SBOL Stack: The One-stop-shop for Storing and Publishing SBOL Data

Curtis Madsen[1], Goksel Misirli[1], Matthew Pocock[1,2], Jennifer Hallinan[1], and Anil Wipat[1]

[1]School of Computing Science, Newcastle University, Newcastle upon Tyne, UK

[2]Turing Ate My Hamster Ltd., Newcastle upon Tyne, UK

August 20, 2014
COMBINE Meeting

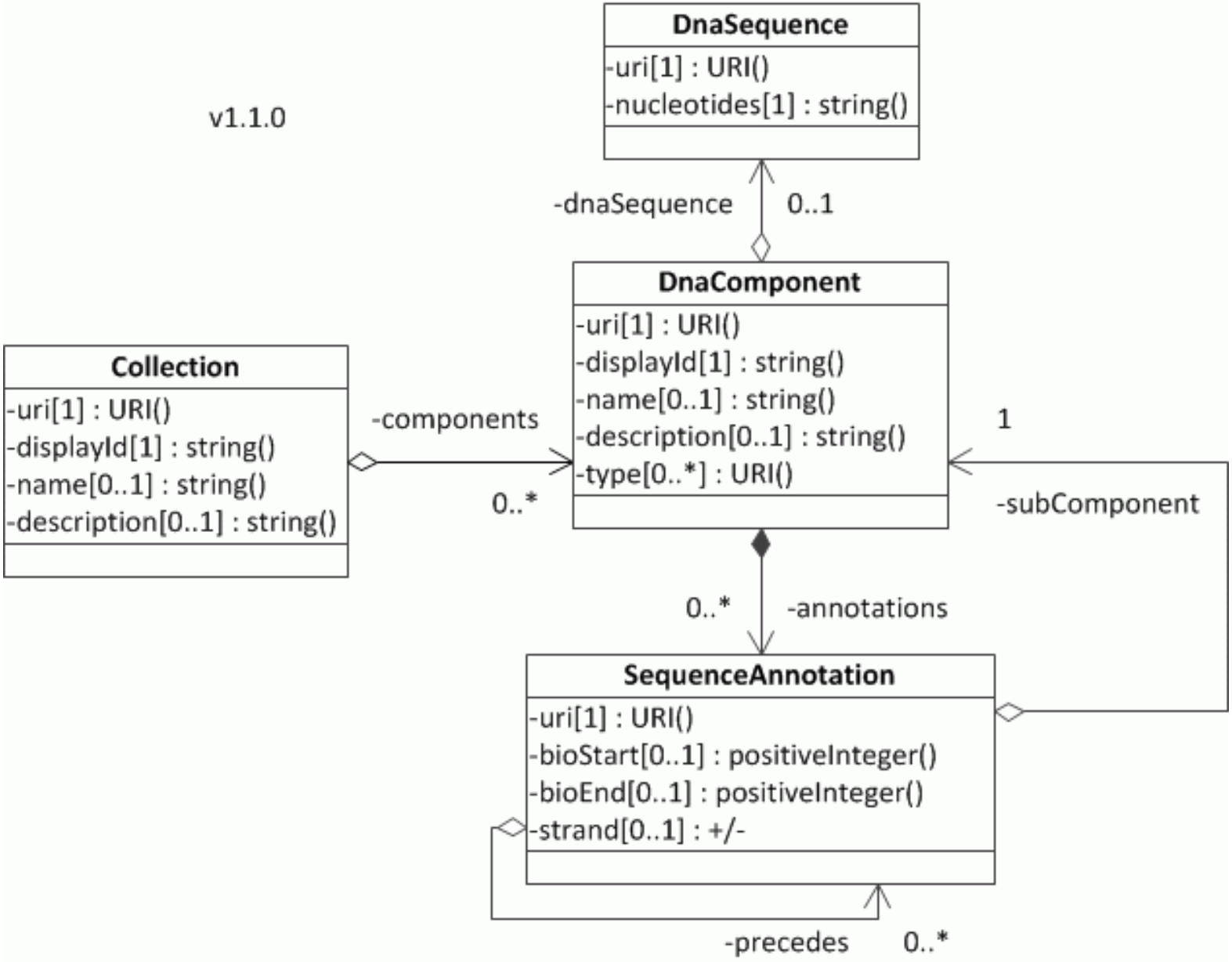
Synthetic Biology

- Synthetic biology is a growing field that combines ideas from biology and engineering.
- The goal of synthetic biology is to design and build new useful biological systems.
- It is, however, often difficult to utilize the extensive amount of biological data for design in synthetic biology:
  - Efforts are usually individual and carried out by teams in different geographic locations.
  - The interests of researchers can vary greatly.
  - Typically, biological data relevant to the design of genetic circuits is not exchanged.

Synthetic Biology Open Language (SBOL)

- To aid in the interpretation and exchange of biological information, standards are necessary.
- An emerging standard in synthetic biology is the *Synthetic Biology Open Language* (SBOL):
  - Designed to allow for the exchange of descriptions of genetic parts, devices, modules, and systems.
  - Facilitates storage of genetic designs in repositories.
  - Allows for designs of genetic parts and systems to be embedded in publications.

Synthetic Biology Open Language (SBOL)

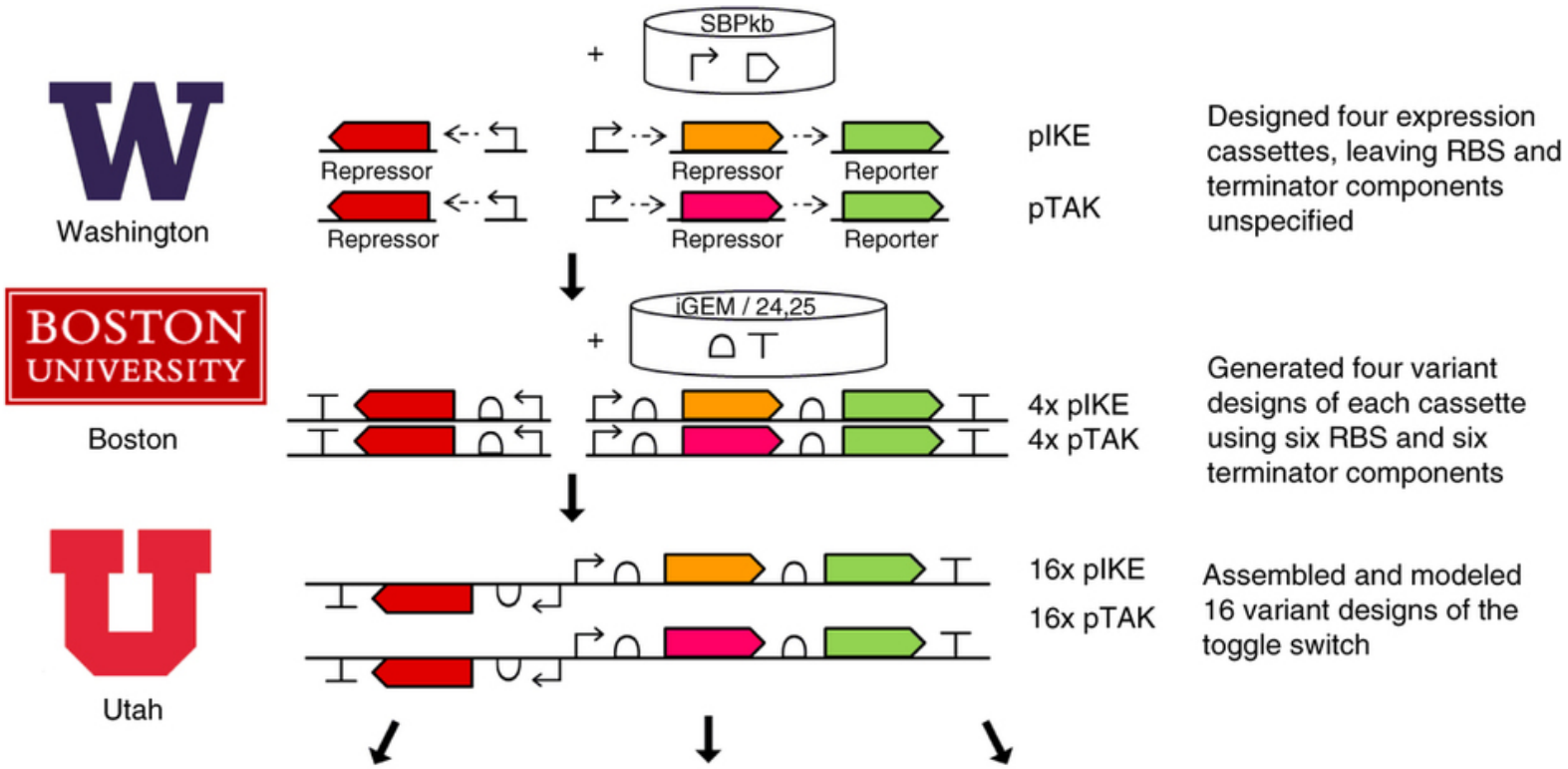


SBOL Workflows

- SBOL can be used to create workflows between different tools and repositories from any number of organizations:
  - CAD tools such as `iBioSim` and `TinkerCell`.
  - Repositories such as JBEI-ICE and the Virtual Parts Repository.
  - Sequencing tools such as `Vector NTI Express Designer`.
- For instance, in a recent *Nature Biotechnology* paper[1], six independent groups collaborated on the design of a set of genetic toggle switches using several SBOL enabled tools.

[1] Galdzicki *et al.*, "The Synthetic Biology Open Language (SBOL) provides a community standard for communicating designs in synthetic biology," Nature Biotechnology, vol. 32, iss. 6, pp. 545-550, 2014.

SBOL Workflows



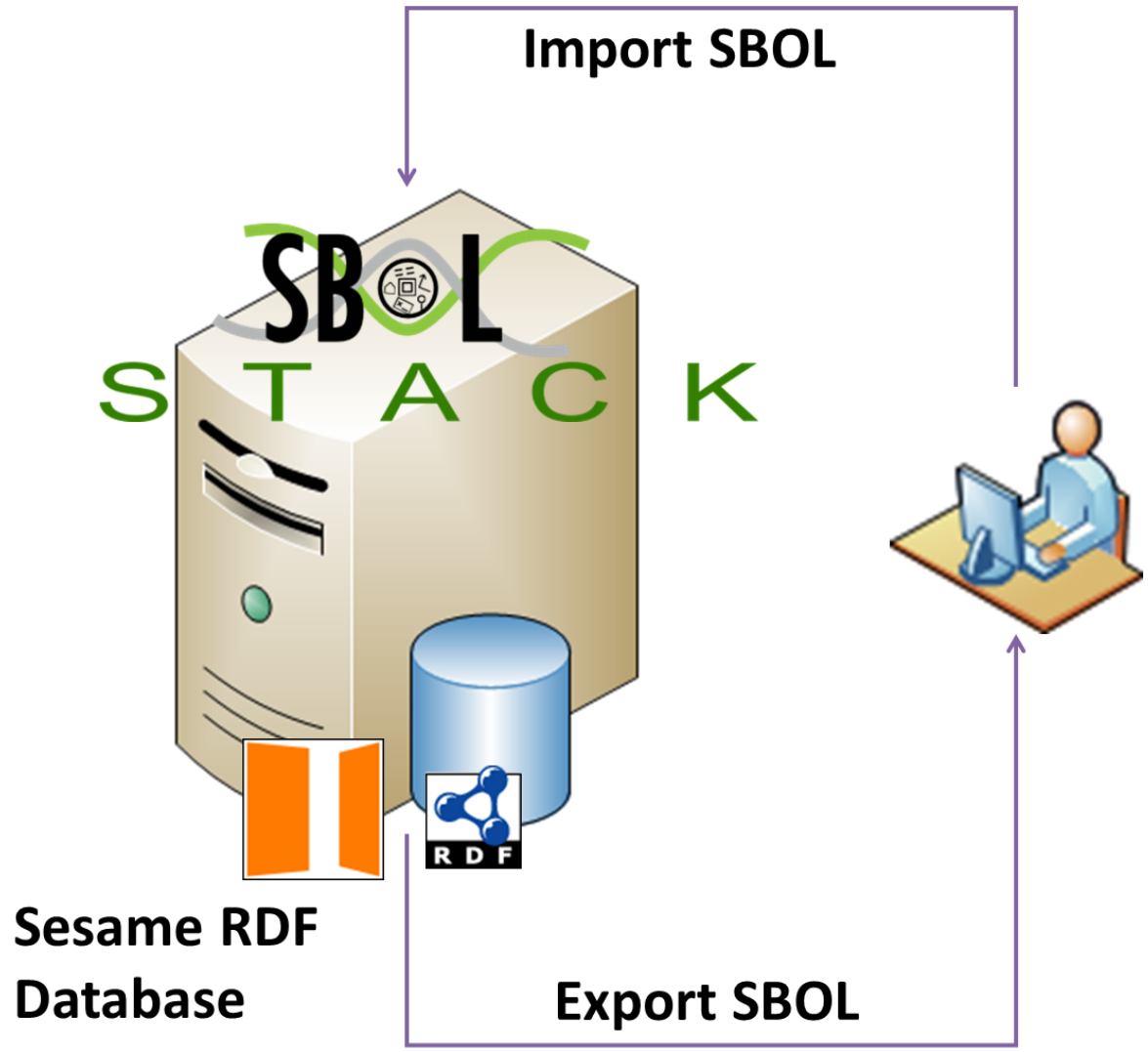
Semantic Web

- The SBOL language is based on RDF/XML and can be stored in triplestore repositories where it can be searched using SPARQL queries.
- Additionally, SBOL automatically integrates into Semantic Web technologies allowing for linking to other biological data not contained within SBOL.
- However, most existing repositories do not take advantage of this infrastructure and simply store individual SBOL files.

Newcastle SBOL Stack

- We have developed the SBOL Stack to allow researchers to better store, retrieve, exchange, and publish SBOL data.
- The SBOL Stack is a Sesame RDF database specifically designed for:
  - Publishing a library of synthetic parts and designs as a service.
  - Sharing SBOL with collaborators.
  - Storing designs of biological systems locally.
- Additionally, it includes a Web client that allows for uploading, downloading, and visualizing SBOL data using SPARQL queries.

SBOL Stack

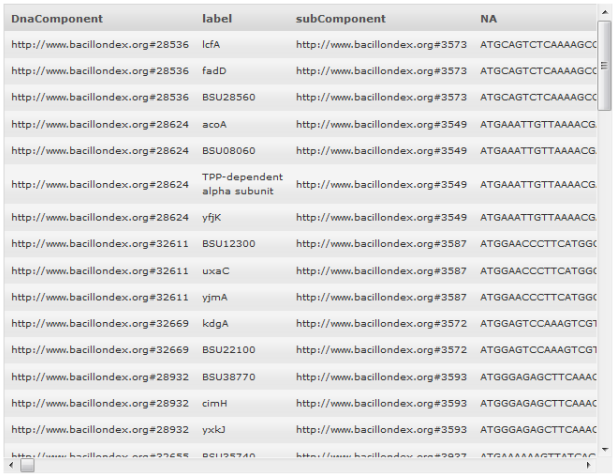


Uploading/Downloading SBOL Data

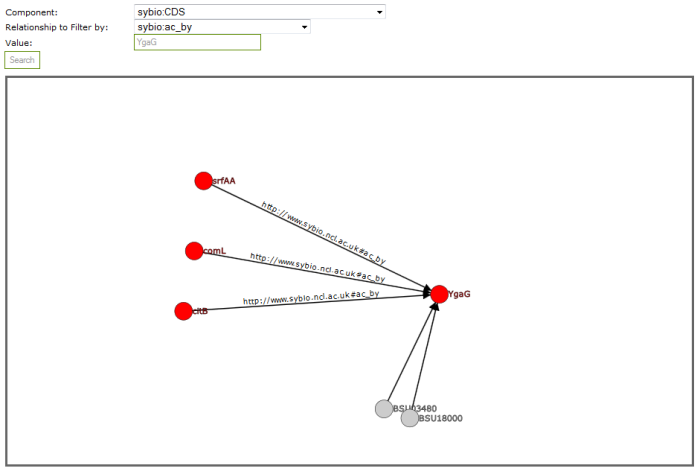
- SBOL data described in XML can be uploaded to the SBOL Stack through the Web interface.
- Queries can be performed to retrieve and download the SBOL data using either the Web interface or the SPARQL endpoint.
- These queries allow users of the SBOL Stack to retrieve only desired parts of the SBOL data.

SPARQL Queries

- SPARQL queries can produce either tables or graphs.
- A table query results in a set of tuples (or variable bindings) and is commonly used to get specific values (URIs, blank nodes, literals) from the stored RDF data.
- A graph query, on the other hand, returns a graph of RDF triples in the form: *Subject* $\rightarrow$ *Predicate* $\rightarrow$ *Object*.



DnaComponent	label	subComponent	NA
http://www.bacillondex.org#28536	lcfA	http://www.bacillondex.org#3573	ATGCAGTCTCAAAAGCC
http://www.bacillondex.org#28536	fadD	http://www.bacillondex.org#3573	ATGCAGTCTCAAAAGCC
http://www.bacillondex.org#28536	BSU28560	http://www.bacillondex.org#3573	ATGCAGTCTCAAAAGCC
http://www.bacillondex.org#28624	acoA	http://www.bacillondex.org#3549	ATGAAATTGTTAAAACG
http://www.bacillondex.org#28624	BSU08060	http://www.bacillondex.org#3549	ATGAAATTGTTAAAACG
http://www.bacillondex.org#28624	TPP-dependent alpha subunit	http://www.bacillondex.org#3549	ATGAAATTGTTAAAACG
http://www.bacillondex.org#28624	yfjK	http://www.bacillondex.org#3549	ATGAAATTGTTAAAACG
http://www.bacillondex.org#32611	BSU12300	http://www.bacillondex.org#3587	ATGGAACCCTTCATGGC
http://www.bacillondex.org#32611	uxaC	http://www.bacillondex.org#3587	ATGGAACCCTTCATGGC
http://www.bacillondex.org#32611	yjmA	http://www.bacillondex.org#3587	ATGGAACCCTTCATGGC
http://www.bacillondex.org#32669	kdgA	http://www.bacillondex.org#3572	ATGGAGTCCAAAGTCGT
http://www.bacillondex.org#32669	BSU22100	http://www.bacillondex.org#3572	ATGGAGTCCAAAGTCGT
http://www.bacillondex.org#28932	BSU38770	http://www.bacillondex.org#3593	ATGGGAGAGCTTCAAAC
http://www.bacillondex.org#28932	cimH	http://www.bacillondex.org#3593	ATGGGAGAGCTTCAAAC
http://www.bacillondex.org#28932	yxkJ	http://www.bacillondex.org#3593	ATGGGAGAGCTTCAAAC



Example Table Query

SPARQL Query:

```
PREFIX  rdfs:  <http://www.w3.org/2000/01/rdf-schema#>
PREFIX  owl:   <http://www.w3.org/2002/07/owl#>
PREFIX  rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX  xsd:   <http://www.w3.org/2001/XMLSchema#>
PREFIX  sbol:  <http://sbols.org/sbol.owl#>
PREFIX  so:    <http://purl.org/obo/owl/SO#>
PREFIX  sybio: <http://www.sybio.ncl.ac.uk#>

SELECT ?DnaComponent  ?label ?subComponent ?NA
WHERE
{
?DnaComponent rdf:type sbol:DnaComponent ;
              rdf:type so:SO_0000316 ;
              rdfs:label ?label ;
              sbol:dnaSequence ?dnaSequence ;
              sbol:annotation ?SequenceAnnotation.
?dnaSequence sbol:nucleotides ?NA ;
             a sbol:DnaSequence.
?SequenceAnnotation sbol:subComponent ?subComponent.
}
```

SPARQL Query

Example Table Query

DnaComponent	label	subComponent	NA
http://www.bacillondex.org#28536	lcfA	http://www.bacillondex.org#3573	ATGCAGTCTCAAAAGCC
http://www.bacillondex.org#28536	fadD	http://www.bacillondex.org#3573	ATGCAGTCTCAAAAGCC
http://www.bacillondex.org#28536	BSU28560	http://www.bacillondex.org#3573	ATGCAGTCTCAAAAGCC
http://www.bacillondex.org#28624	acoA	http://www.bacillondex.org#3549	ATGAAATTGTTAAAACG
http://www.bacillondex.org#28624	BSU08060	http://www.bacillondex.org#3549	ATGAAATTGTTAAAACG
http://www.bacillondex.org#28624	TPP-dependent alpha subunit	http://www.bacillondex.org#3549	ATGAAATTGTTAAAACG
http://www.bacillondex.org#28624	yfjK	http://www.bacillondex.org#3549	ATGAAATTGTTAAAACG
http://www.bacillondex.org#32611	BSU12300	http://www.bacillondex.org#3587	ATGGAACCCTTCATGGC
http://www.bacillondex.org#32611	uxaC	http://www.bacillondex.org#3587	ATGGAACCCTTCATGGC
http://www.bacillondex.org#32611	yjmA	http://www.bacillondex.org#3587	ATGGAACCCTTCATGGC
http://www.bacillondex.org#32669	kdgA	http://www.bacillondex.org#3572	ATGGAGTCCAAAGTCGT
http://www.bacillondex.org#32669	BSU22100	http://www.bacillondex.org#3572	ATGGAGTCCAAAGTCGT
http://www.bacillondex.org#28932	BSU38770	http://www.bacillondex.org#3593	ATGGGAGAGCTTCAAAC
http://www.bacillondex.org#28932	cimH	http://www.bacillondex.org#3593	ATGGGAGAGCTTCAAAC
http://www.bacillondex.org#28932	yxkJ	http://www.bacillondex.org#3593	ATGGGAGAGCTTCAAAC
http://www.bacillondex.org#32655	BSU35740	http://www.bacillondex.org#3937	ATGAAAAAGTTATCAC

SPARQL Graph Queries

- Since graph queries return well formed RDF, the graph can be downloaded and used in another tool that supports RDF data.
- As the SBOL language is RDF, these types of queries can return SBOL data.
- The SBOL Stack has been optimized for these types of queries and contains a search option that automatically performs graph queries without the need to write SPARQL directly.
- All a user needs to specify is a type of Subject, a type of Predicate, and an optional Object.

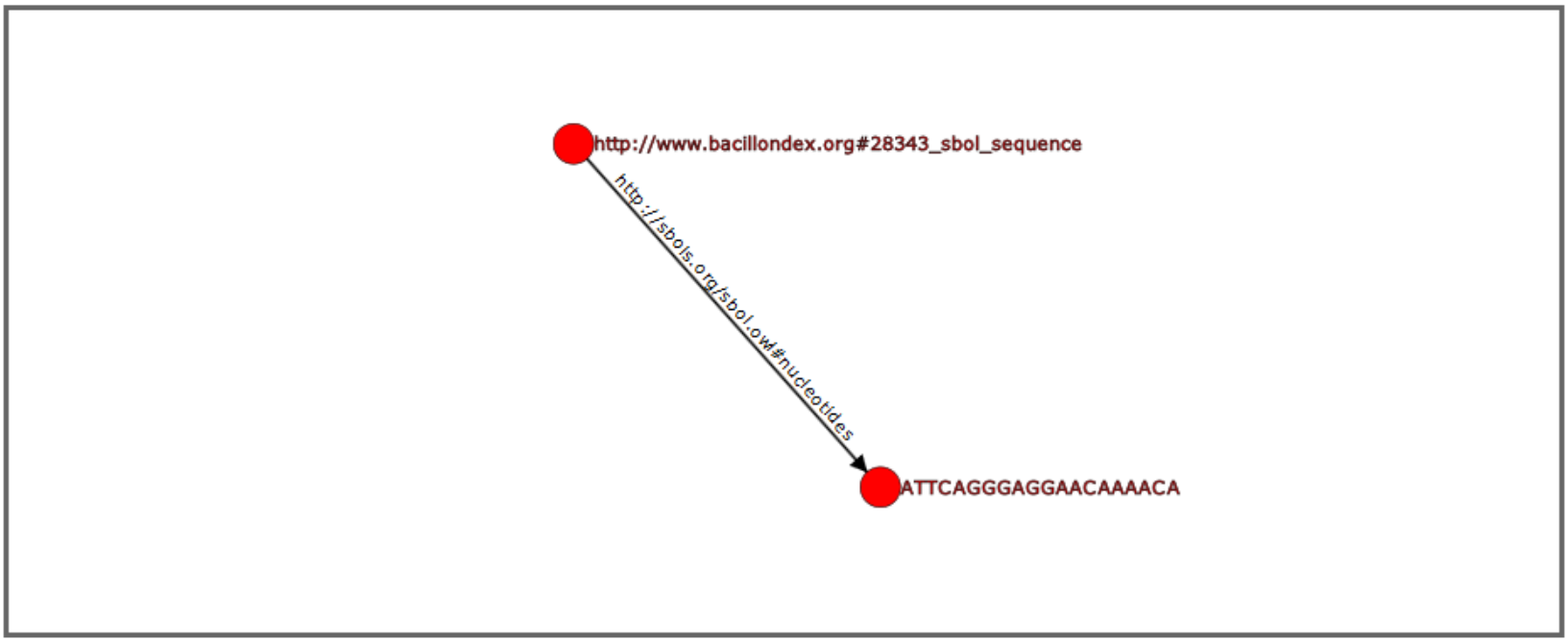
Example Graph Query

Component: sbol:DnaSequence

Relationship to Filter by: sbol:nucleotides

Value (Optional): ATTCAGGGAGGAACAAAACA

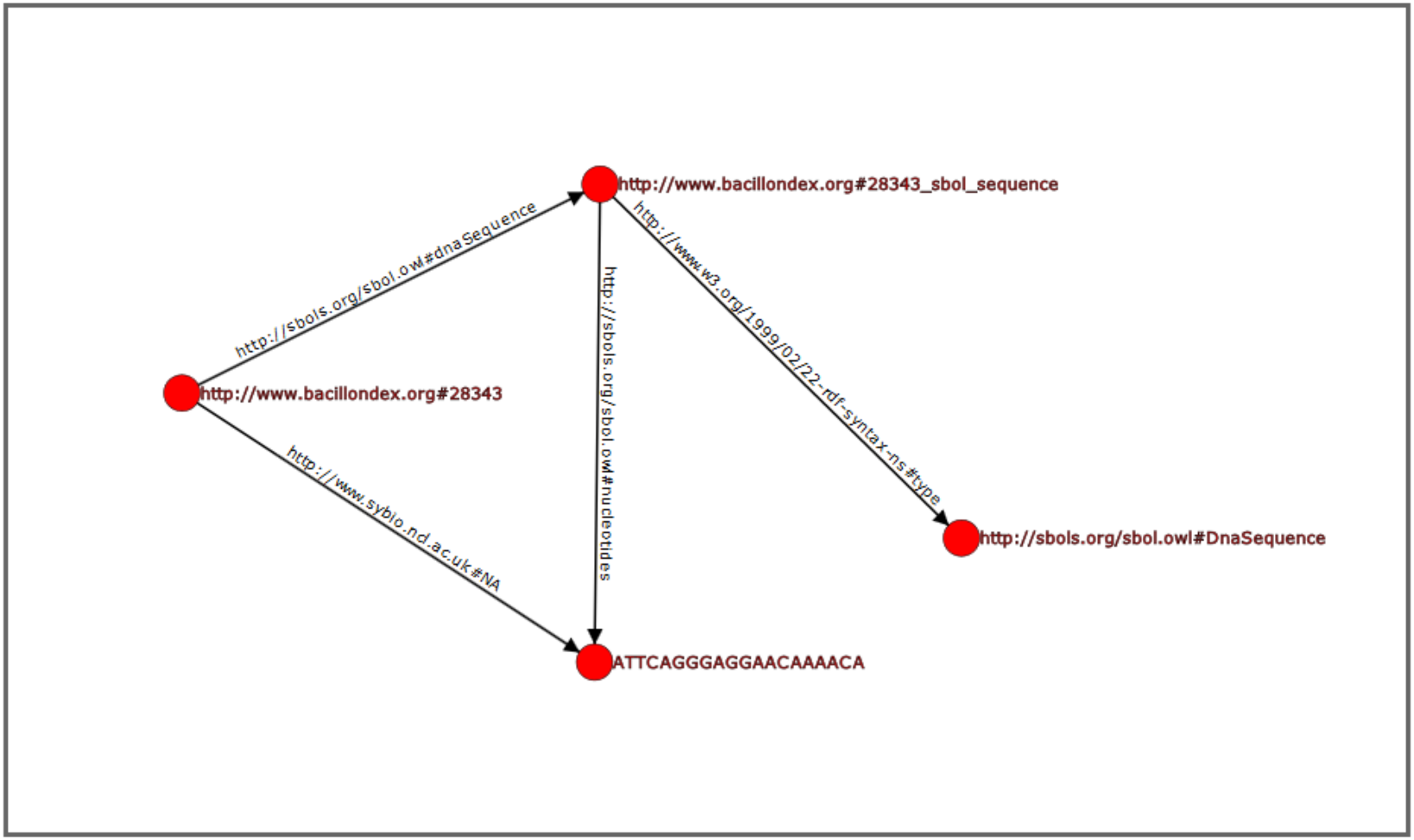
Search



Searching Selected Nodes

- The SBOL Stack contains the option to further search selected nodes in the results of a graph query.
- This option is very useful for someone who is interested in searching for more data about a particular element.
- For example, a user may be interested in:
  - Interactions with other elements;
  - The DNA sequence on the element; or
  - Components the element is contained within among others.

Example Searching Selected Nodes

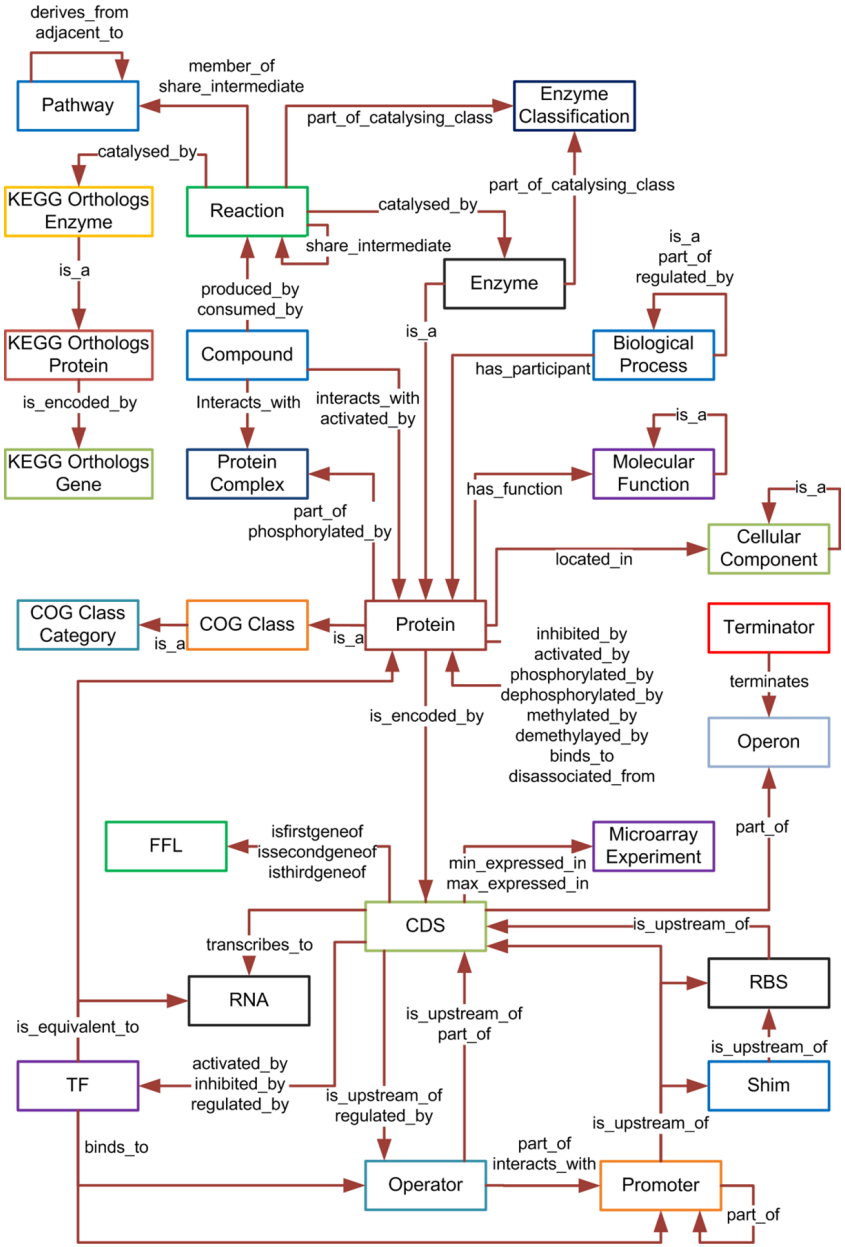


Data Integration

- It can be difficult to identify the mapping between a nucleotide sequence and information such as biological function.
- Information about genetic features and their biological constraints is usually spread amongst many databases.
- Since SBOL is based on RDF, it is ideal for data integration and can easily be linked to other RDF data.
  - Some examples include integrating with ontologies such as the Sequence Ontology and the Gene Ontology.

Custom Ontologies

- In addition to SBOL, the SBOL Stack includes an ontology about genetic features, gene products and their annotations, gene regulatory networks, metabolic pathways, and so on.
  - It is possible to include other custom ontologies in the SBOL Stack such as BioPAX.
- Biological entities can be mapped to SBOL objects using the ontology to enrich the data.
- The data model from the ontology can be used to automate the identification of biological parts via SPARQL queries.


derives_from
adjacent_to
Pathway
member_of
share_intermediate
part_of_catalysing_class
Enzyme Classification
catalysed_by
KEGG Orthologs Enzyme
Reaction
catalysed_by
share_intermediate
part_of_catalysing_class
Enzyme
is_a
part_of
regulated_by
is_a
produced_by
consumed_by
KEGG Orthologs Protein
Compound
is_a
has_participant
Biological Process
is_encoded_by
Interacts_with
interacts_with
activated_by
is_a
KEGG Orthologs Gene
Protein Complex
has_function
Molecular Function
part_of
phosphorylated_by
is_a
located_in
Cellular Component
COG Class Category
is_a
COG Class
is_a
Protein
Terminator
inhibited_by
activated_by
phosphorylated_by
dephosphorylated_by
methylated_by
demethylayed_by
binds_to
disassociated_from
terminates
is_encoded_by
Operon
FFL
isfirstgeneof
issecondgeneof
isthirdgeneof
Microarray Experiment
part_of
min_expressed_in
max_expressed_in
CDS
is_upstream_of
transcribes_to
RBS
RNA
is_upstream_of
part_of
is_equivalent_to
is_upstream_of
TF
activated_by
inhibited_by
regulated_by
is_upstream_of
regulated_by
Shim
is_upstream_of
binds_to
part_of
interacts_with
Operator
Promoter
part_of

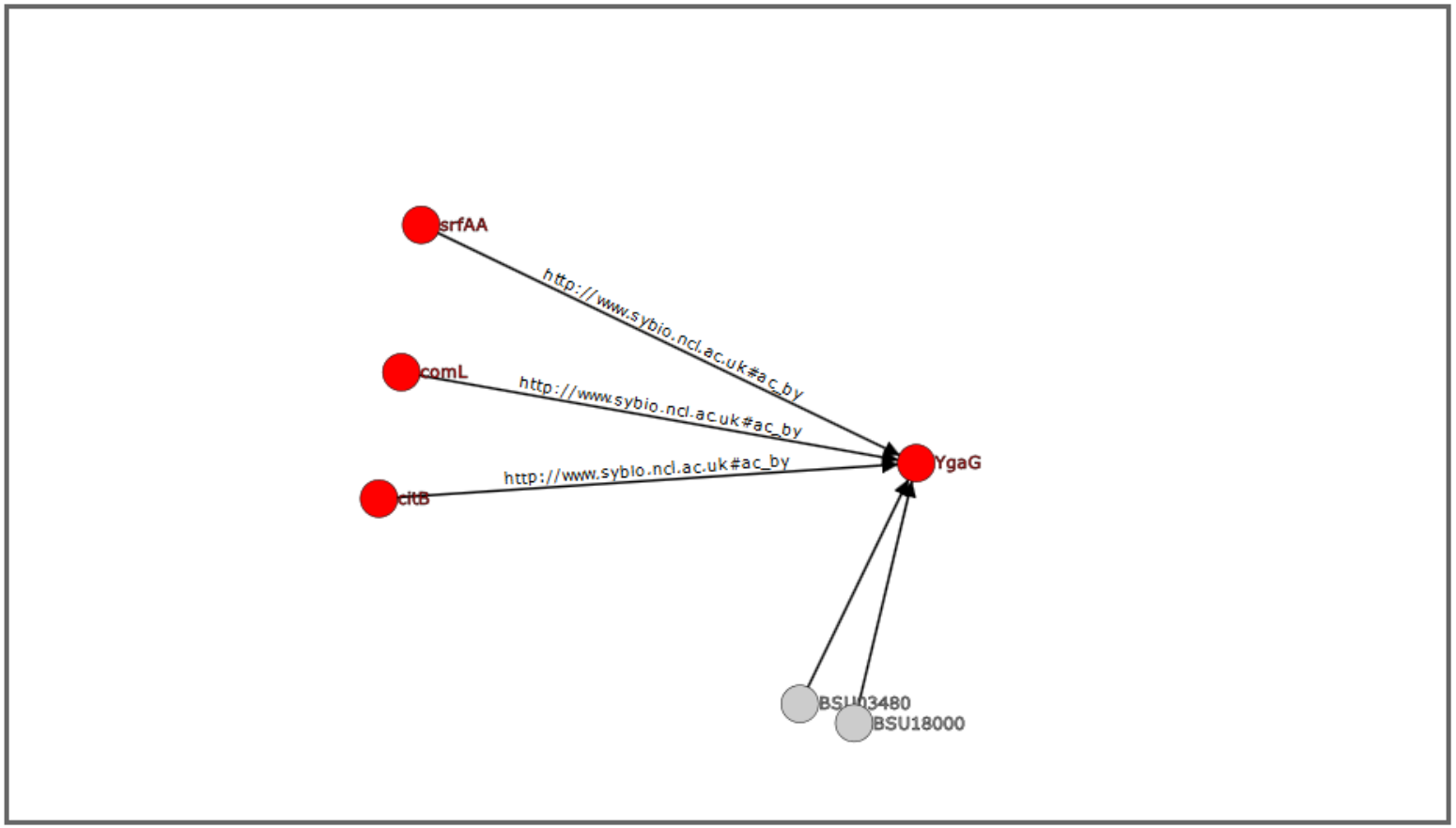
Example SynBiOnt Query

Component: sybio:CDS

Relationship to Filter by: sybio:ac_by

Value: YgaG

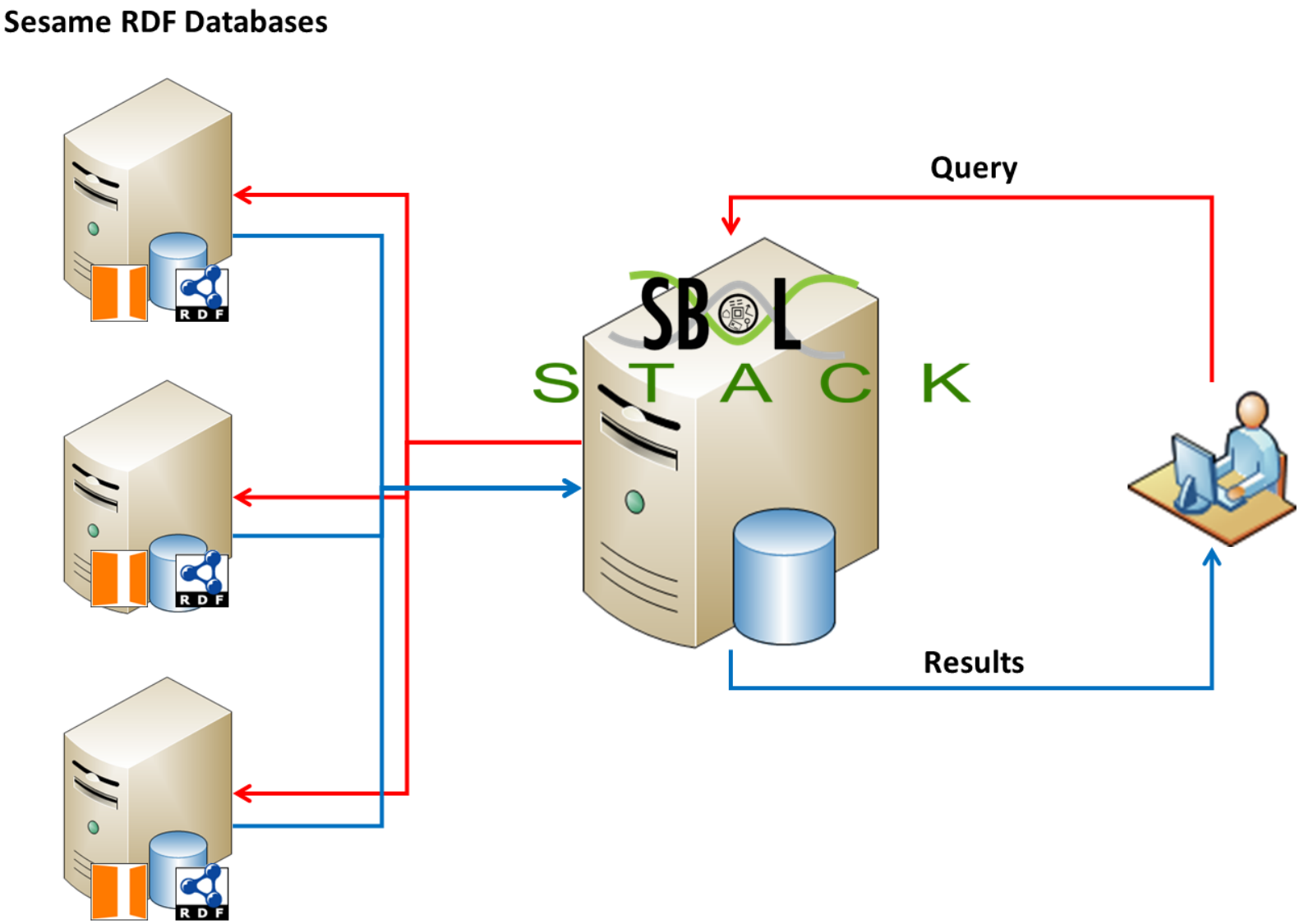
Search



Federated Querying

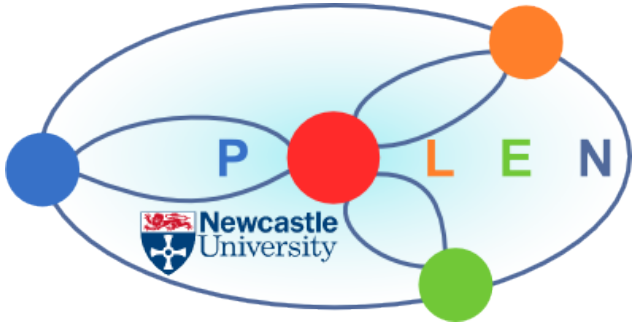
- One of the strengths of the SBOL Stack is its ability to register many Sesame RDF databases and perform federated queries for data integration.
- These queries allow for the retrieval and compilation of more complete data from multiple databases without the need to manually query each individual repository.
- Repositories that contain any RDF information about biological parts can be included in the federated queries.
- The SBOL Stack is compliant with and integrates seamlessly with Web2.0 resources.

Federated Querying



Automatic Updated Data Integration

- The SBOL Stack automatically retrieves the most up-to-date data each time a query is made.
- The SBOL Stack can also take advantage of the cloud-based messaging system, POLEN (PrOtocol for Linking External Nodes) developed for the Flowers Consortium[1], to receive a notification when data are updated.



[1] http://www.synbiuk.org/

SBOL Stack Webpage

Discussion

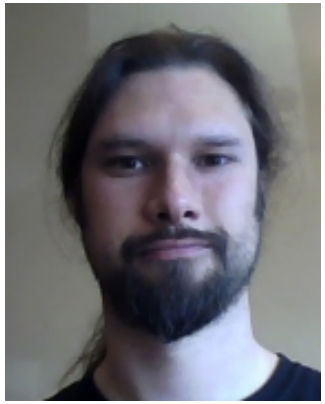
- As more biological data are generated, it will become essential to adopt standards and repositories so computer applications can communicate and exchange data efficiently and in an automated manner.
- Tools and repositories that support standards like SBOL will be required to create workflows for design in synthetic biology.
- The automatic retrieval and integration of SBOL data provided by the SBOL Stack makes it a must-have tool for synthetic biology workflows.

Ongoing Work

- We have begun adding SBOL 2.0 triples to the SBOL Stack.
  - We plan to fully upgrade the SBOL Stack to store SBOL 2.0 data once the specification has been finalized.
- The SBOL Stack can be accessed computationally as a SPARQL endpoint.
  - However, we are extending the API to allow for computational tools to access the SBOL Stack using the direct search interface that is utilized by the Web client.
  - This API will eliminate the need to write raw SPARQL queries.

Acknowledgements

Goksel Misirli



Matthew Pocock

Jennifer Hallinan

Anil Wipat

I(C$_2$)$_2$S

FLOWERS Consortium

Newcastle University