TOPICAL: Linked and semantic data retrieval

Dan Berrios
Space Biosciences Research Branch
1 (650) 604 0470
KBR/NASA Ames Research Center
daniel.c.berrios@nasa.gov

Sylvain V. Costes
Space Biosciences Research Branch
1 (650) 604.5343
NASA Ames Research Center
sylvain.v.costes@nasa.gov

Daniel Jamieson
CEO, Biorelate
+44 7984 436 284
dan@biorelate.com

The continuing exponential increase in scientific data generation has accentuated the lack of investment in tools for researchers to discover and access all data relevant to a given hypothesis quickly, accurately, and completely. They are typically faced with a wide range of challenges related to gathering and comparing multiple sets and types of data from many different sources and databases. Yet research into the biological processes underlying health risks of space environments will necessarily involve comparing findings from experiments on model organisms with one another, and ultimately corroborating them with those of human studies. For example, an investigator wishing to probe the effects of CO_2 on the human eye may wish to construct the following query:

Query 1. "Find transcription data on all mammals studied in space with gravity between 0 and 0.2 and exposed to pCO_2 levels higher than 0.35 mmHg, that also have proteome profiles and measurements of intra-ocular pressure".

NASA-funded scientific data discovery systems often will fail to discover all the data relevant to a complex query like Query 1, and/or include too much irrelevant data. This is primarily because these systems most commonly employ natural language term indexing. Natural language term indexing technologies, the most prevalent kind of indexing used by scientific data systems today, index terms devoid of any context or meaning. The result is too often a set of search results with precision and recall that are low, particularly when executing a complex query like Query 1 due in part to the lack of uniformly applied structuring, understanding and indexing of both the data and the query itself. Thus, for an ambiguous term such as "space" in Query 1, a system will be unaware of whether it refers to an anatomic region of an organism (such as "epidural space" [1]), a quality of swarming by organisms ("space swarm"), or to the vacuum of "outer space [2]." Ontological logic can be used to discriminate among these concepts when data are indexed and queried so that queries specifying the concept of "outer space" do not yield results that include data associated only with epidural space. In the case of Query 1, the word 'space', is indexed against one or more ontologies to specify the correct interpretation of what is written: the concept "outer space".

Conversely, while many natural language query systems suffer from returning wrong results, they are also prone to missing many results, particularly where the concept being searched is represented in a different way to the word searched. For example, many data discovery systems do not leverage the power of synonymy, hypernymy, and hyponymy information available through resources developed by the biomedical community (such as [3] and [4]). Query 1 contains a tremendous amount of domain knowledge, both explicit and implicit, that is essential to the proper execution of this kind of query: 1) transcription data include those data from micro-array, "single-cell RNA-seq", spatial transcriptomics and other assays; 2) mammals includes rodents, mini pigs, dogs, humans, and monkeys (and rodents includes mice, rats and others); 3) a "space environment" must include a location (approximately) outside the earth's atmosphere, etc. Capturing and representing this knowledge when indexing data for retrieval or when generating queries of these data have long presented significant challenges for scientific data systems designers [5, 6].

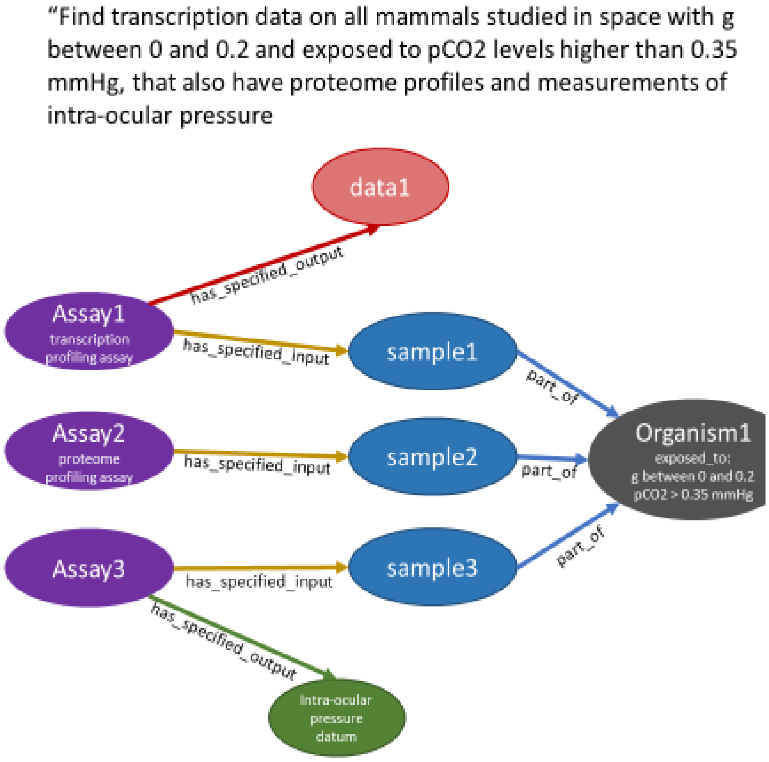
We highly recommend space agencies fund systems that can build and maintain knowledge graphs linking the wide range of data collected through space biology research. Such a knowledge graph would aim to resolve common challenges of data sourcing by vastly improving the way it is both queried and indexed. We recommend building the knowledge graph using current state-of-the-art technologies, and by doing so, space agencies will set a new standard for data collection and retrieval in the biological sciences.

Most scientific data repositories, particularly the largest ones in the biological and physical sciences, rely on relational database technologies. Implementations of the relational database model have evolved into various highly efficient and powerful data management systems capable of rapidly searching and accessing millions of records using modest equipment. However, as knowledge expands in a given domain (application area), maintaining systems that rely on relational database systems (RDBS) is all too often expensive and painstaking, largely because lexical domain knowledge is almost always inextricably embedded in the table schemas at the core of these systems. For example, while the introduction of a new attribute for an object type represented in an RDBS is often fairly straightforward (as simple as adding a new column to an existing table describing the object type's attributes), adding a new object type that has only some of the attributes of an existing object, plus additional new attributes can be quite challenging, and result in large numbers of sparsely populated tables. What this means is that, particularly for a large, established RDBS, these systems cannot evolve to accommodate changes in data structures and become static and difficult to maintain.

Furthermore, the ability of Structured Query Language (SQL), the most-commonly implemented standard for searching RDBS, to leverage domain knowledge easily and automatically is limited. Designers of scientific data repository systems must develop and incorporate bespoke methods for expanding or otherwise enhancing user-generated queries based on domain semantics prior to execution, which are often cost-prohibitive. For example, to expand the query term "mammals", developers might choose to develop software that leverages the NIH taxonomy of living organisms [7], that can transform the term into a SQL phrase consisting of a union of terms describing all sub-types of mammals. This query expansion capability, while critically important, not only requires significant investments in software development, but supervision of the development by those with specialized domain knowledge. In addition, term expansion strategies, when naively used, can lead to large reductions in search precision.

These and other weaknesses of RDBS have led to the development of several alternative formulations for storing and querying data, including those that have come to be labeled "NoSQL" systems. By de-coupling data schemas from the structures used to store the data (to form schema-less systems), NoSql systems feature a robustness to changes in domain knowledge, and, more importantly, have the ability to leverage this knowledge in query generation and execution. Many NoSql systems support the SPARQL query language, a standard for representing queries as a series of triples to filter knowledge or other kinds of graph representations of data. SPARQL can also be used to query knowledge in the form of ontologies, and a large and growing body of biomedical knowledge is currently represented in ontologies through world-wide open source ontology-building efforts such as the OBO Foundry. Using the features of NoSql, SPARQL, and the burgeoning knowledge available in the form of ontologies together, systems can achieve a much higher level of incorporation of domain knowledge into data retrieval.

However, NoSQL systems that are "aggregate orientated" suffer from problems due to the way data is stored. Aggregate orientated NoSQL systems only group data based on a single dedicated view, meaning that to realize new projections and perspectives of the data that it must then be crunched and duplicated. These issues make aggregate orientated NoSQL systems inefficient when needing to compute graph like queries, such as those shown in figure 1. On the other hand, NoSQL native graph databases, such as Neo4J and OrientDB, offer a much better alternative for both graph processing and graph storage. Systems like these are designed specifically to query only the proportion of the data (graph) traversed in the query. Graph databases, like aggregate-orientated NoSQL databases, are schema-free, naturally additive, and do not suffer from query issues such as "join-pain" commonly encountered in RDBS alternatives.



"Find transcription data on all mammals studied in space with g between 0 and 0.2 and exposed to pCO2 levels higher than 0.35 mmHg, that also have proteome profiles and measurements of intra-ocular pressure

Figure 1. Query 1 represented as a graph of linked data.

There are several ways that scientific data discovery systems can and should leverage more suitable query systems to empower users. One such way involves systems dynamically querying a repository of domain semantics like OBO Foundry for terms related in pre-defined way to submitted query terms, and then searching the index of the repository for these additional terms, to enhance recall ("query expansion" [6, 8]). For example, if a user submits a search to OBO Foundry using the terms "eukaryote" and "space radiation", a system could pre-process the query by determining that "solar cosmic radiation" is a kind of "space radiation", and that "Mus musculus" is a type of "eukaryote". The system could then augment the user's original query with the terms "solar cosmic radiation" and "Mus musculus", likely yielding more, relevant query results. Alternatively, the system could have indexed data sets with these additional terms prior to any query ("index expansion"), yielding similar results.

These simple query and index expansion examples indicate the power of leveraging domain semantics to improve data discovery using one kind of domain semantics (subclass relationships), and there are more complex ways to transform user queries using other kinds of domain semantics. For example, by leveraging consider the query that includes the terms "radiation" and "gene expression". Using existing knowledge of domain semantics like that found in the OBO Foundry in the form, this query could automatically be expanded to include the term "RNA-seq", even though that term is not a synonym, hyponym or hypernym of either of the two query terms. In the OBO Foundry knowledge base, the class "gene expression" is identified as an output of the "assay"

“Find transcription data on all mammals studied in space with g between 0 and 0.2 and exposed to pCO_2 levels higher than 0.35 mmHg, that also have proteome profiles and measurements of intra-ocular pressure”

```
PREFIX OBI:   <http://somewhere/peopleInfo#>
PREFIX rdf:   <ww.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX BFO:   < http://www.ifomis.org/bfo
PREFIX NCBITaxon:   <http://purl.obolibrary.org/obo/ncbitaxon>

SELECT ?data1
WHERE
{
    ?assay1 OBI:has_specified_output  ?data1 .
    ?assay1 rdf:type  OBI:’transcription profiling assay’ .
    ?assay1 OBI:has_specified_input ?sample1 .
    ?sample1 BFO:part_of ?organism1 .
    ?organism1 BFO:part_of ?organism1 .
    ?organism1 RBO:exposed_to ‘g between 0 and 0.2’ .
    ?organism1 RBO:exposed_to ‘pCO2 > 0.35 mmHg’ .
    ?sample1 NCBITaxon:order “Mammalia” .
    ?assay2 OBI:has_specified_input ?sample2 .
    ?sample2 BFO:part_of ?organism1 .
    ?assay2 rdf:type  OBI:’proteome profiling assay?’ .
    ?assay3 OBI:has_specified_input ?sample1 .
    ?sample3 BFO:part_of ?organism1 .
    ?assay3 OBI:has_specified_output OBI:’intra-ocular pressure datum’
.
}
```

Figure 2. Query 1 represented as a SPARQL query.

subclass “RNA-seq” (specifically, the class RNA-seq includes the axiom “RNA-seq” “has_specified_output” some [is about some gene expression]”, meaning the assay outputs something that is, at least in part, about the class “gene expression”). The result of using this kind of readily available, community vetted, continuously maintained domain knowledge would be increased search recall. More specifically, it is likely some data sets exist which are identified in repositories as employing RNA-seq, but that make no mention of the concept “gene expression”, the subject of the output of this assay.

The application of the above-described approaches that leverage domain semantics to enhance retrieval precision and recall is certainly a step forward, compared to commonly used, keyword retrieval methods. However, these methods still yield suboptimal results when given the kind of query complexity inherent in Query 1. Consider a system that transformed Query 1 into the disjunction of a set of semantically expanded terms; the transformed query is likely to yield too many irrelevant studies in search results, as it would include any and all studies involving transcription profiling, proteome profile or measuring intra-ocular pressure. And using instead a conjunction of the expanded set of terms could exclude relevant data, when no single investigation used all these kinds of assays. The only way to properly execute Query 1 is if **all** the knowledge implicit in the query is represented in both the query and/or index generated.

Query 1 contains important specifications for how elements of the query *relate* to each other (Figure 1). It seeks data from the transcription profiling assay of samples from an organism from which other samples were taken that were subjected to different kinds of assays. It also stipulates that the organism from which the samples were collected was exposed to certain environmental conditions. This kind of knowledge that the query includes as part of its constraints on the data it seeks forms the basis of "linked data" [9], a vision for using relations to represent and query information as an interconnected graph of objects. Complex queries that reference linked data and objects can only be successfully executed when there is communication of this type of knowledge from users to data retrieval systems.

Linked data can be represented by using any of a variety of technologies, but over the past decade a few have been developed for the specific purpose of efficiently and easily representing linked data. Chief among these technologies are graph databases [10], and the more specialized triplestores [11]. The knowledge graph implicit in the natural language of Query1 could be represented by the example SPARQL query shown in Figure 2. Note that in this example, there are 14 separate constraints on the data sought, and these constraints involve 7 different object types, three instances of samples, three assays, and 1 whole organism, with six different properties linking them all together. These properties each provide powerful and meaningful context for the objects in the query; for example, by specifying that the organism for which transcriptome data are sought also had samples in which intra-ocular pressure was measured (as opposed to searching for the term "intra-ocular pressure" in any metadata context). While the complexity of Query1 as represented either as a graph or in the SPARQL language may seem dauntingly complex, there has been progress in the last decade in the development of interfaces that support users generating these kinds of complex, graph-based queries that include domain semantics [12]. In addition, commercial-grade technologies like Neo4j include visualization modules that support users inspecting SPARQL results and allowing them to validate their results match their query's intent.

Each of the properties in the example query in Figure 2 is specified as current, actual relationship type as specified by ontologies of the OBO Foundry. While this global knowledgebase has yielded broad models in many areas of biomedicine, it currently lacks deep models for specific application areas like the space sciences. It will be imperative for the space science communities to further these models to leverage this knowledge in all kinds of scientific data systems. Augmenting these models using the process and principles recommended by the OBO Foundry, in which ontologies are transparently and cooperatively developed and interconnected by the scientific community themselves, should also be among top priorities for the space agencies, as well as academia and industry involved in space science research.

The challenges to building a deep and broad understanding of the biological effects of off-world environments on organisms and their ecosystems are great. But humankind is still relatively early in this process. However, now is the time for building processes and systems that support the creation and querying of knowledgebases that can support an ever-evolving understanding of space biology. Further delay would doom future investigators to using outdated, brittle data retrieval paradigms and limit their ability to corroborate and augment this knowledge rapidly and accurately.

REFERENCES

[1] UBERON. *Epidural Space*. Available from: http://purl.obolibrary.org/obo/UBERON_0003691.

[2] ENVO. *Outer Space*. Available from: http://purl.obolibrary.org/obo/ENVO_01000637.

[3] Tuttle, M.S., et al. *The semantic foundations of the UMLS metathesaurus*. in *MEDINFO 92. Proceedings of the Seventh World Congress on Medical Informatics*. 1992. Amsterdam, Netherlands: North-Holland.

[4] Smith, B., et al., The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*, 2007. 25(11): p. 1251-5.

[5] Berrios, D.C. and R.M. Keller. *Developing A Web-based User Interface for Semantic Information Retrieval*. in *Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data: ISWC 2003*. 2003. Sanibel Island, Florida, USA.

[6] Voorhees, E.M. *Query expansion using lexical-semantic relations*. in *SIGIR '94. Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. 1994. Berlin, Germany: Springer-Verlag.

[7] National Center for Biotechnology Information, U.S.N.L.o.M. *Taxonomy Database*. Available from: https://www.ncbi.nlm.nih.gov/taxonomy.

[8] Aronson, A.R. and T.C. Rindflesch, Query expansion using the UMLS Metathesaurus. *Proc AMIA Annu Fall Symp*, 1997: p. 485-9.

[9] Wikipedia. *Linked Data*. Available from: https://en.wikipedia.org/wiki/Linked_data.

[10] Wikipedia. *List of Graph Databases*. Available from: https://en.wikipedia.org/wiki/Graph_database#List_of_graph_databases.

[11] Wikipedia. *Triplestore*. Available from: https://en.wikipedia.org/wiki/Triplestore.

[12] Dastgheib, S., et al., *SPARQLing: A Graphical Interface for SPARQL*. 2015.