



Creating Knowledge Graphs with Trust

Brian Ulicny

@bulicny

Director, Data Innovation Lab

Thomson Reuters

**METHOD 2015: 4th Int'l Workshop on Methods for
Establishing Trust of Open Data, Oct 11, 2015**

Who is Thomson Reuters?



FINANCIAL & RISK

Critical news, information & analytics, enables transactions, and connects trading, investing, financial and corporate professionals.



LEGAL

Critical information, decision support tools, software & services to legal, investigation, business and government professionals.



TAX & ACCOUNTING

Integrated tax compliance and accounting information, software & services for professionals in accounting firms, corporations, law firms and government.



INTELLECTUAL PROPERTY & SCIENCE

Comprehensive IP & scientific information, decision support tools & services to enable governments, academia, publishers, corporations & law firms.

REUTERS NEWS

Powered by more than 2,800 journalists reporting in 20 languages from bureaus around the world, **Reuters** is the world's largest international news organization



Our Trust Principles (1941)

- That Thomson Reuters shall at no time pass into the hands of any one interest, group or faction;
- That the integrity, independence and freedom from bias of Thomson Reuters shall at all times be fully preserved;
- ***That Thomson Reuters shall supply unbiased and reliable news services to newspapers, news agencies, broadcasters and other media subscribers and to businesses, governments, institutions, individuals and others with whom Thomson Reuters has or may have contracts***;
- That Thomson Reuters shall pay due regard to the many interests which it serves in addition to those of the media; and
- That no effort shall be spared to expand, develop and adapt the news and other services and products of Thomson Reuters so as to maintain its leading position in the international news and information business.



THOMSON REUTERS

Data Overview, Single Company: Boehringer Ingelheim



Financial
& Risk



Legal



Tax &
Accounting



IP &
Science

48269

16268

180

86753 docs

News
Broker Research
Bonds
Fundamentals
Press Releases

Case Law
Admin Decisions
Public Records
Dockets
Arbitration

Editorial Analysis

Scientific Articles
Patents
Trademarks
Domain Names
Clinical Trials
Drugs

Three Vs at TR:

Velocity from fractions of seconds to quarterly filings.

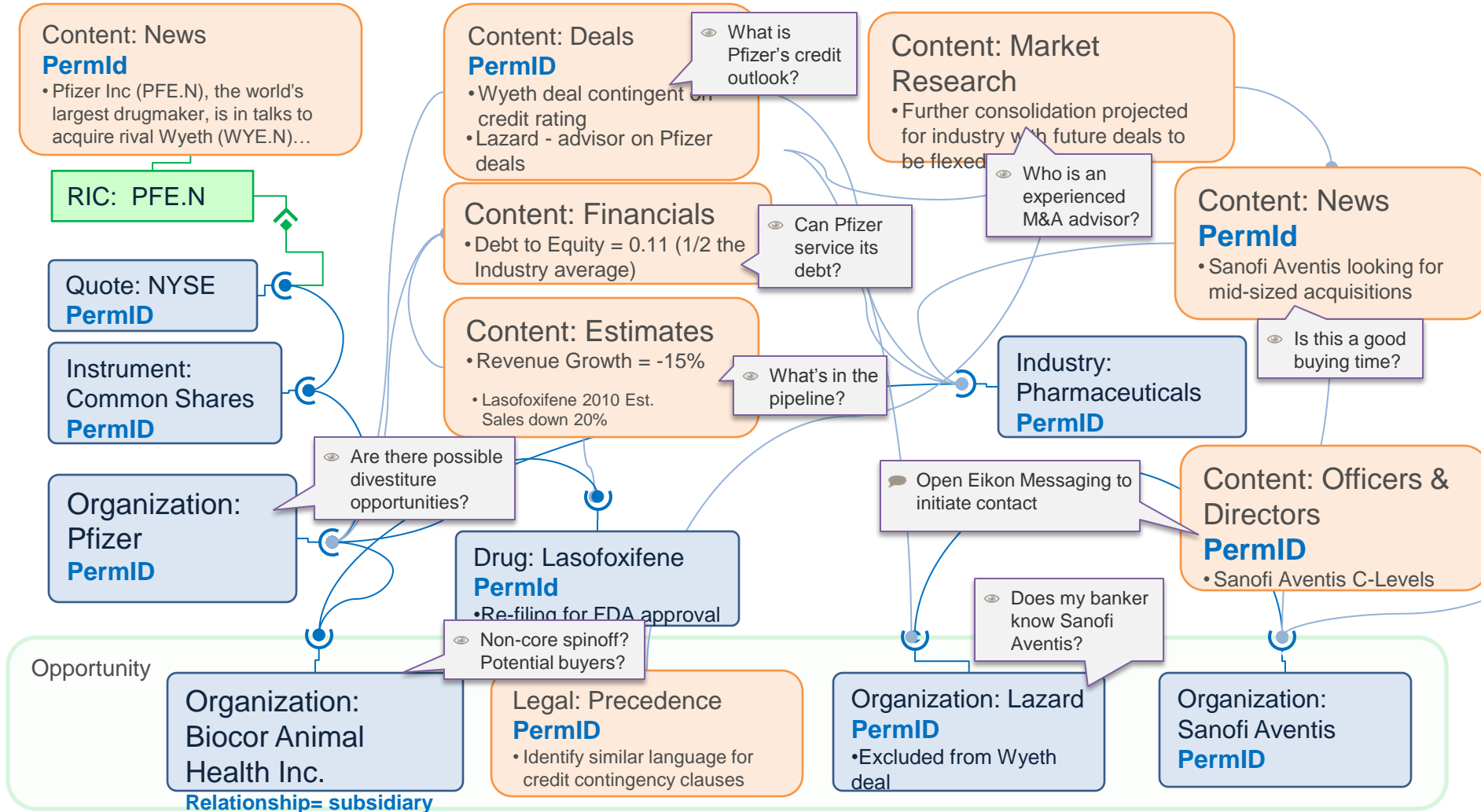
Volume: all the data needed by target professionals

Variety: multiple disparate content, formats, languages.

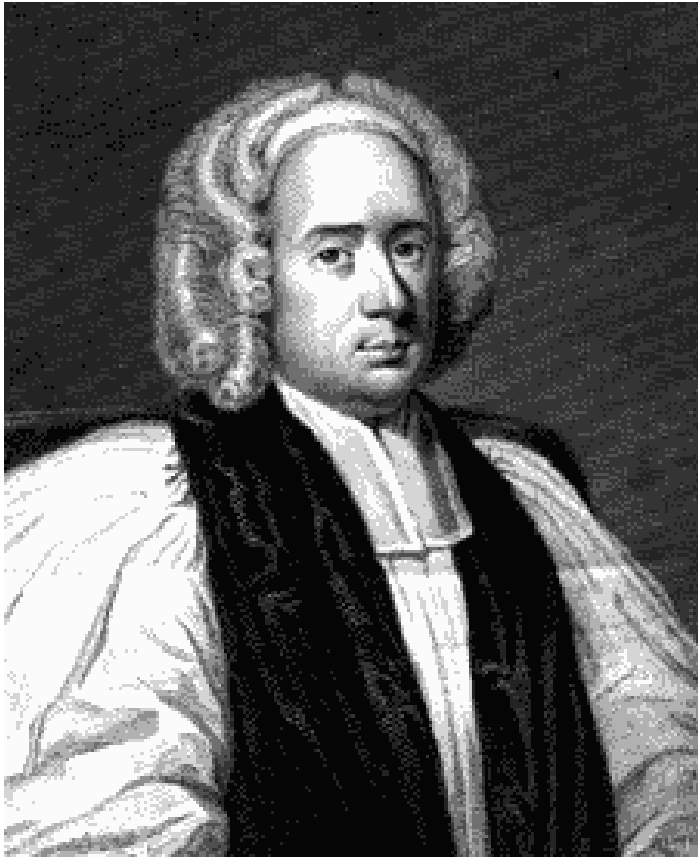


THOMSON REUTERS

Knowledge Graph



How Should We Denote Entities in Graphs?



Joseph Butler (1729): Everything is what it is and not another thing.

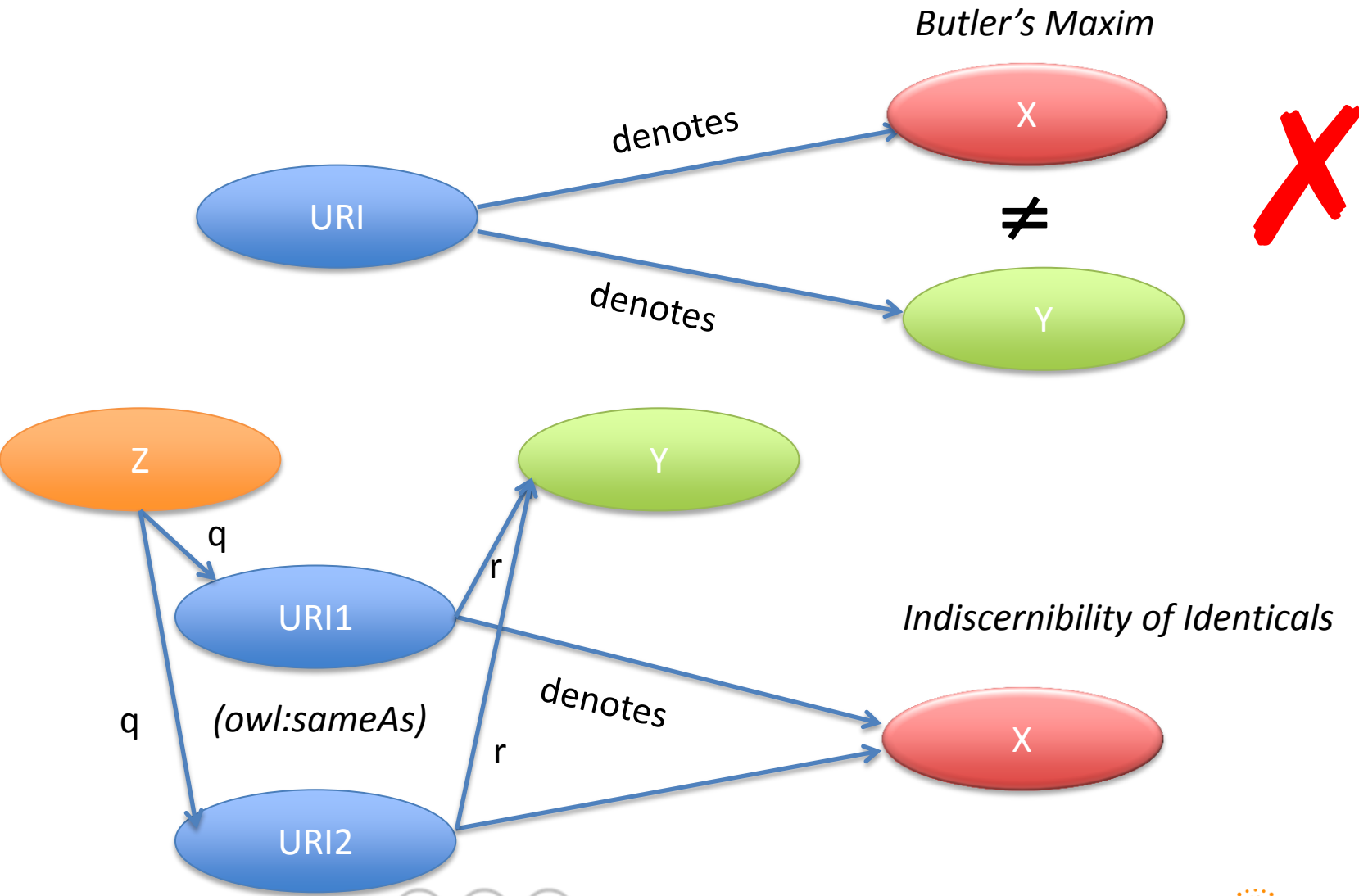
G. W. Leibniz (1686): For any x and y , if x is identical to y , then x and y have all and only the same properties.



$$\forall x \forall y [x = y \rightarrow \forall P (Px \leftrightarrow Py)]$$



In Semantic Web Context




















































Some Candidate Company Identifiers

Identifier	Problem?
Reuters Instrument Code (RIC) e.g. IBM.N	No RICs for private companies like Boehringer Ingelheim
DBpedia URLs	Multiple owl:sameAs URIs (e.g. across languages); can't guarantee consistency (per Ind of Identicals)
Dun & Bradstreet DUNS numbers	Correspond to operational locations. Union of URIs correspond to company. To choose any one DUNS invites inconsistency
Company Website URI	Contra Butler, don't correspond 1:1 to legal entities; so can't represent, e.g. merger of Fiat S.p.A. into Fiat Investments N.V
Tax Identifiers	Not openly accessible; also, potentially multiple for int'l companies, so potentially inconsistent



PermiDs vs Other Symbolologies

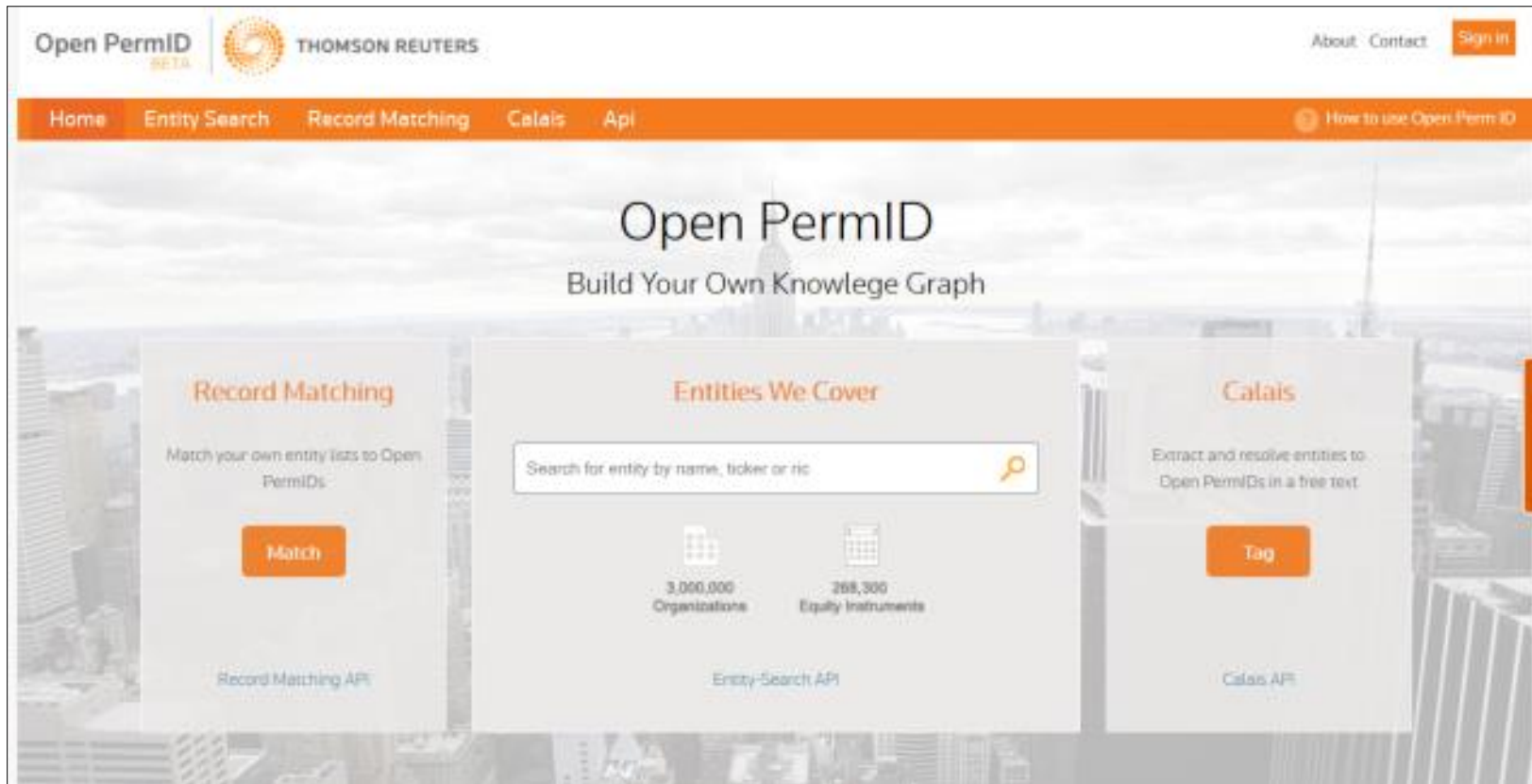
Legend:  Weak  Moderate  Strong

Feature	Description	TR RICs	TR PERMiDs	Typical TR Client	Company Website	Tax IDs	D&B DUNS IDs	Dbpedia URIs
Compre-hensiveness	Covers every financial entity, instrument, and transaction.							
Butler's Maxim	There are no ambiguous symbols.							
Indiscernibility of Identicals	Everything asserted about X and Y = X is true and consistent							
Temporality	Uniqueness of values over time							
Openness	Identifiers are accessible by anyone without any major constraints.							
Third Party Minting	Identifiers can be created by anyone and related information easily linked.							
Information Model	Identifiers are associated with rich info model that provides context to link and understand content.							



THOMSON REUTERS

Open PerMID Site & License



The Open PerMID database is licensed under the [Creative Commons](#) with Attribution license, version 4.0 (CC-BY). A [plain language summary of this license](#) is available on the [Creative Commons website](#).



Permid Dereferencing: Boehringer Ingelheim

@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 .



THOMSON REUTERS

THOMSON REUTERS INTELLIGENT TAGGING
MAKING DATA INTELLIGENT

Content uploaded from news articles, blog postings, proprietary data, catalogs, social media, and more



Our unique identifiers leverage the deep knowledge in Thomson Reuters' professional data, creating semantic metadata to enrich your own content – and also mapping it to Thomson Reuters' content to give the best of both worlds.



Thomson Reuters' key advantage is assigning unique identifiers, or **PermiDs**, which go beyond keywords, returning the right connections you'd otherwise miss.

WHAT IS OPEN CALAIS?

- **Open Calais** is a free service currently accessible via a public website (opencalais.com) and will also be available via a Thomson Reuters sponsored public website, PermID.org.
- This free service provides document tagging using basic fields including companies, people, geography, industry classifications, topics, social tags and events. The service is hosted by Thomson Reuters and allows users to upload up to **5,000 documents per day** (or a maximum upload size of 500MB a day).
- Currently we have about **1,400 active users** of the opencalais.com with the most popular document being tagged as news stories with blog posts close behind.



Calais Output

Entities: ⓘ

- ☒ **City**
 - ☒ Amsterdam, North
- ☒ **Company**
 - ☒ Boehringer Ingelheim
- ☒ **Country**
 - ☒ United States
- ☒ **Medical Condition**
 - ☒ asthma
 - ☒ chronic bronchitis
 - ☒ chronic obstructive pulmonary
 - ☒ COPD
 - ☒ emphysema
- ☒ **Organization**
 - ☒ International Congress
- ☒ **Person**
 - ☒ Bill Fallon
 - ☒ Danny McBryan
- ☒ **Position**
 - ☒ leader
 - ☒ vice president for clinical

September 29, 2015 No Comment

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 pulmonary disorder**.

Stiolto Respimat was approved for use in the **U.S.** in May for the long-term, once-daily maintenance treatment of airflow obstruction in patients with **COPD**, including **chronic bronchitis** and/or **emphysema**. It is not indicated to treat **asthma** or acute deterioration of **COPD**, **Boehringer Ingelheim** said in announcing the results.

The data were presented at the European Respiratory Society **International Congress**, Sept. 26 to Sept. 30, in **Amsterdam**.

Social Tags: ⓘ

Respimat ★★★★★
Boehringer Ingelheim ★★★★★
Chronic lower respiratory diseases ★★★★★
Obstructive lung disease ★★★★★
Olodaterol ★★★★★
Asthma ★★★★★
Bronchitis ★★★★★
Chronic obstructive pulmonary disease ★★★★★

Industries: ⓘ

Bio Therapeutic Drugs (B:1737) 80%

Topics: ⓘ

Health Medical Pharma 100%
Pharmaceuticals & Medical Research (TRBC) 72%
Western Europe (G:3) 54%
Health / Medicine (M:P) 50%



Open Calais: Instances

```
<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 Expected&lt;/Title&gt;</c:suffix>
  <c:offset>40</c:offset>
  <c:length>4</c:length>
</rdf:Description>
```



Confidence Metrics

```
<rdf:Description rdf:about="http://d.opencalais.com/er/company/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/company/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>
```

```
<rdf:Description rdf:about="http://d.opencalais.com/comphash-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>
```



THOMSON REUTERS

Conclusion

- Thomson Reuters's Open PermlD 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 permlD 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.

