Part of SPRINGER NATURE

Editors in your field

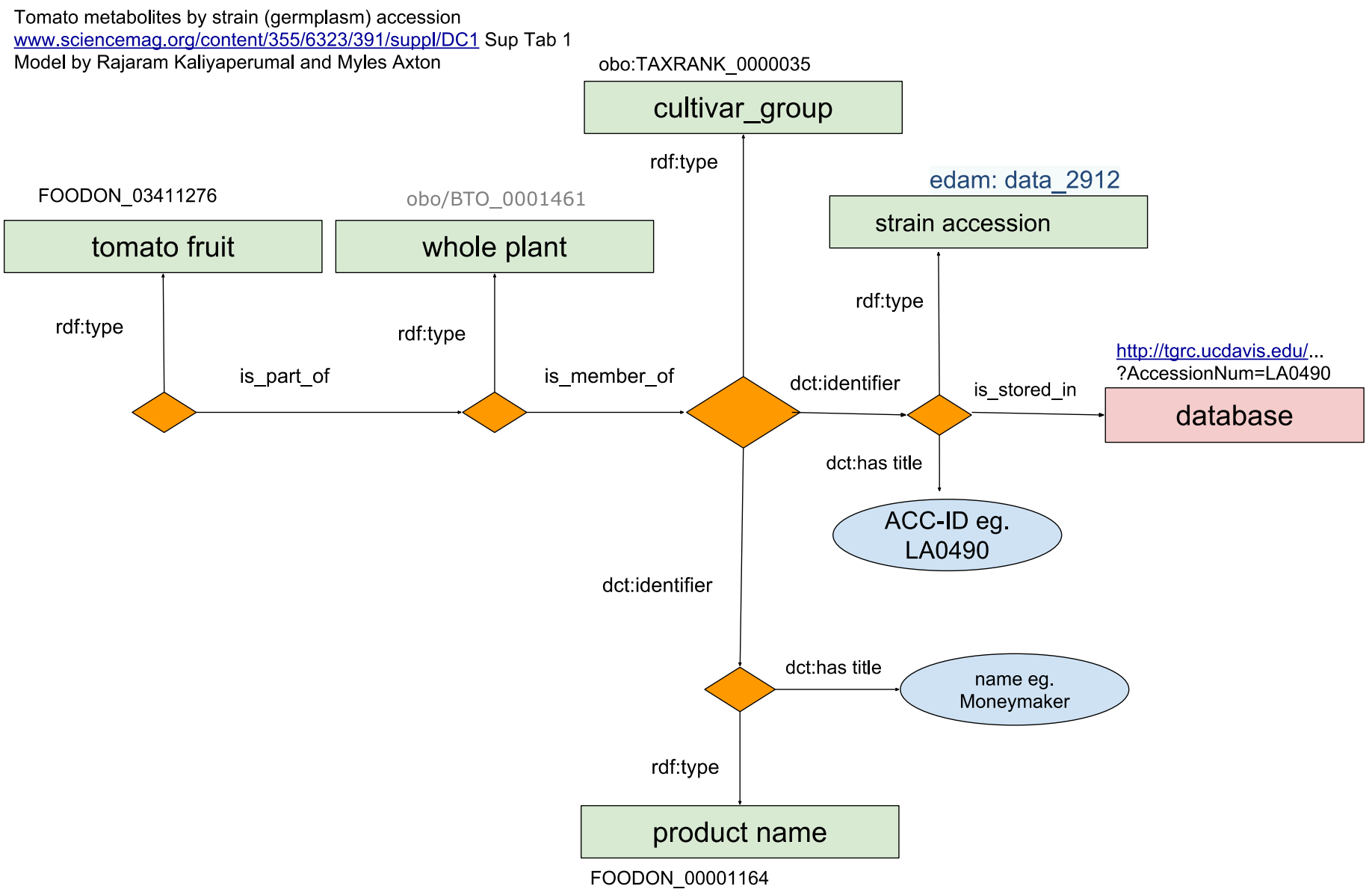
nature research

Automating assessment of FAIR project architecture

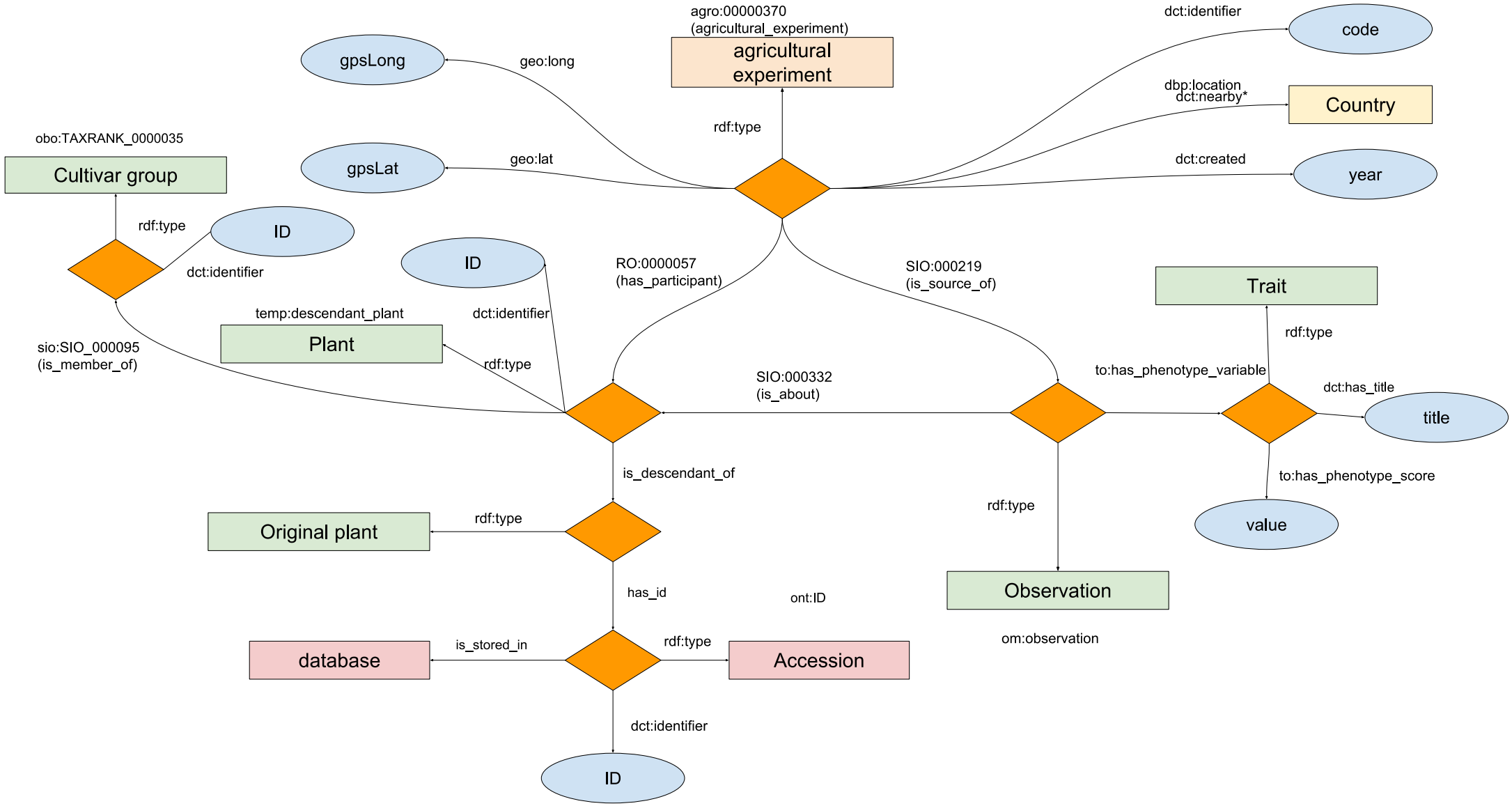
Daniel Kohn

Myles Axton
Chief Editor
Nature Genetics
Nov 1st, 2017

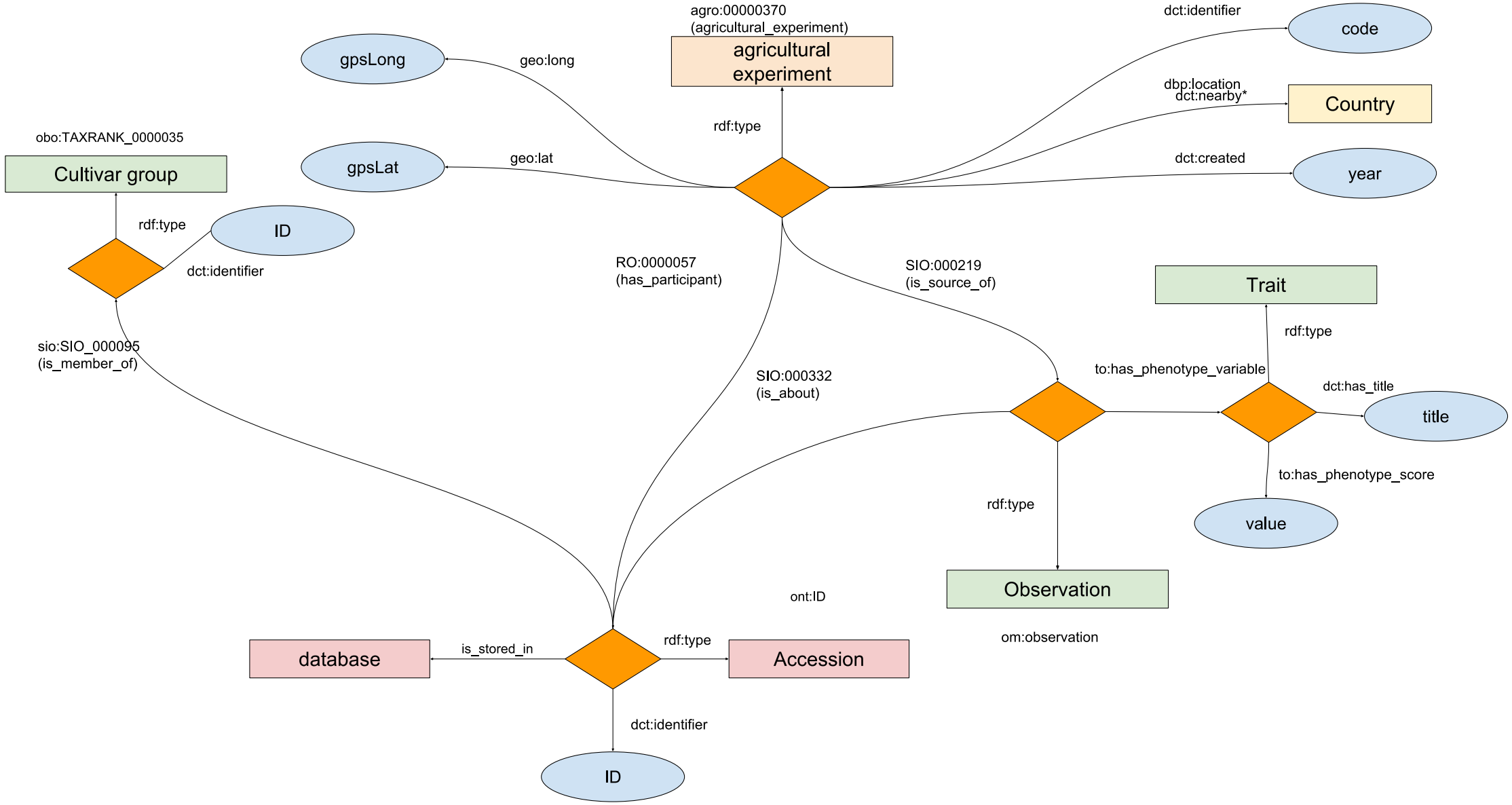
Domain experts overuse key terms to mean many things: unpacking “accession” according to plant geneticists



Eliana, Guangtao and Patrick's complete data model



Model peer reviewed by removing the subtleties of the key “accession” concept



SPARQL query on the intact model

```
prefix ds1: <http://92.222.88.166/tomato.ttl>
prefix ds2: <http://92.222.88.166/grouping-mockup-csv_v2.ttl>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix WUR_ont: <http://www.wur.nl/purl/ontology#>
prefix dct: <http://purl.org/dc/terms/>
prefix sio: <http://semanticscience.org/resource/>
prefix to: <http://purl.obolibrary.org/obo/TO#>
prefix WUR_acc: <http://www.wur.nl/purl/accession#>


select
  (?acc1 as ?AccessionID)
  (?phenotitle as ?Trait)
  (?phenoscore as ?TraitValue)
  (?gtitle as ?GroupName)
from ds2:#
from ds1:#
where {
  ?acc1 a WUR_ont:id .
  ?orPlant1 WUR_ont:has_id ?acc1 .
  ?desPlant1 WUR_ont:descendant_of ?orPlant1 .

  ?acc2 a WUR_ont:id .
  ?orPlant2 WUR_ont:has_id ?acc2 .
  ?desPlant2 WUR_ont:descendant_of ?orPlant2 .

  ?desPlant2 sio:SIO_000095 ?cgroup .
  ?cgroup dct:identifier "2" .
  ?cgroup dct:identifier ?gtitle .
  ?obs sio:SIO_000332 ?desPlant .
  ?obs to:has_phenotype_variable ?phenovar .
  ?phenovar a <http://purl.obolibrary.org/obo/SP_0000009> ;  # fruit
colour
    dct:title ?phenotitle .
  ?phenovar to:has_phenotype_score ?phenoscore .
  filter (?acc1 = ?acc2)
}
```

acc	phenoscore	phenotitle	gtitle
WUR_acc:EA01982	"ripening"	"Fruit color"	"2"
WUR_acc:EA01802	"light red"	"Fruit color"	"2"
WUR_acc:EA01903	"red"	"Fruit color"	"2"
WUR_acc:EA01953	"red"	"Fruit color"	"2"
WUR_acc:EA01804	"light red"	"Fruit color"	"2"
WUR_acc:EA01960	"light red"	"Fruit color"	"2"
WUR_acc:EA02304	"red"	"Fruit color"	"2"
WUR_acc:EA01371	"apricot / Juans Flame"	"Fruit color"	"2"
WUR_acc:EA01915	"tangerine"	"Fruit color"	"2"
WUR_acc:EA01684	"red"	"Fruit color"	"2"
WUR_acc:EA01712	"red"	"Fruit color"	"2"
WUR_acc:EA01784	"yellow"	"Fruit color"	"2"
WUR_acc:EA01356	"tangerine"	"Fruit color"	"2"
WUR_acc:EA01989	"ripening"	"Fruit color"	"2"

SPARQL query on the deleted model produces false positive results

```
prefix ds1: <http://92.222.88.166/tomato.ttl>
prefix ds2: <http://92.222.88.166/grouping-mockup-csv_v2.ttl>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix WUR_ont: <http://www.wur.nl/purl/ontology#>
prefix dct: <http://purl.org/dc/terms/>
prefix sio: <http://semanticscience.org/resource/>
prefix to: <http://purl.obolibrary.org/obo/TO#>
prefix WUR_acc: <http://www.wur.nl/purl/accession#>


select
  (?acc1 as ?AccessionID)
  (?phenotitle as ?Trait)
  (?phenoscore as ?TraitValue)
  (?gtitle as ?GroupName)
from ds2:#
from ds1:#
where {
  ?acc1 a WUR_ont:id .
  ?orPlant1 WUR_ont:has_id ?acc1 .
  ?desPlant1 WUR_ont:descendant_of ?orPlant1 .

  ?acc2 a WUR_ont:id .
  ?orPlant2 WUR_ont:has_id ?acc2 .
  ?desPlant2 WUR_ont:descendant_of ?orPlant2 .

  ?desPlant2 sio:SIO_000095 ?cgroup .
  ?cgroup dct:identifier "2" .
  ?cgroup dct:identifier ?gtitle .
  ?obs sio:SIO_000332 ?desPlant .
  ?obs to:has_phenotype_variable ?phenovar .
  ?phenovar a <http://purl.obolibrary.org/obo/SP_0000009> ;  # fruit
colour
    dct:title ?phenotitle .
  ?phenovar to:has_phenotype_score ?phenoscore .
  filter (?acc1 = ?acc2)
}
```

acc1	phenoscore	phenotitle	gtitle
WUR_acc:EA04236	"red"	"Fruit color"	"2"
WUR_acc:EA01270	"red"	"Fruit color"	"2"
WUR_acc:EA01982	"ripening"	"Fruit color"	"2"
WUR_acc:EA05612	"red"	"Fruit color"	"2"
WUR_acc:EA03126	"orange"	"Fruit color"	"2"
WUR_acc:EA00744	"red"	"Fruit color"	"2"
WUR_acc:EA01802	"light red"	"Fruit color"	"2"
WUR_acc:EA01155	"light red"	"Fruit color"	"2"
WUR_acc:EA04001	"light red"	"Fruit color"	"2"
WUR_acc:EA02959	"red"	"Fruit color"	"2"
WUR_acc:EA03463	"orange"	"Fruit color"	"2"
WUR_acc:EA02669	"light red"	"Fruit color"	"2"
WUR_acc:EA01237	"red"	"Fruit color"	"2"
WUR_acc:EA01185	"red"	"Fruit color"	"2"
WUR_acc:EA00448	"dark red"	"Fruit color"	"2"
WUR_acc:EA03083	"red"	"Fruit color"	"2"
WUR_acc:EA00240	"red"	"Fruit color"	"2"
WUR_acc:EA01230	"red"	"Fruit color"	"2"
WUR_acc:EA02764	"light red"	"Fruit color"	"2"
WUR_acc:EA02895	"red"	"Fruit color"	"2"
WUR_acc:EA02728	"red"	"Fruit color"	"2"
WUR_acc:EA03174	"dark red"	"Fruit color"	"2"
WUR_acc:EA03533	"light red"	"Fruit color"	"2"
WUR_acc:EA03648	"red"	"Fruit color"	"2"
WUR_acc:EA03650	"red"	"Fruit color"	"2"
WUR_acc:EA02660	"red"	"Fruit color"	"2"
WUR_acc:EA01903	"red"	"Fruit color"	"2"

FAIR metadata architecture

How do we connect the data model to the metadata?

Pedro Vieira https://www.flickr.com/photos/pppedro/14557981/

dcat:Repository

dcat:Catalog

dcat:Dataset

dcat:Distribution

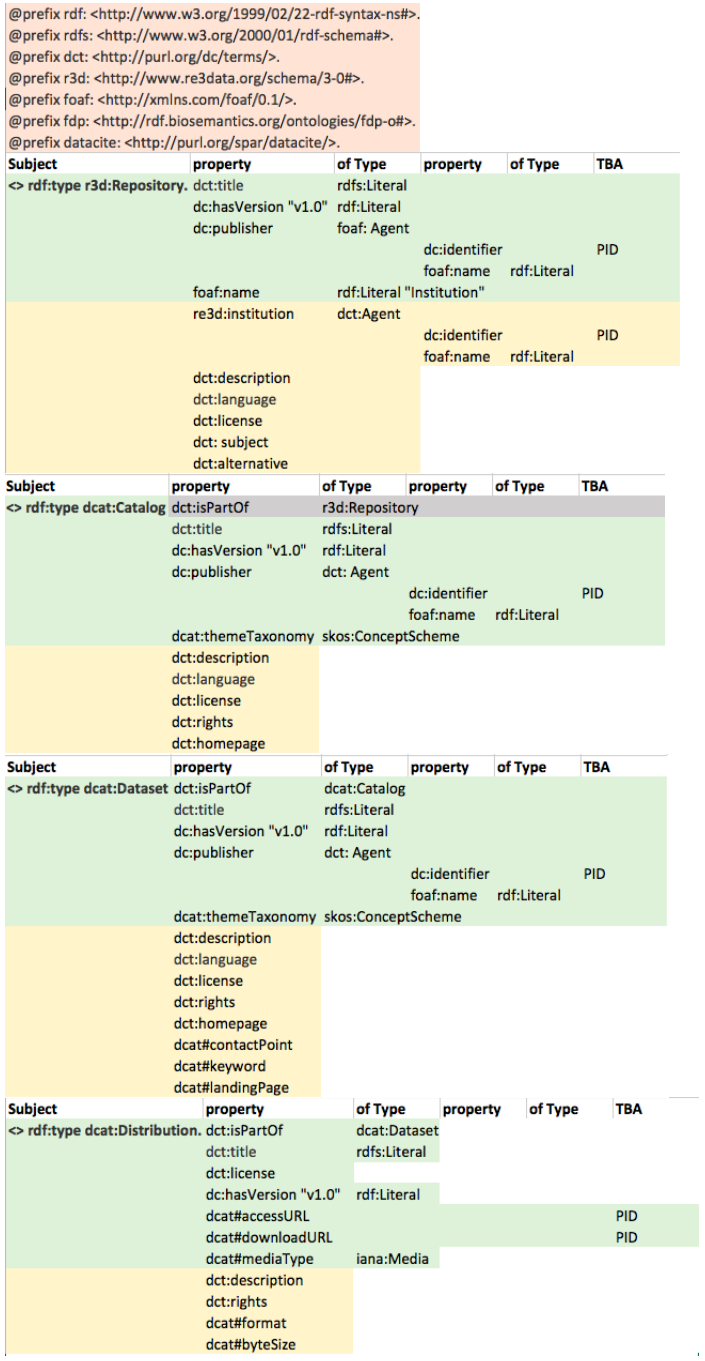
RDF, DCAT and FDP provide FAIR now

I2) RDF and ontologies throughout

A2) Catalog persists even if data link is lost outside

R1.3) community peer review of data models is possible via SPARQL

A1, 1.1, 1.2) Data stored elsewhere can have license and be downloaded via downloadURL



```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix dct: <http://purl.org/dc/terms/>.
@prefix r3d: <http://www.re3data.org/schema/3-0#>.
@prefix foaf: <http://xmlns.com/foaf/0.1/>.
@prefix fdp: <http://rdf.biosemantics.org/ontologies/fdp-o#>.
@prefix datacite: <http://purl.org/spar/datacite/>.
```

Subject	property	of Type	property	of Type	TBA
<> rdf:type r3d:Repository.	dct:title	rdfs:Literal			
	dc:hasVersion "v1.0"	rdf:Literal			
	dc:publisher	foaf: Agent			
			dc:identifier		PID
			foaf:name	rdf:Literal	
	foaf:name	rdf:Literal "Institution"			
	re3d:institution	dct:Agent			
			dc:identifier		PID
			foaf:name	rdf:Literal	
	dct:description				
	dct:language				
	dct:license				
	dct: subject				
	dct:alternative				

Subject	property	of Type	property	of Type	TBA
<> rdf:type dcat:Catalog	dct:isPartOf	r3d:Repository			
	dct:title	rdfs:Literal			
	dc:hasVersion "v1.0"	rdf:Literal			
	dc:publisher	dct: Agent			
			dc:identifier		PID
			foaf:name	rdf:Literal	
	dcat:themeTaxonomy	skos:ConceptScheme			
	dct:description				
	dct:language				
	dct:license				
	dct:rights				
	dct:homepage				

Subject	property	of Type	property	of Type	TBA
<> rdf:type dcat:Dataset	dct:isPartOf	dcat:Catalog			
	dct:title	rdfs:Literal			
	dc:hasVersion "v1.0"	rdf:Literal			
	dc:publisher	dct: Agent			
			dc:identifier		PID
			foaf:name	rdf:Literal	
	dcat:themeTaxonomy	skos:ConceptScheme			
	dct:description				
	dct:language				
	dct:license				
	dct:rights				
	dct:homepage				
	dcat#contactPoint				
	dcat#keyword				
	dcat#landingPage				

Subject	property	of Type	property	of Type	TBA
<> rdf:type dcat:Distribution.	dct:isPartOf	dcat:Dataset			
	dct:title	rdfs:Literal			
	dct:license				
	dc:hasVersion "v1.0"	rdf:Literal			
	dcat#accessURL				PID
	dcat#downloadURL				PID
	dcat#mediaType	iana:Media			
	dct:description				
	dct:rights				
	dcat#format				
	dcat#byteSize				

More for GO-FAIR in the future?

A2) mirroring between FDPs to ensure persistence of metadata and models

F1) agreement to use eg. trusty URIS and PURLs. High value distributions may need DOI or handles.

A1) automated query filtering for sensitive data (returns only metadata and access protocol if not authorized)

R1.3) community Catalogs to share standards

I1) automate matching of structural constraints and vocabularies (eg. SHACL)

F2,R1) few high level Classes and many precise properties

R1.3) community libraries of standard study models (eg. linkedISA)

R1.2) dct:provenance actually refers to changes in creator or publisher, so provenance file should include license and version

R1.1) CC0 or CC-BY licenses at least for Distributions of: basic models, intrinsic metadata and data access statements

SHACL query for FAIR architecture is like an antibody interrogating this structure.
SPARQL query interrogates function.

Yes, we are totally going to store data models on DNA!

dcat:Catalog

dcat:Dataset

skos:ConceptScheme

dcat:Distribution

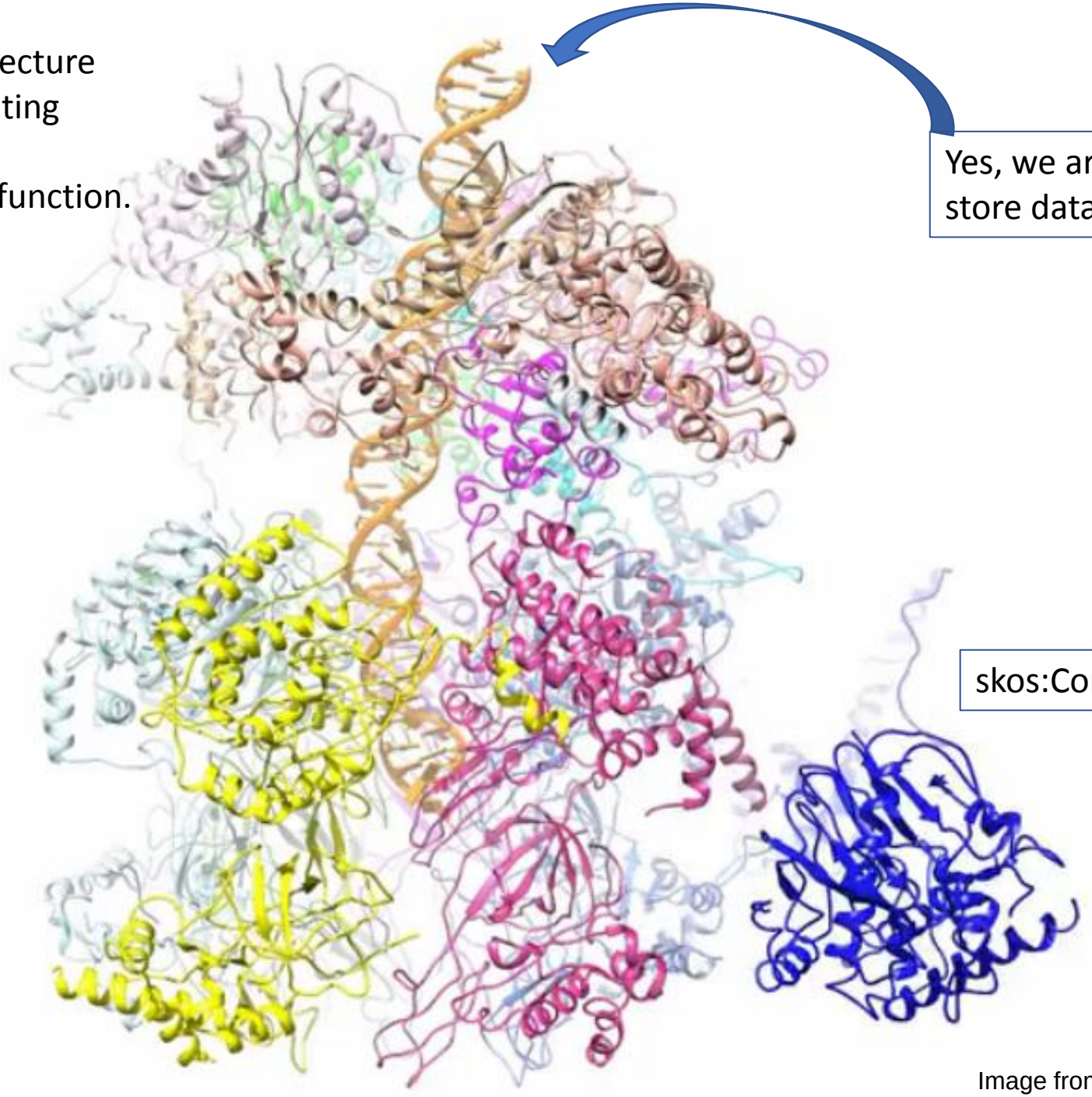
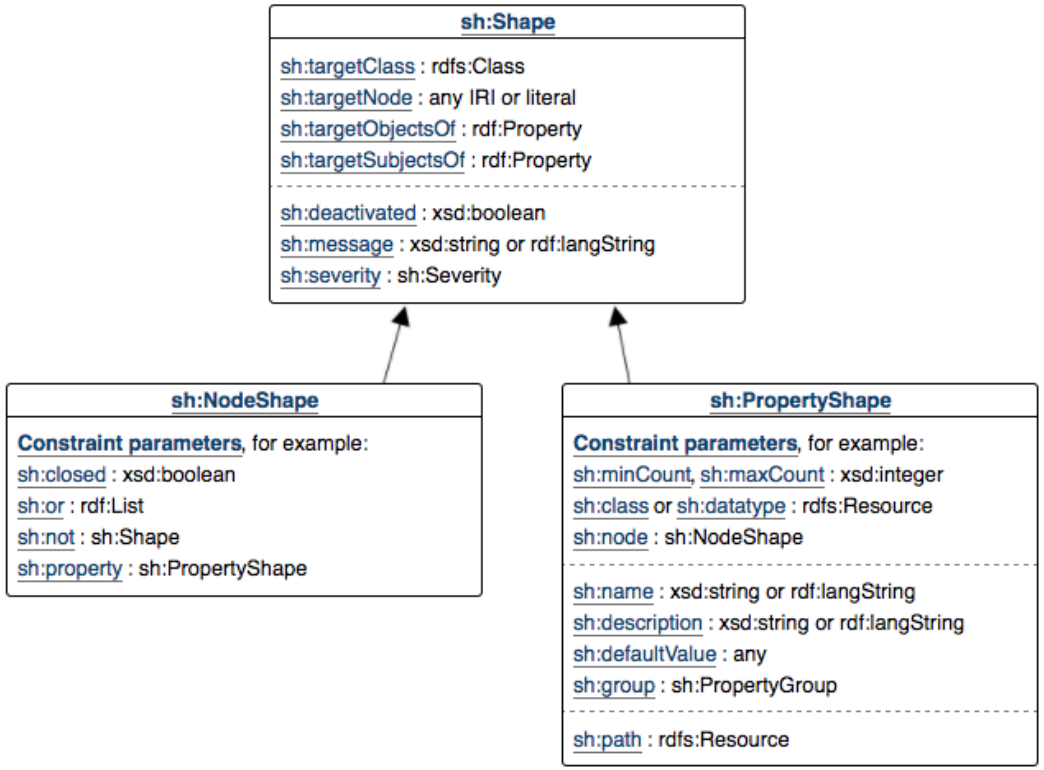


Image from doi:10.1038/nsmb.3372 SV3.htm

SHACL interrogates rdf structures

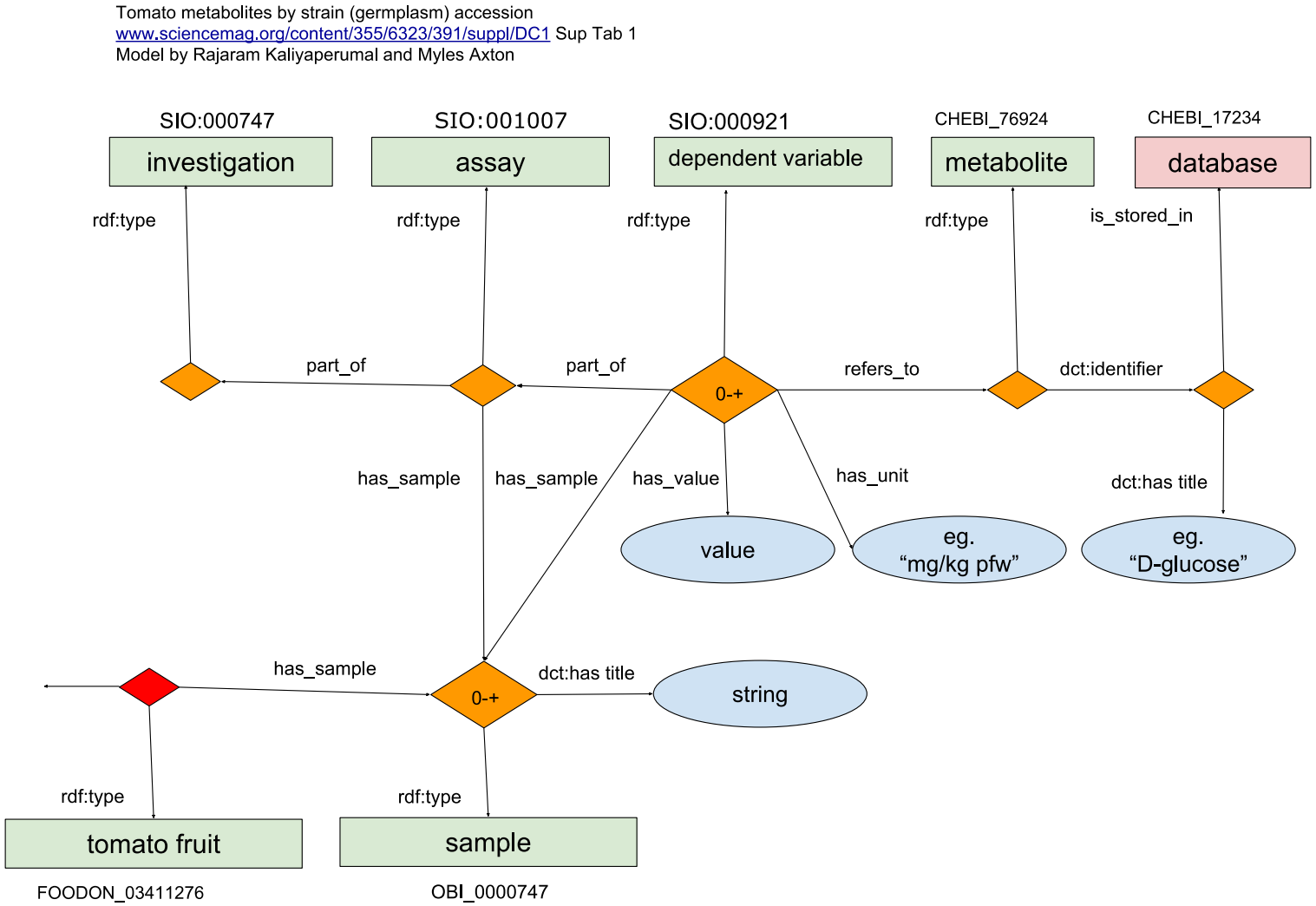


Possible uses in FAIR Data Repository

1. Instantiate community standards for data types and models
2. Instantiate constraints on interoperability and access
3. SHACL queries could filter by levels of FAIR implementation
4. With SPARQL, could be used to peer review structure and function of data models
5. Remap inverse relations in models to match FDP standard

https://www.w3.org/TR/shacl/
https://www.w3.org/TR/shacl-ucr/

Experimental designs contain networks of references and should be hierarchic to include multiple instances



FAIRifying I3: Improving cited cross references between (meta)data with explicit intent

recognition sites[19] or nickase sites[21]), which can provide sufficient contextual information to scaffold assembled contigs[22] or correct existing reference assemblies[23]. Both optical mapping[21] and Hi-C[15] yield excellent scaffold continuity metrics[15,17,18,24]. However, both methods have limited ability to scaffold small contigs in fragmented short-read assemblies[25].

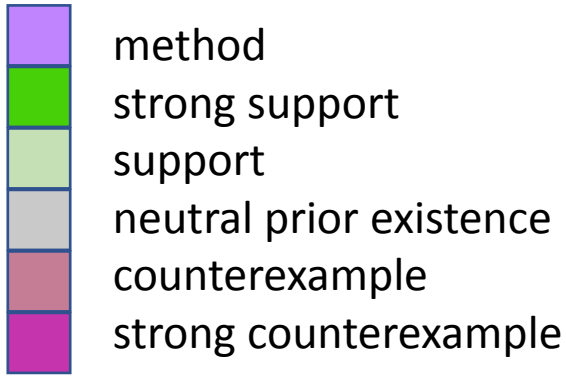
Single-molecule sequencing[26] can now produce reads tens of kilobases in size, albeit with relatively high error rate. The Pacific Biosciences PacBio RSII sequencing platform achieves an average read length of 14 kb, with maximum read lengths >60 kb[27], and is routinely used to reconstruct complete bacterial genomes[28,29] and highly continuous eukaryotic genomes[27,30,31]. When maximum read length exceeds the maximum repeat size, it is theoretically possible to assemble complete mammalian chromosomes. However, the read depth required to ensure that all repeats are spanned by such reads is currently prohibitive, so mammalian assemblies will continue to comprise thousands of pieces[27,30] until average read lengths exceed ~30 kb. Currently, combinations of long-read sequencing and long-range scaffolding represent the most efficient approach to produce near-finished reference assemblies. For example, a recent study using long-read sequencing and optical mapping assembled a human genome *de novo* into 4,007 contigs and 202 scaffolds that covered the entire reference assembly[31].

Here we present a near-finished reference genome for the domestic goat (*C. hircus*) using a combination of long-read single-molecule sequencing, high-fidelity short-read sequencing, optical mapping, and Hi-C-based chromatin interaction maps. Unlike cattle, which are derived from two different subspecies[32], extant domestic goats appear to derive from a single wild ancestor, the bezoar[33]. Owing to this singular domestication event, creation of a polished reference genome for goat could enable easier identification of adaptive variants in sequence data from descendent breeds. The most recent goat assembly was generated via short-read sequencing and optical mapping and is highly fragmented[18]. Our new assembly strategy achieves animal owing to tissue storage complications. Assembly of these complementary data types proceeded in a stepwise fashion (Online Methods), producing progressively improved assemblies (**Table 1** and **Fig. 1**). Initial assembly of the PacBio data alone resulted in a contig NG50 (the minimum length of contigs accounting for half of the haploid genome size) of 3.8 Mb. PacBio contigs were first scaffolded using optical mapping data, and the resulting scaffolds were clustered using Hi-C data into chromosome-scale scaffolds. To assess quality, the resulting assembly was validated via statistical methods and comparison to a radiation hybrid (RH) map[34] (**Supplementary Table 1**) and previous assemblies (**Supplementary Note**). To maximize accuracy of the final reference assembly, the RH map was used to correct 21 inversions (consisting of 83 scaffolds) and 4 misplacements before final gap filling and polishing[35,36]. Our final assembly, ARS1, totaled 2.92 Gb of sequence with a contig NG50 of 18.7 Mb, a scaffold NG50 of 87 Mb, and an estimated quality value (QV)[37] of 34.5 (**Table 1**, **Fig. 2** and **Supplementary Note**). After error correction and validation, ARS1 contained four major disagreements with the RH map (**Fig. 3**), which will require further investigation to confirm. Considering that ARS1 comprises just 31 scaffolds and 649 gaps covering 30 of the 31 haploid, acrocentric goat chromosomes[38] (excluding only the Y chromosome), our assembly compares favorably with the current human reference (GRCh38), which has 24 scaffolds, 169 unplaced or unlocalized scaffolds, and 832 gaps in the primary assembly[39].

Scaffolding technology comparisons

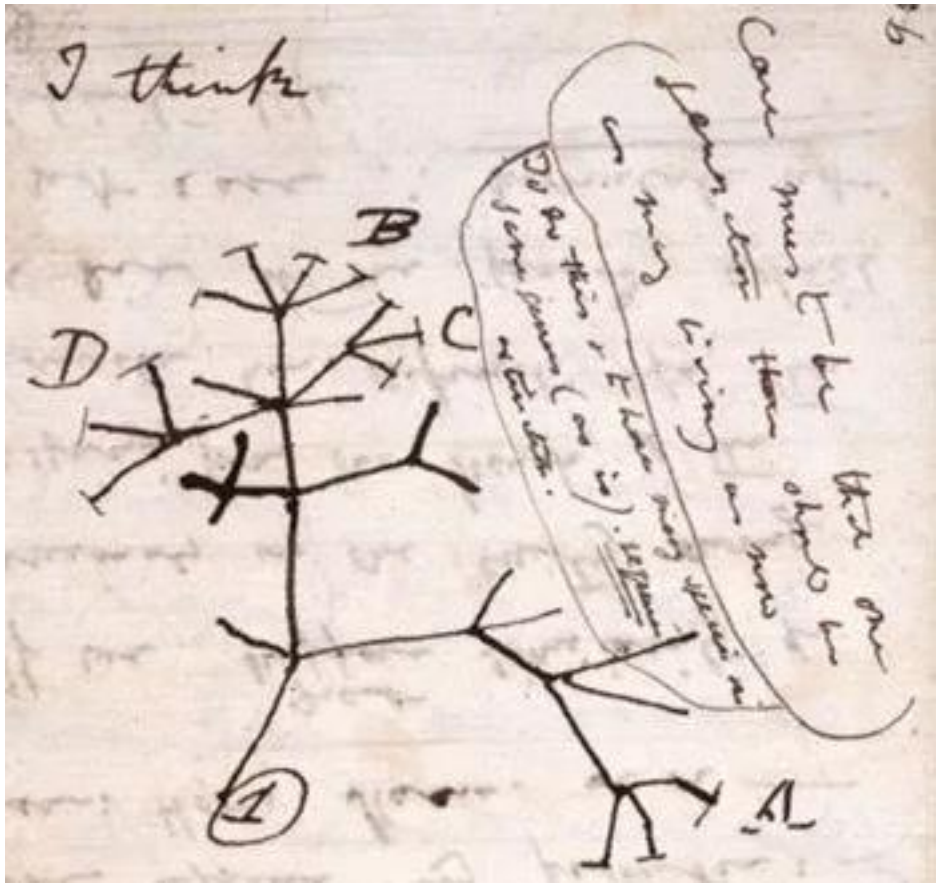
We compared initial *de novo* optical map and Hi-C scaffolds to our final validated reference assembly to evaluate the independent performance of the two scaffolding strategies. The optical map consisted of 2,944 scaffolds with an NG50 of 1.487 Mb. It is likely that optical map fragment sizes (**Supplementary Fig. 1**) were limited by double-strand breaks caused by close proximity of Nt.BsqI sites on opposing DNA strands, as reported previously[21]. Optical map scaffolding of PacBio contigs produced an assembly of 333 scaffolds, containing

Technical Report discusses alternative methods critically. Note the strongest claim of a technical advance at the end of the first results section.
doi:10.1038/ng.3802



A strong conclusion expressed as negation of the current theoretical framework of the field

the data nor the process that gave rise to the data. Evidence for genome-wide contradictions between gene trees and species trees is rapidly accumulating in a wide range of species[15,34–44], raising the question of whether this should lead to a reevaluation of the utility of the tree as model for speciation[45,46]. We think it should, but time will tell.



Charles Darwin 1837
Syndics of Cambridge University Library

Strong counterexamples with a soft center of qualified counterexample. “False modesty” to deter criticism of a strong claim of a revolutionary conclusion?
doi:10.1038/ng.2617