

Semantics for Cyberinfrastructure: The iPlant Collaborative's Semantic Web Platform



**VIRTUAL WORKSHOP
SEMANTICS IN GEOSPATIAL AND OTHER ARCHITECTURES:
DESIGN AND IMPLEMENTATION**

May 2013

**DAMIAN GESSLER, Ph.D.
SEMANTIC WEB ARCHITECT
UNIVERSITY OF ARIZONA**

dgessler (at) iplantcollaborative (dot) org

The iPlant Collaborative



NSF-funded, multi-institutional, collaborative project
similar in scope to EarthCube.

Mission:

To build a cyberinfrastructure for the nation's plant scientists.

Successfully completed its first five years
and is anticipatory of the next (May 2013)

www.iplantcollaborative.org

All-too-Familiar Challenges



1. Disparate data
2. Varying research agendas
3. Lack—or proliferation—of standards
4. Disruptive technologies (NGS)
5. Rapidly changing science
6. Global externalities
7. Hierarchical, non-linear systems
8. Finite time, money, and resources

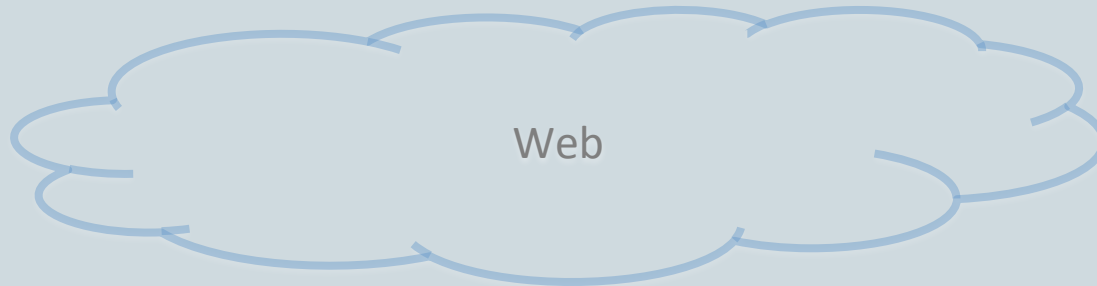
A Constrained Solution Space



Perfect solutions either do not exist,
or are not feasibly attainable under the constraints.

Yet even good solutions can be powerfully transformative
(*e.g.*, WWW in 1991)

Enterprise and HPC Assets



Enterprise Class

Discovery Environment, Atmosphere

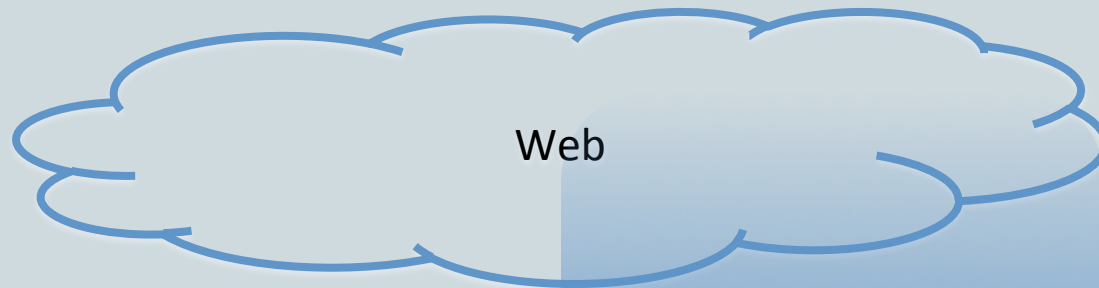
Enterprise Class
Virtual Workdesk
Cloud, Virtual machines

Foundational Infrastructure

iPlant Data Store, HPC, etc.

High Performance Computing
PFLOPS-scale compute power
Petabyte scale storage

Bridging Enterprise, HPC, and Web Assets



Enterprise Class

Discovery Environment, Atmosphere

Semantic Integration

Semantic Pipelining

Foundational Infrastructure

iPlant Data Store, HPC, etc.

Distributed
Semantic Web Services
Logic-driven semantics

Enterprise Class
Virtual Work desk
Cloud, Virtual machines

High Performance Computing
PFLOPS-scale compute power
Petabyte scale storage

The Actors



The World



You



Your lab

Semantic Mediation Layer



The World

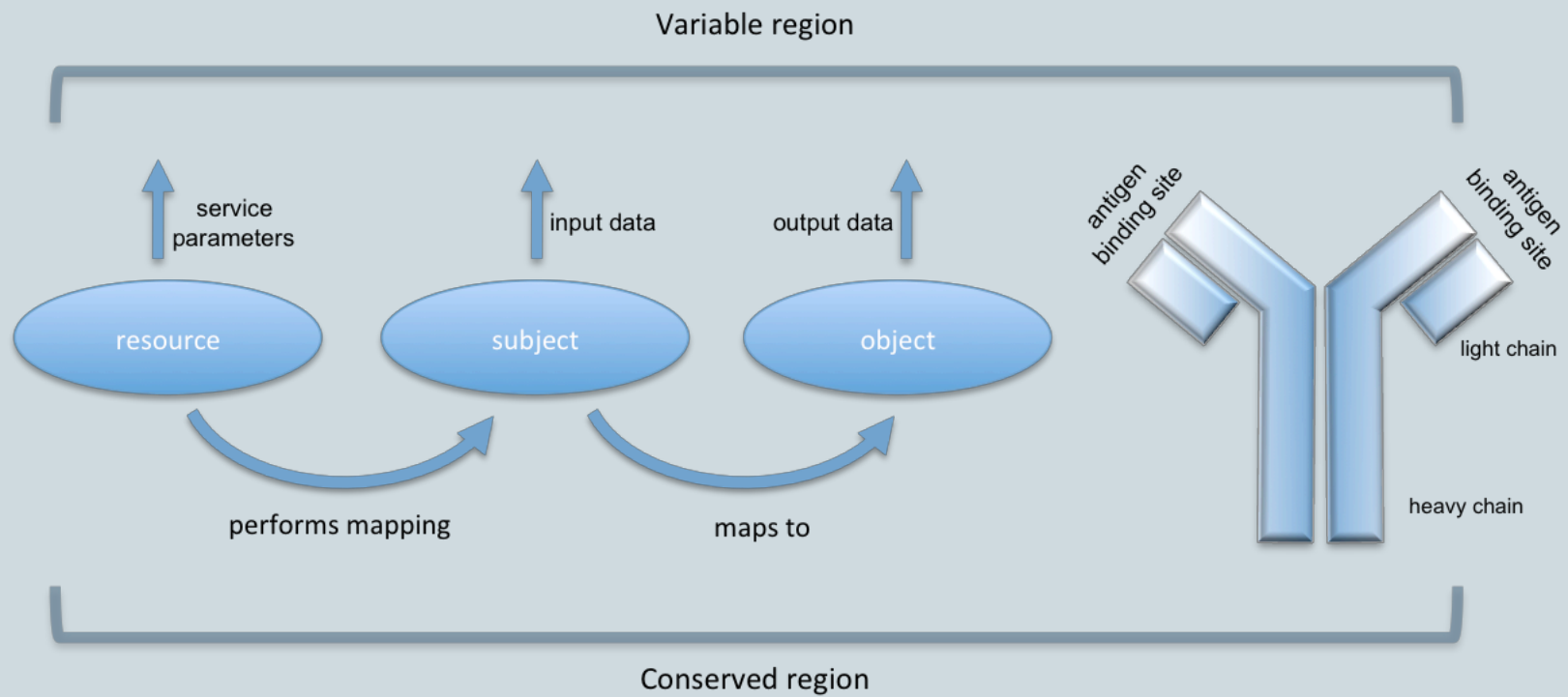


**iPlant Computational
Resources**



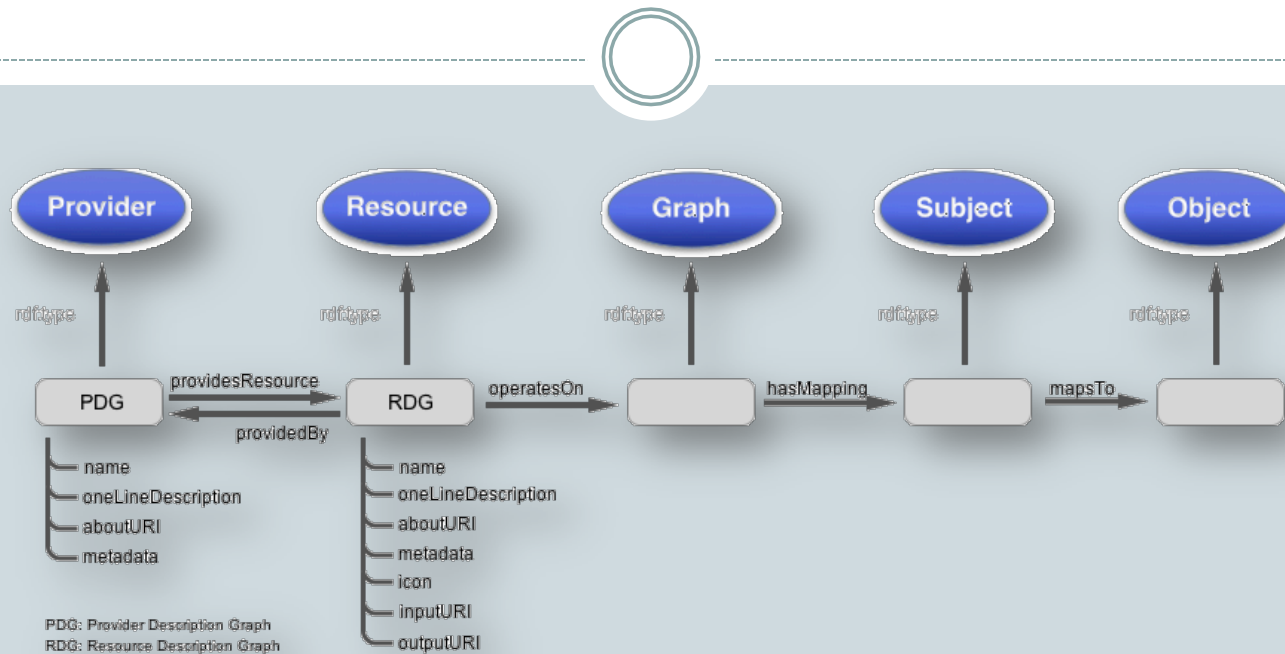
**Community
MODS (Model Organism Databases)
and
CODS (Clade Oriented Databases)**

The Antibody Analogy as a Semantic Mediation Layer



Simple Semantic Web Architecture and Protocol

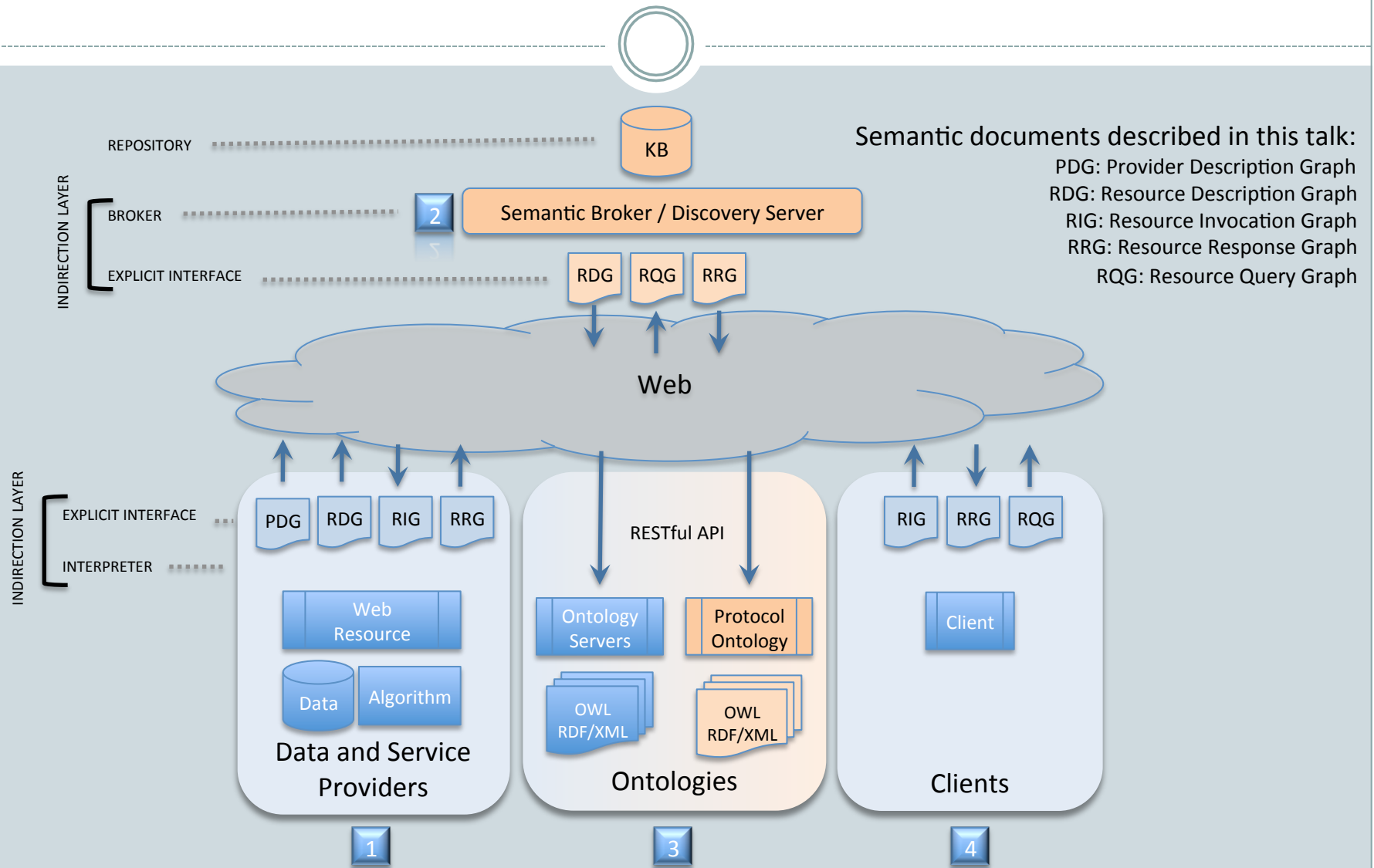
W3C OWL RDF/XML



- Establish the framework for web resources to describe themselves and their offerings
- Establish the framework for ontological integration
- Engage first-order, description logic reasoning
- Provide a semantically enabled Discovery Server for service and pipeline coordination

<http://sswap.info/protocol>

iPlant Semantic Architecture



sswap.info/example



Semantic Integration from Third-party Web sites

TreeGenes' DiversiTree



The screenshot shows the DiversiTree web application interface. The search query is 'rs|t_contig_||||copper ion binding|||Abal,Lade,Paab,Pice,Piel,Pila,Pimg,Pipn,Pira,Pisy,Pita,Psmc'. The results table is as follows:

ID	Contig Name	Species	GO Term	Has Alignments	Has SNPs
75	0_10068	Pinus taeda (Loblolly Pine)	copper ion binding	Yes	Yes
1697	0_11582	Pinus taeda (Loblolly Pine)	copper ion binding	Yes	Yes
4147	0_13881	Pinus taeda (Loblolly Pine)	copper ion binding	Yes	Yes
4953	0_14634	Pinus taeda (Loblolly Pine)	copper ion binding	Yes	Yes
8140	0_17607	Pinus taeda (Loblolly Pine)	copper ion binding	Yes	Yes
8866	0_1829	Pinus taeda (Loblolly Pine)	copper ion binding	Yes	Yes
9909	0_2414	Pinus taeda (Loblolly Pine)	copper ion binding	Yes	Yes
10776	0_367	Pinus taeda (Loblolly Pine)	copper ion binding	Yes	Yes
10957	0_3937	Pinus taeda (Loblolly Pine)	copper ion binding	Yes	Yes
11256	0_4342	Pinus taeda (Loblolly Pine)	copper ion binding	Yes	Yes
13739	0_7878	Pinus taeda (Loblolly Pine)	copper ion binding	Yes	Yes
14065	0_8291	Pinus taeda (Loblolly Pine)	copper ion binding	Yes	Yes
14200	0_8419	Pinus taeda (Loblolly Pine)	copper ion binding	Yes	Yes
14320	0_8531	Pinus taeda (Loblolly Pine)	copper ion binding	Yes	Yes
15122	0_9288	Pinus taeda (Loblolly Pine)	copper ion binding	Yes	Yes
15266	0_9423	Pinus taeda (Loblolly Pine)	copper ion binding	Yes	Yes
22822	2_1343	Pinus taeda (Loblolly Pine)	copper ion binding	Yes	Yes
24776	2_3296	Pinus taeda (Loblolly Pine)	copper ion binding	Yes	Yes
30763	2_8946	Pinus taeda (Loblolly Pine)	copper ion binding	Yes	Yes
cg47012	CL1367Contig1	Pinus taeda (Loblolly Pine)	copper ion binding	Yes	Yes
cg49217	CL3147Contig1	Pinus taeda (Loblolly Pine)	copper ion binding	Yes	Yes
cg50235	CL4447Contig1	Pinus taeda (Loblolly Pine)	copper ion binding	Yes	Yes
cg50336	CL4575Contig1	Pinus taeda (Loblolly Pine)	copper ion binding	Yes	Yes
cg50472	CL4726Contig1	Pinus taeda (Loblolly Pine)	copper ion binding	Yes	Yes

The interface includes a search bar, a 'Display' button, and a list of species to filter by. A yellow circle highlights the 'View Selected' button at the bottom of the table.

Javascript snippet to launch data for Web Discovery with the press of a button

Web Discovery into Semantic Pipelines

Reasoner-assisted Web workflows



The screenshot shows the sswap.info web interface. At the top, there is a navigation bar with the logo and the text "sswap.info beta iPlant Collaborative™". Below this, there is a "New pipeline" section with a "TreeGenes Contig" icon and a "Drag service icon here" area. A "Login" link is also present. The "Input Data Set" section shows a "TreeGenes Contig" icon and a "Display Data" link. Below this, there is a search bar with "Clear filter" and "Filter results" buttons. The search results show three services:

Name	Description	Service URI	Provider
TreeGenes' Tree Sample Service	A service that accepts a Tree Sample Id, such as "P1cb-1" or "LA0E0725" and returns a Tree Sample record. The service also accepts Contig. Amplicon	http://sswap.dendrome.ucdavis.edu/resource/treeSampleService/TreeSampleService	TreeGenes' Semantic Web Service Provider
TreeGenes' Contig Service	A service that accepts a Contig ID, such as "0_10022" and returns a Contig record.	http://sswap.dendrome.ucdavis.edu/resource/contigService/ContigService	TreeGenes' Semantic Web Service Provider
Soybase locus Type Service	Returns a non-redundant list of all soybean genetic map loci contained in Soybase.	http://sswap-a.iplantcollaborative.org/sswap-pipeline-test/test/data/pipeline/SoybaseLocusTypeService	Test provider

Reasoner uses first-order, description logic to present services and pipelines that can operate on the data at any given step

Just-In-Time Ontology Hosting

sswap.info/jit



SSWAP Just-In-Time Ontol x

← → ↻ 🏠 sswapmeet.sswap.info/jit/

sswap.info beta
Simple Semantic Web Architecture and Protocol
iPlant Collaborative™

Just-In-Time Ontology Editor [Sign in](#) [New user? Register](#)

Navigation Section

http://sswapmeet.sswap.info/treeGenes/

Browse Type Property

- treeGenes
 - alignment
 - Alignment
 - hasAlignment
 - id
 - length

Term Section

Alignment Add New Tab

http://sswapmeet.sswap.info/treeGenes/alignment/Alignment.json

```
{  
  "api" : "/makeType",  
  "prefix" : {  
    "treeGenes" : "http://sswapmeet.sswap.info/tre  
    "sequence" : "http://sswapmeet.sswap.info/sequ  
  },  
  "treeGenes:alignment/Alignment" : {  
    "rdfs:label" : "Alignment",  
    "rdfs:comment" : "TreeGenes notion of an Align  
    "treeGenes:alignment/id" : { },  
    "treeGenes:alignment/length" : { "treeGenes:align
```

http://sswapmeet.sswap.info/treeGenes/alignment/Alignment

```
<rdf:RDF  
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-  
  syntax-ns#"  
  xmlns:sswap="http://sswapmeet.sswap.info/sswap/"  
  xmlns:owl="http://www.w3.org/2002/07/owl#"  
  
  xmlns:sequence="http://sswapmeet.sswap.info/sequence/  
  "  
  
  xmlns:treeGenes="http://sswapmeet.sswap.info/treeGene  
  s/"  
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"  
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-  
  schema#">  
  <owl:Class
```

RESTful Pipeline Execution



sswap.iplantcollaborative.org/ipc/aa053bd-c804-4f3b-acb9-682478ce9146

sswap.info
Single Semantic Web Architecture and Protocol
iPlant Collaborative™

New pipeline - running Login

TreeGenes Contig → TreeGenes Tree Sample

TreeGenes' Tree Sample Service

Name: TreeGenes' Tree Sample Service
Description: A service that accepts a Tree Sample Id, such as "Fib-1" or "LA0E0725" and returns a Tree Sample record. The service also accepts: Contig, AmplIcon, A
Service URL: <http://sswap.dendrome.usdavis.edu/resources/treesampleService/treeSampleService>
Provider: TreeGenes' Semantic Web Service Provider
Time started: October 10, 2012 3:29:40 PM MST
Time finished: N/A

Clear Filter Filter results

Showing 0 results

Data Tree View

“Ontologized” data and metadata



The screenshot displays the SSWAP Data Tree View interface. At the top, the URL is sswap.iplantcollaborative.org/ipc/b618b478-c94b-4747-9a8e-752f09188d07. The page features the SSWAP logo and navigation options like 'New pipeline - executed' and 'Login / Register'. A workflow diagram shows 'TreeGenes Cards' leading to 'TreeGenes Tree Sample' and then to 'MyData'. The main content is a 'Data Set View' window containing a table of tree samples.

Data tree	Type	Country	County	Elevation	Genotyped	Latitude	Longitude	Provi
um:treeGenes:treeSample.id:01011	Tree Sample	UNITED STATES	MERIWETHER		No			PIED
um:treeGenes:treeSample.id:02040	Tree Sample	UNITED STATES	SUSSEX	1113	No	44.933 N	121.984 W	COA
um:treeGenes:treeSample.id:06014-331	Tree Sample	UNITED STATES	GREENSVILLE		No			Provi
um:treeGenes:treeSample.id:06042	Tree Sample	UNITED STATES	GREENSVILLE		No			PIED
um:treeGenes:treeSample.id:07056	Tree Sample	UNITED STATES	WILLIAMSBURG		No			COA
um:treeGenes:treeSample.id:08103	Tree Sample	UNITED STATES	ONSLOW		No			COA
um:treeGenes:treeSample.id:09010	Tree Sample	UNITED STATES	ANSON		No			PIED
um:treeGenes:treeSample.id:10-Italy	Tree Sample	ITALY			No			
um:treeGenes:treeSample.id:10005	Tree Sample	UNITED STATES	JASPER		No			COA
um:treeGenes:treeSample.id:101029	Tree Sample	UNITED STATES	LIBERTY		No			COA
um:treeGenes:treeSample.id:11011-089	Tree Sample	UNITED STATES	LIBERTY	120	No	42.506 N	124.266 W	Provi
um:treeGenes:treeSample.id:12021-0153	Tree Sample	UNITED STATES	LIBERTY	240	No	44.405 N	123.636 W	Provi
um:treeGenes:treeSample.id:121057	Tree Sample	UNITED STATES	TALLAPOOSA		No			PIED
um:treeGenes:treeSample.id:14xP	Tree Sample				No			
um:treeGenes:treeSample.id:17016	Tree Sample	UNITED STATES	SUMTER		No			COA
um:treeGenes:treeSample.id:18001	Tree Sample	UNITED STATES	EDGEFIELD		No			PIED
um:treeGenes:treeSample.id:19002	Tree Sample	UNITED STATES	TISHOMINGO		No			PIED
um:treeGenes:treeSample.id:191-Finland	Tree Sample	FINLAND			No			
um:treeGenes:treeSample.id:2002019	Tree Sample	FRANCE			No			
um:treeGenes:treeSample.id:2005014	Tree Sample	FRANCE			No			
um:treeGenes:treeSample.id:219	Tree Sample	FINLAND			No			
um:treeGenes:treeSample.id:22005	Tree Sample	UNITED STATES	LEVY		No			FLOP
um:treeGenes:treeSample.id:23009	Tree Sample	UNITED STATES	JASPER		No			COA
um:treeGenes:treeSample.id:25	Tree Sample	POLAND			No			
um:treeGenes:treeSample.id:25624	Tree Sample	UNITED STATES	MADISON		No			MAPI

SSWAP enables on-the-fly Data Tree views

Semantic Integration into Third-party Web sites

TreeGenes' CartograTree



CartograTree
A Forest Tree Map Utility

Search for ID [Search]
*Currently only works for ID searches

Taxa: All

Tree Samples
Sequenced: Yes No
Genotyped: Yes No
Phenotyped: Yes No
Phenotype Search: Search traits here
GPS Resolution: Coordinates Approximation
TRY-DB Markers: TRY-DB:

Environmental
 AmeriFlux Sites
 WorldClim Points

Legend
Selection Tools: [Color Swatches] Delete Selected Shape

Layers
Repository: Layer: --Select Layer-- Layer Info
Opacity: 0.8 Add Layer

Type	ID	Lat	Lng	Sequenced	Genotyped	Species
------	----	-----	-----	-----------	-----------	---------

Third-party web sites can engage as renderers on result sets ...

CartograTree

Custom user interface data selection and analysis



CartograTree
A Forest Tree Map Utility

Search
Search for ID: [] Search

Display
Taxa: All

Tree Samples
Sequenced: Yes No
Genotyped: Yes No
Phenotyped: Yes No

Phenotype Search
Search traits here

GPB Resolution
 Coordinates Approximation

TRY-DB Markers
TRY-DB:
Trait Search: Search traits here

Environmental
 Ameriflux Sites

Legend
Selection Tools: [] [] [] [] [] [] Delete Selected Shape

Layers
Repository: [] Select Layer
Opacity: 0.8

Selected Markers Common Amplicon

Summary/Guidelines View

Number	Amplicon Id	Top Blast Description (BLAST nr)	Species-Specific BLASTs	GO Annotations	Interpro Annotations	PFAM Annotations	ExPASy EC Annotations	thmm / SignalP
	UMN_CL228Con	Filter by Keyword	Filter by Keyword	Filter by Keyword	Filter by Keyword	Filter by Keyword	Filter by Keyword	Filter by Keyword
<input checked="" type="checkbox"/> 405	UMN_CL228Contig_03	PREDICTED: 405 ribosomal protein S26-2-like [Vitis vinifera] (7.50E-68) [Vitis vinifera]	PREDICTED: 405 ribosomal protein S26-2-like [Vitis vinifera] (7.50E-68) [Vitis vinifera]	GO:0006412 translation GO:0005840 ribosome GO:0003735 structural constituent of ribosome	IPR000892 Ribosomal protein S26e	PF01283 Ribosomal protein S26e	3.4.5.3 Protein-synthesizing GTPase.	thmm: (no) SignalP: (no)

... enabling directed user experiences back into Web Discovery

Semantic Integration from Third-party Web sites

Data slicing and contextual augmentation

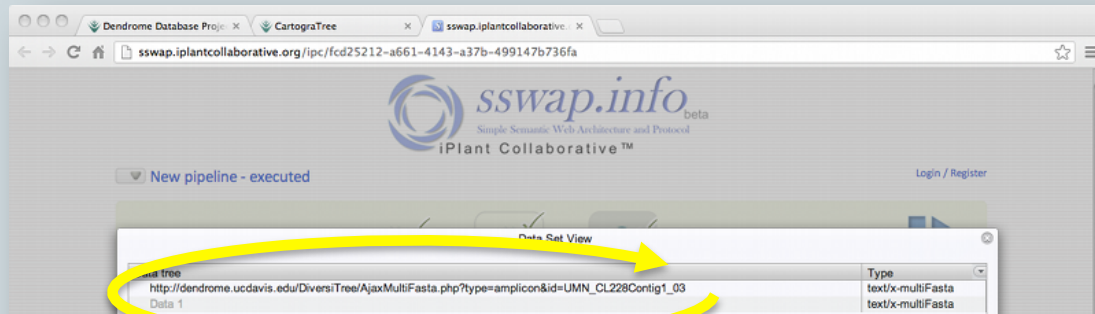


The screenshot displays the sswap.info website interface. At the top, the logo for sswap.info beta (Simple Semantic Web Architecture and Protocol) and iPlant Collaborative™ is visible. Below the logo, there is a "New pipeline" section with a visual workflow diagram showing "TreeGenes Amplicon" and "TreeGenes MultiPars" connected by arrows, followed by a dashed box labeled "Drag service icon here" and a blue play button. To the right of this section are "Login / Register" links. Below the pipeline builder is an "Input Data Set" section with a "TreeGenes Amplicon" icon and text: "Display Data: Click here to send the output data to a viewer (opens in a new window)", "Data Set URI: http://sswap.iplantcollaborative.org/ipc/rrg?token=30f8e273-f7c2-40ee-b9b6-9cceb939750", and "Produced by: http://sswap.dendrome.ucdavis.edu/resources/ampliconService/AmpliconService". Below this is a search bar with the placeholder text "Enter a phrase or user:<login> to filter results" and a "Filter results" button. The search results section shows "Showing 7 results" and lists several services:

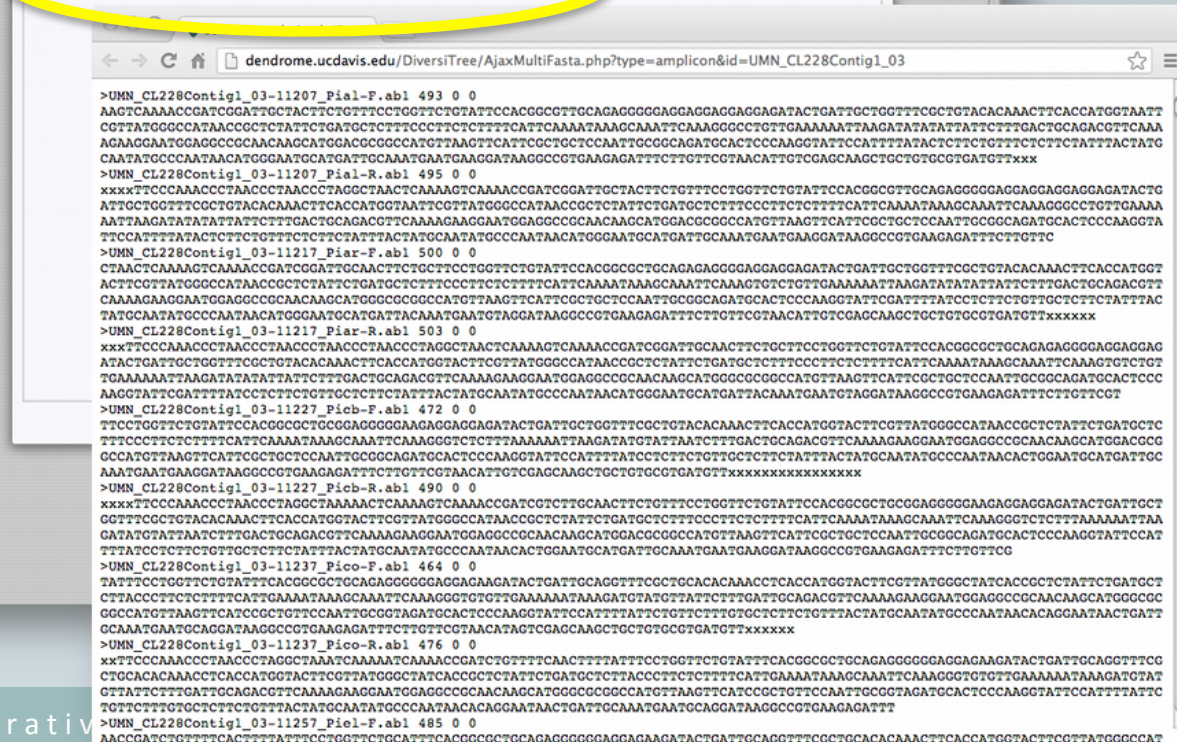
- CBRC**
Name: iPlant MAFFT (Multiple Sequence Alignment) service
Description: Multiple alignment program for amino acid or nucleotide sequences
Service URI: http://sswap.info/iplant/svc/mafft-ionestar-6.864
Provider: The iPlant Collaborative
- MUSCLE**
Name: iPlant Muscle (Multiple Sequence Alignment) service
Description: MUSCLE is a program for creating multiple alignments of amino acid or nucleotide sequences
Service URI: http://sswap.info/iplant/svc/muscle-ionestar-3.8.31
Provider: The iPlant Collaborative
- W2 CLUSTAL**
Name: iPlant ClustalW2 (Multiple Sequence Alignment) service
Description: Multiple alignment of nucleic acid and protein sequences
Service URI: http://sswap.info/iplant/svc/clustalw2-ionestar-2.1
Provider: The iPlant Collaborative
- TreeGenes**
Name: iPlant FastTree service
Description: FastTree infers approximately-maximum-likelihood phylogenetic trees from alignments of nucleotide or protein sequences.
Service URI: http://sswap.info/iplant/svc/fasttree
Provider: The iPlant Collaborative
- TreeGenes**
Name: iPlant grep service

Direct and Indirect Data Referencing

URI dereferencing of arbitrarily large data sets



*Serialize the data itself,
or a URI to where the
data is located*



High Performance Computing

Services engage like any other Web services

The screenshot displays the sswap.info web interface. At the top, the browser address bar shows the URL `sswap.iplantcollaborative.org/ipc/88b2c347-4cc6-4cf5-86ab-c7f4762f1b51`. The page header features the sswap.info logo and the text "Simple Semantic Web Architecture and Protocol" and "iPlant Collaborative™".

The main content area shows a workflow titled "Amplicon UMN_CL18Contig1_08 alignment, phylogenetic tree, and visualization". The workflow consists of several steps: TreeGenes Amplicon, TreeGenes MultiParsa, MUSCLE, a tree visualization icon, and a service icon. A "Logout Logged as sswap" link is visible in the top right.

Below the workflow, the "iPlant TreeViz service" is detailed:

- Name: iPlant TreeViz service
- Description: Tree Visualization service
- Service URI: `http://sswap.info/iplant/svc/treeviz`
- Provider: The iPlant Collaborative
- Time started: N/A
- Time finished: N/A

A search bar with "Clear filter" and "Filter results" buttons is present. Below it, "Showing 4 results" are listed:

- iPlant grep service**: A service that looks for phrases in occurring files stored on the user's iPlant Data Store cloud. Service URI: `http://sswap.info/iplant/svc/grep-ranger-1.0`. Provider: The iPlant Collaborative.
- Extract images from page**: Scans an HTML page and extracts all images from it. Service URI: `http://sswap-a.iplantcollaborative.org/sswap-pipeline-test/test/data/pipeline/extractImagesFromPage`. Provider: Test provider.
- SoyBase Locus Type Service**: Returns a non-redundant list of all soybean genetic map loci contained in SoyBase. Service URI: `http://sswap-a.iplantcollaborative.org/sswap-pipeline-test/test/data/pipeline/SoybaseLocusTypeService`. Provider: Test provider.
- iPlant Data Store**: Uploads data to a staging area in a user's account in the iPlant Data Store.

Publishing Pipelines

Private data, shared service parameterization



The screenshot shows a web browser window with the URL `sswap.iplantcollaborative.org/ipc/04fd8134-1d43-451e-82d4-91fb1c78da85`. The page features a search bar with the text "Enter a phrase or user:<login> to filter results" and a "Filter results" button. Below the search bar, it says "Showing 8 results" and "CLUSTAL".

The main content area displays a list of services, each with an icon, name, description, service URI, and provider:

- iPlant FastTree service**: FastTree infers approximately-maximum-likelihood phylogenetic trees from alignments of nucleotide or protein sequences. Service URI: `http://sswap.info/iplant/svc/fasttree-ranger-2.1.4`. Provider: The iPlant Collaborative.
- iPlant grep service**: A service that looks for phrases in occurring files stored on the user's iPlant Data Store cloud. Service URI: `http://sswap.info/iplant/svc/grep-ranger-1.0`. Provider: The iPlant Collaborative.
- SoyBase Locus Type Service**: Returns a non-redundant list of all soybean genetic map loci contained in SoyBase. Service URI: `http://sswap-a.iplantcollaborative.org/sswap-pipeline-test/test/data/pipeline/SoybaseLocusTypeService`. Provider: Test provider.
- iPlant Data Store**: Uploads data to a staging area in a user's account in the iPlant Data Store. Service URI: `http://sswap.info/iplant/svc/iPlantDataStore`. Provider: The iPlant Collaborative.

At the bottom of the list, there is a "View pipelines" button and a list of pipeline components:

1. iPlant ClustalW2 (Multiple Sequence Alignment) service
2. iPlant FastTree service

Below the list, there are three icons: ClustalW2, iPlant FastTree, and iPlant Data Store. A "View pipelines" button is also present.

At the bottom of the page, there are navigation links: [Home](#), [iPlant](#), [Protocol](#), [Ontologies](#), [API](#), [Wiki](#), and [Contact](#).

Logos for THE UNIVERSITY OF ARIZONA, CLARK PARSIA, and a globe icon are visible at the bottom.

Phylogenetics

Pipeline runs are persisted in OWL and can start new pipelines



The screenshot displays the sswap.info web interface. At the top, the browser address bar shows the URL `sswap.iplantcollaborative.org/ipc/88b2c347-4cc6-4cf5-86ab-c7f4762f1b51`. The page header features the sswap.info logo with the tagline "Simple Semantic Web Architecture and Protocol" and "iPlant Collaborative™".

The main content area shows a completed pipeline titled "Amplicon UMN_CL18Contig1_08 alignment, phylogenetic tree, and visualization - executed". The pipeline steps are visualized as a sequence of icons: TreeGenes Amplicon, TreeGenes MultiParts, MUSCLE, a tree visualization icon, a network diagram icon, and MyData. A "Logout Logged as sswap" link is visible in the top right corner.

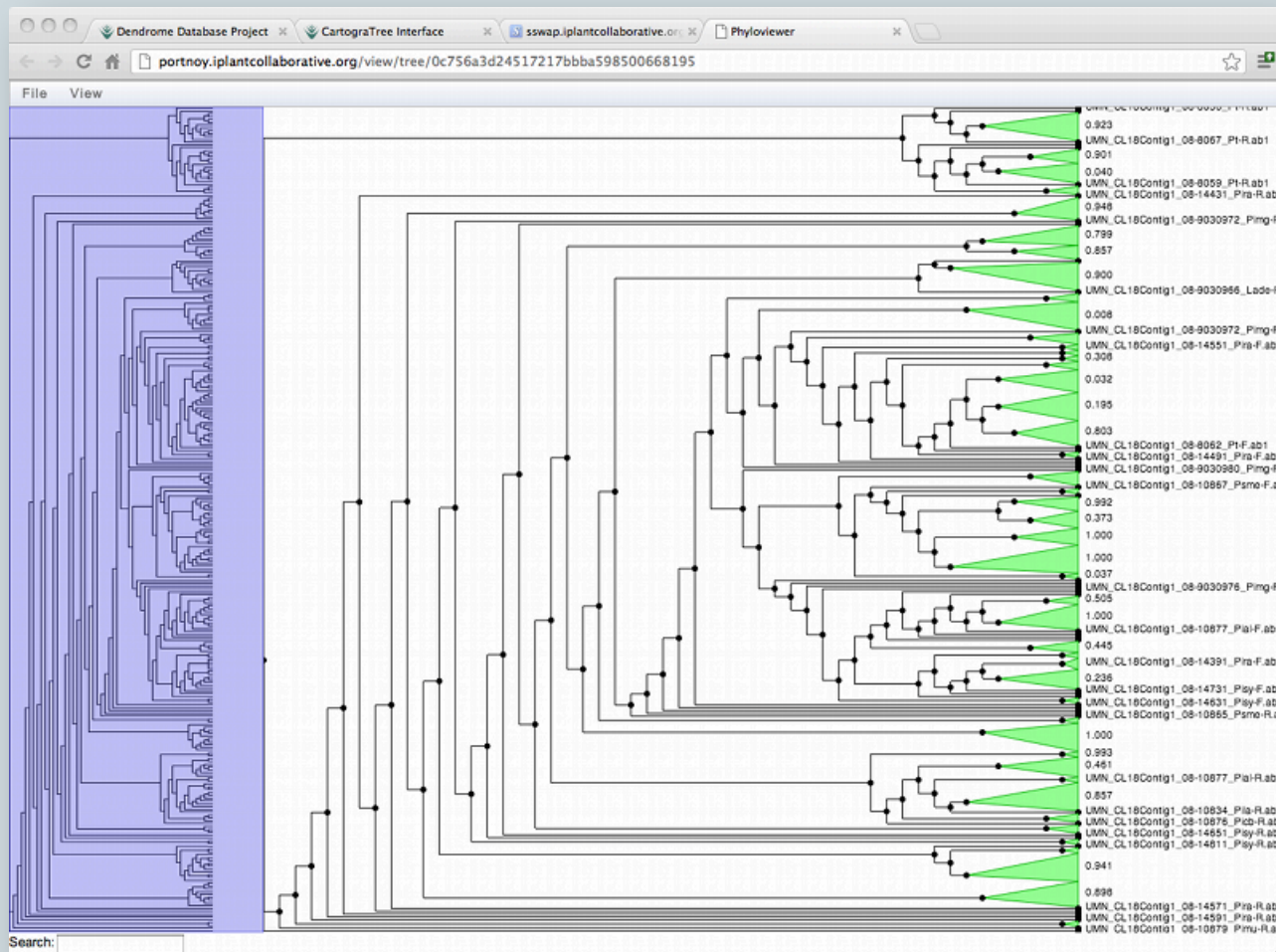
Below the pipeline, an "Output Data Set" section provides the following information:

- Display Data: [Click here to send the output data to a viewer \(opens in a new window\)](#)
- iPlant Path: `/sswap/Amplicon_UMN_CL18Contig1_08_alignment_phylogenetic_tree_and_visualization-2012-10-11-07-53-10.rrg.owl`
- Produced by: `http://sswap.info/iplant/svc/treeviz`

At the bottom of the interface, there is a search bar with a "Clear Filter" button, a text input field containing "Enter filter phrase", and a "Filter results" button. Below the search bar, it indicates "Showing 0 results".

TreeViz

Multi-state, multi-institution, web/HPC run



Quick Vitals



- 185,000 lines of code
- 100+ libraries
- Open source and freely available
- Three 48 GB RAM servers
- More info: sswap.info/wiki

Acknowledgements



Special thanks to:

- iPlant Collaborative
- Semantic Web engineering by Clark and Parsia, LLC
- NSF grants #0943879 and #EF-0735191