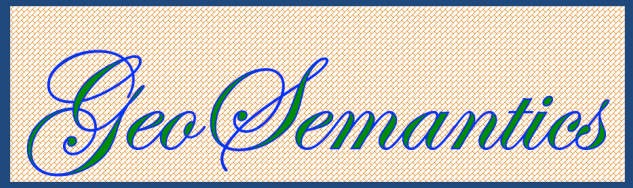
The Geoscience Standard Names Ontology

Scott D. Peckham
Senior Research Scientist at INSTAAR
Lead PI for Earth System Bridge
Former Chief Software Architect for CSDMS
University of Colorado, Boulder
January 10, 2017

EarthCube

EarthCube Projects that have Contributed to the GSN Ontology



EarthCube websites:

Earth S. Bridge: https://www.earthcube.org/group/earth-system-bridge
OntoSoft: https://www.earthcube.org/group/ontosoft
GeoSemantics: https://earthcube.org/group/geosemantics

Main websites:

Earth S. Bridge: https://www.earthsystemcog.org/projects/earthsystembridge/
OntoSoft: http://www.ontosoft.org
GeoSemantics: http://ecgs.ncsa.illinois.edu

The Big Problem: Our Motivation

If you have worked to serve a community of geoscientists, or if you have studied a large number of cross-domain geoscience "use cases", sooner or later you come to realize that:

(1) The big, generic problem facing geoscientists today stems from *lack of interoperability* across a huge number of *heterogeneous resources*.

(2) While discovery and access could certainly be improved (especially for "dark resources"), the real time sink for geoscientists comes when they try to use, understand and connect resources into workflows. Analogy: You shop online, find some pre-fab furniture or vehicle parts and have these shipped to your house. Then the real work begins. Discovery & citation well-served by Dublin Core & DataCite.

(3) The *only practical way* to "tame" this heterogeneity is to do 2 things:
- (a) Collect standardized, "deep-description" metadata for resources, then
- (b) "Wrap" the resources with standardized APIs that provide callers with access to both the data and the metadata. (Adapter Pattern)

Software written to *utilize* these 2 things will be called a mediator or a broker. The only alternative to this, which is completely impractical when the number of different resources is large, is to write separate software to deal with each individual resource. Standardized metadata => ontologies.

Thinking About Variable Names

Variables Underpin Everything We Do

- Variables underpin everything we do in scientific research.
- We measure their values and store them in data sets.

- They appear in science equations that encapsulate our current state of knowledge and show how different variables are related to one another.

Navier–Stokes equations *(general)*

$$\rho\left(\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v}\cdot\nabla\mathbf{v}\right) = -\nabla p + \nabla\cdot\mathbf{T} + \mathbf{f}$$

- All computational models are driven by values of input variables and produce values for output variables, thereby giving us predictive power.
- At one level, variables are symbols associated with concepts that can be quantified with a numerical value that often has units.

Objects and Attributes

In the broad sense, an object is anything in the physical world that we can observe. It could be a body (e.g. rock), a substance or medium (e.g. air or water), a phenomenon or event (e.g. earthquake, flood), or a place (e.g. NYC).

This is also the definition of object used in ISO 80000, the International System of Quantities, sister to the International System of Units (SI units).

Object names are always nouns. As children, we first learn the names of different things, and then attributes and relationships between them.

Objects have observable attributes and may undergo processes.

Attributes can be divided into two distinct types:

quantities = attributes that can be quantified with a number, often with units, such as your weight and height

string-type = attributes that can be stored as a string, such your eye color, name, address, & favorite food

Object – Attribute –Value

This use of Objects, Attributes and Values is an extremely powerful "data model" that underpins object-oriented programming.

It is also called the Entity-Attribute-Value or EAV data model: see
https://en.wikipedia.org/wiki/Entity–attribute–value_model

Note that:

- It is the *values* of variables that are the "exchange items" that we write to and read from data files, store in computer memory and pass between models and

- A variable name associates a symbol to a value.

If we want to construct unique variable names for the purpose of semantic mediation that are unambiguous, human & machine readable and standardized, it therefore makes sense to construct these variable names as unique pairings of object names and quantity names.

Essential Parts of a Variable Name

Variable names therefore have two *essential* parts:

- an object name that identifies an object in the world that we have some interest in understanding, and

- a quantity name that identifies a measurement concept that can be used to quantify that object in some way (e.g. mass, energy, length)

If either part is omitted, there will be ambiguity. For example:

temperature is an ambiguous variable name, because the object for which the temperature was measured was not specified. In a hydrologic model, it could be snow, soil, air, water, etc.

acetic acid is an ambiguous variable name, because it is the name of a substance that can be associated with many possible quantities:

molar mass = 60.06 g / mole

freezing point temperature = 16.6 C (pure, anhydrous, “glacial”)

mass or molar concentration in air or water = ??

It appears in both atmospheric and aquatic chemistry.

"Seldom Heard" ISO 80000

Everyone has heard of the International System of Units or SI Units.

The International System of Quantities, or ISO 80000, is a companion standard that provides the foundations for understanding quantities.

In ISO 80000 (and in the CSDMS Standard Names) an object is defined as anything in the physical world that we can observe. It could be a body (e.g. rock), a substance or medium (e.g. air or water), a phenomenon or event (e.g. earthquake, flood), a place (e.g. NYC).

In ISO 80000 (and in the CSDMS Standard Names) a quantity is defined as a property of an object that can be quantified with a number and optional units.

$$L^a\, M^b\, T^c\, I^d\, \Theta^e\, N^f\, J^g$$

Table 1: Base Quantities of ISO 80000.

Base Quantity (ISO 80000)	**Dimension Symbol**	**SI Unit**
length	**L**	meter
mass	**M**	kilogram
time	**T**	second
electric current	**I**	ampere
thermodynamic temperature	Θ	kelvin
amount of substance	**N**	mole
luminous intensity	**J**	candela

Starting Point: The CSDMS Standard Names

Rules-based, cross-domain, unambiguous, standard names for variables, quantities, processes, assumptions, etc. Needed for "deep description" metadata.

Semantic Matching for Model Variables

Hydro Model A

Output variables:

- streamflow
- rainrate

Hydro Model B

Input variables:

- discharge
- precip_rate

CSDMS Standard Names

- channel_exit_water_x-section__volume_flow_rate
- atmosphere_water__rainfall_volume_flux

Goal: Remove ambiguity so that the framework can automatically match outputs to inputs.

The CSDMS Standard Names

The EAV data model and object-oriented programming use:

Entity/Object + Attribute + Value

CSDMS Standard Names use a similar pattern for creating unambiguous and easily understood standard *variable names or "preferred labels"* according to a set of rules. These are then used to retrieve values and metadata. The pattern is:

Object name + [Operation name] + Quantity name

Simple examples:

atmosphere_carbon-dioxide__partial_pressure
atmosphere_water__rainfall_volume_flux
earth_ellipsoid__equatorial_radius
land_surface__time_derivative_of_elevation
soil__saturated_hydraulic_conductivity

The CSN also includes a large set of standard Assumption & Process Names.

Five Delimiters in CSDMS Standard Names

Double underscore – separates the object and quantity parts
Single underscore – separates distinct words
Hyphen – binds words into single object, e.g. carbon-dioxide
Tilde – separates adjectives from noun in object names
The word "of" – at the end of every operation name

Examples:

sea_water_phosphorous~dissolved~inorganic__time_derivative_of_mole_concentration

atmosphere_air_flow__elevation_angle_of_gradient_of_potential_vorticity

The CSDMS Standard Names

The CSDMS Standard Names can be viewed as a ***lingua franca*** that provides a bridge for mapping variable names between models. They play an important role in the Basic Model Interface (BMI). Model developers are asked to provide a BMI interface that includes a mapping of their model's internal variable names to CSDMS Standard Names and a Model Coupling Metadata (MCM) file that provides model assumptions and other information.

IMPORTANT: Model developers continue to use whatever variable names they want to in their code, but then "map" each of their internal variable names to the appropriate CSDMS standard name in their BMI implementation.

Main Page:	csdms.colorado.edu/wiki/CSDMS_Standard_Names
Basic Rules:	csdms.colorado.edu/wiki/CSN_Basic_Rules
Object Names:	csdms.colorado.edu/wiki/CSN_Object_Templates
Operation Names:	csdms.colorado.edu/wiki/CSN_Operation_Templates
Quantity Names:	csdms.colorado.edu/wiki/CSN_Quantity_Templates
Process Names:	csdms.colorado.edu/wiki/CSN_Process_Names
Assumption Names:	csdms.colorado.edu/wiki/CSN_Assumption_Names
Metadata Names:	csdms.colorado.edu/wiki/CSN_Metadata_Names
Model Metadata Files:	csdms.colorado.edu/wiki/CSN_MMF_Example

The Geoscience Standard Names: A Formal Ontology Based on The CSDMS Standard Names

Taking CSN to the next level: Extending and repackaging the CSN Using Semantic Web Technologies and Best Practices

geostandardnames.org
geoscienceontology.org

Now available as a SPARQL endpoint v(Apache Jena Fuseki)

The 8 Core Entities of the GSN

Variable Names

Variables are the fundamental currency of science. Values of variables are what scientists measure and save in data sets of all kinds. They are the inputs and outputs of predictive models and the items exchanged between coupled models. They also appear in the equations that summarize our scientific knowledge. But what are they? Variables are symbols, names or labels that refer to the *pairing of an object and one of its attributes*.

Object Names

In our context, an *object* is any physical *thing* that we can observe (*body*, *substance*, etc.). We are often interested in a particular part of something larger, or an object contained in another object. For context and alphabetical grouping, it is therefore helpful to use hierarchical *object names*. Objects may have both numerical and string attributes. In the GSN, a word after a tilde '~' in an object name is an *adjective*.

Quantity Names

A *quantity* is an attribute of an object that has a *numerical value*. It will often have measurement units but can also be *dimensionless* (e.g. [m/m]). It may be represented as a scalar, vector or tensor. Many distinct quantities may have the same root quantity, such as *constant*, *exponent* and *angle*. Good quantity names are *object free* and can then be applied to many different objects. For example, *volume flow rate* is preferable to *streamflow*.

Operation Names

When a mathematical operation is applied to a quantity it simply creates a new quantity, often with new units. So quantity names may contain zero, one or a chain of operations. In the GSN, all operation names end in the word **of**. Examples include: *time_derivative_of*, *area_integral_of*, *x_component_of*, *log_of* and *divergence_of*.

The 8 Core Entities of the GSN

Process Names

A *process* is an action that an object can do or that can happen to it. For example, a glacier can advance, calve, melt, sublimate, slide, or deform. Process names are *nouns derived from verbs*. E.g. water can infiltrate into soil, and this process is called *infiltration*.

Numerical Grids

Variables can be associated with a fixed location or can vary in space and time, such as temperature within a room. As appropriate, they may then be treated as scalar, vector or tensor fields. A *grid* is a subdivision or *discretization* of space into *grid cells*. Grids for geospatial variables require geo-referencing with ellipsoids, datums and map projections.

Assumption Names

In the GSN ontology, the term *assumption* is used broadly to refer to any type of *qualifier*, such as a simplification, limitation, convention, exclusion, condition, approximation, clarification or restriction. Scientists refer to assumptions with standard phrases, such as *incompressible flow*. Any of the other 7 entities in the GSN can be tagged and qualified with an assumption.

Science Domain Names

The GSN is currently using the UNESCO Nomenclature for Fields of Science and Technology, which uses SKOS. This is a hierarchical classification of different science and technology domains. These can be used to tag the other 7 entities, as appropriate, so that they can be filtered based on the most relevant science domain.

What Does the GSN Have So Far?

Ocean and Atmosphere Variables
- ROMS Ocean Model (500+ names)
- WRF Atmosphere Model (268 names)
- CF Standard Names (70% of 2600 names)

Hydrologic Variables
- TopoFlow (120+ names)
  - channel flow, snowmelt, evaporation, infiltration, meteorology, ...
- PIHM (80+ names)
- Glaciology and snow hydrology

Sediment Transport Variables
- Landscape evolution models
- Coastline evolution models
- Seafloor, stratigraphic evolution models
- River delta models

Basic Physics Variables
- Projectile motion
- Electricity and magnetism
- Optics & radiometry (in progress)
- Thermodynamics

Environmental Chemistry Variables
- Atmospheric chemistry (CF names)
- Aquatic chemistry from:
  - NWIS Parameter Code Long Names
  - ODM2 / CUAHSI VarName CV

Earth Interior / Deep Earth Process Vars
- Continuum mechanics
- Rheological stress-strain laws
- Seismology and Electromagnetics

Physical and Mathematical Constants
- Large collection

Dimensionless Numbers
- Large collection

Many Empirical Formula Parameters

The GSN currently has close to 14,000 geoscience variable names.

GSN by the Numbers

The Geoscience Standard Names ontology currently has approximately:

11,533	geoscience variable names
4,723	object names (more, w/ adjectives)
1,501	quantity names
1,300	process names
1,056	assumption names (in 25 categories)
130	operation names

But we are still in the process of adding variable names from the mapping of the CF Standard Names and the ROMS and WRF models.

Example: Fluxes, Flow Rates, Etc.

In physics, there are 7 main "root quantities" that are conserved, and these are used across the geosciences in models and data sets. They are:

charge [C], energy [J], mass [kg], moles [mol],
momentum [kg m s-1], number [1] and volume [m3].

Let **X** be any of these, with units U. We then have associated quantities:

X_flux	[U m-2 s-1]	Vector
X_flow_rate	[U s-1]	Scalar
X_concentration	[U m-3]	Scalar
X_fraction	[U/U]	Scalar
X_ratio	[U/U]	Scalar
X_diffusivity	[m2 s-1]	Scalar

Note: charge_flux = electric current density, charge flow rate = electric current, charge concentration = volume charge density

divergence_of_X_flux	[U m-3 s-1]	Scalar
z_integral_of_X_flux	[U m-1 s-1]	Vector ("unit-width")
gradient_of_X_concentration	[U m-4]	Vector
z_integral_of_X_concentration	[U m-2]	Scalar ("content")

Note: X_flow_rate = area_integral_of_normal_component_of_X_flux
Note: The "volume_flux" of an incompressible 3D fluid flow = its 3D velocity field.

The Importance of Operations

Many of the quantities used by geoscientists are generated by applying some sort of mathematical or other operation to an existing quantity to create a new quantity.

In English, we almost always insert the verbal delimiter of after these operations, which can be chained together. So the GSN uses the word of as its delimiter for operations. The *GSN has a large collection of operations*, such as:

time_derivative_of	(adds “per time” units, [T-1])
area_integral_of	(adds “area” units, [L2])
x_component_of	(does not affect units)
log_of	(has log of original units, [log(U)])
divergence_of	(adds “per length” units, [L-1])

Some operations only apply to a specific “field type” (i.e. Scalar, Vector or Tensor) and this is captured in the GSN ontology.

divergence_of	(applies to: Vector, returns: Scalar)
x_component_of	(applies to: Vector, returns: Scalar)
azimuth_angle_of	(applies to: Vector, returns: Scalar)
gradient_of	(applies to: Scalar, returns: Vector)
laplacian_of	(applies to: Scalar, returns: Scalar)
curl_of	(applies to: Vector, returns: Vector)
x_y_component_of	(applies to: Tensor, returns: Scalar)

Our New MCM App for Collecting Model Metadata that Uses the GSN

Model Component Metadata (MCM) App (v0)

Built with Ionic 2 and Angular 2

Learned about Ionic Framework for Mobile App Development from another EarthCube project called Flyover Country at GSA Meeting.

Ionic 2 is a high-productivity development framework based on Google's Angular 2. Well over a million apps have been developed with Ionic.

Apps run in a desktop browser, tablet or smart phone (iOS and Android)

Each page in the app has three source code files:
(1) JavaScript / TypeScript, (2) HTML and (3) CSS / SCSS.

Near-native performance: Angular 2 is 5 times faster than Angular 1.

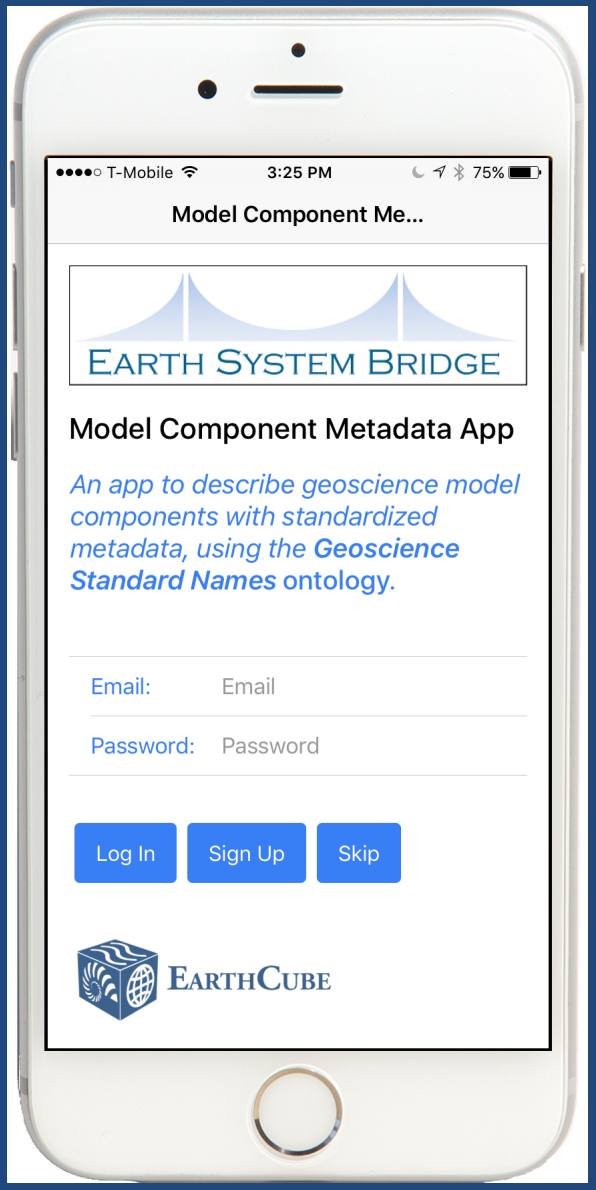
MCM App Communicates with the GSN Server

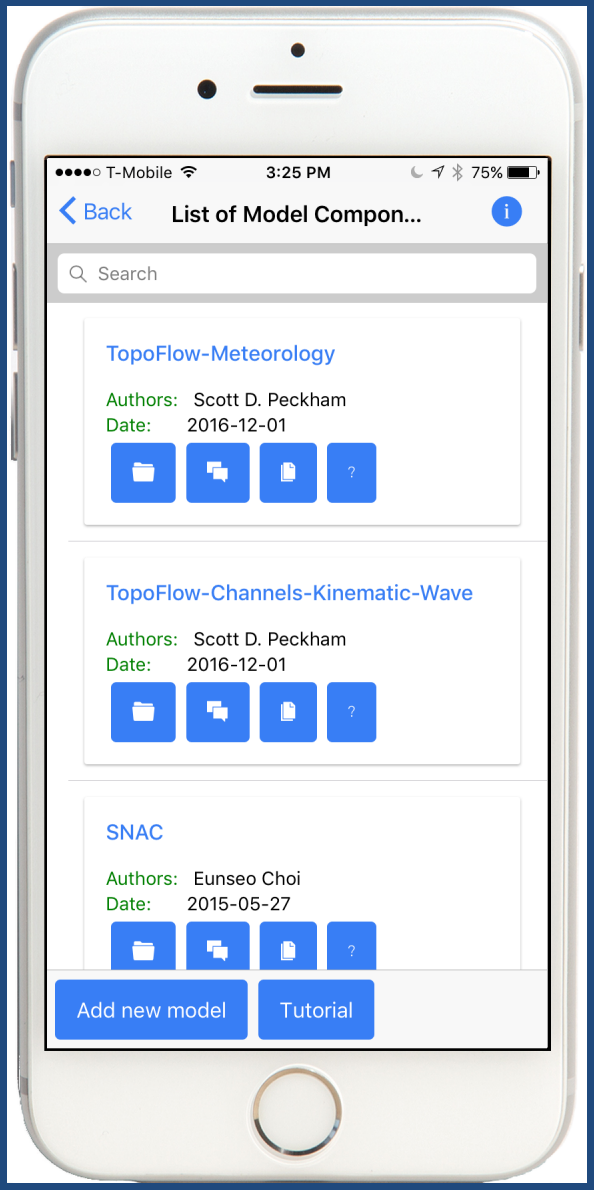
Two-way communication with our server via the "MEAN Stack":
MEAN = MongoDB, Express, Angular 2 (Ionic 2), and Node.js.

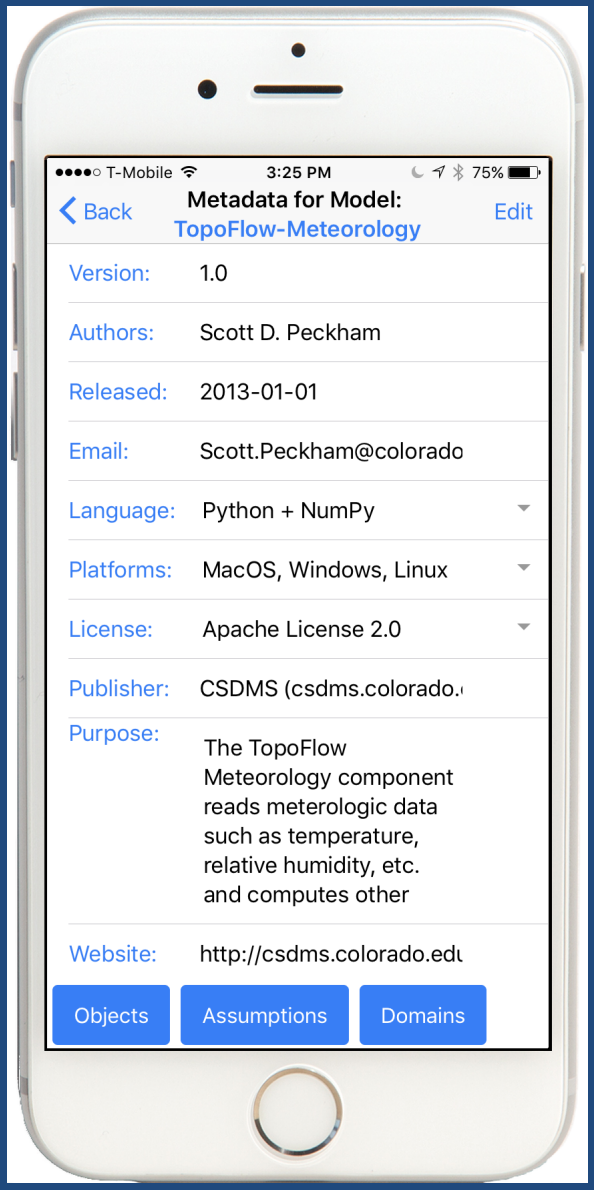
Uses an InAppBrowser to display Wikipedia pages (community-based help).

Uses role-based authentication for different app users.

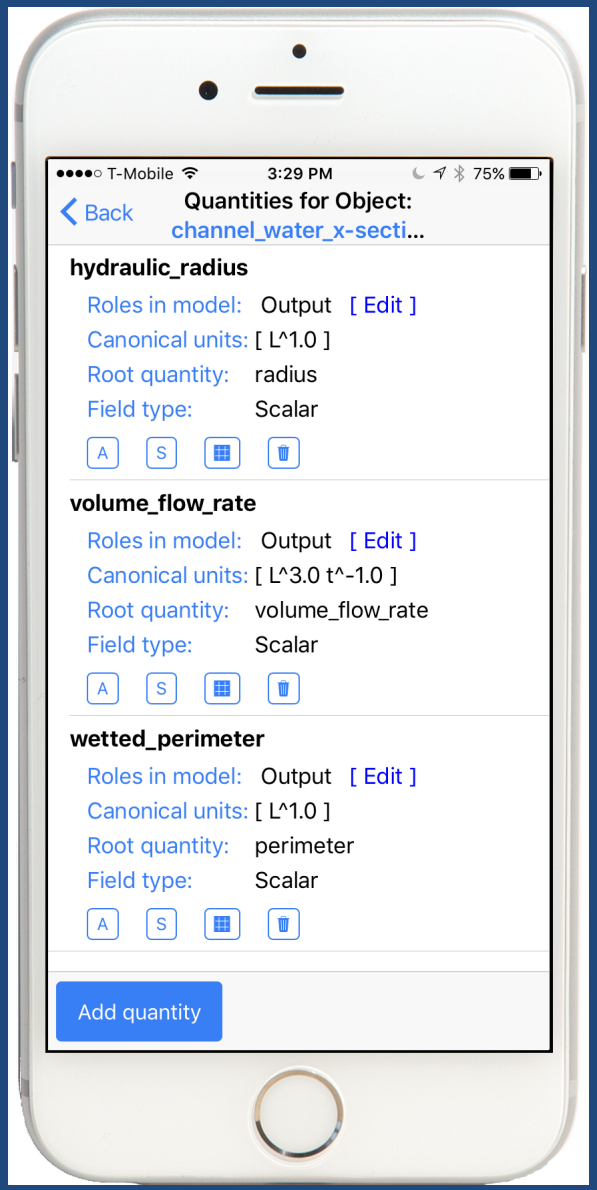
Model Component Metadata (MCM) App (v0)

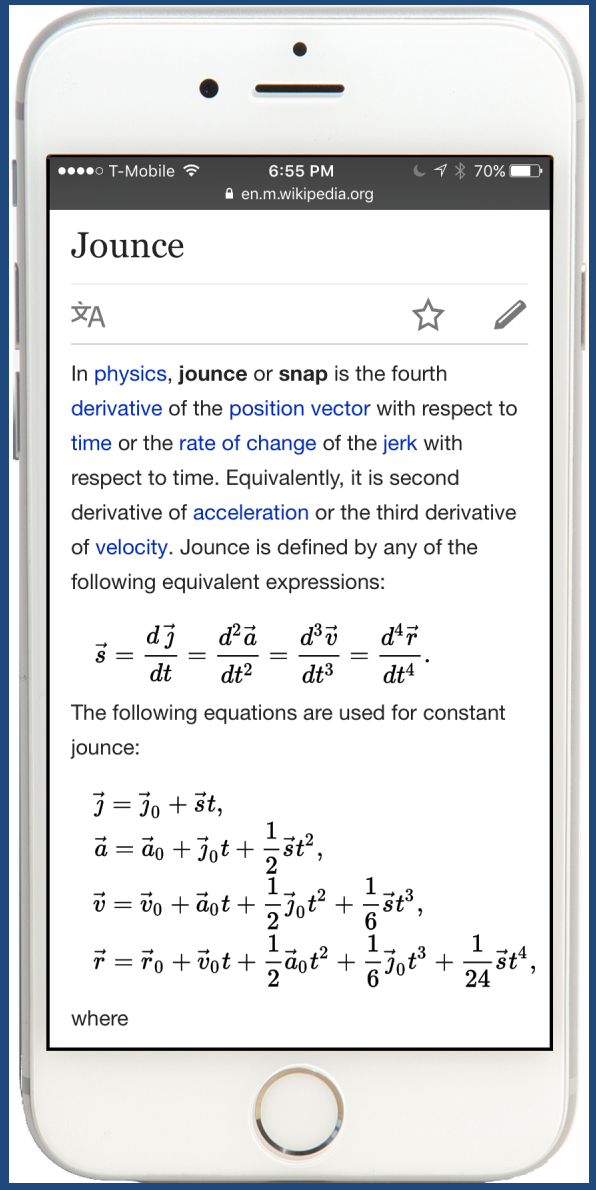


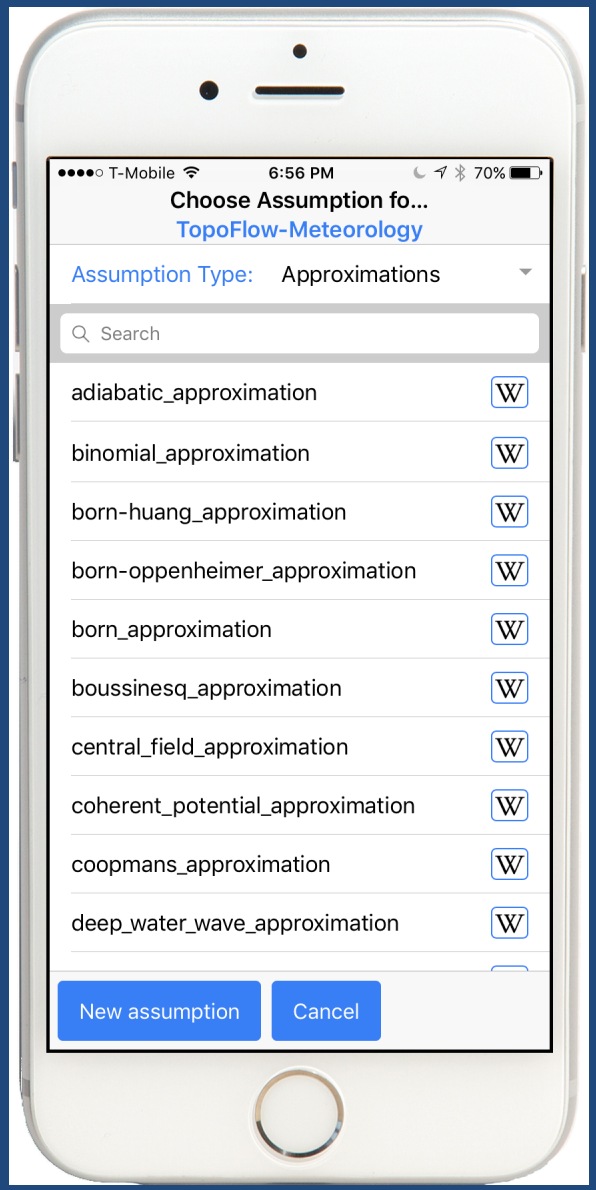




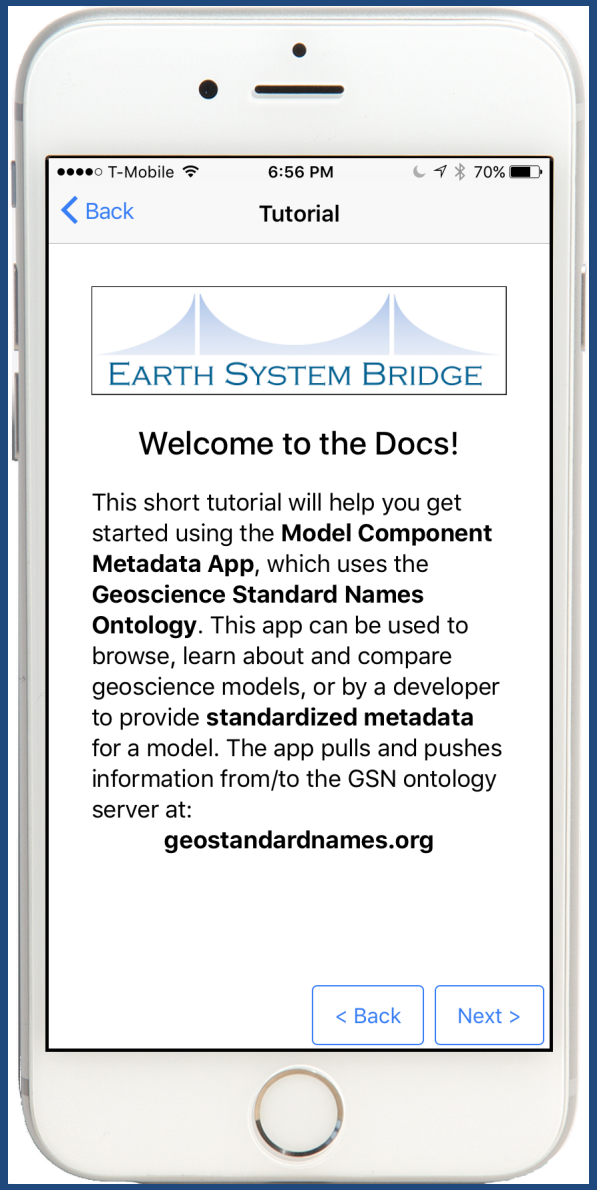
Model Component Metadata (MCM) App (v0)

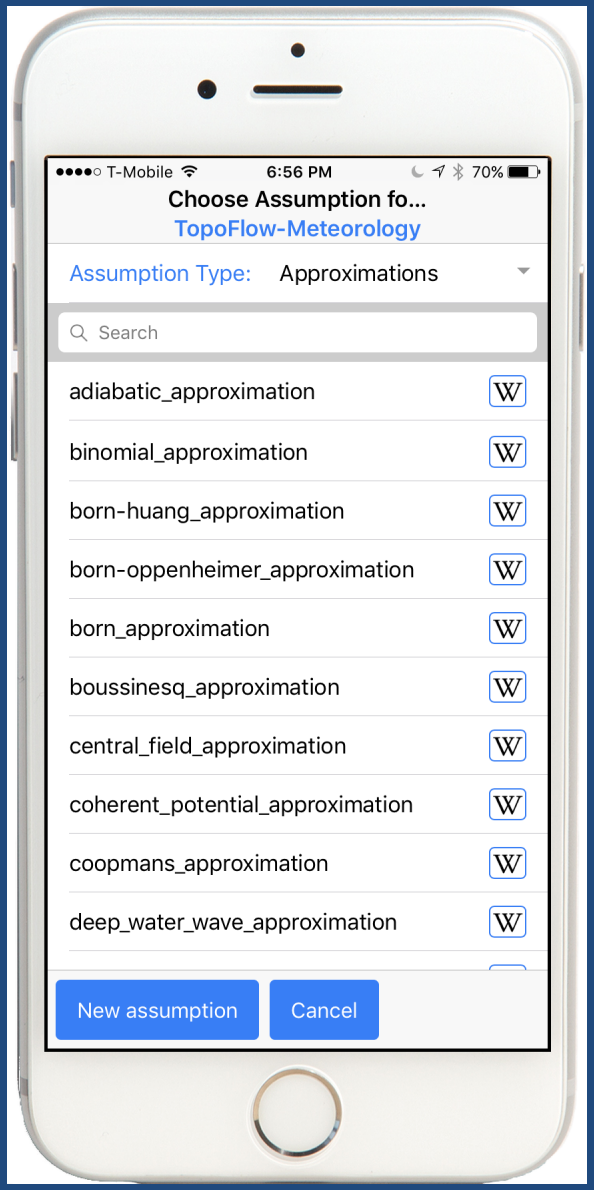


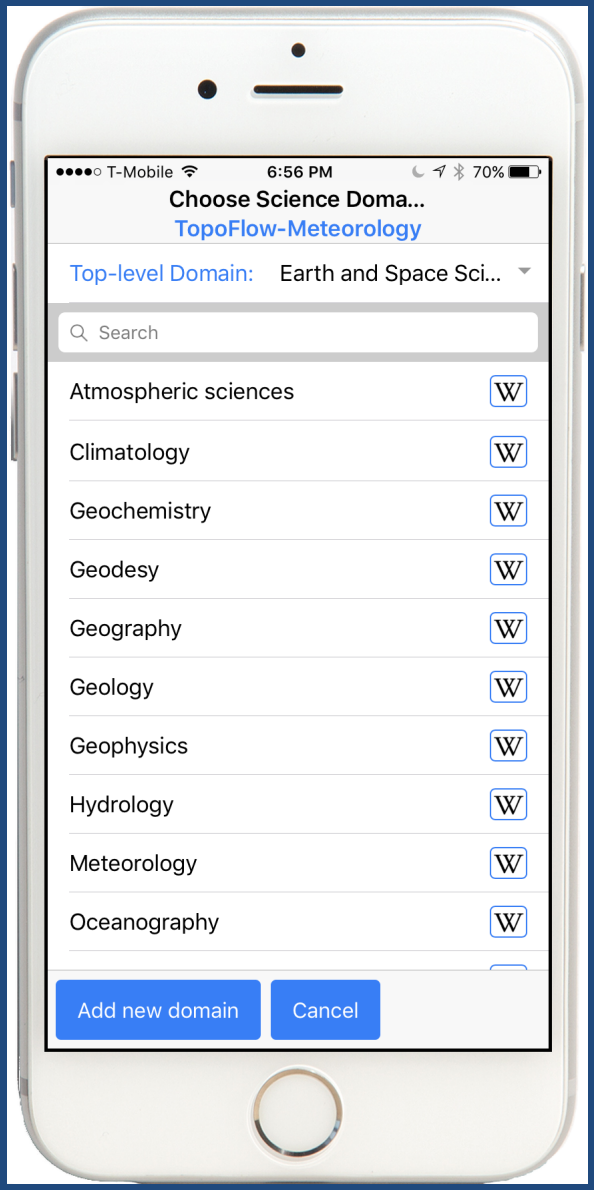




Model Component Metadata (MCM) App (v0)







GSN Ontology and MCM App

Upper Ontology

CSN concept definitions as types/classes, with their predicates/properties (RDFS?)

(e.g. Root Object, Root Quantity, Object, Quantity, Operation, Variable, Process and Assumption)

Lower Ontology

- Object Names List
- Variable Names List
- Quantity Names List
- Process Names List
- Operation Names List
- Assumption Names List

Entity Relationships

- Object Q, P and A
- Quantity Op, P and A

SKOS Crosswalks

- CF Standard Names
- USGS P-code Names

Model Metadata from MCM App

- Model 1 Metadata
- Model 2 Metadata
- Model 3 Metadata

Each has model-specific choices & assumptions, but the ontology is model & data agnostic.

Holding Tank for Newly Proposed Names, Name Associations & Changes

Changes are vetted. Additions integrated continuously. Other changes wait for next release.

Every blue box is a separate RDF file with assertions as S-P-O triples and may import others (TTL).

Closing Thoughts

Minimal Governance by Design: Rules-based, Assisted Name Construction

We learned from the CF Standard Names effort that with only guidelines and no rigorous set of rules for constructing names, the vetting of proposed, new names was a tedious and time-consuming process, requiring a lot of volunteer/committee work and near-endless email discussion. This led to:

(1) restricting the scope to only ocean and atmosphere model names

(2) long delays between when new names were proposed and adopted.

(3) internal inconsistencies or self-contradictions.

Our approach is based on a close examination of the variable names that are currently used in the most sophisticated computational models, a study of prior, related projects such as the CF Standard Names and the NWIS Parameter Code Dictionary Long Names. This led to the identification of common patterns that cut across science domains, so that in most cases new names can be constructed from existing templates.

Can "Deep Description" Ontologies be Created Through Automation?

The short answer is *sort of* or *not really* or *it depends on how much precision is required for the application*.

Good ontologies are very precise things (schemas) that organize knowledge in a manner that is both human-readable and machine-actionable. They make it possible to distinguish between concepts that are similar, related or equivalent.

While it is possible to mine existing controlled vocabularies and online resources to piece together an ontology for describing geoscience resources, the result can only be as good as the best resources that already exist and can be pulled in.

Here is a simple test case to illustrate this point. Note that SWEET contains the concept "heat capacity". Unfortunately, in thermodynamics this broad concept leads to a number of distinct concepts and corresponding variable names that must be resolvable for model-model and model-data coupling. A few of these are:

C_p = isobaric heat capacity
C_v = isochoric heat capacity
c_p = mass-specific, isobaric heat capacity
c_v = mass-specific, isochoric heat capacity

Note: There are many, many examples like this one from across the geosciences.

Thank You!

If you'd like to be a *beta tester* for our new app when it is officially released, please email me at:

Scott.Peckham@colorado.edu

For More Information

- Peckham, S.D., E.W.H. Hutton and B. Norris (2013) A component-based approach to integrated modeling in the geosciences: The Design of CSDMS, *Computers & Geosciences*, special issue: Modeling for Environmental Change, 53, 3-12 .
  http://dx.doi.org/10.1016/j.cageo.2012.04.002.

- Peckham, S.D. (2014) The CSDMS Standard Names: Cross-domain naming conventions for describing process models, data sets and their associated variables, *Proceedings of the 7th Intl. Congress on Env. Modelling and Software*, International Environmental Modelling and Software Society (iEMSs), San Diego, CA. (Eds. D.P. Ames, N.W.T. Quinn, A.E. Rizzoli).
  http://www.iemss.org/sites/iemss2014/proceedings.php

For More Information, cont'd

- Peckham, S.D. (2014) EMELI 1.0: An experimental smart modeling framework for automatic coupling of self-describing models, *Proceedings of HIC 2014*, 11th International Conf. on Hydroinformatics, New York, NY.

  http://academicworks.cuny.edu/cc_conf_hic/464/

- Peckham, S.D., A. Kelbert, M.C. Hill and E.W.H. Hutton (2016) Towards uncertainty quantification and parameter estimation for Earth system models in a component-based modeling framework, *Computers & Geosciences*, special issue: Uncertainty and Sensitivity in Surface Dynamics Modeling, 90(B), 152-161 .

  http://dx.doi.org/10.1016/j.cageo.2016.03.005

For More Information, cont'd

- Peckham, S.D. and J.L. Goodall (2013) Driving plug-and-play models with data from web-services: A demonstration of interoperability between CSDMS and CUAHSI-HIS, *Computers & Geosciences*, special issue: Modeling for Environmental Change, 53, 154-161, http://dx.doi.org/10.1016/j.cageo.2012.04.019

- Laniak, G.F., G. Olchin, J. Goodall, A. Voinov, M. Hill, P. Glynn, G. Whelan, G. Geller, N. Quinn, M. Blind, S. Peckham, S. Reaney, N. Gaber, R. Kennedy and A. Hughes (2013) Integrated environmental modeling: A vision and roadmap for the future, 39, 3-23, *Environmental Modeling & Software*, http://dx.doi.org/10.1016/j.envsoft.2012.09.006

Assumption Names

https://csdms.colorado.edu/wiki/CSN_Assumption_Names

Standard Assumption Names

Assumptions --- interpreted broadly to include:

conditions, *simplifications*, *approximations*, *limitations*, *conventions*, *provisos*, *exclusions*, *restrictions*, etc.

--- are *not included* in CSDMS Standard Variable Names.

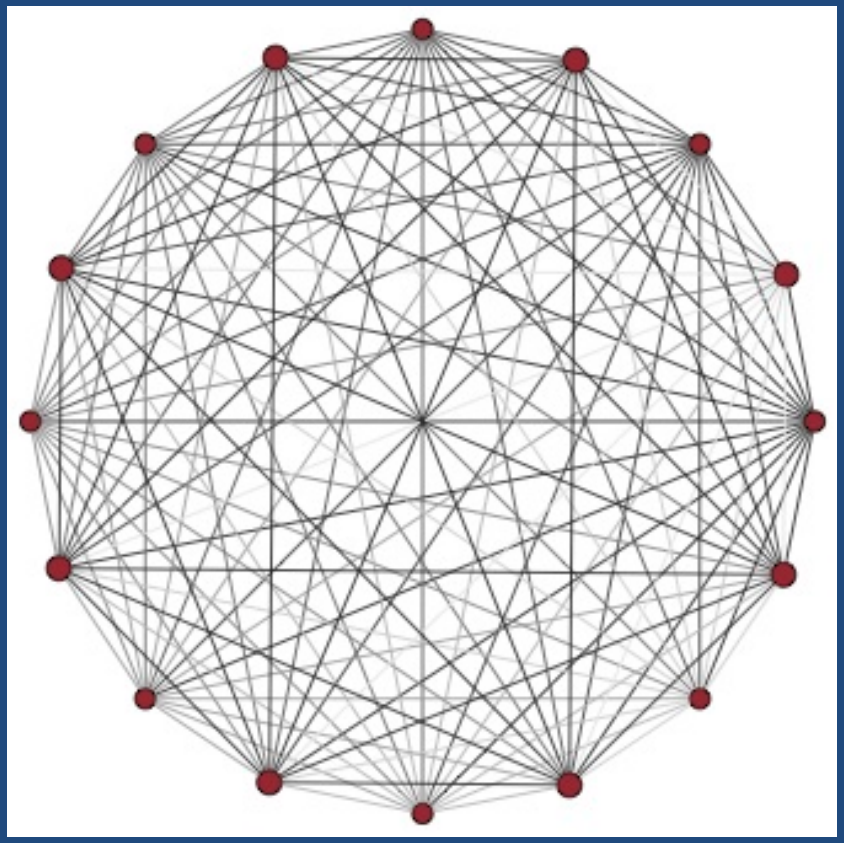
Instead, developers are encouraged to use multiple *<assume> tags* in a *Model Coupling Metadata (MCM)* files to clarify how they are using a CSDMS Standard Name within their model. (Read once at start.)

In order for a *Modeling Framework* to be able to *compare the assumptions* made by different models (about the model or its variables), *standard assumption names* are needed, in addition to the *standard variable names*.

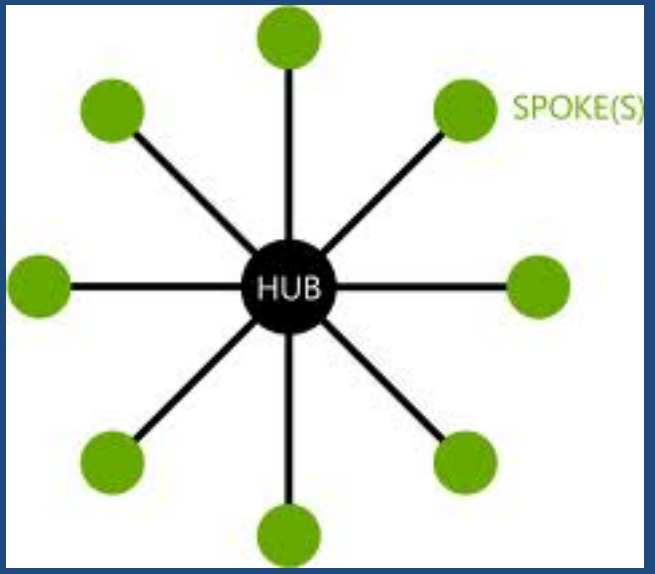
Standard Assumption Names

Assumption Type:	***Example***
Boundary conditions:	no_slip_*boundary_condition*
Conserved quantities:	momentum_*conserved*
Coordinate system:	cartesian_*coordinate_system*
Angle conventions:	clockwise_from_north_*convention*
Dimensionality:	2_*dimensional*
Equations used:	navier_stokes_*equation*
Closures:	eddy_viscosity_turbulence_*closure*
Flow-type assumptions:	laminar_*flow*
Fluid-type assumptions:	herschel_bulkley_*fluid*
Geometry assumptions:	trapezoid_*shaped*
Named model assumptions:	green_ampt_infiltration_*model*
Thermodynamic processes:	isenthalpic_*process*
Approximations:	boussinesq_approximation
Averaging methods:	reynolds_averaged
Numerical methods used:	arakawa_c_grid
State of matter:	liquid_*phase*

Reconciling Differences with Standards



vs.



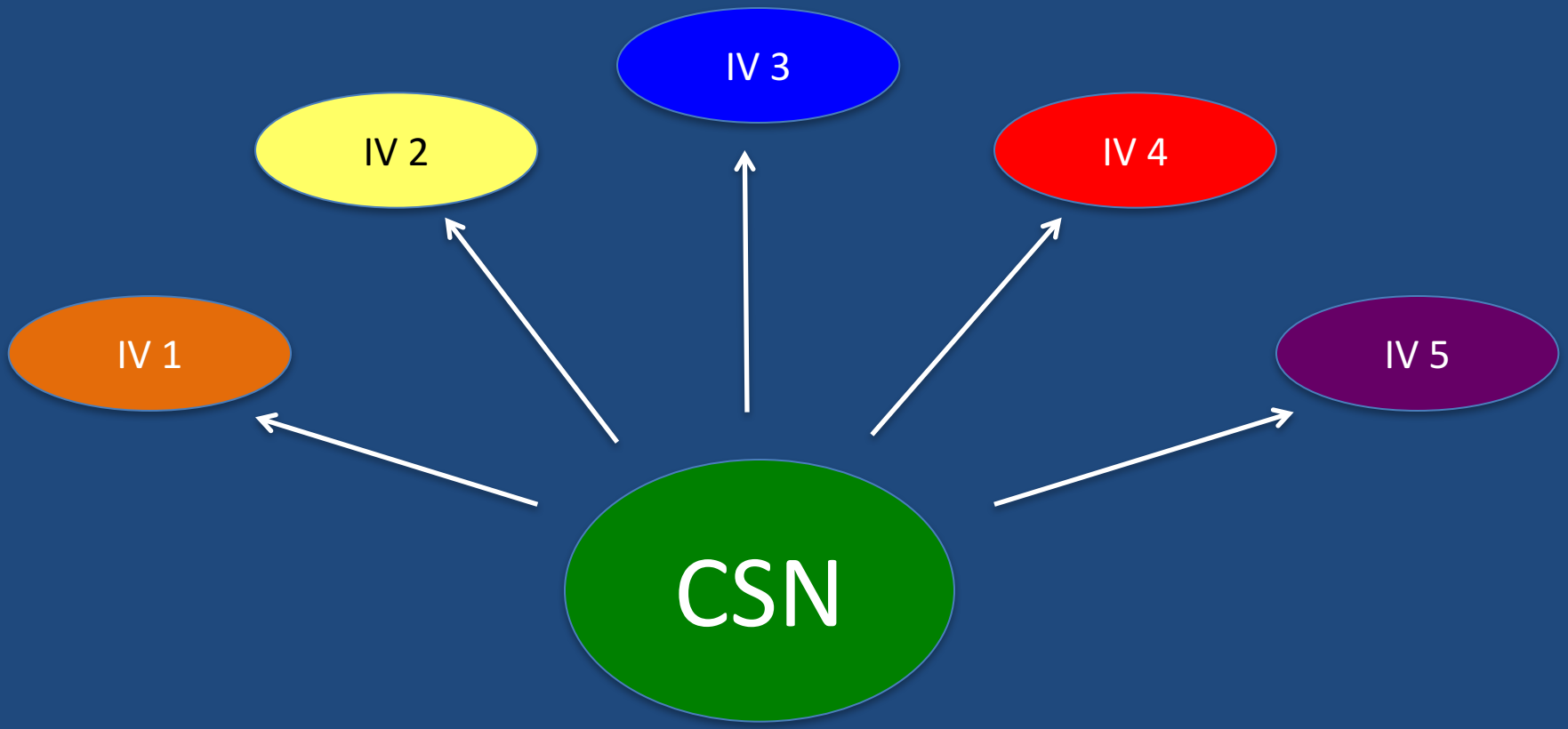
If we reconcile differences between the resources in a pairwise manner, the amount of work, etc. grows fast:

Cost(N) = N (N-1) / 2 ~ N^2.

Introduce a new, generic or standard representation (the "hub"), then map resources to and from it. The amount of work, maintenance, etc. drops to:

Cost(N) = N.

CSDMS Standard Names (CSN)



The *semantic mediation problem* can be solved by mapping resource *internal vocabularies* (IV) to an *expressive*, central vocabulary.

EARTH SYSTEM BRIDGE

CSDMS COMMUNITY SURFACE DYNAMICS MODELING SYSTEM

CIG

CUAHSI UNIVERSITIES ALLIED FOR WATER RESEARCH

ESMF

NOAA

Boulder

GEORGIA TECH

OSU Oregon State UNIVERSITY

THE UNIVERSITY OF MEMPHIS

NUOPC

NCAR


Scott Peckham
CU Boulder


Cecelia DeLuca
CU Boulder/NOAA


David Gochis
UCAR


Anna Kelbert
OSU


Eunseo Choi
Univ. of Memphis


Rocky Dunlap
CU Boulder/NOAA


Rick Hooper
CUAHSI

My Other NSF EarthCube Projects

GeoSemantics Project (Lead PI: Praveen Kumar)

A decentralized framework that combines Linked Data technology and RESTful web services to annotate, connect, integrate, and reason about integration of geoscience resources. This enables the semantic enrichment of web resources and semantic mediation among heterogeneous geoscience resources, such as models and data. *http://earthcube.org/group/geosemantics*

OntoSoft Project (Lead PI: Yolanda Gil)

Building an ontology to describe and classify models according to many different types of standardized metadata, e.g. for publication, sharing, execution, composition. This ontology underpins a set of interlinked portals for different modeling communities. *http://earthcube.org/group/ontosoft*