

# SSWAP – Simple Semantic Web Architecture and Protocol

Damian Gessler, Ph.D.  
University of Arizona  
Tucson, AZ 85721  
dgessler@iplantcollaborative.org

**Biology is rapidly becoming an information science.** Scientific advances require wet lab and field investigations—indeed, that is the *sine qua non* of science. But informatics is rapidly becoming a necessary, but not sufficient, component of the biological scientific endeavor. Given this, we still find ourselves expending thousands of programmer hours and millions of dollars on tying together low-throughput, one-off, integration operations across the thousands of data and service providers on the web. This wastes effort and hinders discovery.

**This problem comes to a front at the interface of the latest data generation technologies.** High throughput sequencing is currently delivering ~20 Gbp of sequence per machine every few days; capital costs are ~\$500K per machine with hundreds of machines across vendors already operational in the field. Industry estimates are placing throughput estimates up to 95 Gbp per 3-day run by 2010. High throughput sequencing is being used in large projects such as the 1000 Genome Project ([www.1000genomes.org](http://www.1000genomes.org)), as well as small labs resequencing a swath of biology (soy, maize, humans, and many more). We are entering an age of the *commoditization of data generation*, where the increasing emphasis and bottleneck is moving downstream to data handling, manipulation, and informatics. This data will need to be captured and managed, and more importantly, it will need to be contextualized with information from other sources in order to deliver translational value.

**The need for high-throughput data and service integration in biology is imminent, yet technologies to address this are sparse.** "High-throughput" means that we need computers to be able to find data on the web, assess it's relevancy on a per problem basis, and integrate it. "Data and service integration" reflects the synthetic and synergistic processes that must occur in order to transform data into information, and information into knowledge. High-throughput data and service integration automates the upstream data collection and initial analysis processes, while placing humans farther downstream into key creative and integrative knowledge areas. In today's informatic landscape, these steps occur on the backbone of the World Wide Web. Protocols such as HTTP (the web), FTP (file transfer), and SMTP/IMAP/POP3 (email) allow the efficient transfer of data and engagement of services, but they, in-and-of-themselves, do not enable high-throughput discovery, assessment, and integration.

**Neither traditional web service nor the semantic web technologies deliver a sufficient pure-play technology for high throughput integration.** Web services address high throughput engagement and data transfer, but are deficient in a high-throughput semantic that is needed for computers to be able to discover and assess data and services based on their contextual relevancy. Semantics (from the Greek *semaino* "to mean"; *semantikos* "significant") is aimed at allowing data and services on the web to self-describe themselves, such that a computer can discover and engage them based on their contextual suitability for purpose. Semantic web technologies, such as the W3C-sanctioned OWL, have a rich semantic, but are deficient in service protocols. What is needed is a hybrid technology that brings a rich, high-throughput semantic to a semantic web service model to allow high throughput semantic discovery and engagement. SSWAP—an acronym for Simple Semantic Web Architecture and Protocol—is a technology that addresses these deficiencies.

**SSWAP (Simple Semantic Web Architecture and Protocol) combines web service functionality with an extensible semantic framework to satisfy the conditions for high-throughput integration.** SSWAP (see <http://sswap.info>; pronounced "swap" as in to "swap info") is a semantic web services technology. As an application of OWL to the web services model, it is built upon a layered approach to distributed web information management. SSWAP defines an OWL *ontology* (a *controlled vocabulary* where some terms describe the relationships of other terms to each other) specifically designed to allow web resources to describe themselves; to enable you to query on those resources; to engage them; and to semantically encode the result. Because SSWAP is based on OWL, SSWAP resources are amenable to automated reasoning, including a powerful feature called semantic searching. SSWAP defines terms such as what it means to be a web resource, who provides that resource, and how the resource maps its input to its output data.

**SSWAP allows the community to use peer-reviewed ontological standards such as the Gene Ontology, the Sequence Ontology, and virtually any Open Biomedical Ontology or even third-party custom ontologies in a semantic web services framework.** SSWAP allows resources—data and services—to describe themselves in terms of publicly available, third-party ontologies. For example, SSWAP allows the use of refactored OBO (Open Biomedical Ontology) ontologies in semantic web services, while also allowing third-party extension of those ontologies or *de novo* introduction of new ontologies. SSWAP's use of the first-order description logic formalism of OWL gives a structured method for new concepts to be introduced without breaking old concepts—an important property for extensibility on the web. As new concepts are introduced under a formal semantics (e.g., `rdfs:subClassOf`), it is explicitly, logically clear to a reasoner (or a human) exactly how the new concept is an extension of the old. This aids in classifying both data and services.

**Reasoners can deduce subclass relations (so-called *subsumption* statements) based on the properties of data and services, providing a critical enhancement over lexically-based searching.** SSWAP allows logical integration over the web based on the properties and class membership of data and services. We use the open-source reasoner Pellet (<http://pellet.owdl.com>) to generate a set of inferred statements, allowing us to perform semantic searching based on subsumption relations, even across unclassified web resources. Because SSWAP derives subsumption based on logical deductions, rather than lexical associations and heuristics, we extend the search scope while mitigating false positives.

**SSWAP is running at <http://sswap.info>.** SSWAP defines a protocol (<http://sswap.info/protocol.jsp>), offers a semantic web service ontology portal (<http://sswapmeet.sswap.info>), has on-line examples (<http://sswap.info/examples/README.jsp>) and developer tools (<http://www.sswap.info/developer.jsp>), and a "Semantic Google"-like interface for resource discovery (<http://sswap.info>). SSWAP is open-source (<http://sourceforge.net/projects/sswap>) and completely free.

**SSWAP has significant broader impacts.** Think of SSWAP as: SSWAP = Semantics + Web Services. The ability to semantically search for resources and engage them based on the contextual relations of data and service types has broad value across the sciences and the web. The technology moves the web towards a distributed, logical network amenable to machine reasoning. For more information please email Damian Gessler at [dgessler@iplantcollaborative.org](mailto:dgessler@iplantcollaborative.org).