Seven Ontologies for Publishing the Scientific Record on the Web

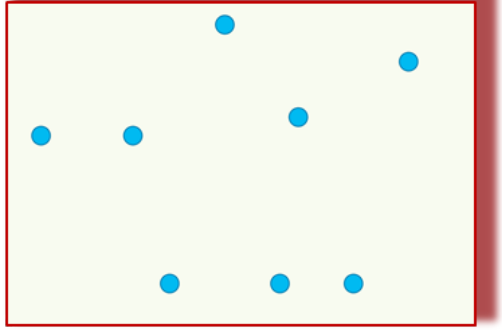
Yolanda Gil
Information Sciences Institute
and Department of Computer Science
University of Southern California
gil@isi.edu

Q52353442

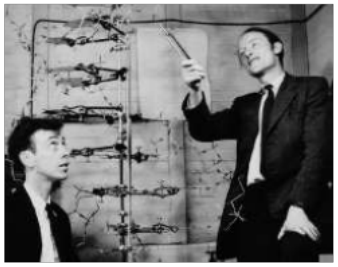
NIST Ontology Summit, 27 May 2020

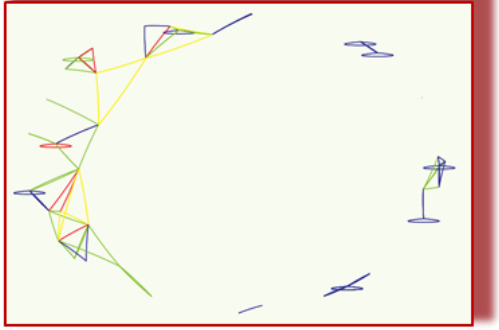
Increasing Complexity of the Scientific Enterprise

Single authorship

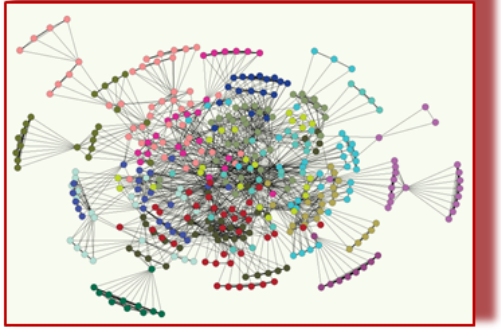


Co-authorship

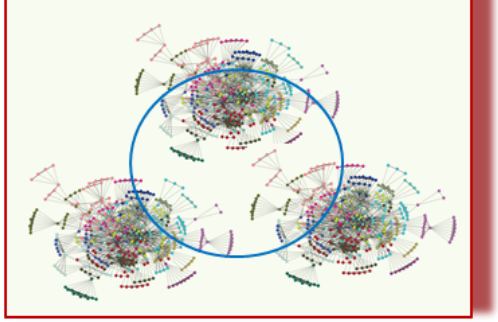




Large number of co-authors



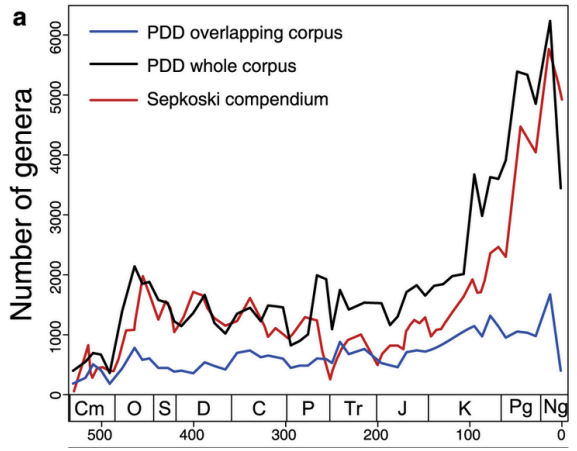
Community as author

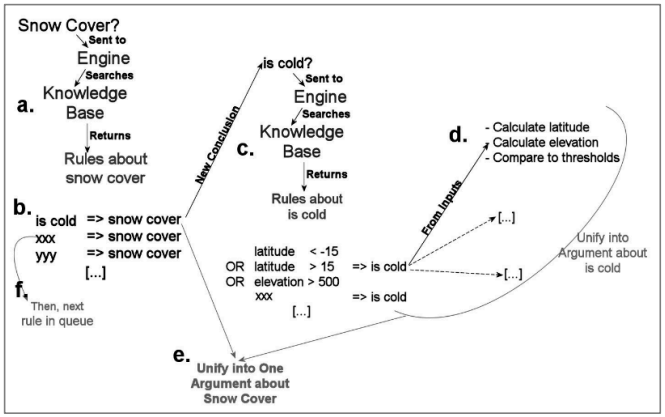


Evolution of the scientific enterprise from [Barabasi, 2005] extended with the ATLAS Detector Project at the Large Hadron Collider [The ATLAS Collaboration, 2012].

Human Limitations Curb Scientific Progress [Gil DSJ'17]

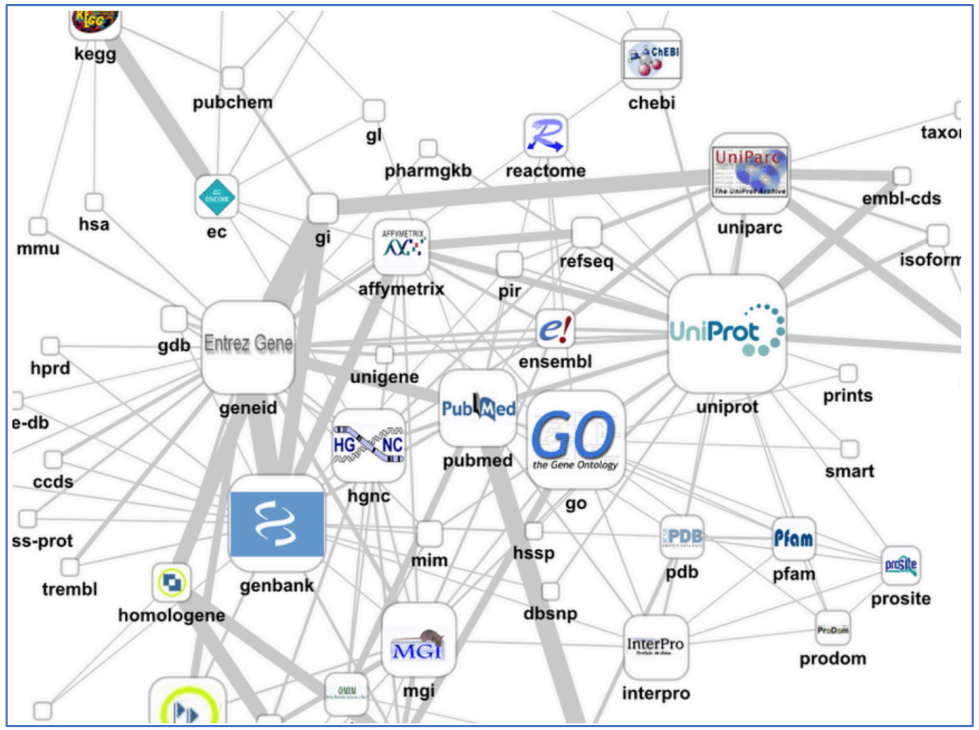
- Not systematic
  - e.g., [Peters et al PLOS 2014]
- Errors
  - e.g., [Herndon et al CJE 2013]
- Biases
  - e.g., [Rassbach et al IAAI 2010]
- Poor reporting
  - e.g., [Garijo et al PLOS 2013]



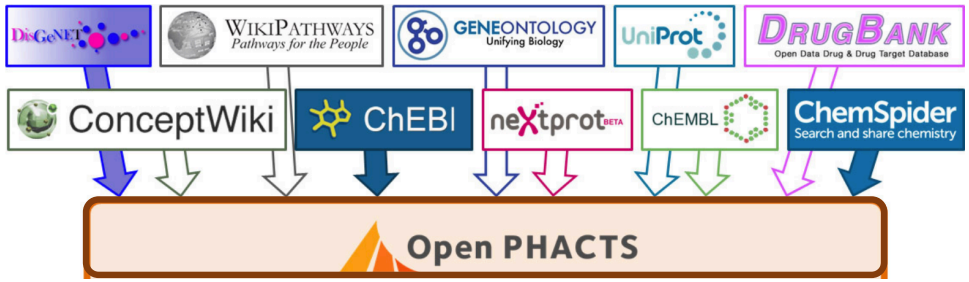


Science Ontologies and Linked Data on the Web

http://bio2rdf.wiki.sourceforge.net



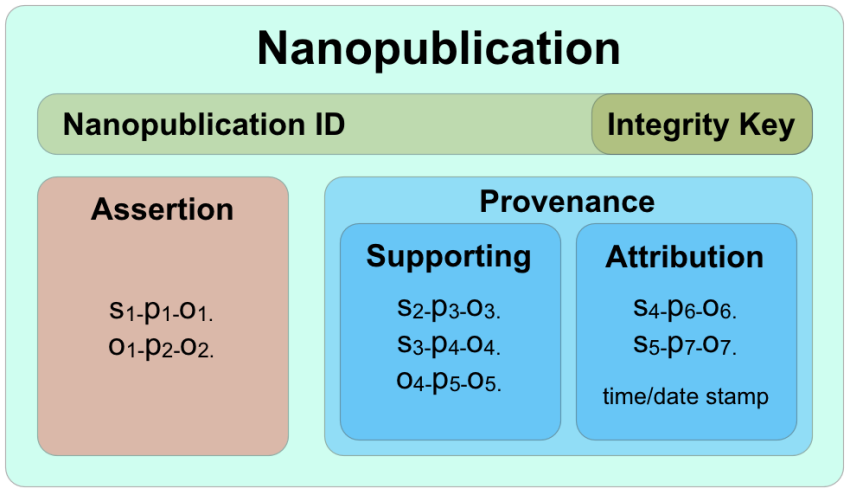
https://www.openphacts.org

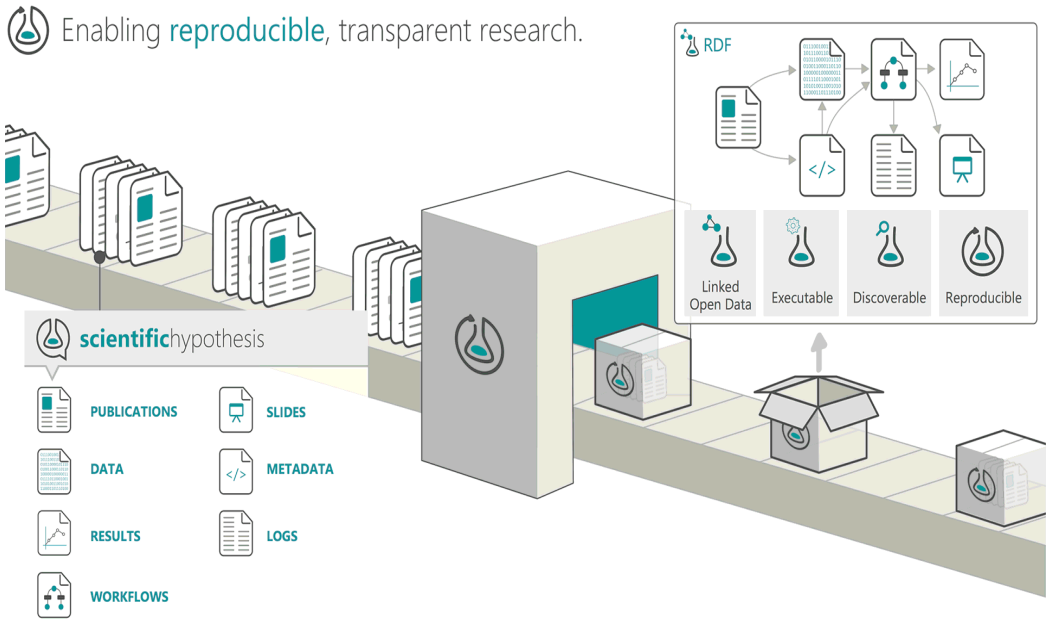


Web Publication of Science Products

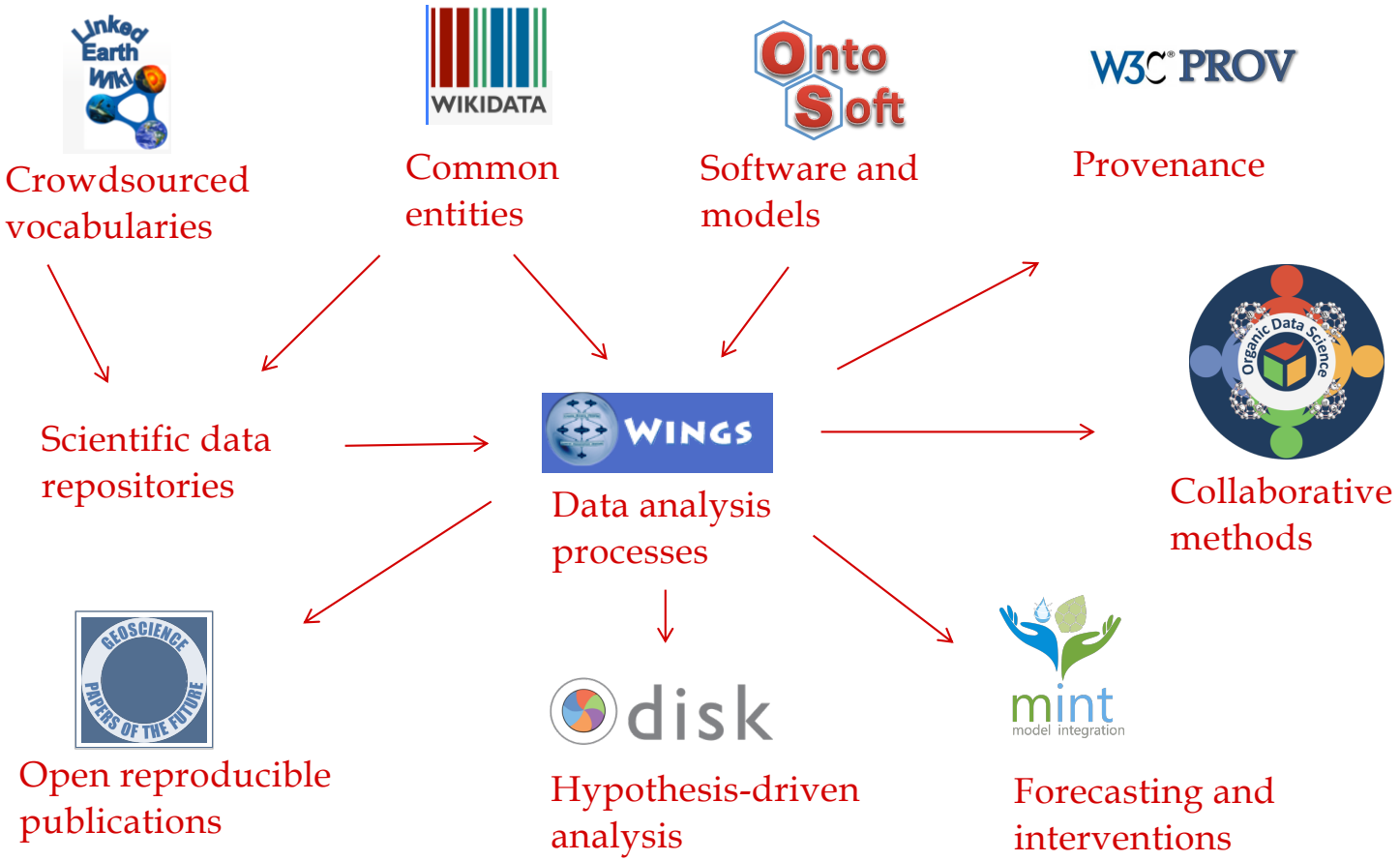
http://www.nanopub.org/

http://www.researchobject.org/

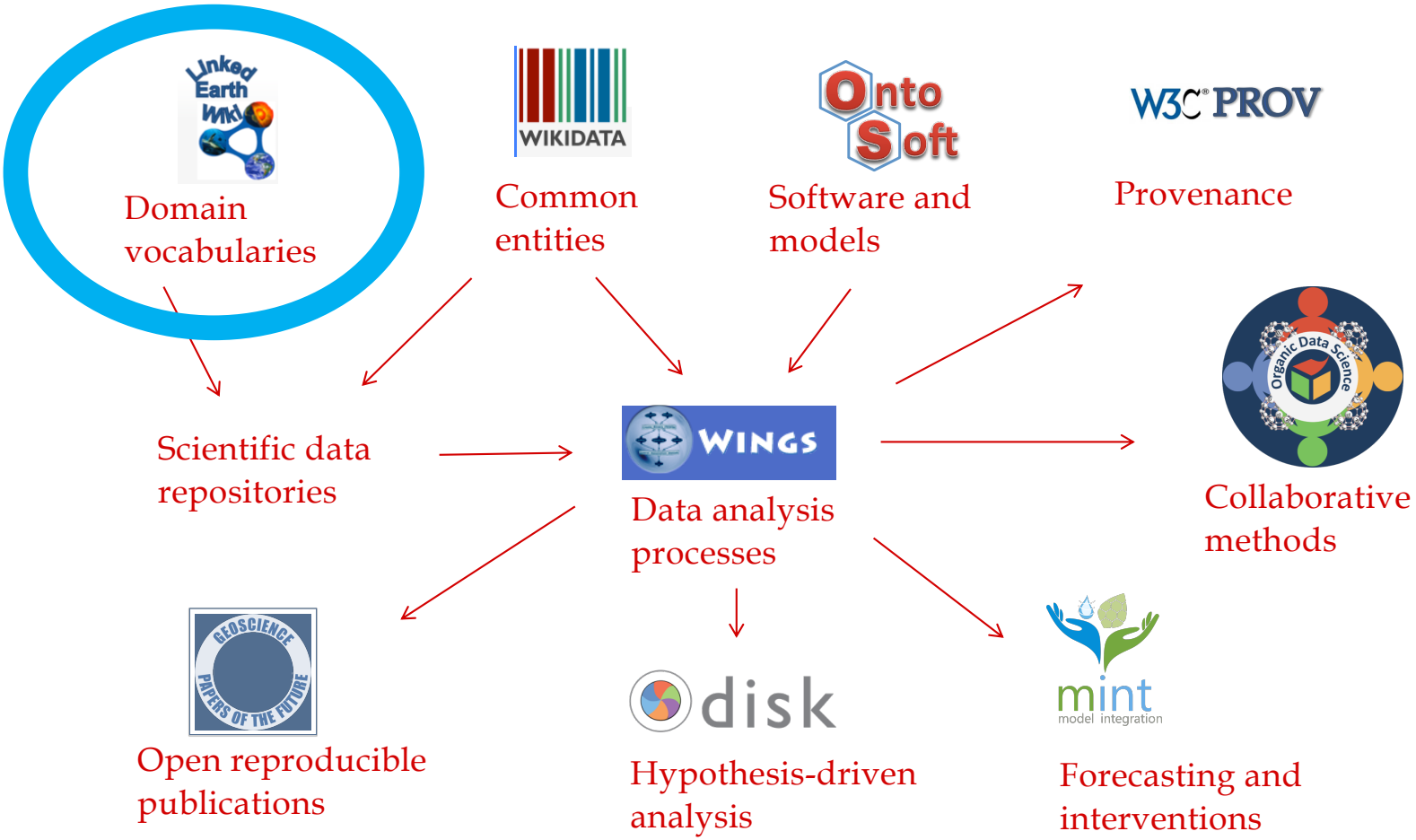




Capturing Scientific Knowledge

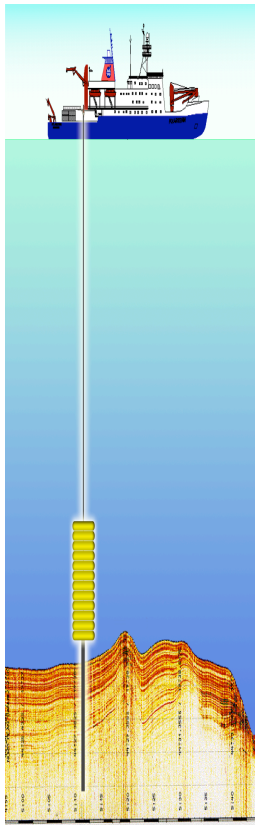


Capturing Scientific Knowledge



Low-Cost Creation of Scientific Vocabulary Standards

[Gil et al ISWC 2017; Khider et al PP 2019; Emile-Geay et al PAGES 2018]



https://commons.wikimedia.org/wiki/File:Gravity-corer_hg.png

https://commons.wikimedia.org/wiki/File:An_ice_core_segment.jpg

AGU100 ADVANCING EARTH AND SPACE SCIENCE

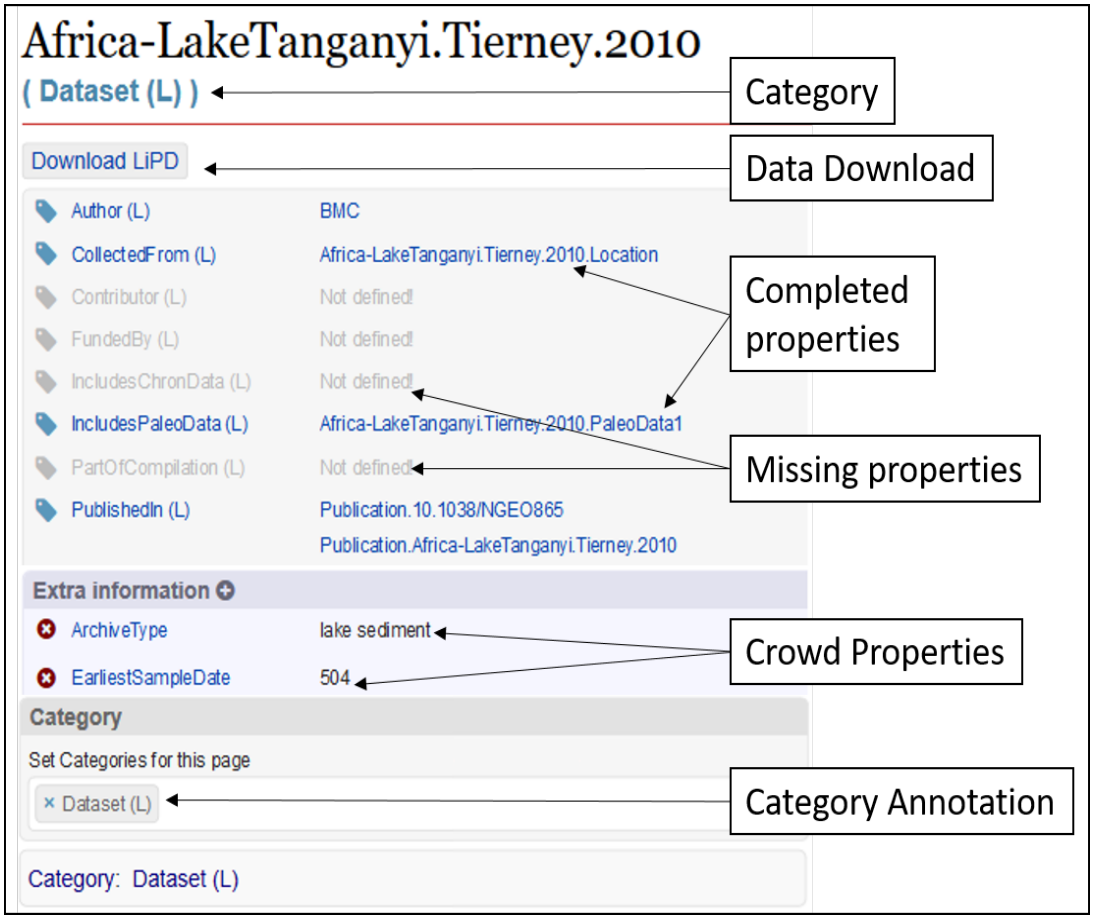
Problem: Diversity of requirements for metadata

Approach: Semantic technologies used for controlled crowdsourcing facilitate creation of community standards to describe highly heterogeneous scientific data

- Organic growth: As scientists annotate their datasets, they propose new metadata properties
- Crowdsourcing: Scientists proposed properties for reuse, vote on priorities
- Editorial oversight: Editors decide what properties will be in future versions

Results: A new standard for paleoclimate (PaCTS 1.0) with one (!!) single initial face-to-face meeting

Controlled Crowdsourcing to Support Continuous Ontology Growth



Linked Earth Ontology Specification Draft

The Linked Earth Ontology

Release 10 August 2016

This version:
http://linkedearth.github.io/Or

Latest version:
http://linked.earth/ontology#

Previous version:
http://linkedearth.github.io/Or

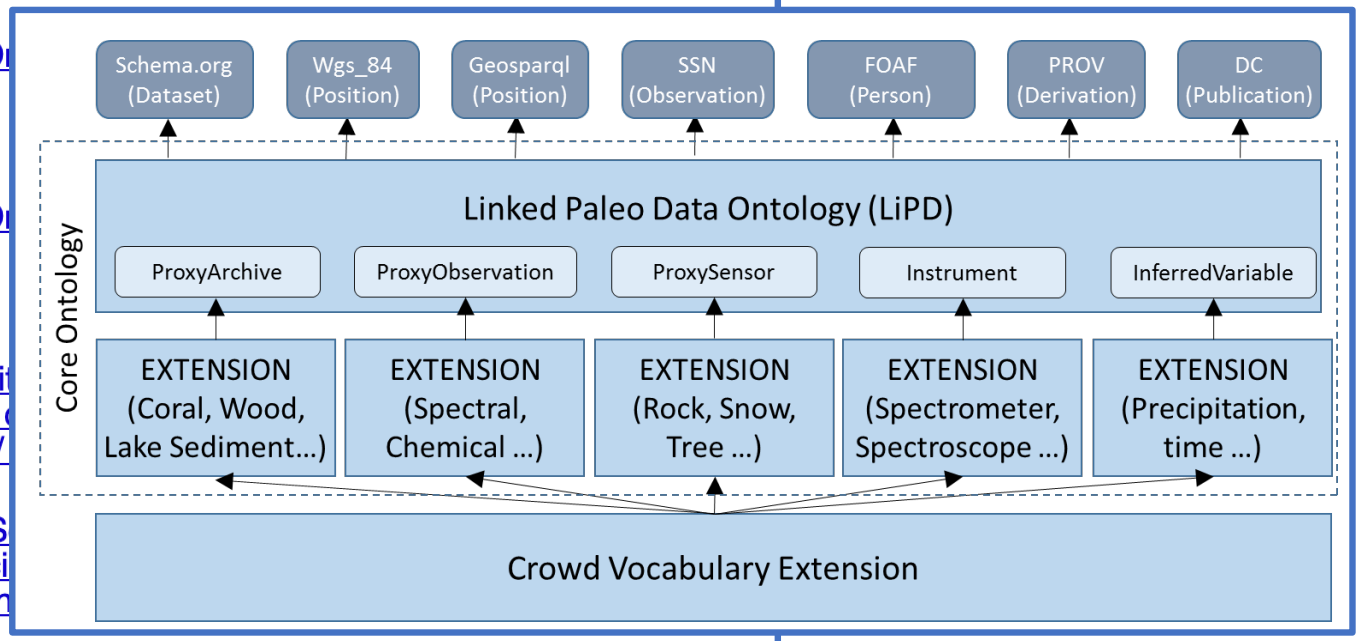
Revision:
1.1.0

Authors:
Julien Emile-Geay, (Universit
Deborah Khider, (University
Nick McKay, (http://nau.edu/

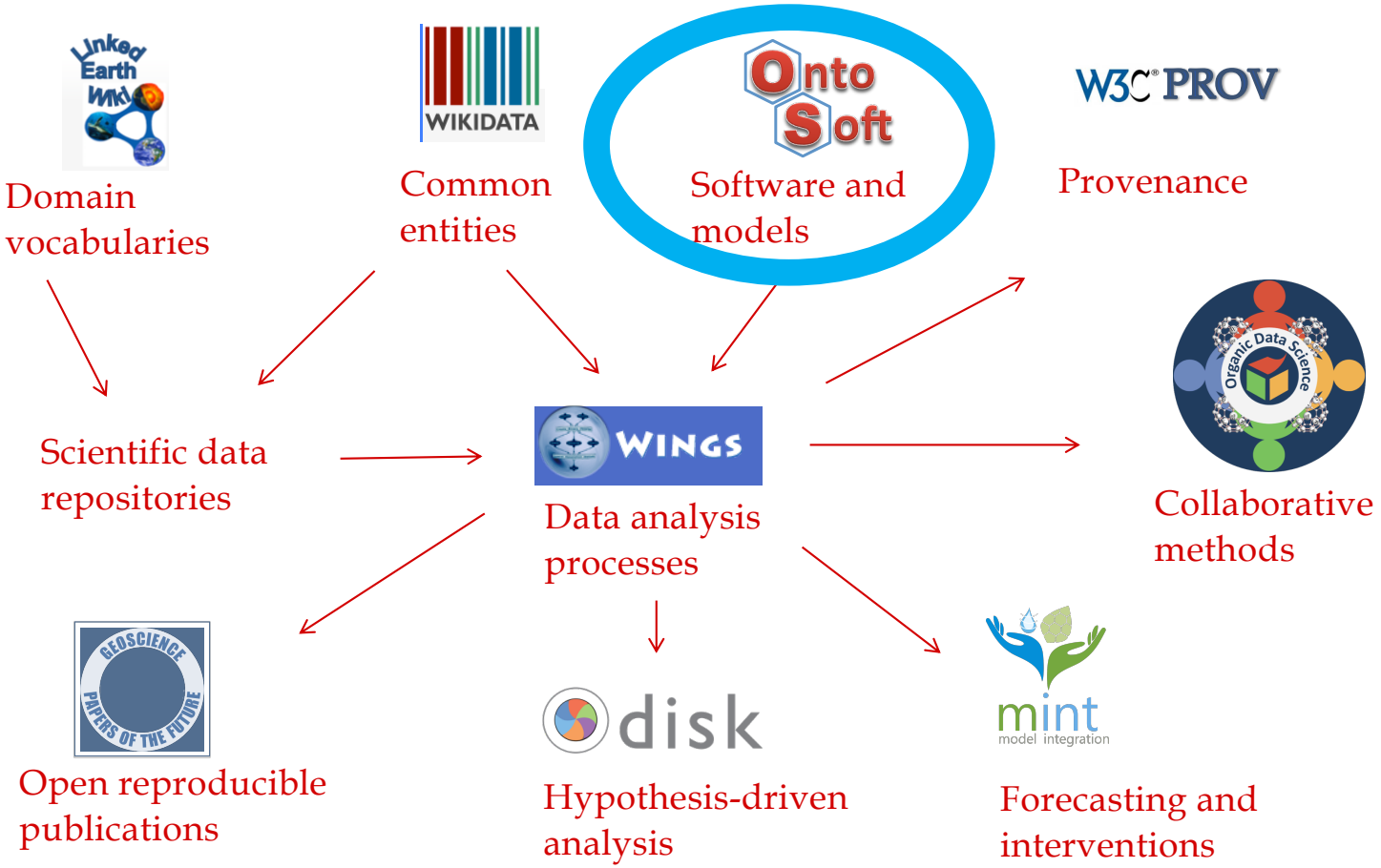
Contributors:
Daniel Garijo, (Information S
Yolanda Gil, (Information Sci
Varun Ratnakar, (Information

Download serialization:
Format RDF/XML Format N Triples Format TTL

License:



Capturing Scientific Knowledge



Representing Scientific Software Metadata

[Gil et al eScience'16; Carvalho et al eScience'18]

Collaboration with Daniel Garijo and Varun Ratnakar (USC/ISI); Lucas Carvalho and Claudia Medeiros (Unicamp)

Locate
unique description

Trust
quality and ratings

Relate
domain knowledge

Contribute
evolution

Track
versions

Discuss
support and community

Access
download

Install
execution requirements

Run
testing execution

Experiment
run with other data

Compose
run with other software

Cite
scientific publications

OntoSoft Ontology

Release 01 May 2015

This version:
http://ontosoft.org/software_20150501 (Alternate link)

Latest version:
http://ontosoft.org/software

Previous version:
http://ontosoft.org/software_20150329 (Alternate link)

Revision
1.0.1

Authors:
Varun Ratnakar, USC/ISI
Yolanda Gil, USC/ISI

Contributors:
Daniel Garijo, C

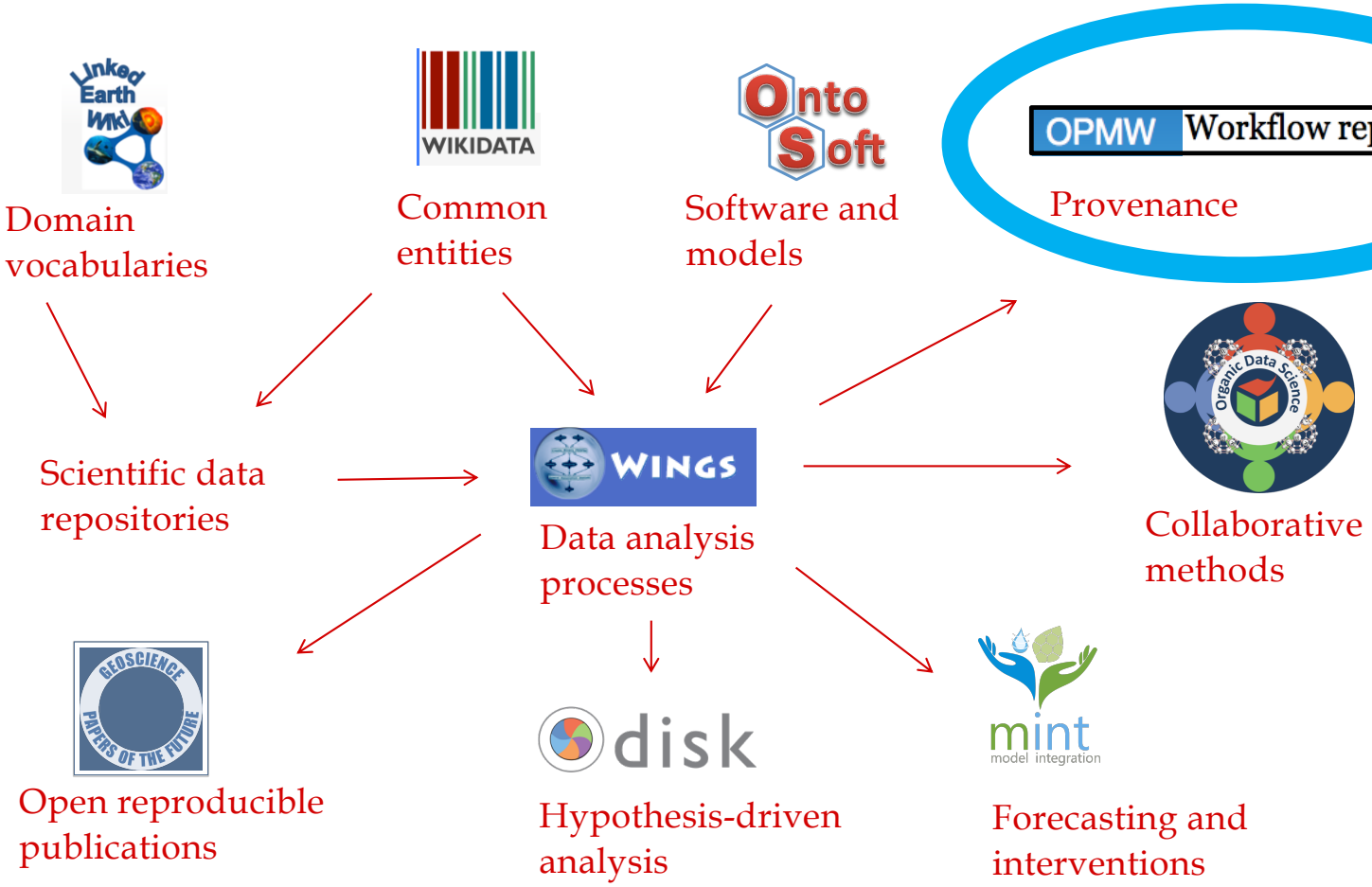
Imported Ontologies
OntoSoft Categ

Extended Ontologie
Dublin Core Ter
PROV

Classes

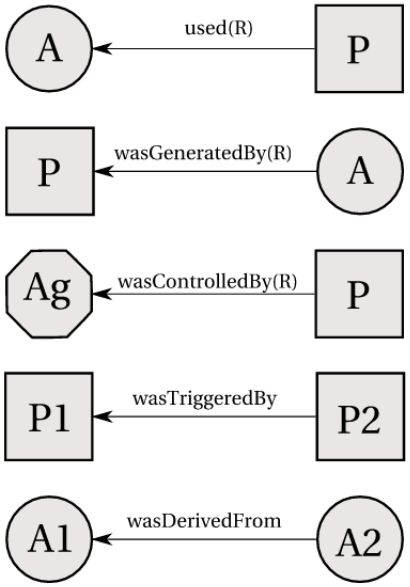
agent Benchmark Information Citation Citation Text Community Information Complex Entity Composition Date Entity Development Information entity Enumeration Entity File I/O Identifier Implementation Details Keywords License Location Measurement Entity Numeric Entity Operating System Parameter Person Programming Language Rating Software Software Description Software Version SoftwareCategory Test Data Test Instructions TestData Description Text Entity Usage Information Usage Statistics Use Limitations Uses and Assumptions Workflow Description

Capturing Scientific Knowledge



The Open Provenance Model Core Specification (v1.1)

Luc Moreau (Editor)[a,*], Ben Clifford[b], Juliana Freire[c], Joe Futrelle[d], Yolanda Gil[e], Paul Groth[f], Natalia Kwasnikowska[g], Simon Miles[h], Paolo Missier[i], Jim Myers[d], Beth Plale[j], Yogesh Simmhan[k], Eric Stephan[l], Jan Van den Bussche[g]



W3C Recommendation

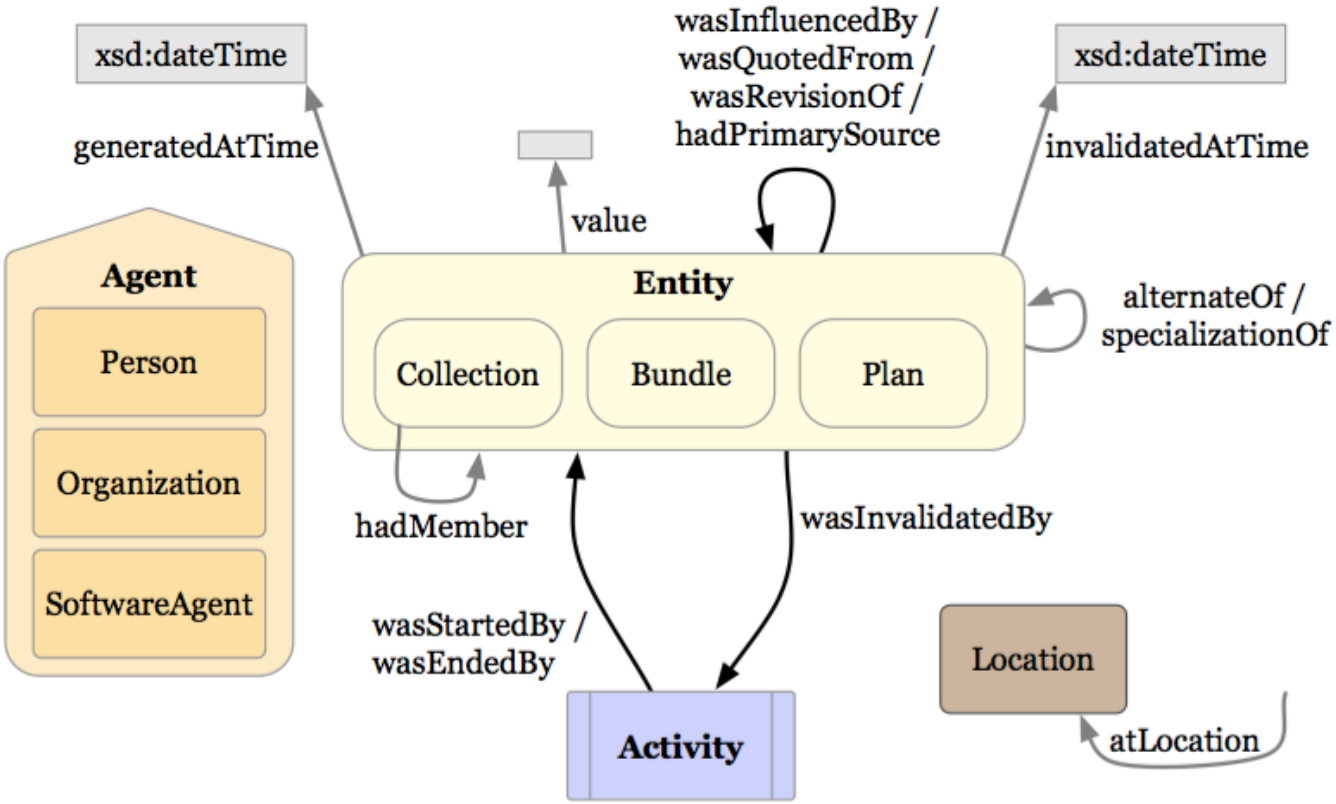
PROV-O: The PROV Ontology

W3C Recommendation 30 April 2013

This version:
http://www.w3.org/TR/2013/REC-prov-o-20130430/
Latest published version:
http://www.w3.org/TR/prov-o/
Implementation report:
http://www.w3.org/TR/2013/NOTE-prov-implementations-20130430/
Previous version:
http://www.w3.org/TR/2013/PR-prov-o-20130312/
Editors:
Timothy Lebo, Rensselaer Polytechnic Institute, USA
Satya Sahoo, Case Western Reserve University, USA
Deborah McGuinness, Rensselaer Polytechnic Institute, USA
Contributors:
(In alphabetical order)
Khalid Belhajjame, University of Manchester, UK
James Cheney, University of Edinburgh, UK
David Corsar, University of Aberdeen, UK
Daniel Garijo, Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
Stian Soiland-Reyes, University of Manchester, UK
Stephan Zednik, Rensselaer Polytechnic Institute, USA
Jun Zhao, University of Oxford, UK

The W3C PROV Provenance Standard

[Gil and Miles 2013; Groth and Moreau 2013; Moreau et al 2014]

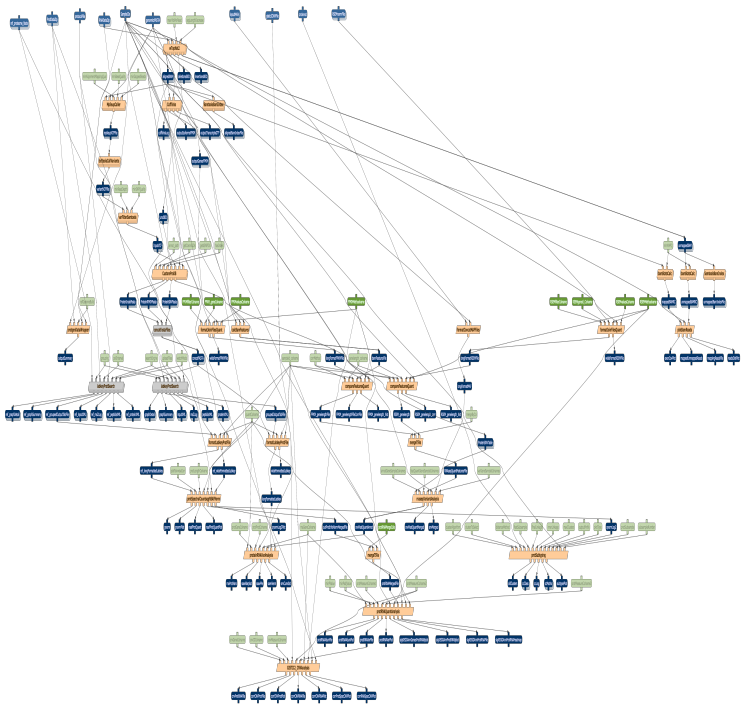


https://www.w3.org/2011/prov/

Publishing Provenance as Linked Data on the Web
[Garijo et al FCGS'17]

Work with Daniel Garijo and Oscar Corcho (UPM)

OPMW Workflow repository



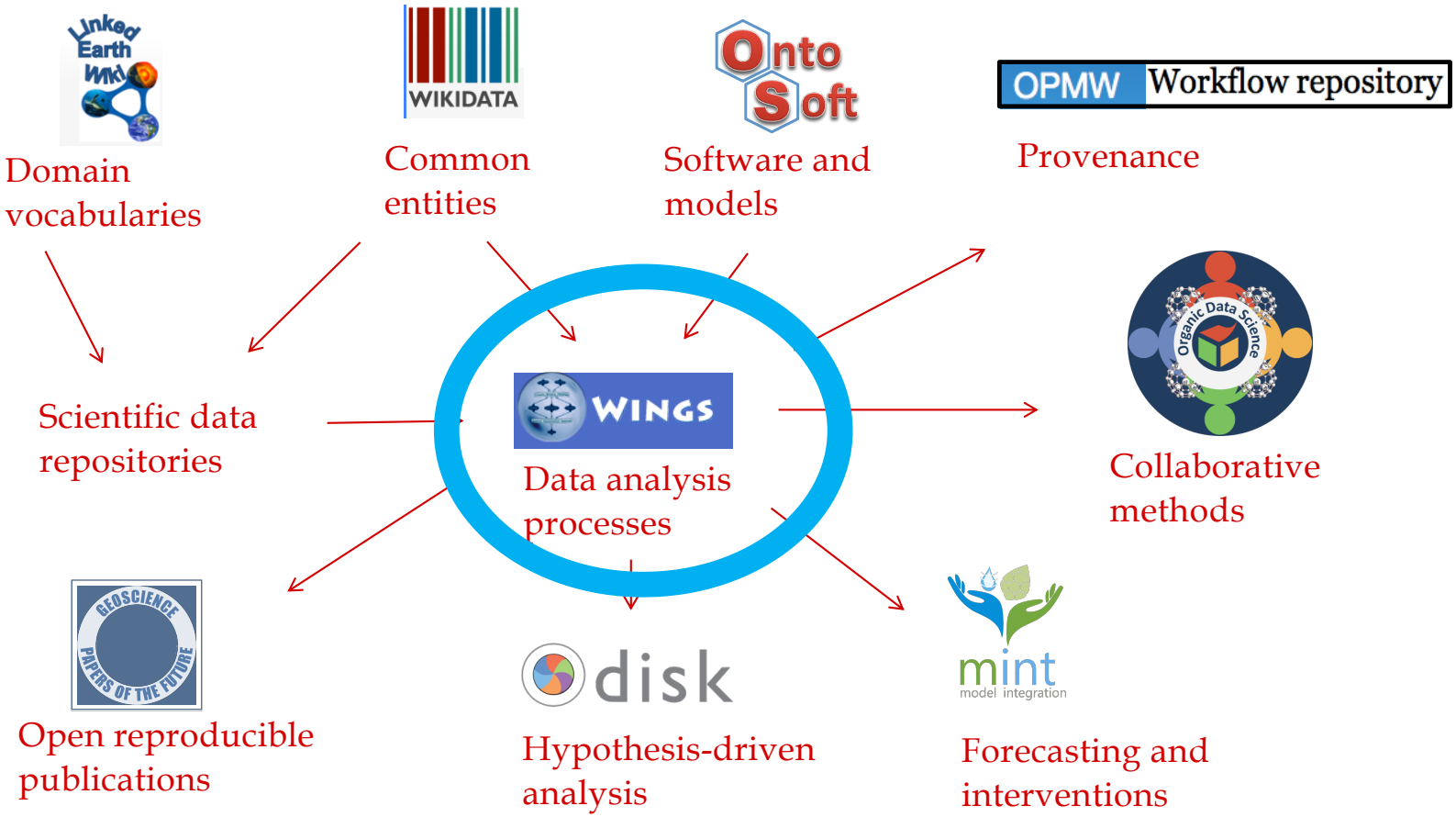
ParenTDT_Workflow at http://seagull.isi.edu/wings-drugome/index.php/Main_Page

http://www.opmw.org/export/resource/WorkflowTemplate/PARENTDT_WORKFLOW

Property	Value
dcterms:contributor	<http://www.opmw.org/export/resource/Agent/GIL>
is ompw:correspondsToTemplate of	<http://www.opmw.org/export/resource/WorkflowExecutionAccount/ACCOUNT1348875527527>
ompw:createdInWorkflowSystem	http://wings.isi.edu (xsd:anyURI)
prov:hadPrimarySource	http://wind.isi.edu/marbles/assets/components/workflow_portal/users/2/genomics/ontology/genomics/ParenTDT_Workflow.owl (xsd:anyURI)
ompw:hasDocumentation	This is variant of the basic family-based association testing for disease traits that also incorporates parental phenotype information, the ParenTDT function in PLINK. *Input: [PEDFile], [MAPFile]* *Output: [OutputTDT], [OutputTDTPerm]* **CreateBinaryPEDFile:** This component converts hapmap format into efficient binary format used by Plink. plink --file [InputPEDFile] --make-bed --out [OutputBPEDFile] **ParenTDT:** This is the same as a basic TDT test except the permutation output is based not on the standard TDT, but the parenTDT. plink --bed [BPEDFile] --bim [BIMFile] --fam [FAMFile] --parentdt1 --out [OutputTDT]
ompw:hasNativeSystemTemplate	http://wind.isi.edu/marbles/assets/components/workflow_portal/users/2/genomics/ontology/genomics/ParenTDT_Workflow.owl (xsd:anyURI)
ompw:hasTemplateDiagram	http://wind.isi.edu/marbles/assets/components/workflow_portal/users/2/genomics/ontology/genomics/ParenTDT_Workflow.owl.png (xsd:anyURI)
is ompw:isStepOfTemplate of	<http://www.opmw.org/export/resource/WorkflowTemplateProcess/PARENTDT_WORKFLOW_CREATE_BINARY_PEDFILENODE> <http://www.opmw.org/export/resource/WorkflowTemplateProcess/PARENTDT_WORKFLOW_PARENTDTNODE>
is ompw:isVariableOfTemplate of	<http://www.opmw.org/export/resource/DataVariable/PARENTDT_WORKFLOW_BIMFILE> <http://www.opmw.org/export/resource/DataVariable/PARENTDT_WORKFLOW_BPEDFILE> <http://www.opmw.org/export/resource/DataVariable/PARENTDT_WORKFLOW_FAMFILE> <http://www.opmw.org/export/resource/DataVariable/PARENTDT_WORKFLOW_MAPFILE> <http://www.opmw.org/export/resource/DataVariable/PARENTDT_WORKFLOW_OUTPUTTDT> <http://www.opmw.org/export/resource/DataVariable/PARENTDT_WORKFLOW_OUTPUTTDTPERM> <http://www.opmw.org/export/resource/DataVariable/PARENTDT_WORKFLOW_PEDFILE>
rdfs:label	ParenTDT_Workflow
dc:rights	http://creativecommons.org/licenses/by-sa/3.0/ (xsd:anyURI)
rdf:type	ompw:WorkflowTemplate prov:Plan
ompw:versionNumber	2 (xsd:int)

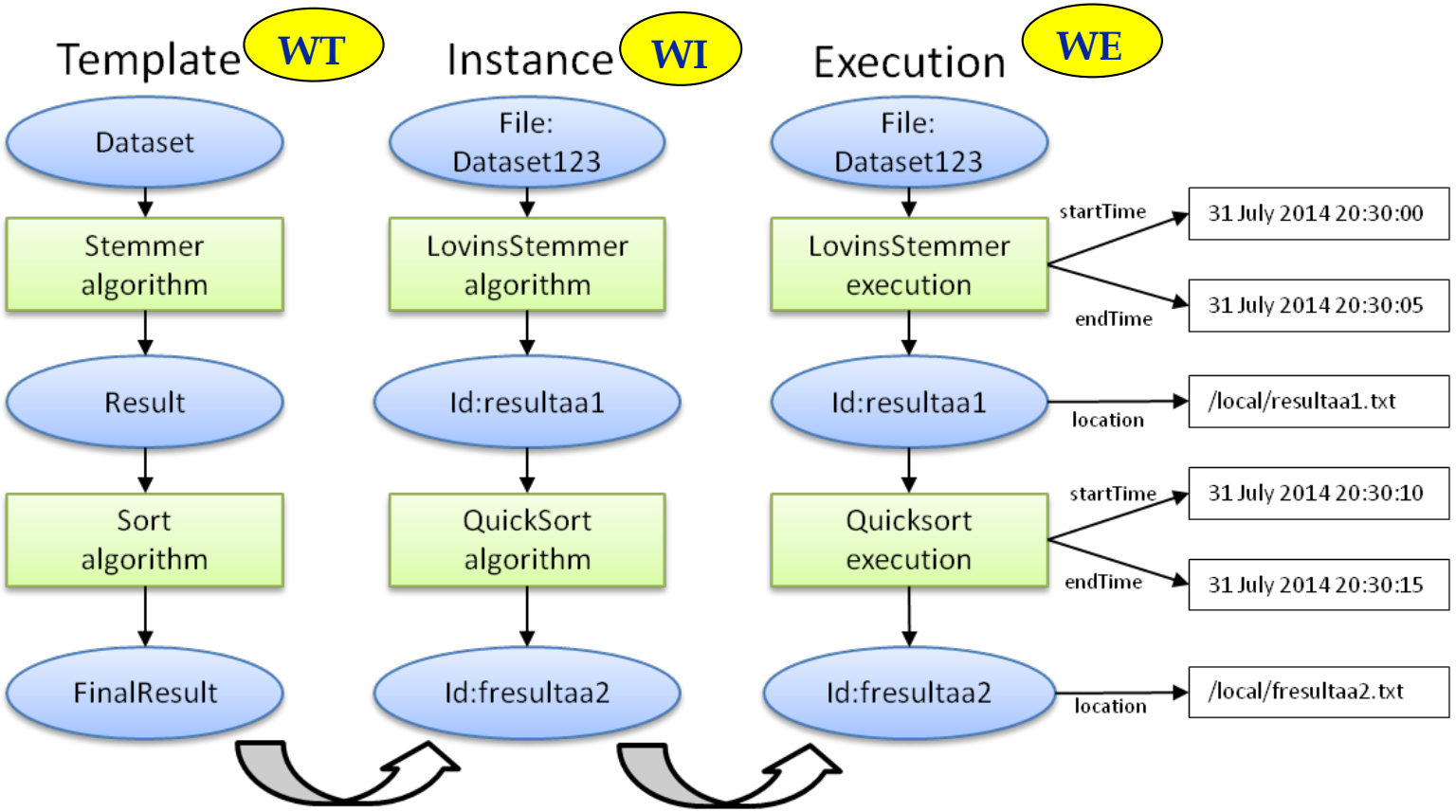
This page shows information obtained from the SPARQL endpoint at http://wind.isi.edu:8890/sparql.
As Turtle | As RDF/XML | Browse in Disco | Browse in Tabulator | Browse in OpenLink Browser

Capturing Scientific Knowledge



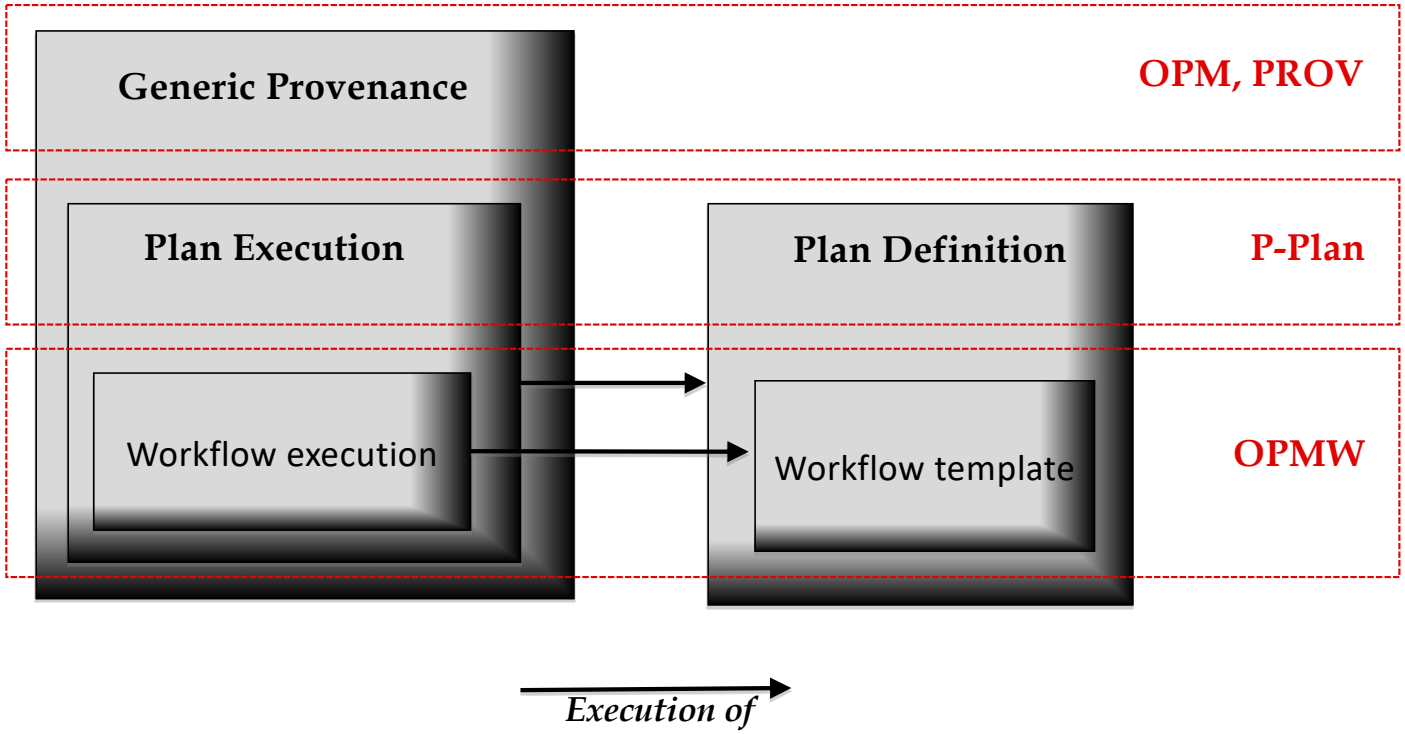
Execution Provenance vs Reusable Workflow

[Gil et al IEEE-IS'11; Gil et al JETAI'11]



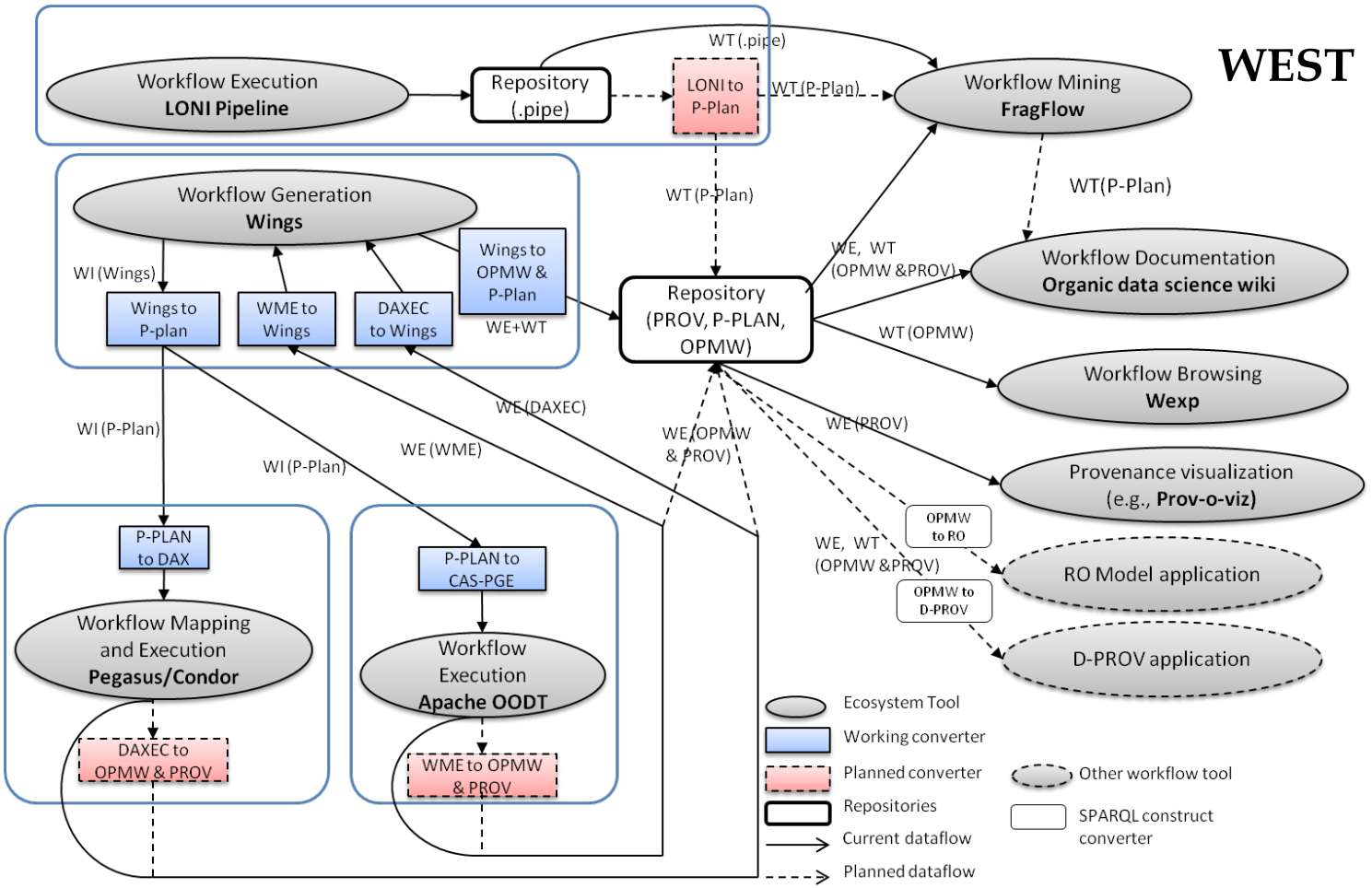
Workflows as Web Objects: PROV, P-PLAN, OPMW
[Garijo et al FGCS'17]

Work with D. Garijo and O. Corcho (UPM)



WEST (Workflow Ecosystems through STandards)
[Gil et al WORKS'14]

- 9 workflow functionalities
- 6 different research groups
- 2-5 consumer systems per exchange
- 4 workflow representations provide different granularity for consumer systems



Functional Heterogeneity in WEST

Generation

USC/ISI ISD

Execution

NASA/JPL & Apache

Visualization

VUA

Mining

UPM

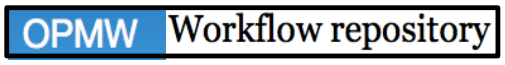
Execution

USC/ISI CST

Browsing

UPM

Repository



USC/ISI ISD

Execution

UCLA/USC

Documentation

USC/ISI ISD

The P-PLAN Ontology

Release 12 March 2014

This version:
http://vocab.linkeddata.es/p-plan/version/03062014/

Latest version:
http://purl.org/net/p-plan#

Previous version:
http://vocab.linkedda

Revision
1.3

Authors:
Daniel Garijo, Ontolo
Yolanda Gil, Informat

Extended Ontologies:
PROV-O: The PROV

W3C Recommendation

W3C

PROV-O: The PROV Ontology

W3C Recommendation 30 April 2013

0130430/

plementations-20130430/

30312/

titute, USA
versity, USA
chnic Institute, USA

er, UK
K

o, Universidad Politécnica de Madrid, Spain
ester, UK
nstitute, USA

The OPMW-PROV Ontology

Release 22 December 2014

This version:
http://www.opmw.org/model/OPMW_20141222

Latest version:
http://www.opmw.org/model/OPMW

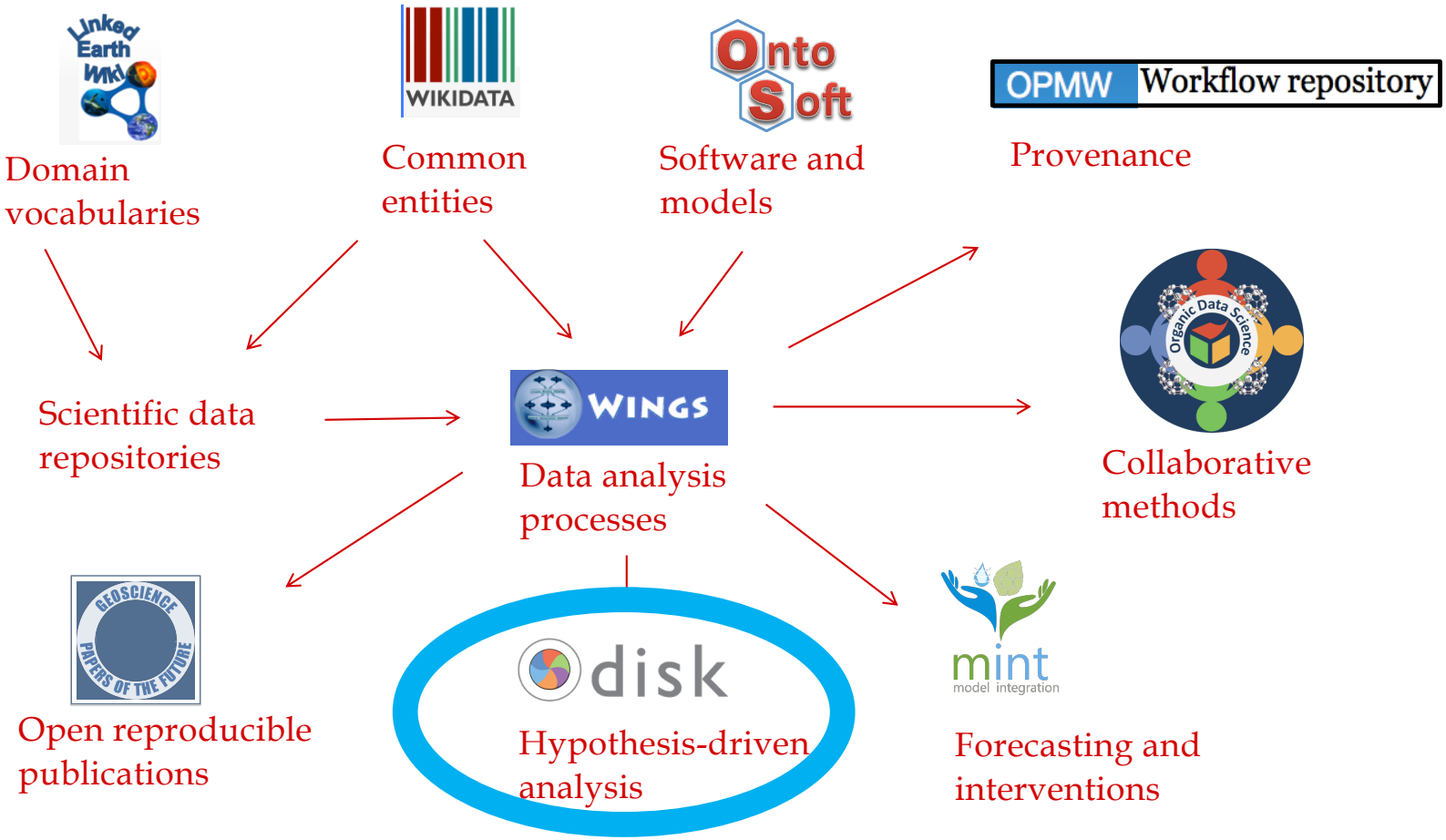
Previous version
http://www.opmw.org/model/OPMW_20140711

Revision
Revision 6

Authors:
Daniel Garijo, Universidad Politécnica de Madrid
Yolanda Gil, Information Sciences Institute, University of Southern California, US

Imported Ontologies:
The Open Provenance Model Ontology (OPMO)
The Open Provenance Model Vocabulary (OPMV) (indirect import)
The PROV-O Ontology
The P-plan Ontology

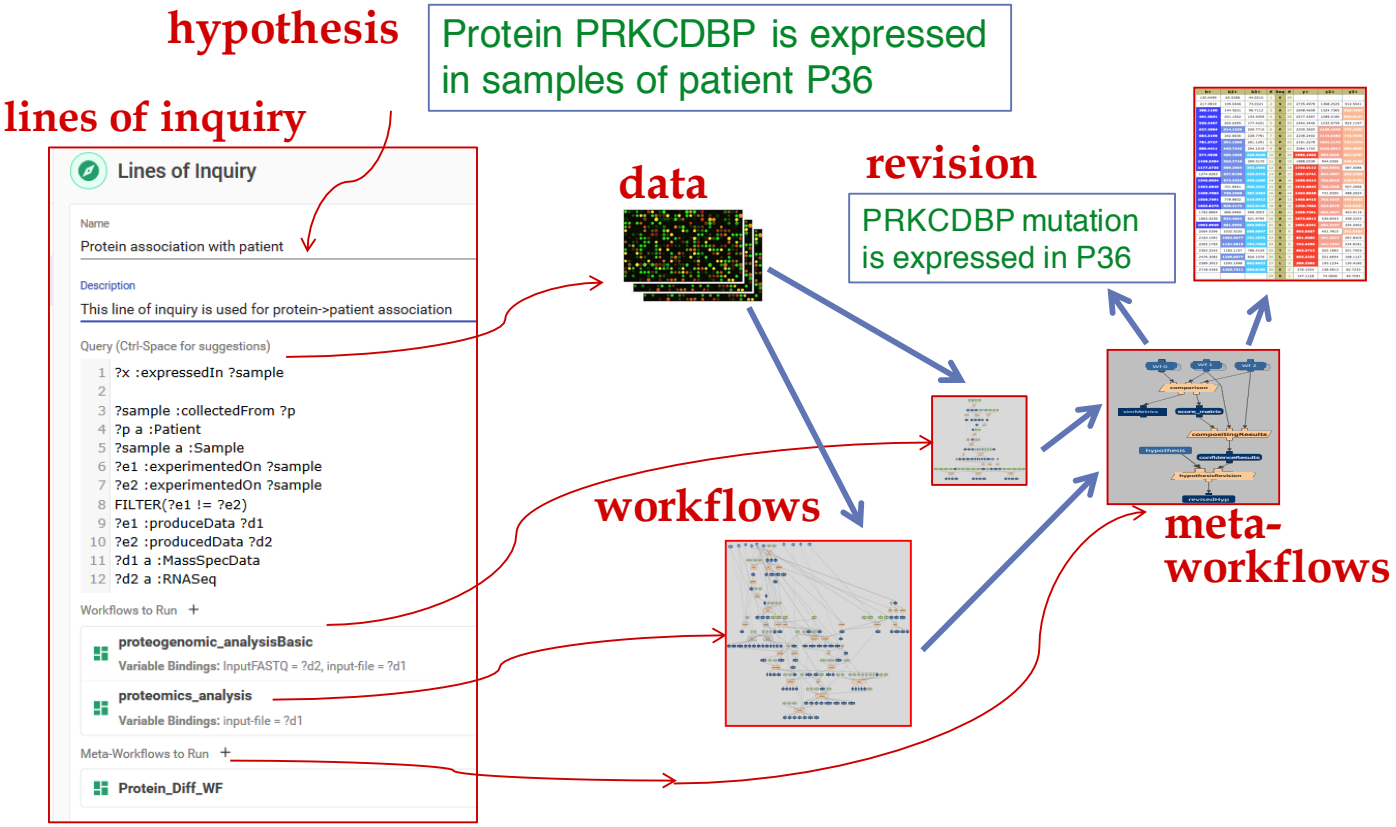
Capturing Scientific Knowledge



Automated Discovery in DISK

[Gil et al AAAI'17; Gil et al ACS'16]

With Parag Malick, Ravali Adusumilli, Hunter Boyce (Stanford); Arunima Srivastava (OSU); Daniel Garijo, Varun Ratnakar, Rajiv Mayani (USC/ISI); Thomas Yu (Sage Bionetworks)



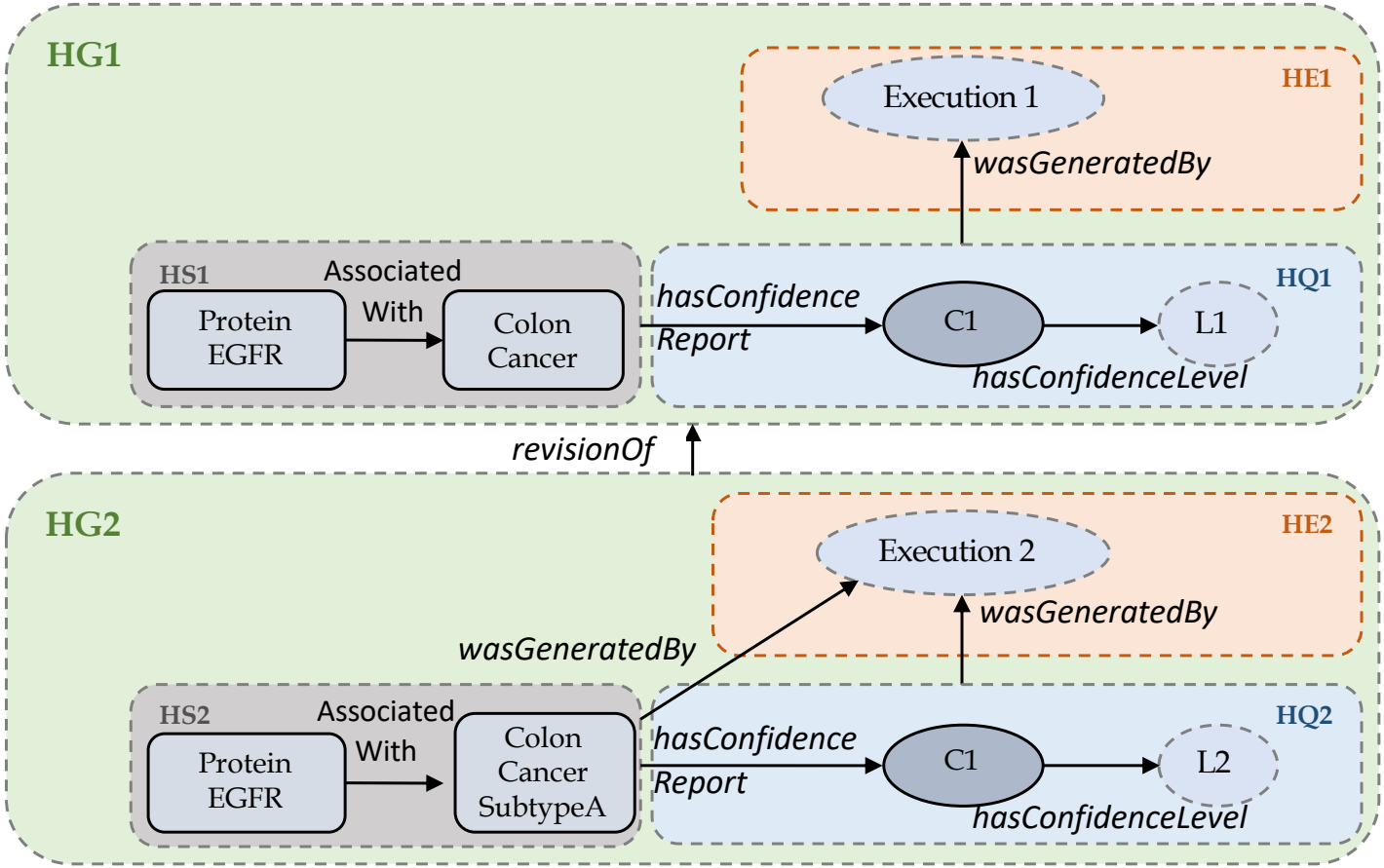
Representing Hypotheses

Statement

Qualifier

Evidence

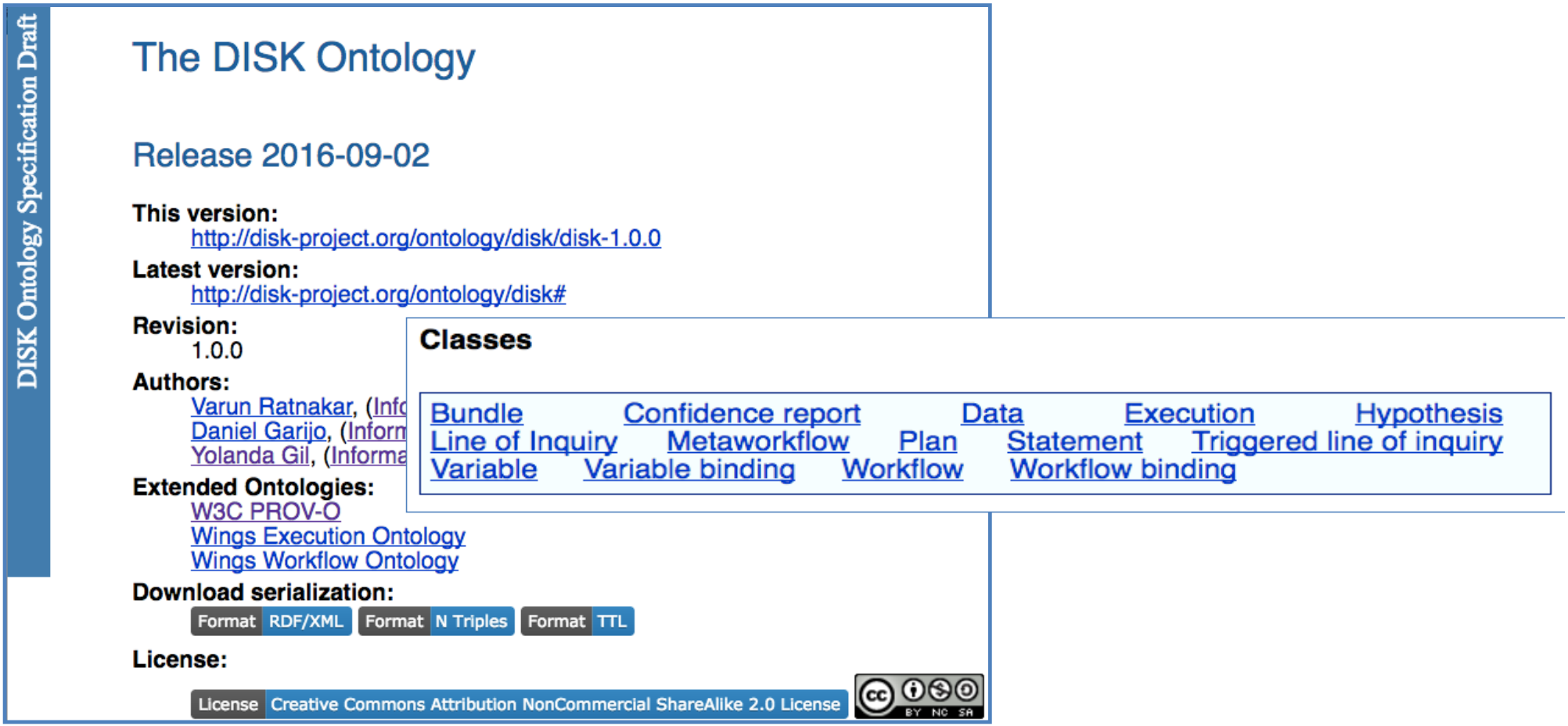
Evolution



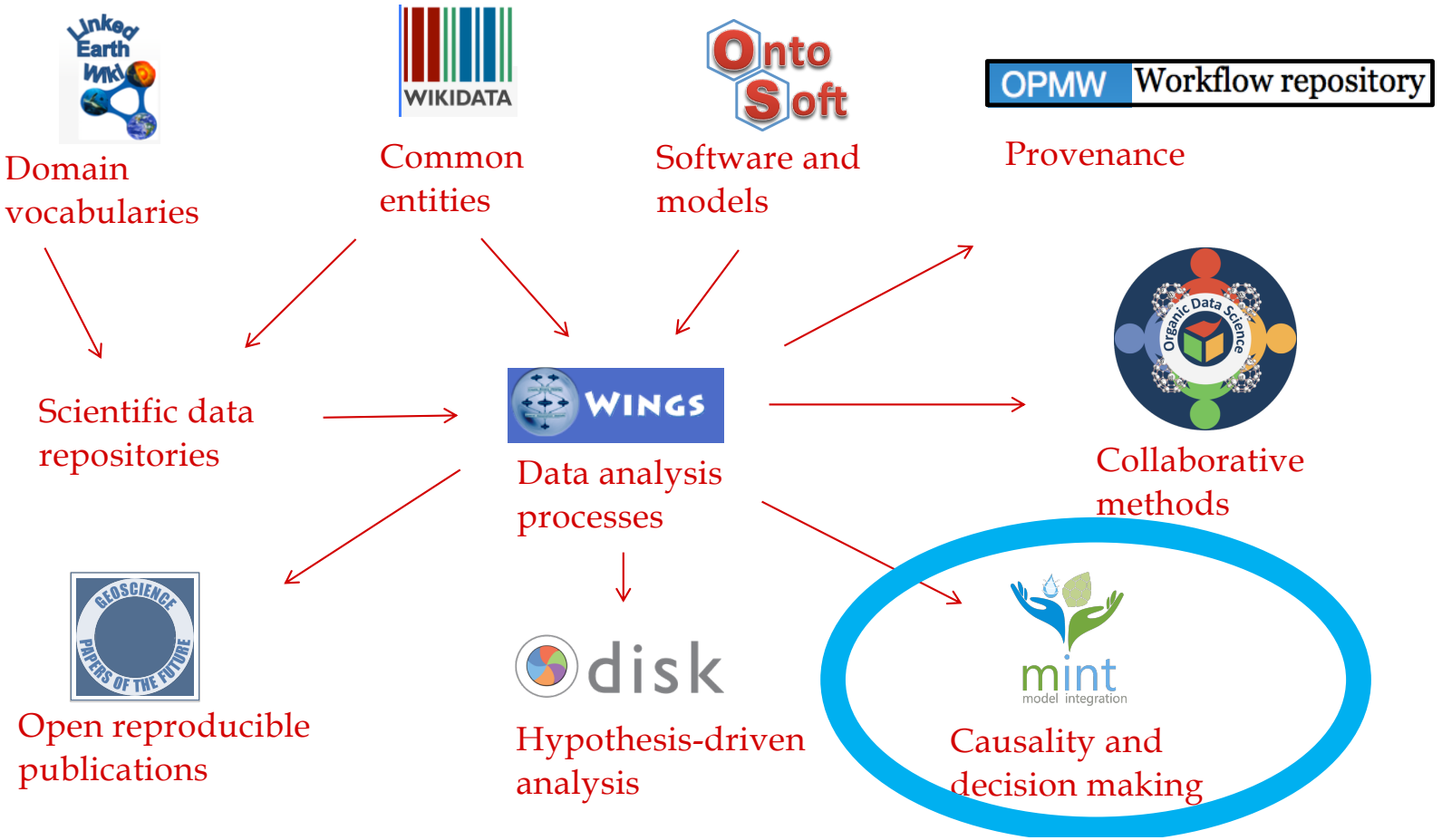
DISK Hypothesis Ontology:

http://disk-project.org/ontology/disk

The DISK Hypothesis Ontology
[Garijo et al SciKnow'17]



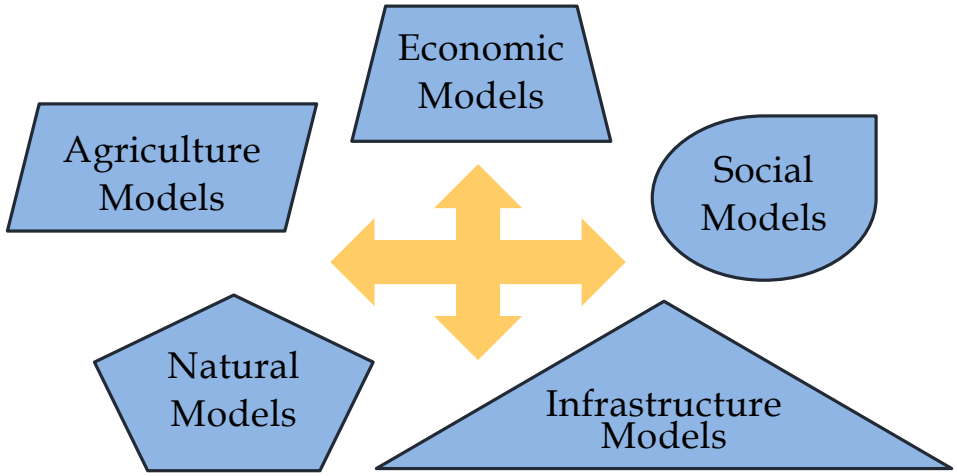
Capturing Scientific Knowledge

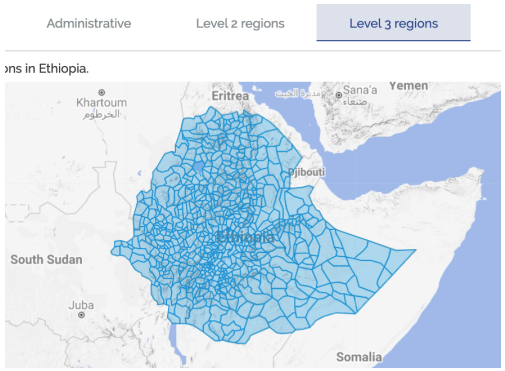


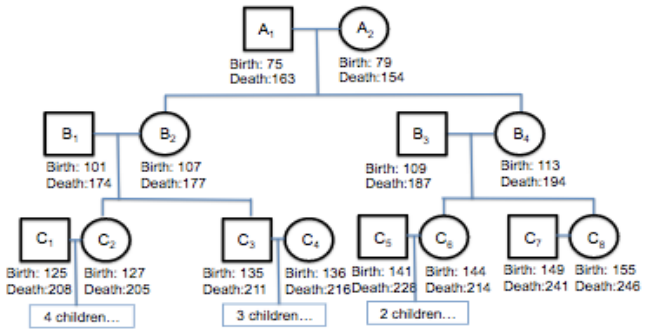
MINT: Model INTegration

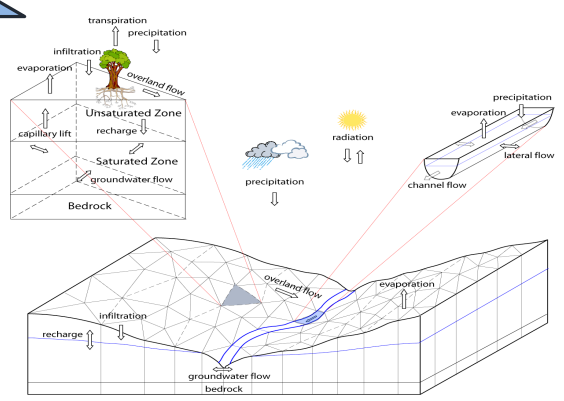
[Gil et al IEMS 2018; Garijo et al eScience 2019]

Collaboration with Daniel Garijo, Deborah Khider, Craig Knoblock, Ewa Deelman, Rafael Ferreira (USC/ISI), Vipin Kumar (UM), Scott Peckham (CU), Chris Duffy & Armen Kemanian (PSU), Kelly Cobourn (VT), Suzanne Pierce (UT)





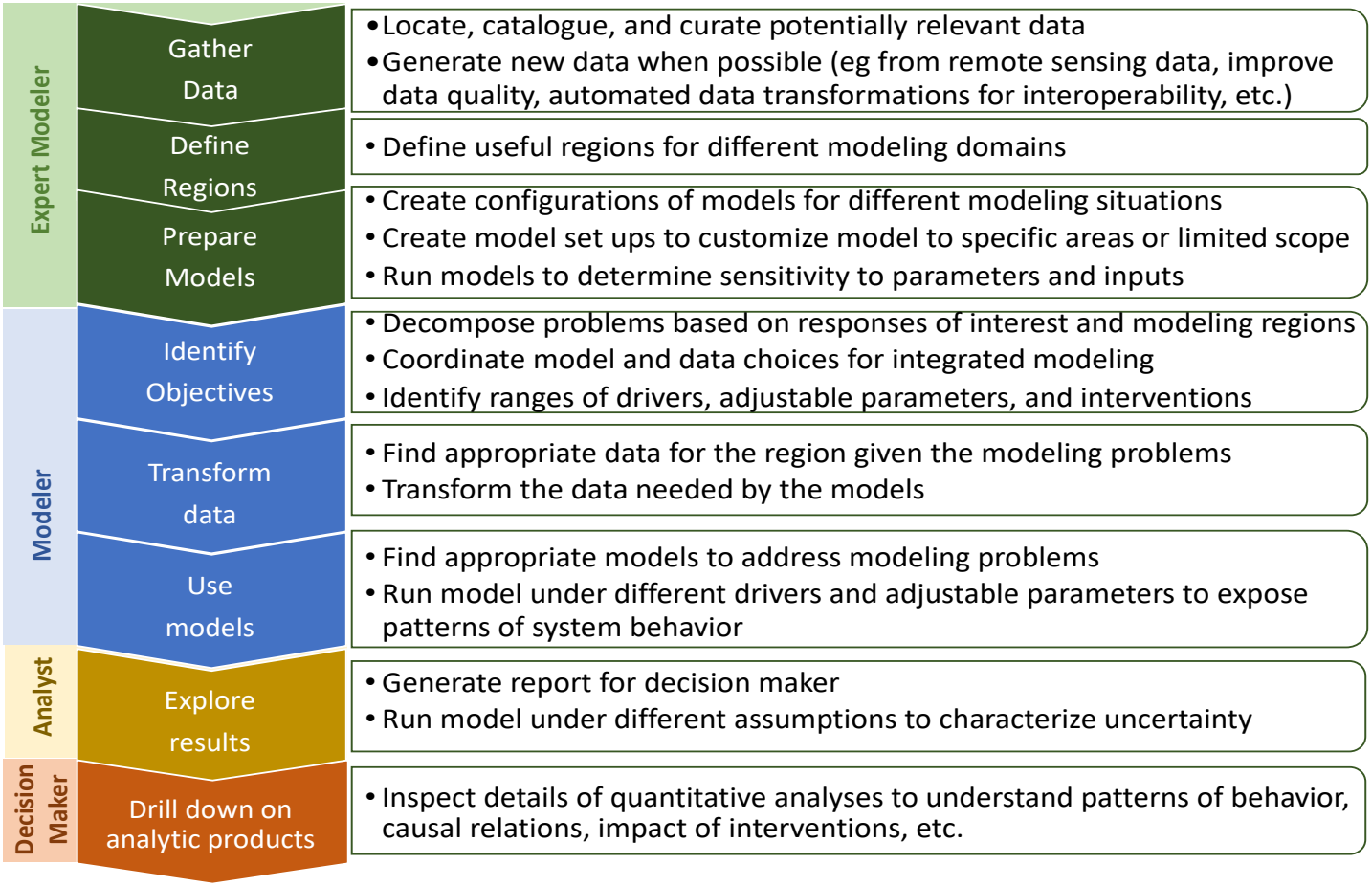




Mediation at many levels

- Modeling scope
- Model set up
- Variable mapping
- Data ingestion
- Spatial gridding

From Modeling to Decisions



Ontology Specification Draft

The Software Description Ontology for Models

Release March 31st, 2020

This version:
https://w3id.org/okn/o/sdm/1.4.0

Latest version:
https://w3id.org/okn/o/sdm

Previous version:
https://w3id.org/okn/o/sdm/1.3.0

Revision:
1.4.0

Authors:
Daniel Garijo
Deborah Khider
Yolanda Gil

Contributors:
Armen Kemanian
Christopher Duffy
Kelly Cobourn
Scott Peckham

Imported Ontologie
sd

Download serializat
Format JSON LD

License:
License http://cre

Classes

Configuration Setup data structure definition Dataset Specification Empirical model Emulator Farming practices Funding Information geo coordinates GeoCoordinates GeoShape Grid Hybrid model ICASA Variable Image Index Intervention Model Model configuration Model configuration setup organization Organization Parameter person Person Point Based Grid region Sample Collection Sample Execution Sample Resource Software Software Configuration Software Image Software Version Source Code Spatially Distributed Grid Standard Variable Subsidy SVO Variable Theory guided model Time Interval unit Unit variable Variable Variable presentation Visualization

Ontology Specification Draft

The Software Description Ontology

language en

Release March 31st, 2020

This version:
https://w3id.org/okn/o/sd/1.5.0

Latest version:
https://w3id.org/okn/o/sd

Previous version:
https://w3id.org/okn/o/sd/1.4.0

Revision:
1.5.0

Authors:
Daniel Garijo
Varun Ratnakar
Yolanda Gil
Deborah Khider

Contributors:
Hernan Vargas
Maximiliano Osorio

Download serialization:
Format JSON LD Format

License:
License http://creativecor

Visualization:
Visualize with WebVowl

Cite as:
Daniel Garijo,Varun R
Revision: 1.5.0. Retrie

Classes

Configuration Setup data structure definition Dataset Specification Funding Information ICASA Variable Image NumericalIndex organization Organization Parameter person Person Sample Collection Sample Execution Sample Resource Software Software Configuration Software Image Software Version Source Code Standard Variable SVO Variable unit Unit variable Variable Variable presentation Visualization

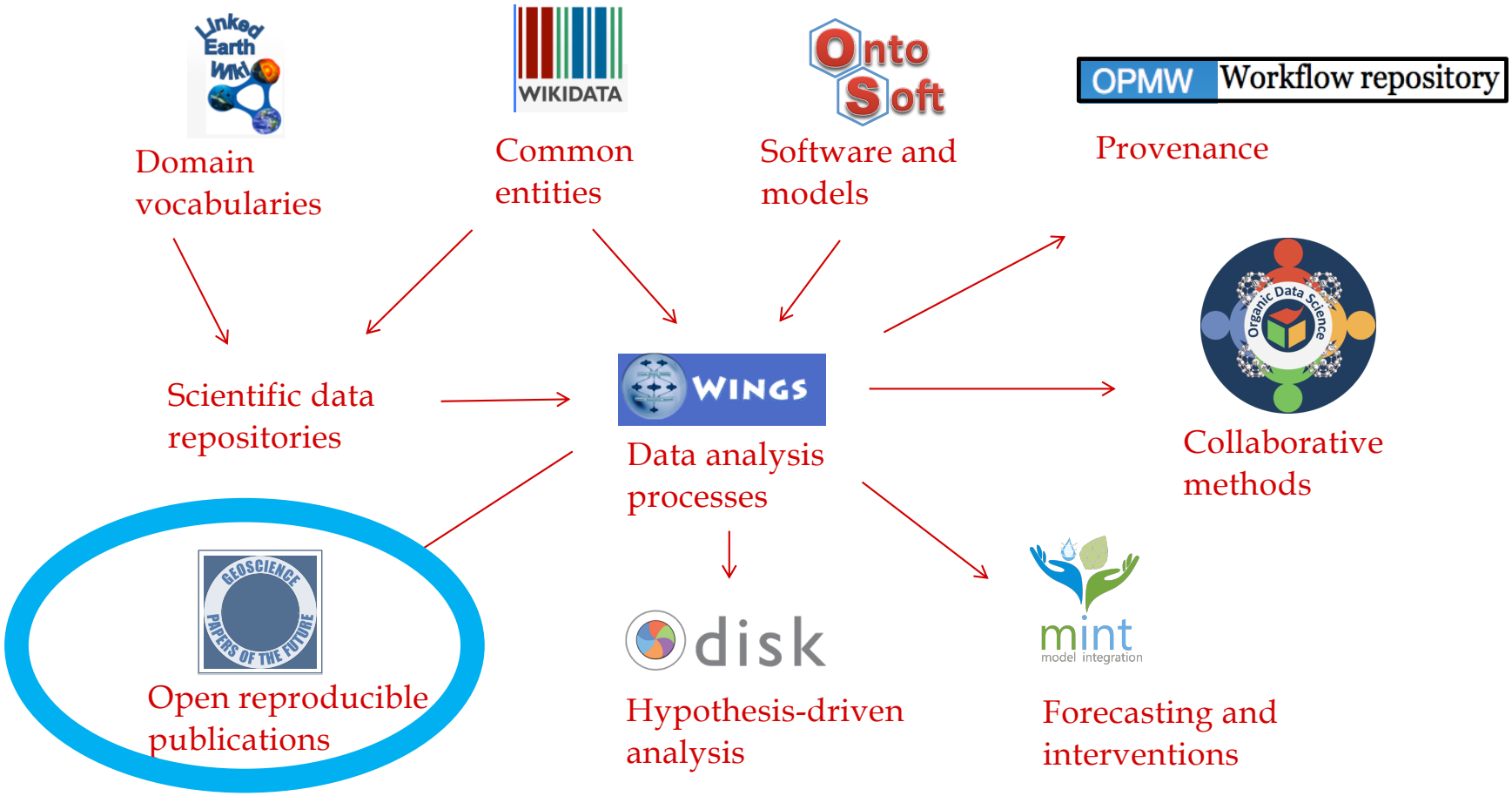
Object Properties

adjustable parameter adjusts variable author author compatible visualization software contributor contributor copyright holder copyright holder copyright year copyright year funding source had primary source had primary source has configuration has contact person has file structure has fixed resource has funding information has input has output has parameter has part has presentation has sample execution has sample result has sample visualization has setup has software image has software version has source code has standard variable logo part of dataset publisher publisher screenshot screenshot used useful for calculating index uses unit was derived from was derived from software was generated by

Data Properties

citation citation code repository code repository data catalog identifier date created date created date published date published description description email email funding funding grant has accepted values has assumption has component location has constraint has data type has default value has dimensionality has documentation has download URL has example has execution command has FAQ has fixed value has format has implementation script location has installation instructions has long name has maximum accepted value has minimum accepted value has purpose has short name has step size has support script location has typical data source has usage notes has version id identifier identifier keywords keywords license license memory requirements memory requirements name name operating systems operating systems position processor requirements processor requirements programming language programming language recommended increment reference publication reference publication short description software requirements software requirements support details tag value value website

Capturing Scientific Knowledge



www.scientificpaperofthefuture.org

[Gil et al ESS 2016; Essawy et al EMS 2017; Goodman et al PLOS CB 2014]

Scientific Paper of the Future

Modern Paper

Text:
Narrative of the method, some data is in tables, figures/plots, and the software used is mentioned

Data:
Include data as supplementary materials and pointers to data repositories

Reproducible Publication

Software:
For data preparation, data analysis, and visualization

Provenance and methods:
Workflow/scripts specifying dataflow, codes, configuration files, parameter settings, and runtime dependencies

Open Science

Sharing:
Deposit data and software (and provenance/workflow) in publicly shared repositories

Open licenses:
Open source licenses for data and software (and provenance/workflow)

Metadata:
Structured descriptions of the characteristics of data and software (and provenance/workflow)

Digital Scholarship

Persistent identifiers:
For data, software, and authors (and provenance/workflow)

Citations:
Citations for data and software (and provenance/workflow)

AGU American Geophysical Union

Earth and Space Science
AN OPEN ACCESS AGU JOURNAL
Special Section: Geoscience Papers of the Future

NATURE REVIEWS | NEUROSCIENCE

Scanning the horizon: towards transparent and reproducible neuroimaging research

Russell A. Poldrack[1], Chris I. Baker[2], Joke Durnez[1,3], Krzysztof J. Gorgolewski[1], Paul M. Matthews[4], Marcus R. Munafò[5,6], Thomas E. Nichols[7], Jean-Baptiste Poline[8], Edward Vul[9] and Tal Yarkoni[10]

Towards the neuroimaging paper of the future

In this Analysis article, we have outlined a number of problems with current practice and made suggestions for improvements. Here, we outline what we would like to see in the neuroimaging paper of the future, inspired by related work in the geosciences[71].

SEG SOCIETY OF EXPLORATION GEOPHYSICISTS

GEOPHYSICS Call for Papers
Reproducible Research:
Geophysics Papers of the Future

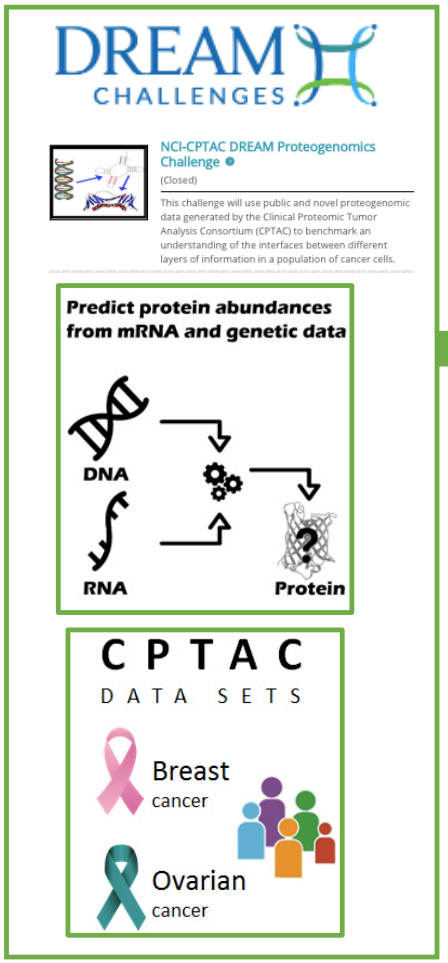
AI magazine

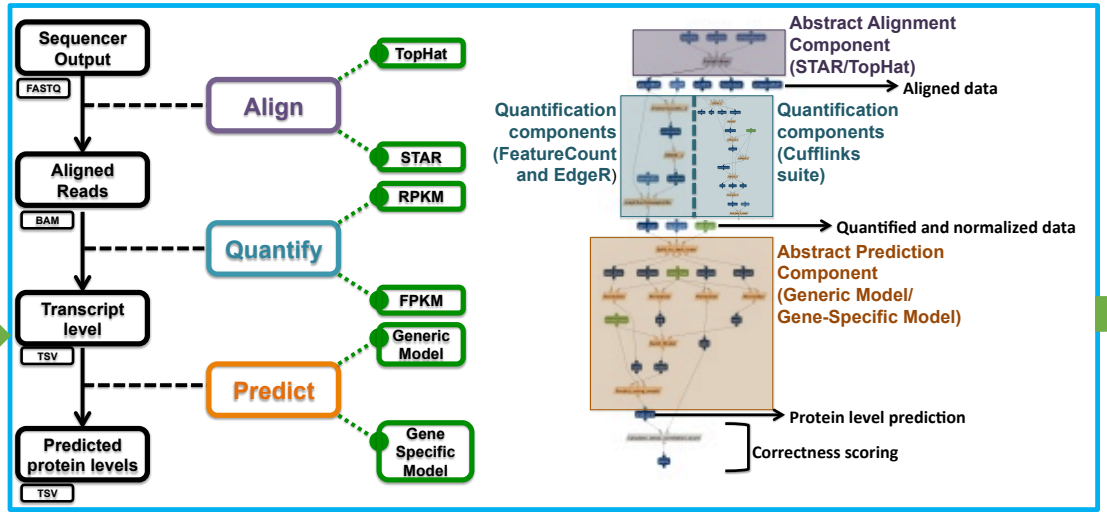
On Reproducible AI: Towards Reproducible Research, Open Science, and Digital Scholarship in AI Publications

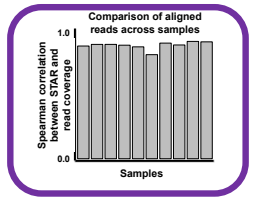
Odd Erik Gundersen, Yolanda Gil, David W. Aha

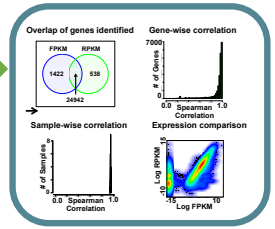
Benchmarking for DREAM Challenges
[Srivastava et al PSB 2019]

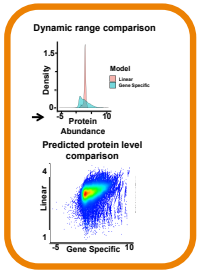
disk





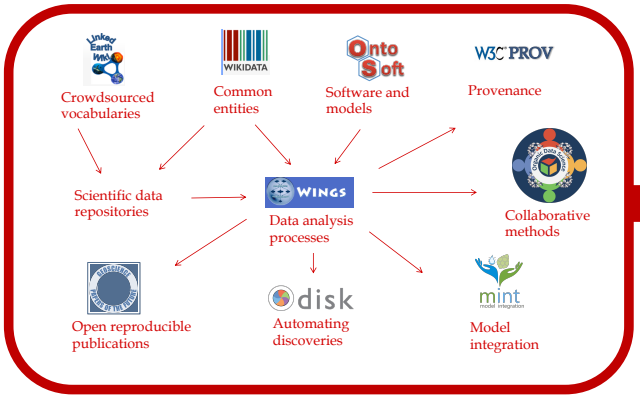


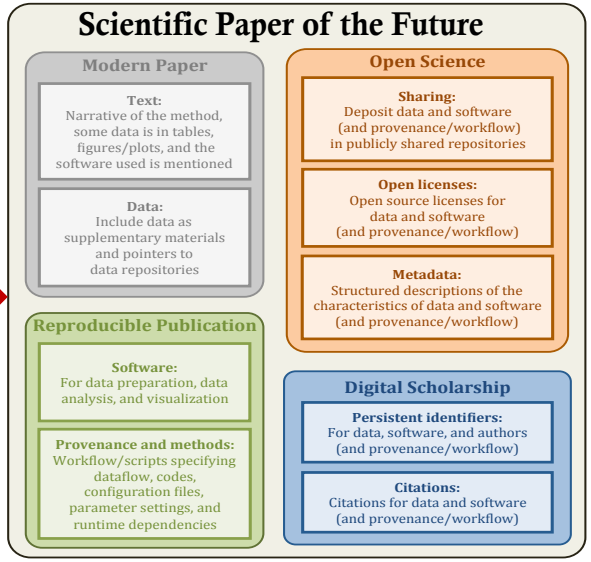




Alignment	Quantification	Predictive Model	Correctness Score	Time Taken
STAR	FPKM	Linear	0.2161	~29 hrs
STAR	RPKM	Linear	0.2155	~20 hrs
STAR	FPKM	Gene-Specific	0.9064	~29 hrs
STAR	RPKM	Gene-Specific	**0.9124**	~20 hrs
TopHat	RPKM	Linear	0.2053	~103 hrs
TopHat	RPKM	Gene-Specific	0.9080	~103 hrs

Will AI Write the Scientific Papers of the Future?





AAAI Presidential Address, February 2020
https://vimeo.com/400177695

Thank you!

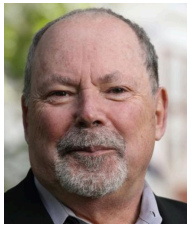












- Varun Ratnakar, Daniel Garijo, Deborah Khider, Maximiliano Osorio, Hernan Vargas (USC)
- *Workflows*: Jihie Kim, Ewa Deelman, Karan Vahi; Rafael Ferreira, Rajiv Mayani, Hyunjoon Jo, Yan Liu, Dave Kale (USC); Ralph Bergmann (U Trier); William Cheung (HKBU); Oscar Corcho (UPM); Pedro Gonzalez, Gonzalo Castro (UCM); Paul Groth (UA); Ricky Sethi (FSU); Carole Goble (UM); Chris Mattmann, Paul Ramirez, Dan Crichton, Rishi Verma (JPL); Natalia Villanueva (UTEP)
- *Linked Earth and Organic Data Science*: Julien Emile-Geay, Deborah Khider (USC); Nick McKay (NAU); Felix Michel and Matheus Hauder (TUM); Chris Duffy (PSU); Paul Hanson, Hilary Dugan, Craig Snortheim (U Wisconsin); Jordan Read (USGS); Neda Jahanshad (USC)
- *Biomedical workflows*: Phil Bourne, Sarah Kinnings (UCSD); Chris Mason (Cornell); Joel Saltz, Tahsin Kurk (Emory U.); Jill Mesirov, Michael Reich (Broad); Shannon McWeeney, Christina Zhang (OHSU); Parag Mallick, Ravali Adusumilli, Hunter Boyce (Stanford U.)
- *Geosciences workflows*: Paul Hanson (U Wisconsin), Tom Harmon & Sandra Villamizar (U Merced), Tom Jordan & Phil Maechlin (USC), Kim Olsen (SDSU); Suzanne Pierce (UT); Chris Duffy & Armen Kemanian (PSU); Scott Peckham & Maria Stoica (CU)
- *And many others!*