Standards for Data Integration in Synthetic Biology

James Alastair McLaughlin; Göksel Mısırlı; Anil Wipat

11 Oct 2017

What do I want to build?

Which parts do I need?

How do those parts interact with each other?

How do those parts interact with the host chassis?

Is there experimental evidence for the interactions?

Where can I find structural information about the parts?

Where can I find functional information about the parts?

How can I integrate this information into my design?

Do I need to convert between data formats?

How can I convert between data formats?

Are the differences syntactic or semantic?

disparity

unify syntax

unify semantics

distribution

data warehouse

federation - **scale**

complexity

data mining

visualization

intractability

linked data

graph queries

disparity

unify syntax

unify semantics

distribution

data warehouse

federation - **scale**

complexity

data mining

visualization

intractability

linked data

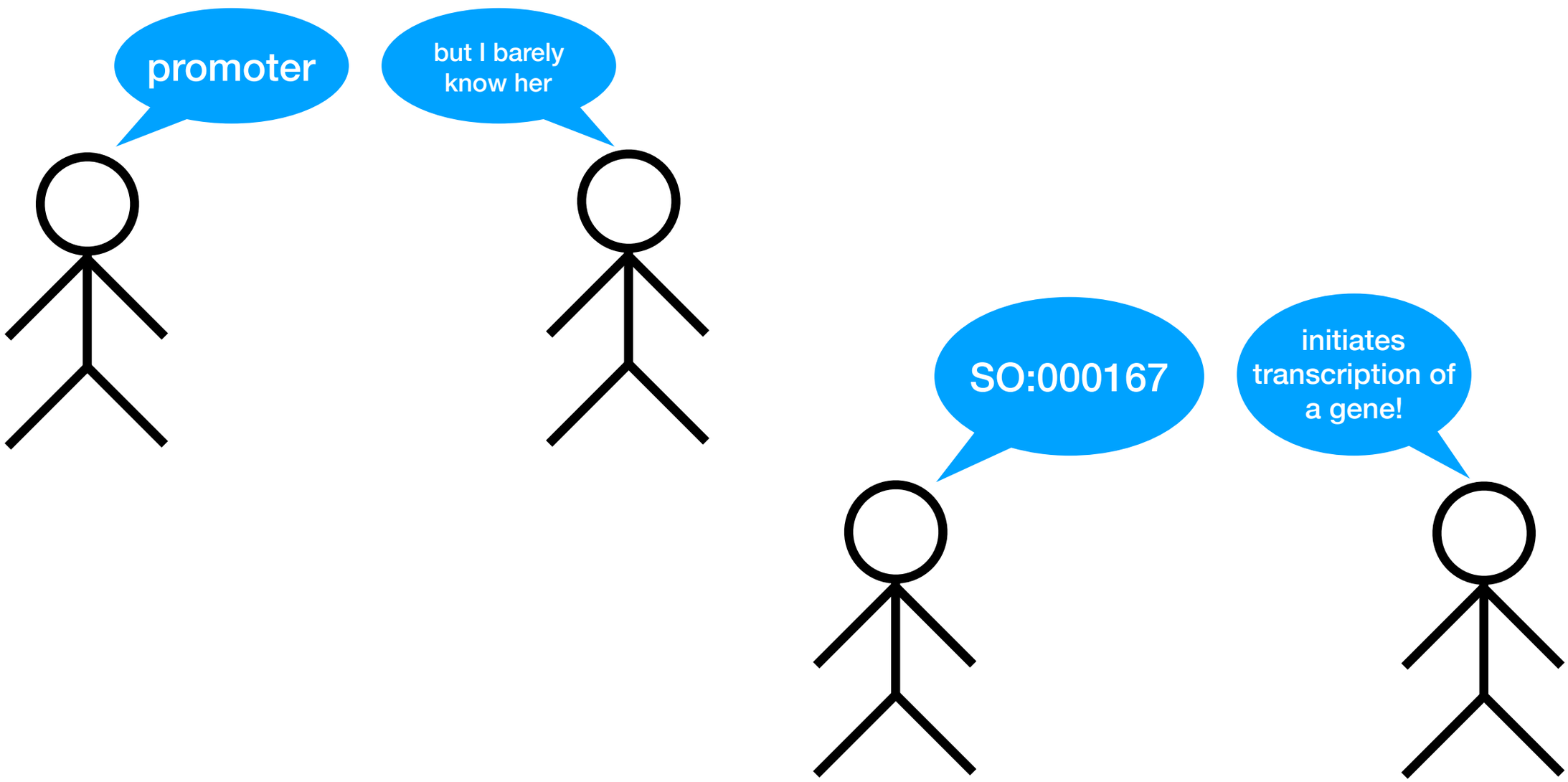
graph queries

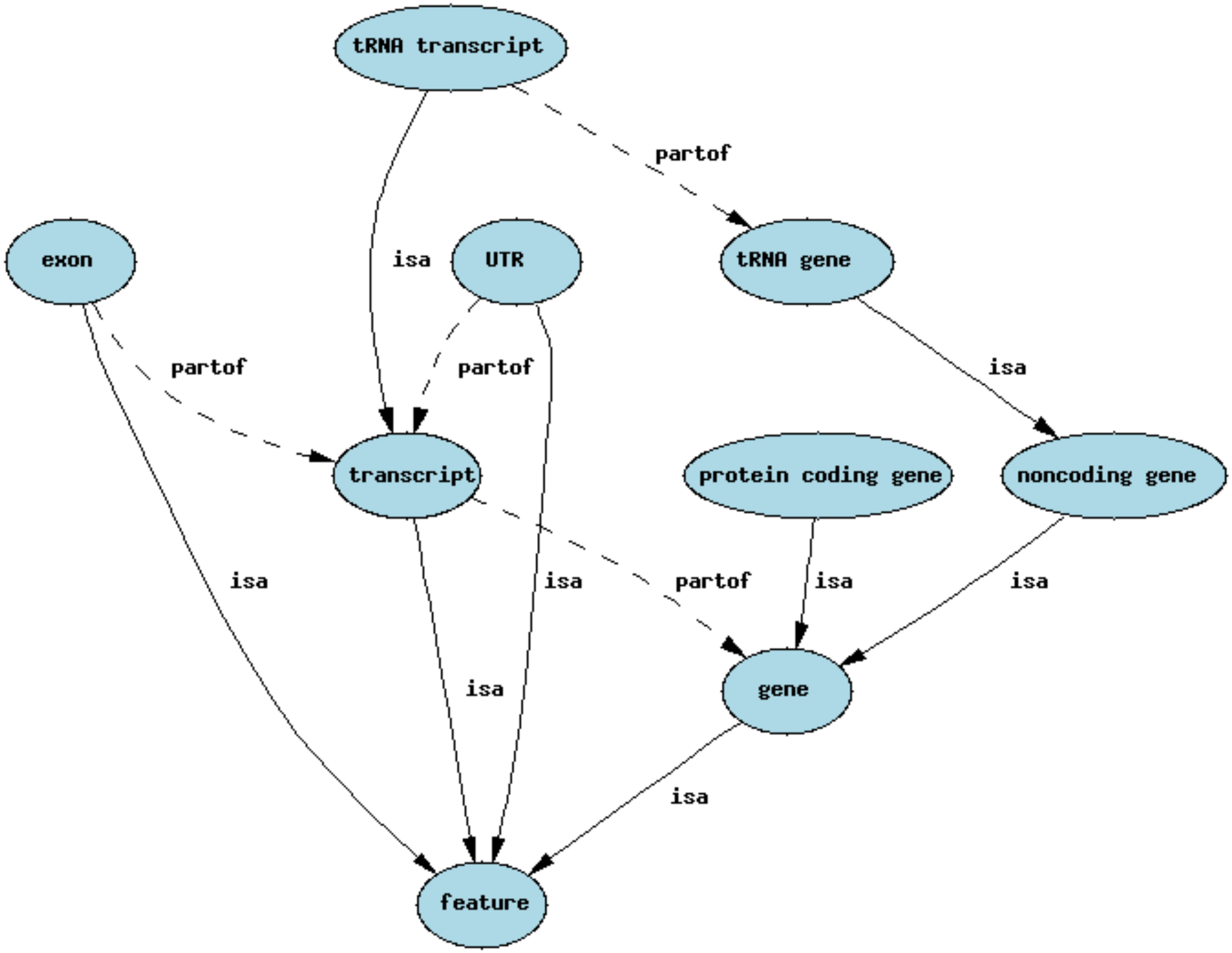
Sometimes human language is ambiguous….

So for complex domains like biology, we create **ontologies**

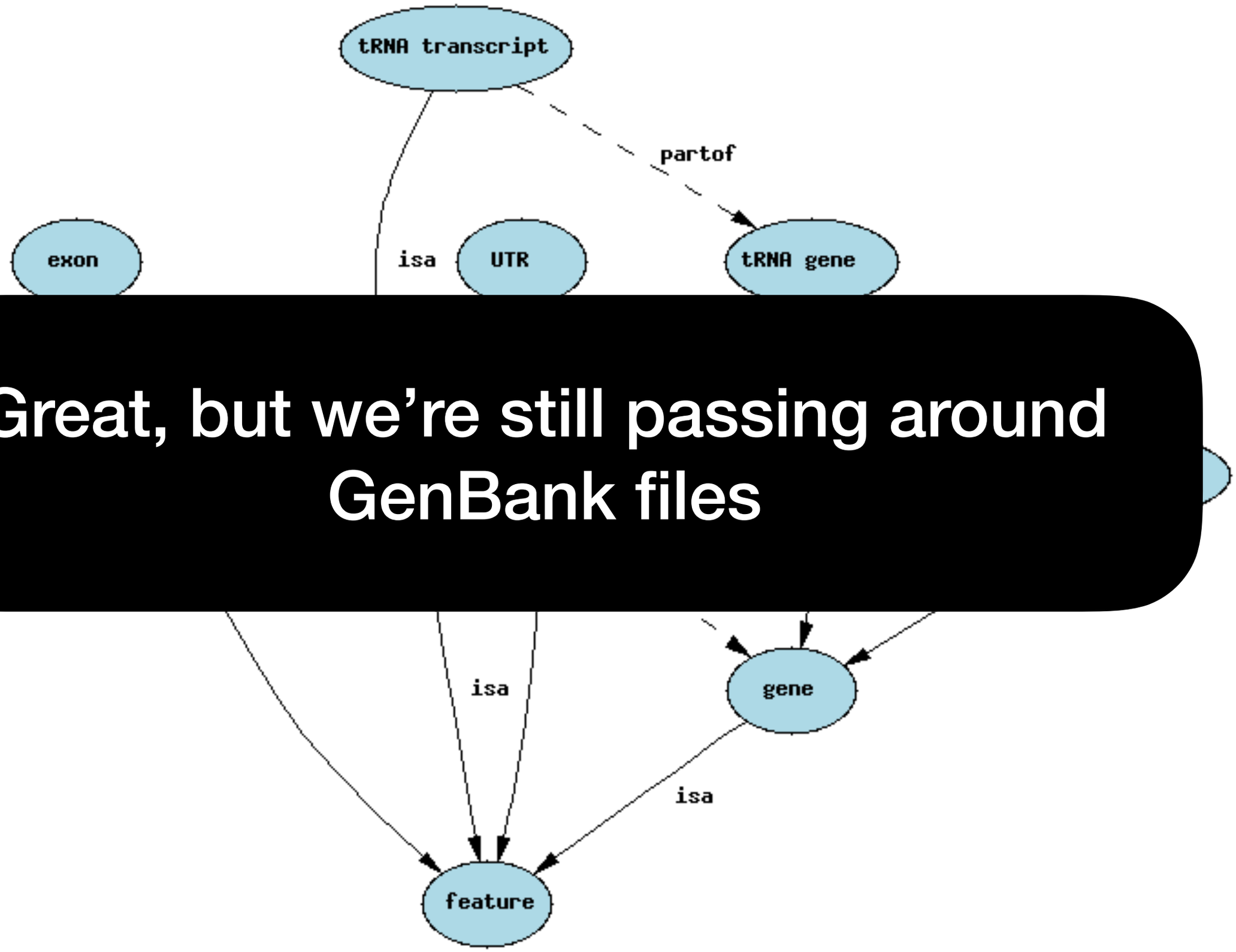
promoter (CURRENT_SVN)	
SO Accession:	SO:0000167 (SOWiki)
Definition:	A regulatory_region composed of the TSS(s) and binding sites for TF_complexes of the basal transcription machinery.
Synonyms:	promoter sequence
DB Xrefs:	SO: regcreative
Parent:	transcriptional_cis_regulatory_region (SO:0001055)
Children:	RNA_polymerase_promoter (SO:0001203)
	bidirectional_promoter (SO:0000568)
	retinoic_acid_responsive_element (SO:0001653)
	constitutive_promoter (SO:0002050)
	PSE_motif (SO:0000017)
	inducible_promoter (SO:0002051)

Ontologies make sure everyone is talking about the same thing





http://www.gmod.org/mediawiki/images/1/16/So-slim-example.png

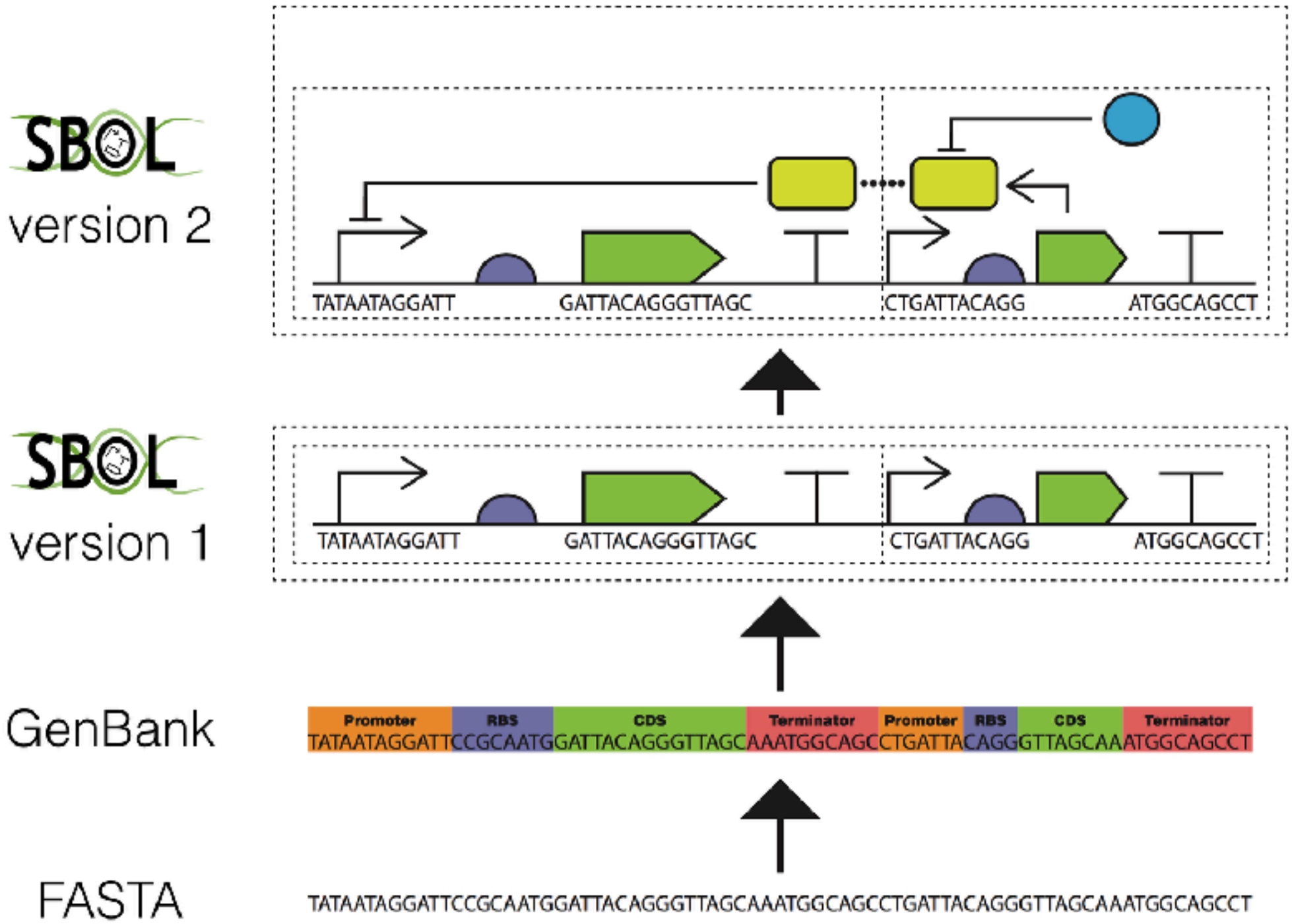


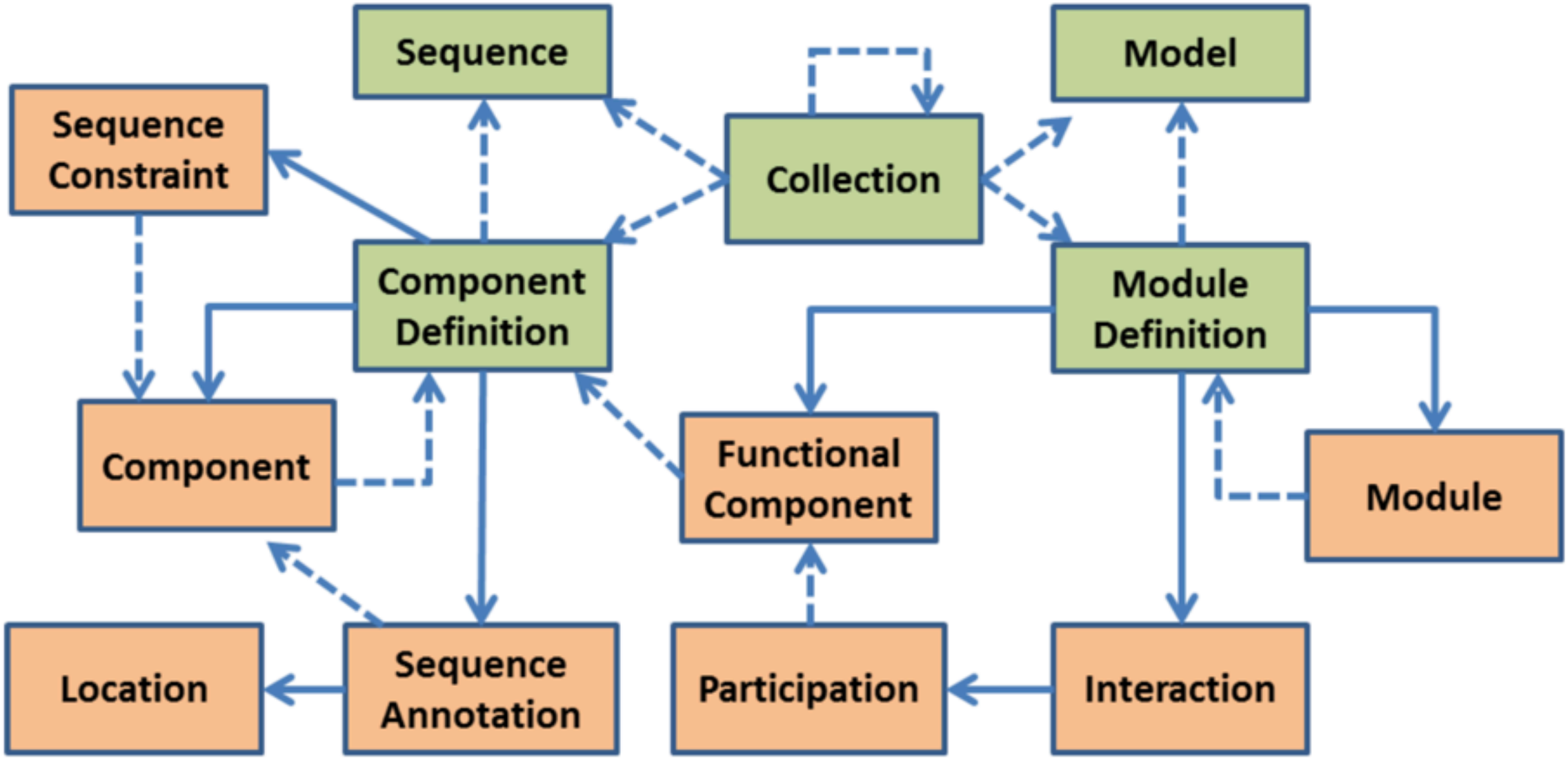
Great, but we're still passing around GenBank files

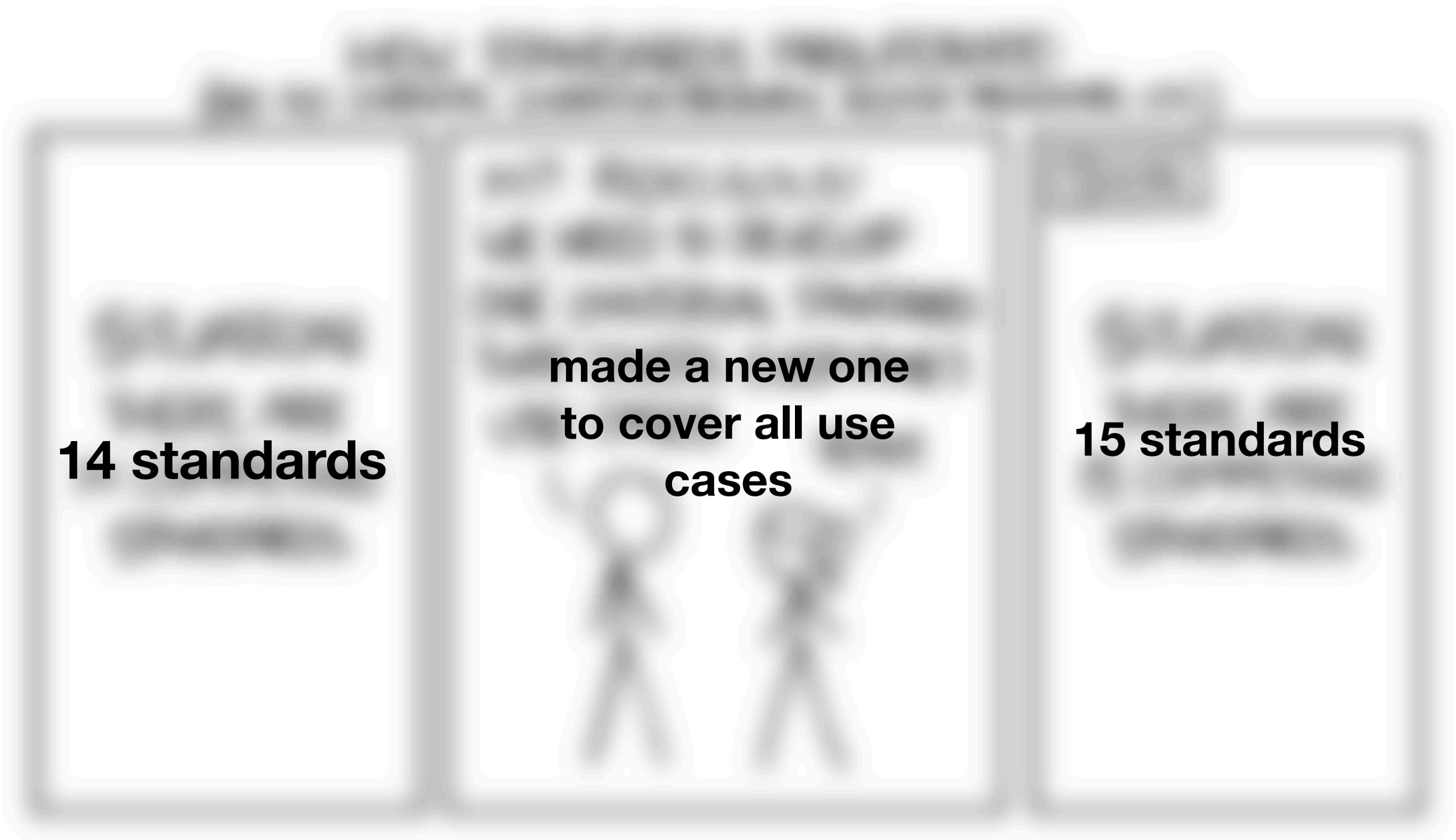
http://www.gmod.org/mediawiki/images/1/16/So-slim-example.png

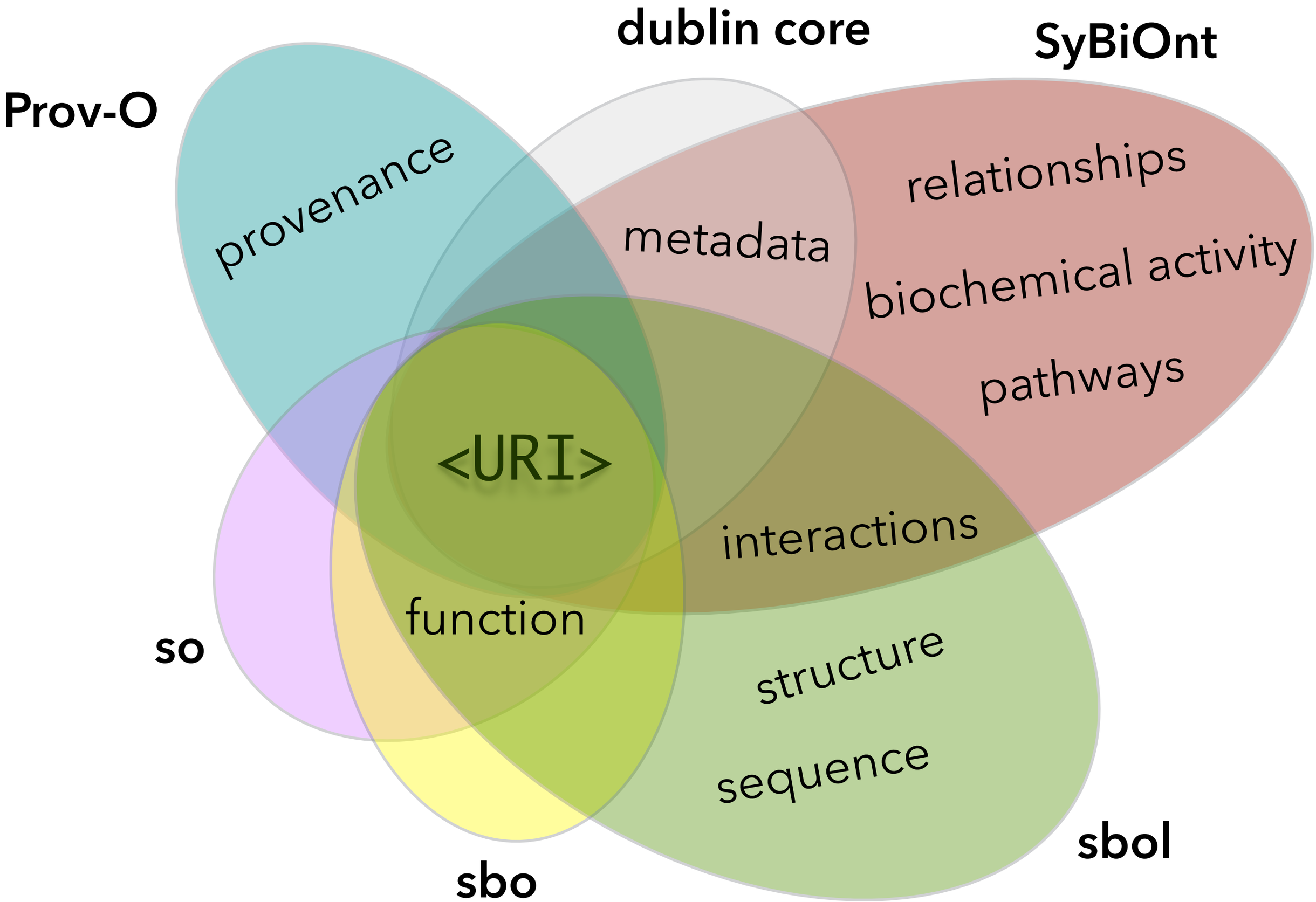
“The SBOL data standard is a data exchange representation for synthetic biology designs. Its goal is to improve the efficiency of data exchange and reproducibility of synthetic biology research.”

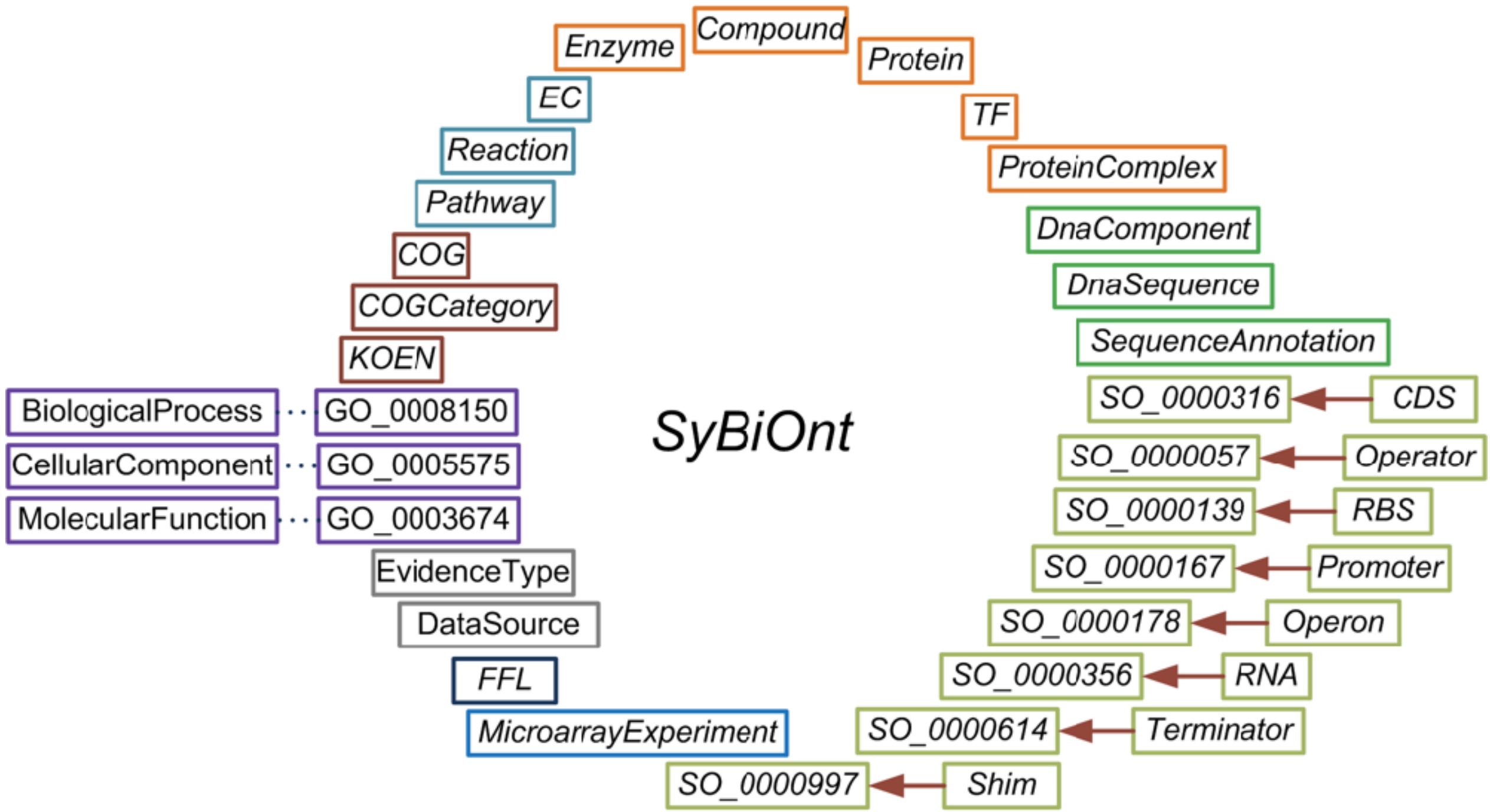
http://sbolstandard.org/data/


SBOL
version 2
TATAATAGGATT
GATTACAGGGTTAGC
CTGATTACAGG
ATGGCAGCCT
SBOL
version 1
TATAATAGGATT
GATTACAGGGTTAGC
CTGATTACAGG
ATGGCAGCCT
GenBank
Promoter
RBS
CDS
Terminator
Promoter
RBS
CDS
Terminator
TATAATAGGATTCCGCAATGGATTACAGGGTTAGCAAATGGCAGCCTGATTACAGGGTTAGCAAATGGCAGCCT
FASTA
TATAATAGGATTCCGCAATGGATTACAGGGTTAGCAAATGGCAGCCTGATTACAGGGTTAGCAAATGGCAGCCT

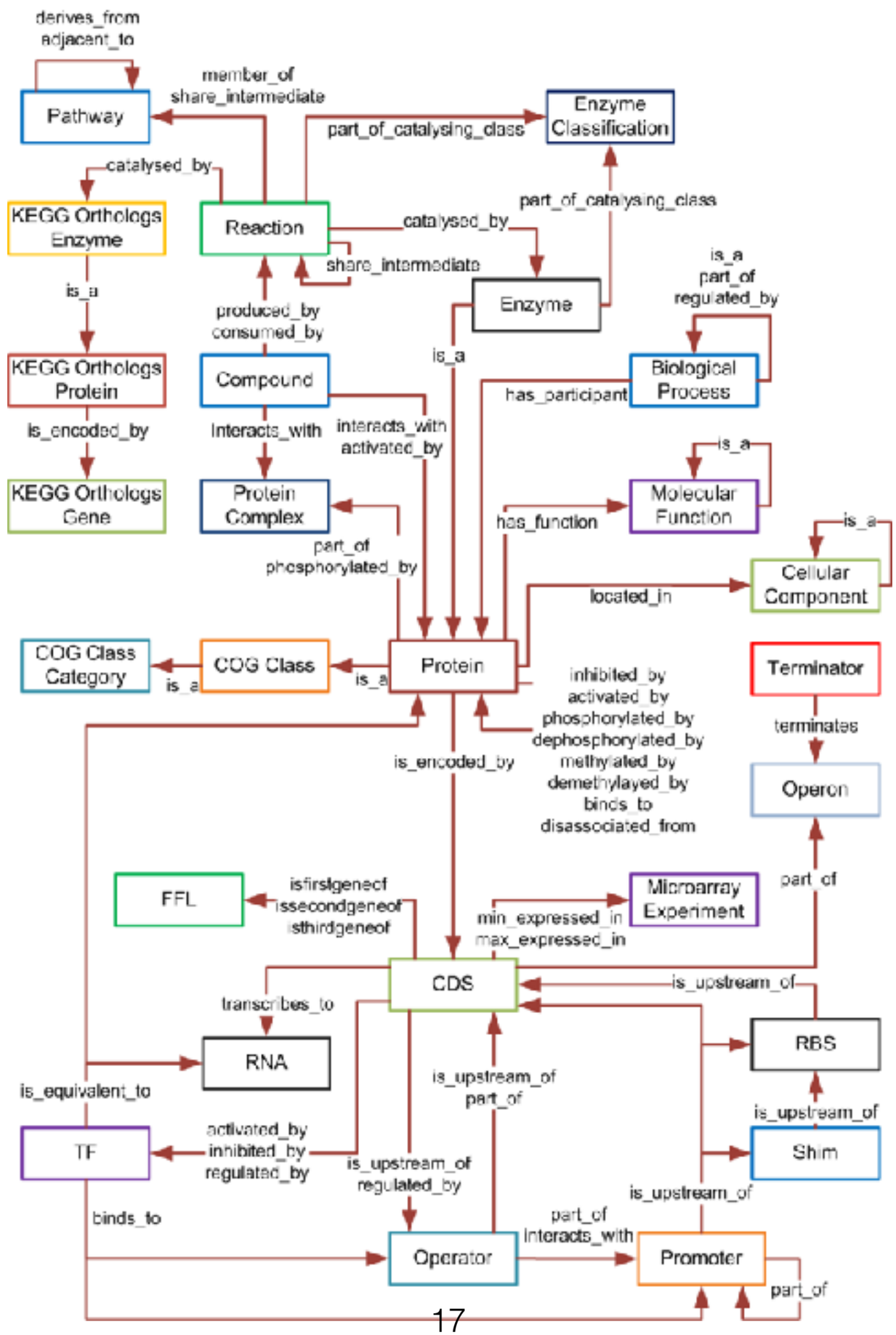

Sequence
Model
Sequence Constraint
Collection
Component Definition
Module Definition
Component
Functional Component
Module
Location
Sequence Annotation
Participation
Interaction

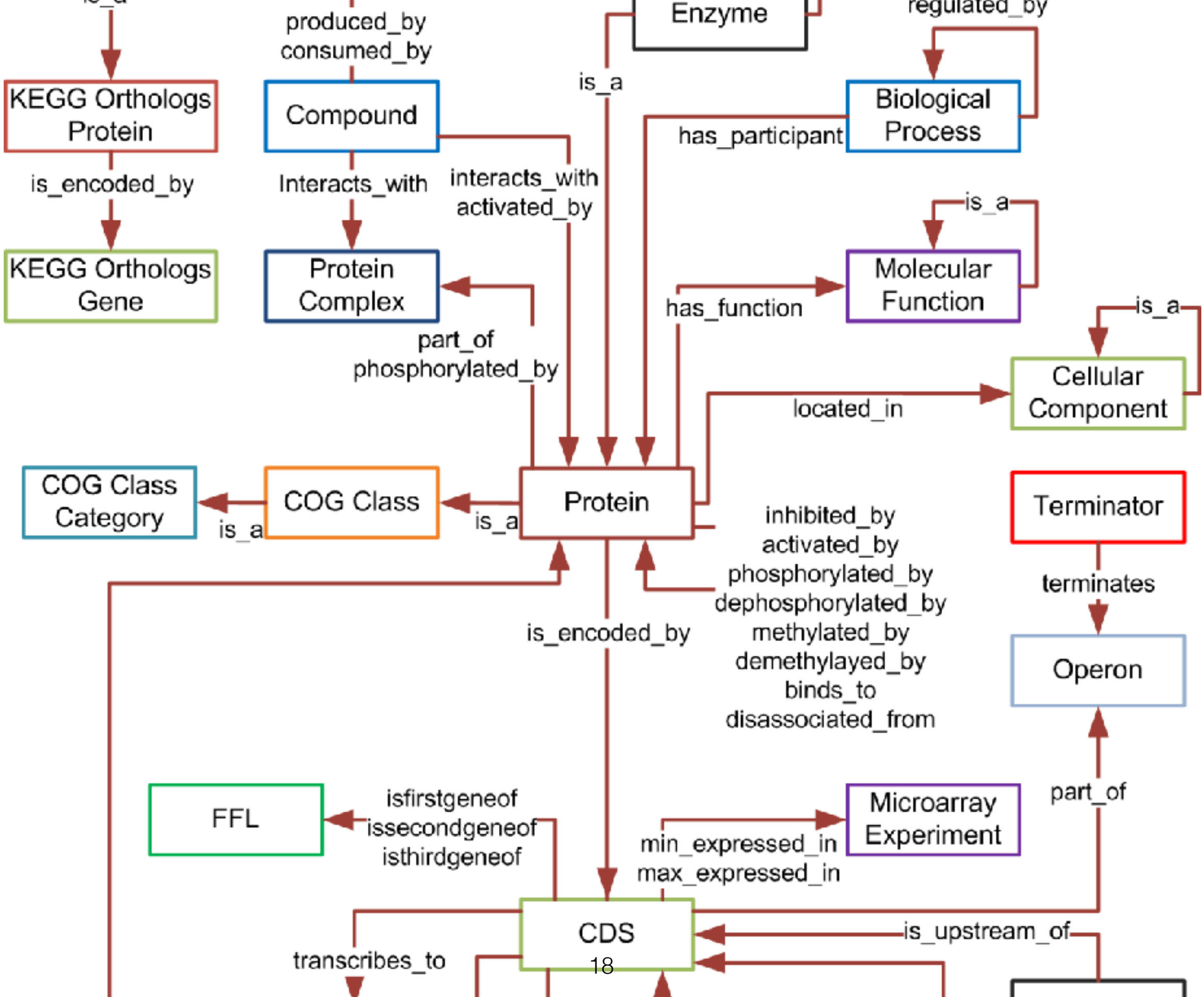

14 standards
made a new one
to cover all use
cases
15 standards


dublin core
SyBiOnt
Prov-O
provenance
metadata
relationships
biochemical activity
pathways
<URI>
interactions
function
so
structure
sequence
sbol
sbo



Misirli, Goksel, et al. "Data Integration and Mining for Synthetic Biology Design." *ACS synthetic biology* (2016).


derives_from
adjacent_to
Pathway
member_of
share_intermediate
Enzyme Classification
part_of_catalysing_class
catalysed_by
KEGG Orthologs Enzyme
Reaction
catalysed_by
part_of_catalysing_class
share_intermediate
is_a
Enzyme
is_a
part_of
regulated_by
produced_by
consumed_by
KEGG Orthologs Protein
Compound
is_a
Biological Process
has_participant
is_encoded_by
Interacts_with
interacts_with
activated_by
is_a
KEGG Orthologs Gene
Protein Complex
Molecular Function
has_function
is_a
part_of
phosphorylated_by
Cellular Component
located_in
COG Class Category
COG Class
is_a
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
RNA
RBS
is_equivalent_to
is_upstream_of
part_of
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


is_a
KEGG Orthologs Protein
is_encoded_by
KEGG Orthologs Gene
produced_by
consumed_by
Compound
Interacts_with
Protein Complex
interacts_with
activated_by
Enzyme
is_a
regulated_by
Biological Process
has_participant
is_a
Molecular Function
has_function
is_a
Cellular Component
located_in
part_of
phosphorylated_by
COG Class Category
is_a
COG Class
is_a
Protein
inhibited_by
activated_by
phosphorylated_by
dephosphorylated_by
methylated_by
demethylayed_by
binds_to
disassociated_from
Terminator
terminates
Operon
is_encoded_by
FFL
isfirstgeneof
issecondgeneof
isthirdgeneof
Microarray Experiment
min_expressed_in
max_expressed_in
part_of
CDS
is_upstream_of
transcribes_to

Integrated *Bacillus subtilis* example dataset

“SyBiOntKB”

```
Class: <http://www.bacillondex.org#12689>

    Annotations:
        rdfs:label "4-Hydroxyphenylacetic acid"@en,
        sybio:url "http://www.kegg.jp/dbget-bin/www_bget?C00642"@en,
        rdfs:comment "C8H8O3"@en,
        <http://www.purl.org/ontolink/tawny#name> "12689"@en,
        rdfs:label "4-Hydroxyphenylacetate"@en,
        sybio:elementOf "http://www.bacillondex.org/cv/KEGG"@en,
        rdfs:label "C8H8O3"@en,
        sybio:evidence "http://www.bacillondex.org/evidenceType/IMPD"@en

    SubClassOf:
        sybio:accession value "B00153"^^xsd:string,
        sybio:producedBy some <http://www.bacillondex.org#22325>,
        sybio:accession value "C00642"^^xsd:string,
        sybio:consumedBy some <http://www.bacillondex.org#22328>,
        sybio:consumedBy some <http://www.bacillondex.org#22442>,
        sybio:producedBy some <http://www.bacillondex.org#22332>,
        sybio:accession value "156-38-7"^^xsd:string,
        sybio:accession value "18101"^^xsd:string,
        sybio:Compound,
        sybio:producedBy some <http://www.bacillondex.org#22535>,
        sybio:producedBy some <http://www.bacillondex.org#22450>,
        sybio:producedBy some <http://www.bacillondex.org#22365>,
        sybio:consumedBy some <http://www.bacillondex.org#22418>,
        sybio:accession value "3915"^^xsd:string
```

Which parts can be used to upregulate the production of ammonium?

The Compound ammonia with the accession of “C00014” is producedBy Reaction RN:R00131, which consumes the Compound carbamide (C00086). Carbamide is producedBy a Reaction that is catalyzedBy an Enzyme, which is a subclassOf a Protein encodedBy the *argI* CDS with the accession BSU40320.

Misirli, Goksel, et al. "Data Integration and Mining for Synthetic Biology Design." *ACS synthetic biology* (2016).

http://sybiont.org

sybiont.org

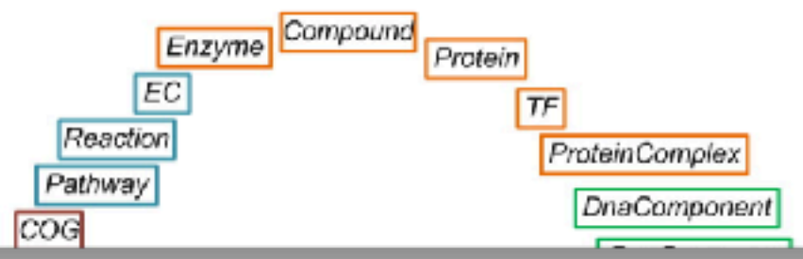
SyBiOnt: The Synthetic Biology Ontology

View on GitHub | Download .zip | Download .tar.gz

SyBiOnt: The Synthetic Biology Ontology

Göksel Mısırlı, Jennifer Hallinan, Matthew Pocock, Phillip Lord, James Alastair McLaughlin, Herbert Sauro, and Anil Wipat. (2016), Data Integration and Mining for Synthetic Biology Design, ACS Synthetic Biology, 5(10), 1086-1097

SyBiOnt is an application ontology for synthetic biology. We developed this ontology to represent the richset of biological knowledge about biological components and their relationships. We demonstrated the use of this ontology to create the SyBiOntKB knowledge base, incorporating and building upon existing life sciences ontologies and standards. The reasoning capabilities of ontologies were then applied to automate the mining of biological parts from this knowledge base. This approach is be useful to speed up synthetic biology design and ultimately help facilitate the automation of the biological engineering life cycle.



Get things into a standard representation

SBOL, SyBiOnt, SO, GO, etc.

Make the information computationally tractable

RDF, graph stores

Mine for the good parts!

Graph queries, user interfaces

disparity

unify syntax

unify semantics

distribution

data warehouse

federation - **scale**

complexity

data mining

visualization

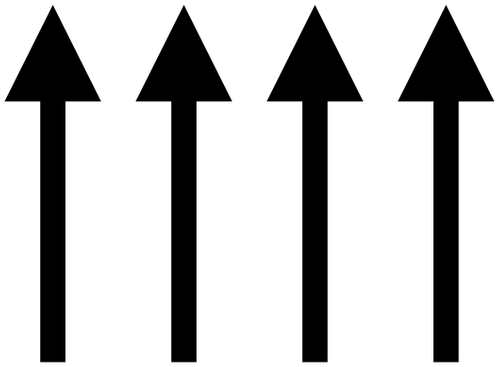
intractability

linked data

graph queries

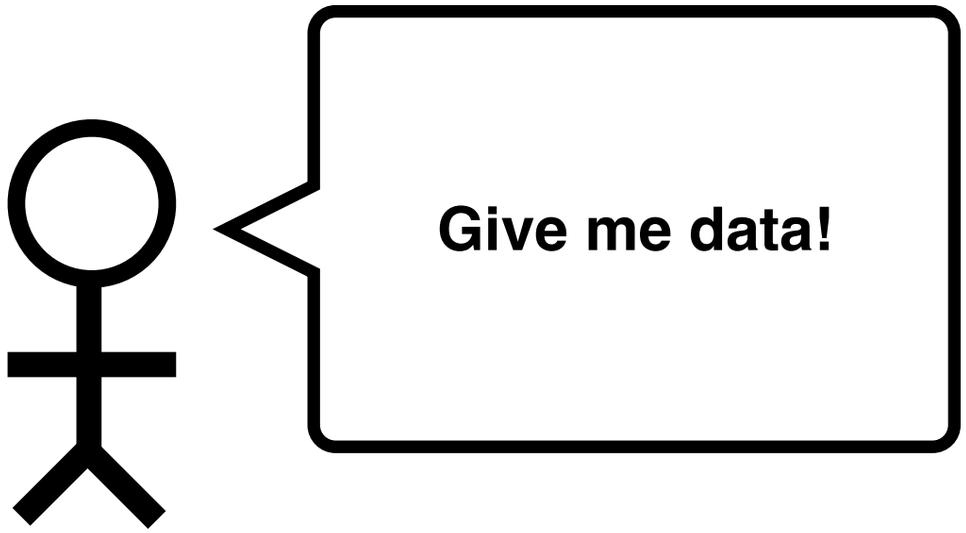
Store on a big server

Unify syntax and semantics



Multiple distributed data sources

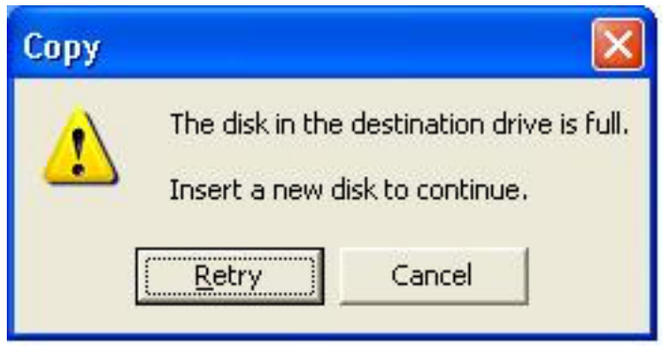
User



Big server

Ok, here you go!

Scale



Synchronization

disparity

unify syntax

unify semantics

distribution

data warehouse

federation - **scale**

complexity

data mining

visualization

intractability

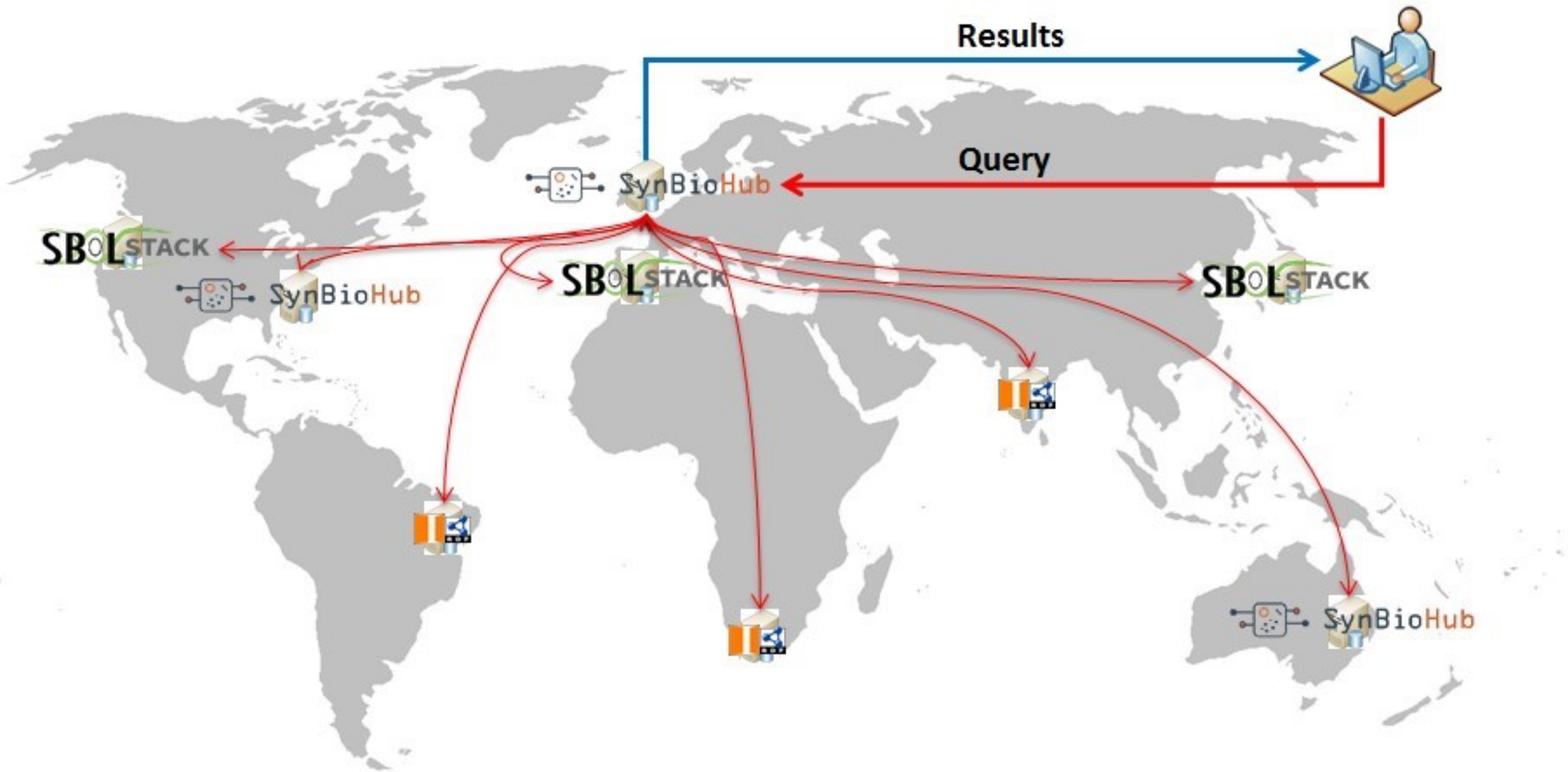
linked data

graph queries

Respond to query

On demand: retrieve, unify syntax and semantics

Multiple distributed data sources



Madsen, Curtis, et al. "The SBOL Stack: A Platform for Storing, Publishing, and Sharing Synthetic Biology Designs." *ACS synthetic biology* (2016).

synbiohub.org/browse

SynBioHub

Submit | About | Login or Register

Search

Search SynBioHub | Search

Advanced Search | SPARQL

ACS Synthetic Biology
version 1
Designs submitted with papers for ACS Synthetic Biology.
PUBLIC

Bacillus subtilis Collection
version 1
This collection includes information about promoters, operators, CDSs and proteins from Bacillus subtilis. Functional interactions such as transcriptional activation and repression, protein production and various protein-protein interactions are also included.
PUBLIC

iGEM Parts Registry
version 1
The iGEM Registry is a growing collection of genetic parts that can be mixed and matched to build synthetic biology devices and systems. As part of the synthetic biology community's efforts to make biology easier to engineer, it provides a source of genetic parts to iGEM teams and academic labs.
PUBLIC

iGEM 2017 Distribution
version 1
Distribution of parts for the 2017 iGEM competition
PUBLIC

SBOL Compliant Software
version 1
A collection of software that supports the Synthetic Biology Open Language (SBOL) standard
PUBLIC

ACS Synthetic Biology
version current
PUBLIC

More about SynBioHub at 10:00

SynBioHub

http://wiki.synbiohub.org

http://synbiohub.org

Acknowledgements

- ICOS group, Newcastle University
- The SBOL community
- Chris Myers
- Zach Zundel