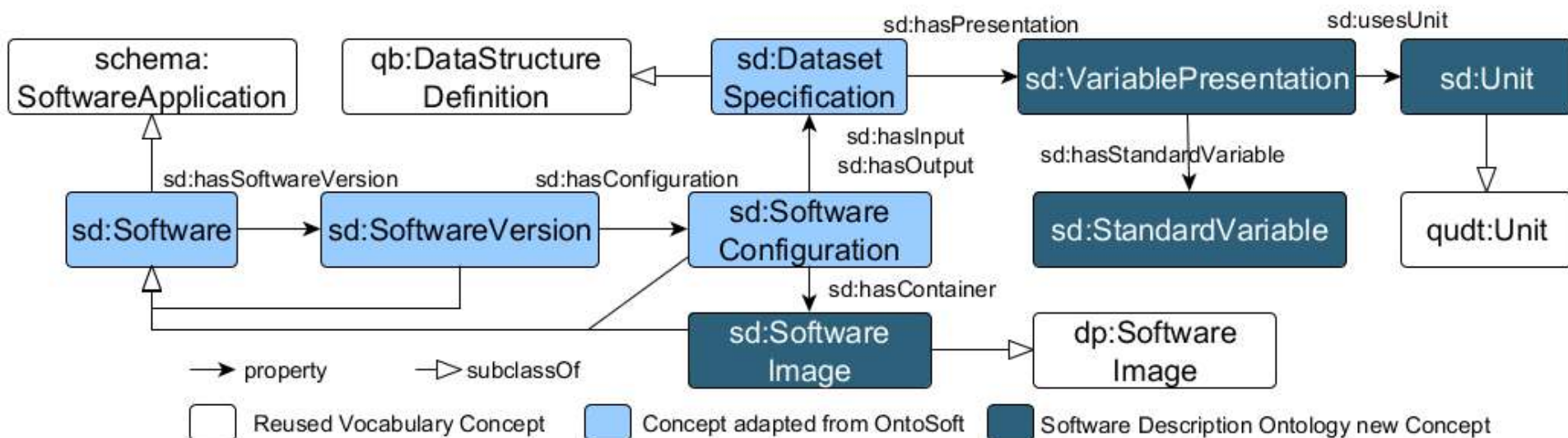
Is data supposed to be in the same folder?

Default arguments/Configuration files?

Volumes?

Do I have to log in in the image

Evolving OntoSoft: Software Description Ontology



Extensions:

- **Schema.org** (software metadata) + **CodeMeta**
- 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#>

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 watercarbon 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 outputs 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

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	PHM model information aggregation file
INPUT	pihm-lc	Vegetation parameters of different land cover types
INPUT	pihm-base	Base file
INPUT	pihm-forc	PHM forcing file with the majority of the relevant variables
INPUT	pihm-soil	Soil parameters for the soil types
INPUT	pihm-att	PHM attribute file with index values of variables for timeseries
OUTPUT	pihm-et0	Evaporation canopy file
OUTPUT	pihm-rivf1d	lateral outflux to the bed beneath river
OUTPUT	pihm-rivf1d4	Baseflow to stream reach from aquifer on the left
OUTPUT	pihm-rech	Recharge Rate file
OUTPUT	pihm-rivf1d10	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

Compare models

<http://models.mint.isi.edu>

Summary

Scientific **Software** is crucial to understand

- Existing **data**
- Published **methods**

Scientific Software Metadata registries help search and understand software

- Enough for **software reusability**?

Requirements for scientific software reusability:

- Describing inputs, outputs, variables and software invocation details

Our approach for **capturing and structuring** scientific software