Constructing Knowledge Graphs with Trust

Conference Paper · October 2015

CITATIONS
0

READS
432

1 author:

Brian Ulicny
Thomson Reuters
50 PUBLICATIONS **174** CITATIONS

SEE PROFILE

Constructing Knowledge Graphs with Trust

Brian Ulicny[1]

[1] Data Innovation Lab, Thomson Reuters,
22 Thomson Pl., Boston, MA 02210

Brian.Ulicny@ThomsonReuters.com

Abstract. This paper argues for an open, fine-grained, freely-accessible scheme for company identifiers in knowledge graphs by means of two general principles that identifiers should comply with and using them to evaluate other candidates. . It goes on to describe how a set of identifiers meeting these criteria has been made available by Thomson Reuters. Thomson Reuters also provides a free web service called Open Calais that enable the construction of knowledge graphs from unstructured text that incorporates these identifiers, along with confidence measures.
Keywords: RDF, ontologies, Open Calais, Thomson Reuters

1 Introduction

Thomson Reuters is a leading source of intelligent information for businesses and professionals in finance and risk, legal, tax and accounting, intellectual property and science and media markets, powered by the world's most trusted news organization. Trust, a theme of this workshop, is a very important concept for Thomson Reuters. In 1941, in the midst of World War II, Thomson Reuters created its Trust Principles, in agreement with the Newspaper Publishers Association and the Reuters shareholders at the time. The Principles imposed obligations on Reuters and its employees, and the employees of its successor company, Thomson Reuters, to act at all times to "supply unbiased and reliable news services" to its customers.[1]

Trusting linked data in knowledge graphs presents unique challenges, since a graph depends on every node and connection. A "knowledge graph", in the generic sense used here, is an extensible connection of links between entities and their attributes. The term is agnostic as to whether the underlying data model is RDF or OWL or some property graph model that allows attribute:value pairs on edges as well as nodes. A knowledge graph representation enables path-based analytics and queries of the underlying data, in order to both find patterns meeting a particular set of constraints within the graph and in order to derive aggregate graph-theoretic metrics of the graph structure as a whole as well as graph-theoretic measures of individual nodes: for example, measures of the centrality of nodes, such as PageRank.

[1] http://thomsonreuters.com/en/about-us/trust-principles.html

The utility of such models is that they allow data to be aggregated without pre-specifying a schema of allowed relations, as in a relational database. New relations between entities and attributes of entities can be added to the graph at any time, without changing anything. Similarly, relations and attributes can be removed without problem. Graphs can be combined freely, since URIs and namespacing guaranteed that connections will be made only where the same entity is involved.

In graph form, this knowledge allows us to determine connections between entities that would be difficult in a standard relational database. In particular, such graphs allow us to identify paths between entities in ways that would be difficult to query for in a standard database context. Relational database queries do not support finding connections or paths between entities that are an unspecified number of links apart.

Because of the connected aspect of knowledge graphs, the integrity of the entire graph depends upon the integrity of every node in the graph to a greater extent than in standard relational databases. In a database, a corrupt or cell of data has limited impact on the integrity of the database as a whole, since it is implicated in only in queries and aggregations that directly involve it. Graph metrics of node centrality and the like, by contrast, depend on evaluating all the paths within a graph: every node is implicated. Hence, a knowledge graph is more vulnerable to incorrect nodes.

To increase trust in knowledge graphs and graphs metrics, it is important that the nodes of a graph denote a single thing. Uniform Resource Identifiers (URIs) or Internationalized Resource Identifiers (IRIs) denote entities in knowledge graphs, and it is important that these URIs/IRIs are unambiguous. This is the Semantic Web correlate of Butler's Maxim: *everything is what it is, and not another thing*.[2] In the context of organizations, this means that a URI/IRI should not denote both a parent company and its subsidiaries, if they are separate legal entities. Nor should the same URI/IRI denote both a company and its product (e.g. BMW makes BMWs), or otherwise denote multiple entities.

Conversely, if two URI/IRIs denote the same entity, then everything that is true of the one must be consistently asserted of the second. That is, every relationship and attribute of the first URI/IRI must be asserted as an attribute/relationship of the second. This is the Semantic Web application of Leibniz's Law or the Law of the Indiscernibility of Identicals.[3] In the context of knowledge graphs, this means that every relation or attribute of a node becomes attributed to any other node that is specified as being *owl:sameAs*.

---

[2] Joseph Buter. Preface to *Fifteen Sermons preached at the Rolls Chapel* (ed. 2, 1729). The maxim was popularized by G. E. Moore in his *Principia Ethica* (1903).

[3] Forrest, Peter, "The Identity of Indiscernibles", *The Stanford Encyclopedia of Philosophy* (Winter 2012 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2012/entries/identity-indiscernible/>.

2 Knowledge Graphs and Company Identifiers

Thomson Reuters provides a great deal of information about the world's corporate entities to its customers. Figure 1 shows the number of documents by category relating to a single company, Boehringer Ingelheim GmbH, a large private pharmaceuticals company based in Germany. This company will be referenced throughout this paper as a running example.

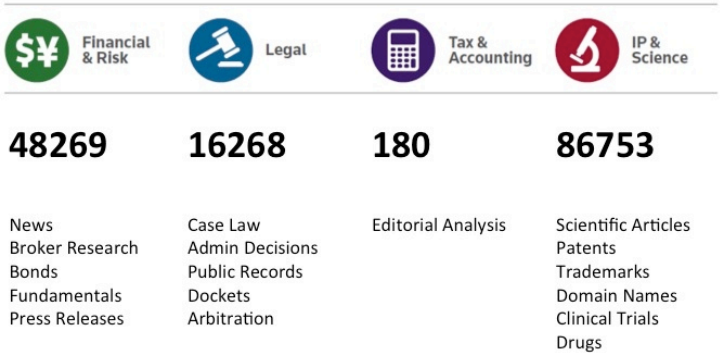


Figure 1 Number of Thomson Reuters documents by Category about one Particular Company, Boehringer Inghelheim, GmbH.

Suppose we would like to construct a knowledge graph from all of this information. It is crucial that the identifiers used are unambiguous and that the semantics of owl:sameAs links do not produce inconsistencies. Figure 2 illustrates various types of entity classes that are important to Thomson Reuters customers. Near the center of this graph is the class 'Company', which is very important to Thomson Reuters' mission

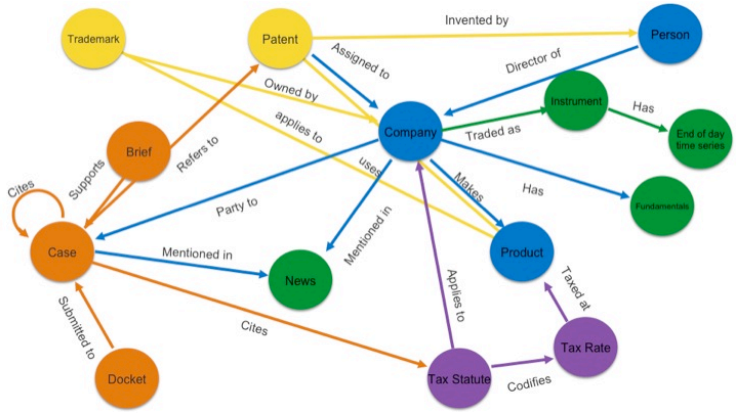


Figure 2 Graph of Node Classes and Relations

What identifiers [1] should be used to for entities such as companies that comply with both Butler's Maxim and the Indiscernibility of Identicals?

Some potential candidates include:

- Thomson Reuters RICs (Reuters Instrument Codes)
- DbPedia URLs
- Dun and Bradstreet DUNS Numbers
- Company website
- Tax identifiers

Reuters Instrument Codes (RICs) are structured and human-readable market-level security identifier, developed and maintained by Thomson Reuters and used extensively by its clients for over 30 years. They are widely used across equities, fixed income, commodities, foreign exchange and money markets as an identifier of both instruments and derivatives. The RIC is made up primarily of the security's ticker symbol, optionally followed by a period and exchange code based on the name of the stock exchange using that ticker. For instance, IBM.N is a RIC that denotes IBM being traded on the New York Stock Exchange. RICs could be transformed into URIs by prepending them with an appropriate URI prefix. However, since it is a private company and not traded, there is no RIC for Boehringer Ingelheim. Thus, RICs are not suitable identifiers for all companies.

DBpedia is an RDF graph derived from Wikipedia info boxes. DBpedia URLs are of the form http://dbpedia.org.org/resource/<resource name>[4]. The URI for Boeheringer Ingelheim would therefore be http://dbpedia.org/page/Boehringer_Ingelheim.

There are many language-specific DBpedia URIs that are specified as being *owl:sameAs* dbr:Boehringer_Ingelheim, such as the German DBpedia URI http://de.dbpedia.org/page/Boehringer_Ingelheim. This fact by itself doesn't violate Butler's Maxim: it is okay for one thing to have several names; it is not okay for one name to have multiple denotations. However, these multiple language URIs impose a high-bar for consistency. By the Indiscernibility of Identicals principle, everything that is said of one of these URIs must be asserted of all of them. In the context of Wiki data, this implies that one is relying on Wikipedia contributors across all languages to make consistent assertions about the same entity. It seems unlikely in practice that this will always be achieved. As it derives from Wikipedia content, DBpedia content is always vulnerable to unreliable or deliberately misleading contributors.[5] The only safeguard is the attention of the crowd. DBpedia URLs are therefore unreliable as company identifiers.

A DUNS number is "[Dun & Bradstreet]'s copyrighted, proprietary means of identifying business entities on a location-specific basis"[6]. The nine-digit number is

[4] dbr: is the namespace prefix.

[5] Julia Carpenter. What we can learn from these less-than-legit Google Knowledge Graph answers. (The Intersect). Washington Post. June 11, 2015.

[6] https://fedgov.dnb.com/webform/pages/dunsnumber.jsp

assigned to each of a business' locations in the D&B database that has a unique, separate, and distinct operation[7]. DUNS numbers may be issued to any business worldwide and are not US-specific. Since 2003, the US Government has required that all grant applicants or awardees must have a DUNS number for identification purposes.[8] The most significant problem in using DUNS numbers as company identifiers is that DUNS numbers essentially denote operational facilities, not companies as legal entities. That is, two persons might work for the same company, but at different locations, with different DUNS numbers. A company could have many factories or other facilities, each of which has a different DUNS number, and none of which has a greater claim to be the unique DUNS number for the company as a whole. Using a DUNS number as an identifier is trying to denote an entity of one type (a legal entity) with an identifier for another type (a physical facility).

Company websites are yet another potential candidate for company identifiers. Nearly every company has a company website these days. However, company websites fail Butler's Maxim because the same website URL can denote different legal entities. The same URL might belong to different companies at different times, for example. To take a more detailed example: there is no adequate way to use website URLs to encode in RDF the merger of Fiat S.p.A. into Fiat Investments N.V., an event described in this document,[9] because it is not clear which URIs should be used for the identifiers. The registrant of the company website http://fiat.it is Fiat S.p.A, it turns out, but it is also the registrant of Fiatspa.com. Either URI/IRI has a claim to denote Fiat S.p.A. However, there is no website corresponding to Fiat Investments N.V. (There is a website http://fcagroup.com that is registered to Fiat Group Automobiles SPA.) In general, then, company websites are not suitable as identifiers for companies because company websites do not necessarily correspond 1:1 to the legal entities that are necessary to represent the events that occur in the world.

Tax identification numbers such as the Employer Identification Number (EIN) used in the US, which is also known as a Federal Tax Identification Number, are another candidate set of identifiers. EINs are usually accessible from publicly available filings for public companies, but in other cases, EINs are not readily accessible and may require subscription fees to look up. Further, a company that does business in multiple countries will have tax identification numbers in all the countries that it is taxed. To guarantee completeness of the graph, based on the completeness of *owl:sameAs* assertions, it would be necessary to look up all sources of tax identifiers for any company in any jurisdiction that it is likely to pay tax and assert the aggregated information about each identifier. As a practical, not theoretical matter, this limits the effectiveness of the tax identification number as company identifier.

What we need, then, is a set of identifiers that is:

1. Comprehensive (cf. Reuters Instrument Codes).

---

[7] https://fedgov.dnb.com/webform

[8] http://www.gpo.gov/fdsys/pkg/FR-2003-06-27/html/03-16356.htm

[9] http://www.fcagroup.com/en-US/investor_relations/merger_of_fiat_spa_with_and_into_FCA_NV /Documents/Equivalent_Document_with_Annexes.pdf

2. Dereferences to reliable and consistent company information (cf. DBpedia URLs)
3. Unambiguously denotes legal entities, as opposed to entities by physical or cyber-location (cf. DUNS numbers and company websites).
4. Is accessible to all (cf. EINs or other subscription-based identifiers).

These considerations have led Thomson Reuters to provide free and open access to URI identifiers for companies based on the permIDs (permanent identifiers) in its internal data model. These URIs are searchable using the publicly available PermID Service at http://permid.org, and they can be bulk downloaded at the same site.[10] The data is continuously updated by human editorial staff at Thomson Reuters.

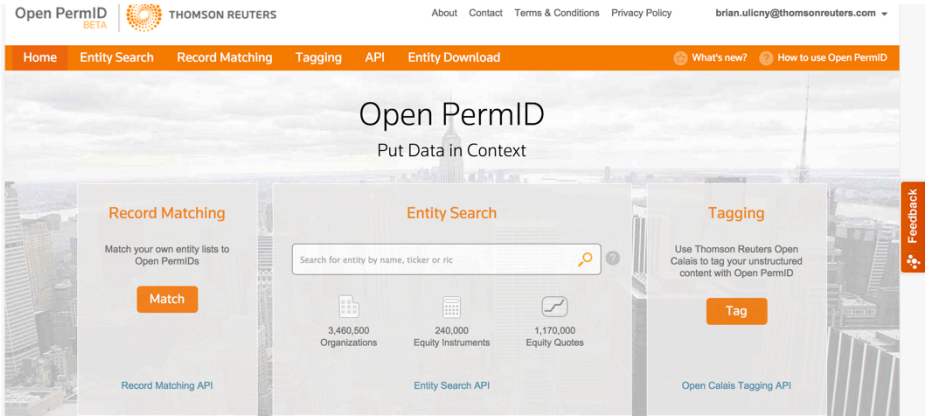


Figure 3 PermID.org Web GUI

PermID.org (Figure 3) is the public-facing portal for Thomson Reuters core entities and metadata, providing the tools that enable users to work with permIDs – unique identifiers for objects in the Thomson Reuters Information Model (TRiM). The PermID Service currently allows users to access the permanent ID of 3.5 million organizations, 240K equity instruments and 1.17 million equity quotes from the Thomson Reuters core entity data set. The PermID Service provides access to Thomson Reuters permanent identifiers (permanent unique IDs formatted as URIs) along with associated descriptive fields that Thomson Reuters exposes to the public. These descriptive metadata fields are provided by dereferencing the company URI and enable the user to verify that a consumed permID represents the entity of interest. They can be easily incorporated directly into a knowledge graph containing the entity permID by dereferencing the permID in Turtle format.

The Entity-Search API enables users to retrieve entities in two ways:

1. By including in the request the PermID Sservice itself.

[10] Bulk downloads of identifiers and associated metadata in Turtle or NTriples format can be accessed at https://permid.org/download.

2. By searching for an entity by name, ticker, or RIC (Reuters Instrument Code).

The data is live; records are updated every 15 minutes. In the future, the PermID Service is expected to support additional entities such as People, Fixed Income Instruments, Fixed Income Quote, and more.

Here is a sample Turtle representation of the metadata associated with Boeheringer Ingelheim International, GMbH:[11]

```
@prefix tr-common: <http://permid.org/ontology/common/> .
@prefix CorporateControl: <http://www.omg.org/spec/EDMC-
FIBO/BE/OwnershipAndControl/CorporateControl/> .
@prefix tr-fin: <http://permid.org/ontology/financial/> .
@prefix fibo-be-oac-cpty: <http://www.omg.org/spec/EDMC-
FIBO/BE/OwnershipAndControl/ControlParties/> .
@prefix mdaas: <http://ont.thomsonreuters.com/mdaas/> .
@prefix fibo-be-le-fbo: <http://www.omg.org/spec/EDMC-
FIBO/BE/LegalEntities/FormalBusinessOrganizations/> .
@prefix xsd:   <http://www.w3.org/2001/XMLSchema#> .
@prefix tr-org: <http://permid.org/ontology/organization/> .
@prefix fibo-be-le-cb: <http://www.omg.org/spec/EDMC-
FIBO/BE/LegalEntities/CorporateBodies/> .
@prefix vcard: <http://www.w3.org/2006/vcard/ns#> .

<https://permid.org/1-4298428312>
         a tr-org:Organization ;
         tr-common:hasPermId          "4298428312"^^xsd:string ;
          tr-org:hasActivityStatus     tr-org:statusActive ;
         tr-org:hasLatestOrganizationFoundedDate   "1958-02-
14T00:00:00Z"^^xsd:dateTime ;
         tr-org:isIncorporatedIn   <http://sws.geonames.org/2921044/> ;
         fibo-be-le-cb:isDomiciledIn http://sws.geonames.org/2921044/ ;
         vcard:organization-name
         "Boehringer Ingelheim International GmbH"^^xsd:string .
```

This metadata incorporates relations and entities from the FIBO (Financial Industry Business Ontology) being developed jointly by OMG and the Enterprise Data Management Council [2] as well as from the W3C's vCard standard and Geonames (for organization country of incorporation and country of domicile) [3].

3 Creating Knowledge Graphs with Trust

An easy way to construct a knowledge graph from a mass of unstructured data is to use the free text tagging service at Permid.org, called Open Calais. Open Calais is a sophisticated Thomson Reuters web service that returns an RDF graph incorporating named-entity metadata-tags to unstructured content, enabling a graph-based representation of the text content that can be integrated with other graphs from structured or

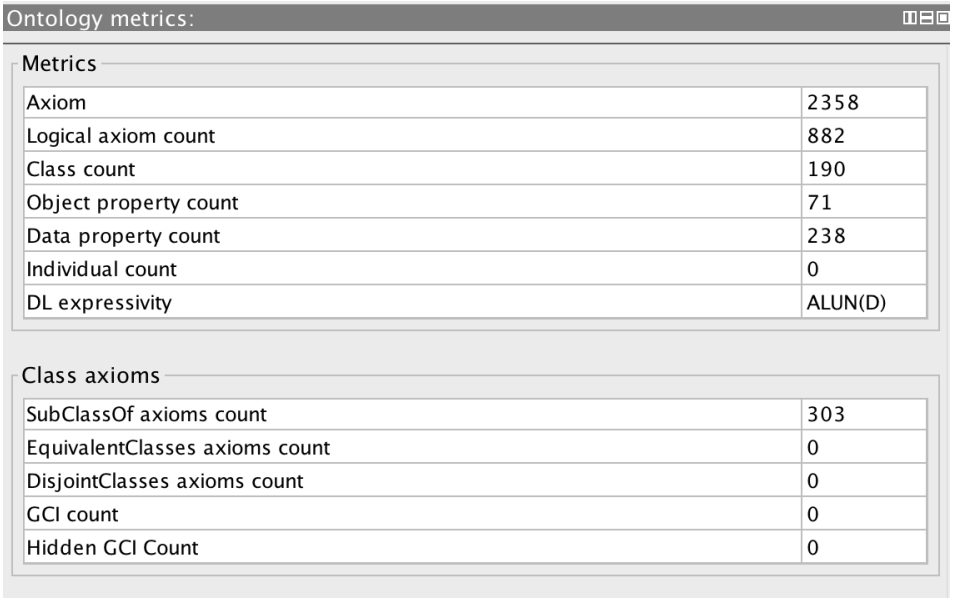
[11] Retrievable at https://permid.org/1-4298428312?format=turtle

unstructured content, enabling powerful analytics. Open Calais analyzes the content of unstructured text files using a combination of statistical, machine-learning, and custom pattern-based methods. Developed by the Text Metadata Services (TMS) group at Thomson Reuters, Open Calais outputs highly accurate and detailed metadata.

The Open Calais ontology to which the RDF graph conforms is available at

http://163.231.4.65/owlschema/8.7/onecalais.owl.allmetadata.xml

Figure 4 details summary statistics for this version (v. 8.7) of the ontology.


Ontology metrics:

Metrics

Axiom	2358
Logical axiom count	882
Class count	190
Object property count	71
Data property count	238
Individual count	0
DL expressivity	ALUN(D)

Class axioms

SubClassOf axioms count	303
EquivalentClasses axioms count	0
DisjointClasses axioms count	0
GCI count	0
Hidden GCI Count	0

Figure 4 OpenCalais Ontology Metrics

Open Calais identifies and tags mentions (text strings) of things like companies, people, deals, geographical locations, industries, physical assets, organizations, products, events, etc., based on the classes and relations in the Open Calais ontology. These are represented in the resultant RDF graph. It also assigns social, category, and industry tags that describe what the input document is about as a whole.

Open Calais classifies mentions of straightforward things like companies, people, cities, telephone numbers, etc. as Entities; more complex mentions that indicate relationships between things are classified as Relations. Some examples of rela-

tions are: deals, IPOs, analyst recommendations, company reorganizations, and product recalls.

An Open Calais output RDF graph contains both textual relations of strings that are mentions of a particular entity or relation as well as metadata about the entity itself. Each mention of a predefined entity or relation type found by Open Calais is expressed as an Instance tag in the output file. The Instance tag describes the mention. It includes the "found" text string itself, the surrounding text, the location and offset of the text string. Each instance is assigned a unique ID.

For example, Open Calais found the following mentions of Tim Cook, the CEO of Apple, Inc., in an article about the anticipated launch of the Apple Watch:

"All Eyes on Apple's **Cook** as Watch Launch Expected"

```
<rdf:Description
rdf:about="http://d.opencalais.com/dochash-1/f4707556-
c36e-39af-b0e6-0103f889be3e/Instance/11">
 <rdf:type
rdf:resource="http://s.opencalais.com/1/type/sys/InstanceInfo"/>
<c:docId rdf:resource="http://d.opencalais.com/dochash-
1/f4707556-c36e-39af-b0e6-0103f889be3e"/>
<c:subject
rdf:resource="http://d.opencalais.com/pershash-
1/e4808181-2cd0-3670-b992-7467229ba691"/>
<!--Person: Tim Cook; -->
<c:detection>[&lt;Title&gt;All Eyes on Apple's ]Cook[ as
Watch Launch Expected&lt;/Title&gt;]</c:detection>
<c:prefix>&lt;Title&gt;All Eyes on Apple's </c:prefix>
<c:exact>Cook</c:exact>
<c:suffix> as Watch Launch Ex-
pected&lt;/Title&gt;</c:suffix>
<c:offset>40</c:offset>
<c:length>4</c:length>
</rdf:Description>
```

Instances of companies with associated permIDs have metadata returned as such in the RDF output graph, along with a confidence metric that the identified company is the company at issue.

For example, a document containing the sentence:[12]

[12] B. Fallon. "Boehringer Ingelheim cites COPD progress". WestFair Communications. September 29, 2015. http://westfaironline.com/74647/boehringer-ingelheim-cites-copd-progress/

```
Global drug company Boehringer Ingelheim Pharmaceuticals
Inc., with its U.S. headquarters in Ridgefield, recently
announced positive data for its new maintenance treatment
called ”Stiolto Respimat” for chronic obstructive pulmo-
nary disorder.
```

Part of the RDF output for this text is as follows:

```
<!— entity -->
<rdf:Description rdf:about="http://d.opencalais.com/er/co
mpany/ralg-oa/4296898441">
<rdf:type rdf:resource="http://s.opencalais.com/1/type/er
/Company"/>
<c:docId rdf:resource="http://d.opencalais.com/dochash-
1/5978c463-325b-39ab-b2a7-2c7943aa7ab8"/>
<c:permid>4296898441</c:permid>
<c:score>0.60709375</c:score>
<!-- Boehringer Ingelheim Pharmaceuticals Inc. -->
<c:subject rdf:resource="http://d.opencalais.com/comphash
-1/b5af4635-b9b5-389d-95bc-f98fb4bec420"/>
<c:legacyid rdf:resource="http://d.opencalais.com/er/comp
any/ralg-tr1r/64cd2908-6aac-3beb-98da-738cf5791239"/>
<c:name>Boehringer Ingelheim Pharmaceuticals Inc</c:name>
<c:commonname>Boehringer</c:commonname>
<c:openpermid rdf:resource="https://permid.org/1-
4296898441"/>
</rdf:Description>

<!— mention -->
<rdf:Description rdf:about="http://d.opencalais.com/comph
ash-1/b5af4635-b9b5-389d-95bc-f98fb4bec420">
<rdf:type rdf:resource="http://s.opencalais.com/1/type/em
/e/Company"/>
<c:forenduserdisplay>true</c:forenduserdisplay>
<c:name>Boehringer Ingelheim Pharmaceuticals
Inc.</c:name>
<c:nationality>N/A</c:nationality>
<c:confidencelevel>0.993</c:confidencelevel>
</rdf:Description>
```

The mention or instance part of the RDF output enables users to construct queries involving company identifiers in the aggregated graphs of Open Calais output where the identification of a company in a text meets or exceeds a specific confidence threshold. It is not the case that the system simply asserts triples for its best guess among identifiers. The confidence with which that assertion is made is also captured directly within the RDF graph itself.

4 Conclusion

Thomson Reuters's Open PermID data and service, along with the free Open Calais tagging tool enables users to construct knowledge graphs from unstructured text easily. These knowledge graphs incorporate company identifiers that are open, free, at the right level of granularity for legal entities and can be dereferenced to retrieve highly reliable, consistent company metadata. Every match for a company with a permID output by the Open Calais engine is marked with a confidence score, enabling users to query relationships between company entities within a specified confidence threshold. As Thomson Reuters proceeds, it expects to make identifiers similarly open and accessible for other important entity types. Knowledge graphs produced using these tools incorporate trust because (1) these knowledge graphs contain unambiguous and consistent identifiers at the right level of granularity, and (2) because they indicate the level of trust the algorithm has that each mention of an entity in the text denotes the associated entity.

References

1. Dodds, L., Phillips, G., Hapuarachchi, T., Bailey, B., Fletcher, A. (2014). *Creating Value with Identifiers in an Open Data World*. Open Data Institute and Thomson Reuters. White Paper. London. October 2014. http://thomsonreuters.com/content/dam/openweb/documents/pdf/corporate/Reports/creating-value-with-identifiers-in-an-open-data-world.pdf
2. The Financial Industry Business Ontology (FIBO) http://www.edmcouncil.org/semanticsrepository/index.html
3. Vatant, B., & Wick, M. (2012). *Geonames ontology*. Chicago .