Towards a stepwise method for unifying and reconciling corporate names in public contracts metadata. The CORFU technique.

Michalis Vafopoulos (speaker)
and
{Jose María Álvarez-Rodríguez, Patricia Ordoñez De Pablos and Jose Emilio Labra-Gayo}

MTSR 2013 | 7th Metadata and Semantics Research Conference
Track on Metadata and Semantics for Open Repositories, Research Information Systems and Data Infrastructures

The Problem...What is the "Big Name"?

"Oracle (Corp) Aust Pty Ltd" "Oracle Corp (Aust) Pty Ltd" "Oracle Corp Aust Pty Ltd" "Oracle Corp. Australia" "Oracle Corp. Australia Pty.Ltd." "Oracle Corpoartion (Aust) Pty Ltd" "Oracle Corporate Aust Pty Ltd" "Oracle Corporation" "Oracle Risk Consultants" "ORACLE SYSTEMS (AUSTRALIA) PTY LTD" "Oracle University" ...	"Oracle"
"Accenture" "Accenture Aust Holdings" "Accenture Aust Holdings" "Accenture Aust Holdings Pty Ltd" "Accenture Australia Holding P/L" "Accenture Australia Holdings P/Ltd" "Accenture Australia Holdings Pty Lt" "Accenture Australia Limited" ...	"Accenture"
...	...

...and if you have 400K corporate names

Scope

Public Procurement

1. e-Procurement is a strategic sector (17 % of the GDP in Europe).
2. Action Plans 2004 and 2020.
3. Projects: E-Certis, Fiscalis 2013, E-Prior, PEPPOL, STORK,etc.
4. Other actions: TED, RAMON metadata server, CPV, NUTS, etc.
5. Legal Framework.
6. Boost participation with special focus on SMEs.s

...but it also requires...

1. Accomplish with Open Data principles.
2. Improve transparency of public bodies.
3. Track where public money goes.
4. ...

Scope

Public Procurement

1. e-Procurement is a strategic sector (17 % of the GDP in Europe).
2. Action Plans 2004 and 2020.
3. Projects: E-Certis, Fiscalis 2013, E-Prior, PEPPOL, STORK,etc.
4. Other actions: TED, RAMON metadata server, CPV, NUTS, etc.
5. Legal Framework.
6. Boost participation with special focus on SMEs.s

...but it also requires...

1. Accomplish with Open Data principles.
2. Improve transparency of public bodies.
3. Track where public money goes.
4. ...

The Problem...

How can we track public procurement processes?

1. Data and information is already out there.
2. Relevant metadata can be (re)used:
   - Normalized product scheme classifications such as the CPV 2008 (Common Procurement Vocabulary) [1].
   - Territorial units (NUTS).
   - Currency.
3. ...

...and "names"?

Both **Payer** and **Payee** names are not usually normalized.

Normalized and unified names ("Big Name")...

...with the aim of tracking both payers and payees.

The Problem...

How can we track public procurement processes?

1. Data and information is already out there.
2. Relevant metadata can be (re)used:
   - Normalized product scheme classifications such as the CPV 2008 (Common Procurement Vocabulary) [1].
   - Territorial units (NUTS).
   - Currency.
3. ...

...and "names"?

Both **Payer** and **Payee** names are not usually normalized.

Normalized and unified names ("Big Name")...

...with the aim of tracking both payers and payees.

The Problem...

How can we track public procurement processes?

1. Data and information is already out there.
2. Relevant metadata can be (re)used:
   - Normalized product scheme classifications such as the CPV 2008 (Common Procurement Vocabulary) [1].
   - Territorial units (NUTS).
   - Currency.
3. ...

...and "names"?

Both **Payer** and **Payee** names are not usually normalized.

Normalized and unified names ("Big Name")...

...with the aim of tracking both payers and payees.

The Problem...

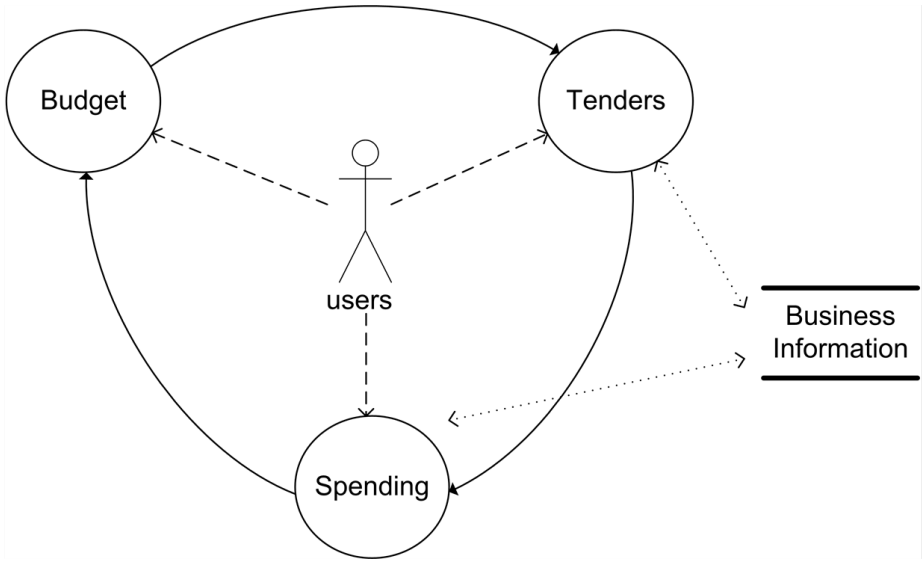
Some remarks...

1. It is **not** a mere problem of reconciling entities (dealing with)...
   - Misspelling errors.
   - Name/acronym mismatches.
   - ...
2. ...but to **create a unique name or link** (n string literals $\rightarrow$ 1 company $\rightarrow$ 1 URI).
3. E.g.
   - "Oracle" and "Oracle University" could be respectively aligned to the entities <Oracle_Corporation> and <Oracle_University>
   - ...but the problem of grouping by a unique (*Big*) name, identifier or resource still remains.

The Public Spending initiative...

...a joint effort trying to answer...

1. Who really gets the public money?
2. For what? From whom?
3. Can we compare them?
4. Is public spending effective?
5. ...



Learn more: `http://publicspending.net/`

Natural Language Processing, Computational Linguistics and Entity Reconciliation.

Existing works and APIs to deal with natural language issues

1. Misspelling errors [13, 8].
2. Name/acronym mismatches [17].
3. APIs such as NLTK for Python, Lingpipe, OpenNLP or Gate for Java and search engines such as Apache Lucene/Solr.
4. Extraction of clinical terms [16] for electronic health records.
5. Creation of bibliometrics [4] or identification of gene names [7, 5].
6. Entity reconciliation processes [6, 2] using the DBPedia [10] or URI comparison [9].
7. Tools such as Google Refine, etc.

Preliminary evaluation...

- Algorithms to deal with natural language heterogeneities are already available.
- Existing works are usually focused in some domain (prototypes cannot be easily customized to other domain).
- ...but methodologies and NLP algorithms can be re-applied to new domains.

Natural Language Processing, Computational Linguistics and Entity Reconciliation.

Existing works and APIs to deal with natural language issues

1. Misspelling errors [13, 8].
2. Name/acronym mismatches [17].
3. APIs such as NLTK for Python, Lingpipe, OpenNLP or Gate for Java and search engines such as Apache Lucene/Solr.
4. Extraction of clinical terms [16] for electronic health records.
5. Creation of bibliometrics [4] or identification of gene names [7, 5].
6. Entity reconciliation processes [6, 2] using the DBPedia [10] or URI comparison [9].
7. Tools such as Google Refine, etc.

Preliminary evaluation...

- Algorithms to deal with natural language heterogeneities are already available.
- Existing works are usually focused in some domain (prototypes cannot be easily customized to other domain).
- ...but methodologies and NLP algorithms can be re-applied to new domains.

Corporate Information.

Corporate Databases

1. Some corporate databases: The Spanish Chambers of Commerce, "Empresia.es" or "Axesor.es" to name a few (just in Spain).
2. The DBPedia and the Orgpedia [3].
3. The CrocTail [12] effort (part of the "Corporate Research Project").
4. "The Open Database Of The Corporate World" [14].
5. Forbes, Google Places, Google Maps, Foursquare, Linkedin Companies or Facebook.
6. Similar initiatives: Openspending.net, LOD2 project e-Procurement, etc.
7. ...

Preliminary evaluation...

- Corporate information is public but access is restricted or under a fee (valuable metadata)...
- Large databases ("infobesity?") but...
- ...the problem of mapping ,(n string literals $\rightarrow$ 1 company $\rightarrow$ 1 URI) as a human would do, still remains.

Corporate Information.

Corporate Databases

1. Some corporate databases: The Spanish Chambers of Commerce, "Empresia.es" or "Axesor.es" to name a few (just in Spain).
2. The DBPedia and the Orgpedia [3].
3. The CrocTail [12] effort (part of the "Corporate Research Project").
4. "The Open Database Of The Corporate World" [14].
5. Forbes, Google Places, Google Maps, Foursquare, Linkedin Companies or Facebook.
6. Similar initiatives: Openspending.net, LOD2 project e-Procurement, etc.
7. ...

Preliminary evaluation...

- Corporate information is public but access is restricted or under a fee (valuable metadata)...
- Large databases ("infobesity?") but...
- ...the problem of mapping ,(n string literals $\rightarrow$ 1 company $\rightarrow$ 1 URI) as a human would do, still remains.

Company, ORganization and Firm name Unifier-CORFU (I)

0
- Load corporate names
  - **Accenture Australia Holding P/L**
  - **Oracle (Corp) Aust Pty Ltd**

1
- Normalize raw text and remove duplicates
  - **Accenture Australia Holding PL**
  - **Oracle Corp Aust Pty Ltd**

2
- Filter the basic set of common stopwords in English
  - **Accenture Australia Holding PL**
  - **Oracle Corp Aust Pty Ltd**

3
- Dictionary-based expansion of common acronyms and filtering
  - **Accenture Australia Holding Proprietary Company Limited**
  - **Oracle Corporation Aust Proprietary Company Limited**

4
- Filter the expanded set of most common words in the dataset
  - **Accenture Australia Holding**
  - **Oracle Aust**

Company, ORganization and Firm name Unifier-CORFU (II)

5
- Identification of contextual information and filtering
  - **Accenture Holding**
  - **Oracle**

6
- Spell checking (optional)
  - **Accenture Holding**
  - **Oracle**

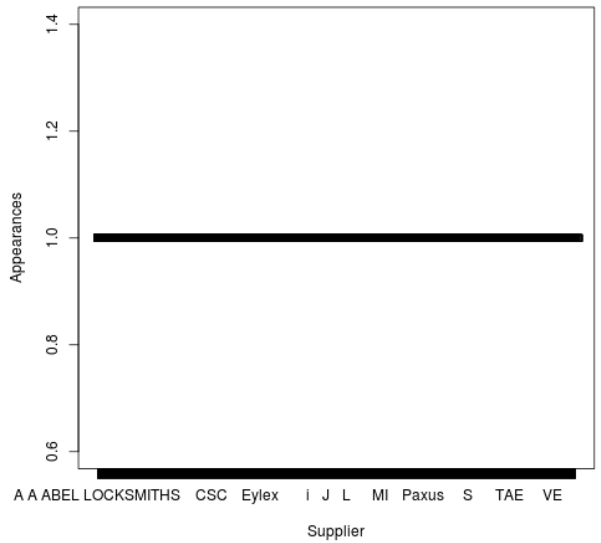
7
- Pos-tagging parts of speech according to a grammar and filtering the non-relevant ones
  - **Accenture**
  - **Oracle**

8
- Cluster corporate names
  - **Accenture**
  - **Oracle**

9
- Validate and reconcile the generated corporate name via an existing reconcile service (optional)
  - **(Accenture, dbpedia-res:Accenture)**
  - **(Oracle, dbpedia-res:Oracle_Corporation)**

Example step by step...

- Scenario: Australian supplier names (400K).
- 1^{st} try: use of Google Refine+Open Corporates reconciliation service (just an 8 % of unified names, see Figure below).
- 2^{nd} try: design of the **CORFU** technique using Python NLTK and other third-party APIs for NLP processing.



Step 0: Load corporate names

Load

Input: a list of corporate names as raw text (one per line).

Output: a set of names represented as strings.

Example:
- "Accenture Australia Holding P/L"
- "Oracle (Corp) Aust Pty Ltd"

Step 1: Normalize raw text and remove duplicates

Normalize

Input: a set of names represented as strings.

Process: this step is comprised of:

1. Remove strange characters and punctuation marks but keeping those that are part of a word avoiding potential changes in abbreviations or acronyms;
2. Lowercase raw text (although some semantics can be lost previous works and empirical tests show that this is the best approach).
3. Remove duplicates.
4. Lemmatize the corporate name.

Output: a set of normalized corporate names.

Example:

- "Accenture Australia Holding PL"
- "Oracle Corp Aust Pty Ltd"

Step 2: Filter the basic set of common stopwords in English

Filter

Input: a set of normalized corporate names and a set of stopwords.

Process: load a set of stopwords (a minimal set of stopwords from the Python NLTK API has been used):

- Including common English stopwords but...
- Avoiding to filter relevant words.

Output: a set of cleaned and normalized corporate names.

Example:

- "Accenture Australia Holding PL"
- "Oracle Corp Aust Pty Ltd"

Step 3: Dictionary-based expansion of common acronyms and filtering

Dictionary-based expansion

Input: a set of cleaned and normalized corporate names and a dictionary of acronyms.

Process: load the dictionary of acronyms and expand.

Output: a set of cleaned, normalized and acronym-expanded corporate names.

Example:

- "Accenture Australia Holding Proprietary Company Limited"
- "Oracle Corporation Aust Proprietary Company Limited"

Step 4: Filter the expanded set of most common words in the dataset

Filter

Input: a set of cleaned, normalized and acronym-expanded corporate names.

Process: extract statistics of "most used words" in the input dataset, expand those words and filter.

Output: a set of cleaned, normalized and acronym-expanded corporate names.

Example:

- "Accenture Australia Holding"
- "Oracle Aust"

Step 5: Identification of contextual information and filtering

Context filtering

Input: a set of cleaned, normalized and acronym-expanded corporate names.

Output: a set of cleaned, normalized and acronym-expanded corporate names without contextual information.

Example:

- "Accenture Holding"
- "Oracle"

Step 6: Spell checking (optional)

Spell checking

Input: a set of cleaned, normalized and acronym-expanded corporate names without contextual information.

Output: the previous set without spelling errors.

Example:

- "Accenture Holding"
- "Oracle"

Step 7: Pos-tagging parts of speech according to a grammar and filtering the non-relevant ones

Pos-tagging

Input: a set of cleaned, normalized and acronym-expanded corporate names without contextual information.

Output: a tree according to a pre-defined grammar for corporate names that only contains nouns.

Example:

- "Accenture"
- "Oracle"

Step 8: Cluster corporate names

Clustering

Input: a set of strings derivate from the aforementioned tree.

Output: a set of clusters for each extracted corporate name.

Example:

- "Accenture"
- "Oracle"

Step 9: Validate and reconcile the generated corporate name via an existing reconcile service (optional)

Validate and reconcile

Input: a set of clusters for each extracted corporate name.

Output: (n string literals $\rightarrow$ 1 company $\rightarrow$ 1 URI).

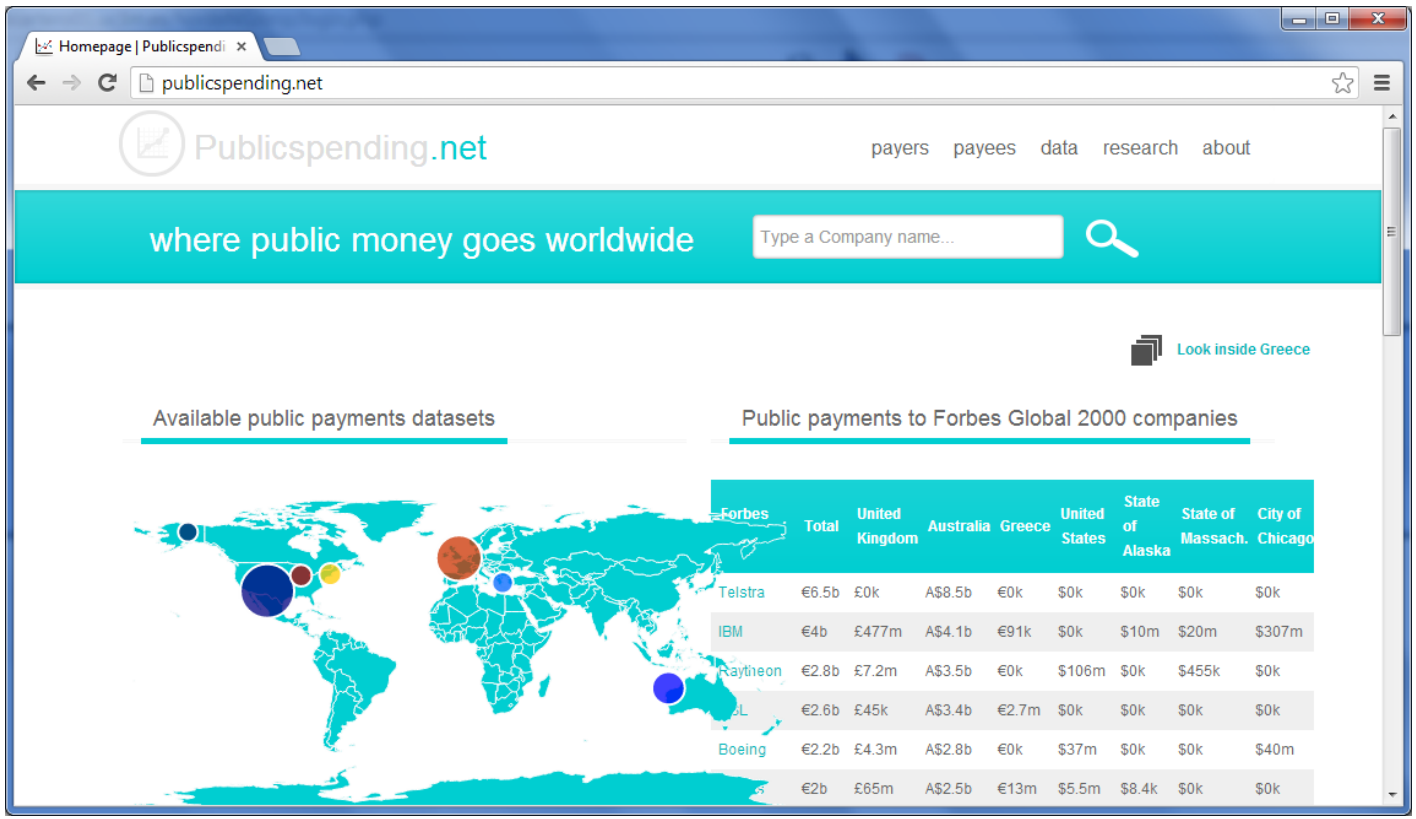
Example:
- ("Accenture", dbpedia-res:Accenture)
- ("Oracle", dbpedia-res:Oracle_Corporation)

Representing and querying the corporate information in RDF...

```
:o1 a org:Organization;
  skos:prefLabel ‘‘Oracle’’;
  skos:altLabel  ‘‘Oracle Corporation’’, ‘‘Oracle (Corp) Aust Pty Ltd’’, ...;
  skos:closeMatch  dbpedia-res: Oracle_Corporation;
   ...
.
```

```
SELECT  str(?label) (COUNT(?org) as ?pCount) WHERE{
  ?ppn :rewarded-to ?org .
  ?org rdf:type org:Organization.
  ?org skos:prefLabel ?label.
  ...
}
GROUP BY str(?label)
ORDER BY desc(?pCount)
```

Demo and application...



Learn more: `http://publicspending.net/`

The case of Australian supplier names...

Step	Name	Customization
0	Load corporate names	430188 full names and 77526 unique names (period 2004-2012)
1	Normalize raw text and remove duplicates	Default
2	Filter the basic set of common stopwords in English	Default
3	Filter the expanded set of most common words in the dataset	Two stopwords sets: 355 words (manually) and words with more than $n = 50$ apparitions (automatically)
4	Dictionary-based expansion of common acronyms and filtering	Set of 50 acronyms variations (manually)
5	Identification of contextual information and filtering	Use of Geonames REST service
6	Spell checking (optional)	Train dataset of 128457 words provided by Peter Norvig's spell-checker [13].
7	Pos-tagging parts of speech according to a grammar and filtering the non-relevant ones	Default
8	Cluster corporate names	Default
9	Validate and reconcile the generated corporate name via an existing reconcile service (optional)	Python client and Google Refine

Research Design

1. Configure and execute the CORFU technique.
2. Validate (manually) the dump of unified names and calculate:
   - **Precision**, see Eq. 1, is "the number of supplier names that have been correctly unified under the same name"
   - **Recall** is, see Eq. 2, "the number of supplier names that have not been correctly classified under a proper name" and **F1** score, see Eq. 3, where...
   - ... *tp* is "the number of corporate names **properly unified**"
   - ... *fp* is "the number of corporate names **wrongly unified**"
   - ... *tn* is "the number of corporate names **properly non-unified**" and
   - ... *fn* is "the number of corporate names **wrongly non-unified**".

$$Precision = \frac{tp}{tp + fp} \quad (1) \qquad Recall = \frac{tp}{tp + fn} \quad (2)$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (3)$$

Results of applying the CORFU approach to the Australian supplier names.

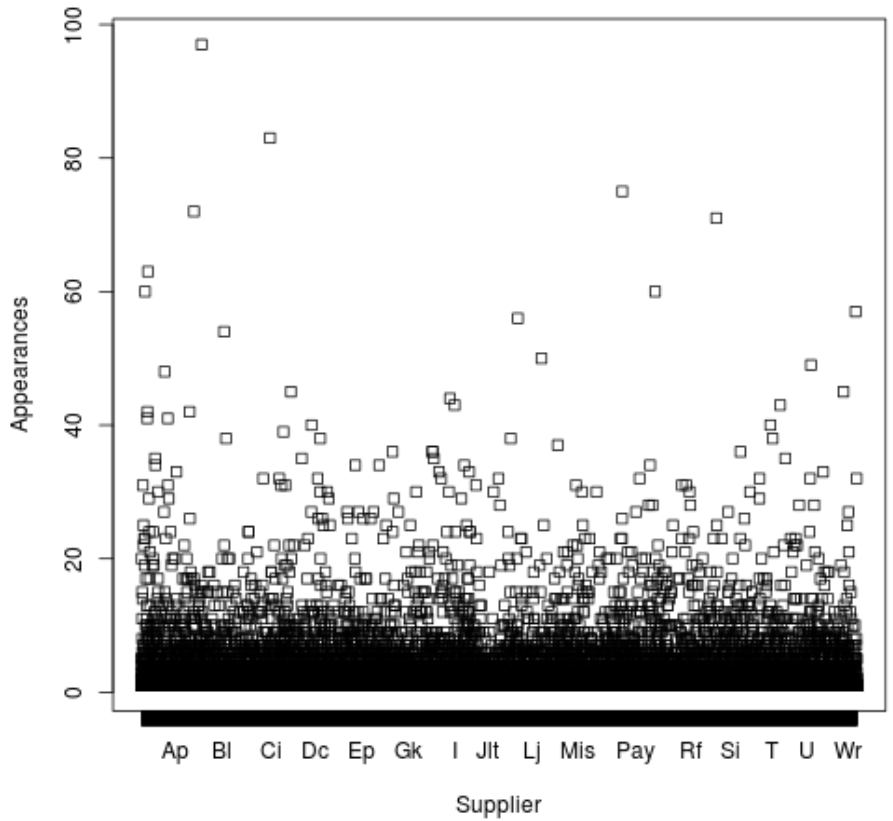
Total number of companies	Unique names	CORFU unified names	% of unified names	Precision	Recall	F1 score
430188	77526	40277	48 %	0,762	0,311	0,441
430188	299 in 77526	68	100 %	0,926	0,926	0,926

Comments

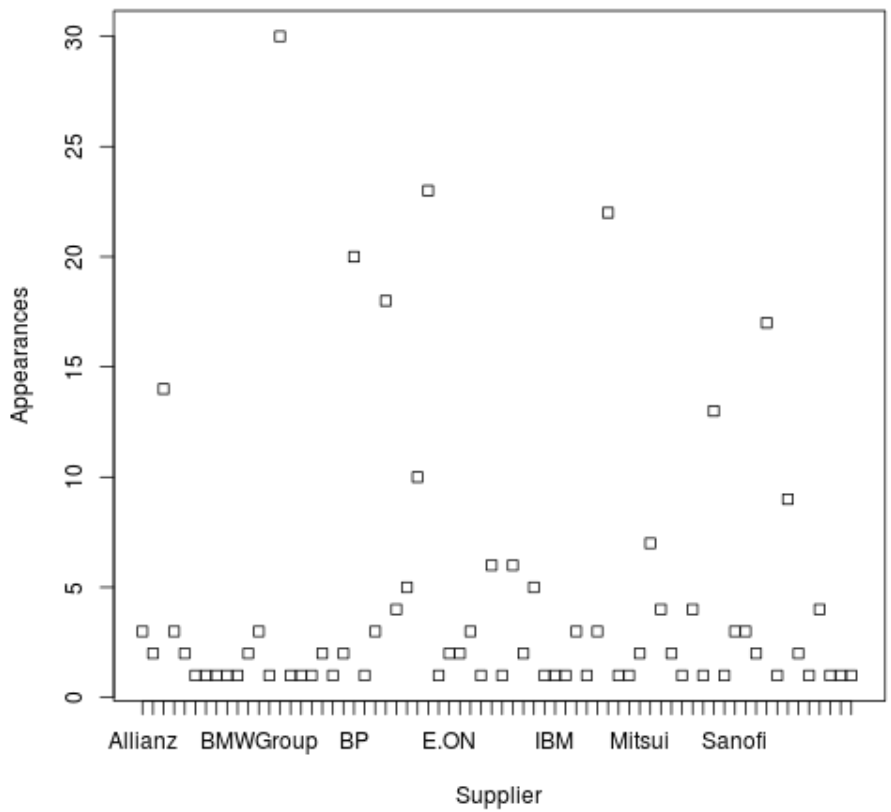
- A 48 % (77526 − 40278 = 37248) of supplier names have been unified with a precision of 0,762 and a recall of 0,311 (best values must be close to 1).
- The first 100 companies in the Forbes list, actually 68 companies were found in the dataset with 299 appearances.
- Results [a], in this second case, show a better performance: precision, 0,926, and recall, 0,926.

[a] Best values must be close to 1.

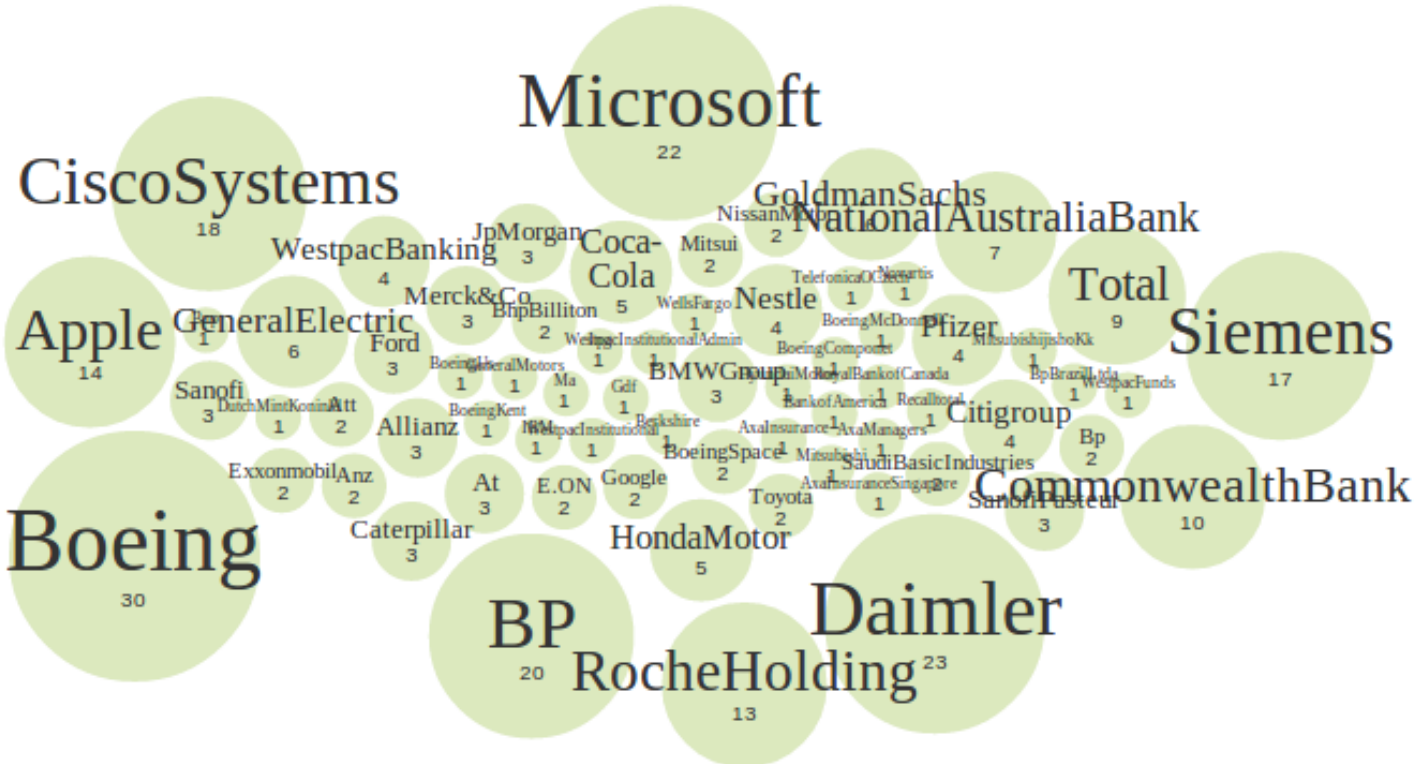
Graphical view after applying the CORFU technique...



Graphical view after applying the CORFU technique to the first 100 Forbes companies...



...the first 100 Forbes companies in bubbles...



Discussion

Advantages

1. A **custom** technique for a particular domain.
2. Unification works pretty nice.
3. It enables the possibility of comparing companies in public procurement.

Drawbacks

1. Execution time (20' for Australian corporate names).
2. It is necessary to test with other datasets.
3. It requires the use of more advanced data mining techniques for machine learning.
4. It should be applied to other domains (Bibliometrics?).

Discussion

Advantages

1. A **custom** technique for a particular domain.
2. Unification works pretty nice.
3. It enables the possibility of comparing companies in public procurement.

Drawbacks

1. Execution time (20' for Australian corporate names).
2. It is necessary to test with other datasets.
3. It requires the use of more advanced data mining techniques for machine learning.
4. It should be applied to other domains (Bibliometrics?).

Conclusions

1. The public e-Procurement sector is seeking for new methods to:
   - ...improve interoperability
   - ...boost transparency
   - ...or increase participation to name a few.
2. The PublicSpending initiative is addressing some of the challenges in the e-Procurement sector.
3. The CORFU technique is a key enabler to ease the comparison of countries, payers, payees, etc.
4. ...a technique that helps to take the most of data.
5. It must be technically improved and extended to cover more datasets and to be "smarter".

Future Work

1. Application to new public procurement datasets.
2. Reuse the Opencorporates reconciliation service (it has been updated).
3. Contribute to the e-Procurement sector and the PublicSpending initiative [15, 11].
4. Add more advanced NLP techniques: $n - grams$.
5. Add machine learning algorithms to automatically classify new corporate names.
6. Improve the execution time (performance).
7. ...


THANK
YOU
GRACIAS
ARIGATO
SHUKURIA
DANKSCHEEN
BİYAN
SHUKRIA
TINGKI
YAQHANYELAY
TASHAKKUR ATU
SUKSAMA
EKHMET
MEHRBANI
GRAZIE
MAAKE
KOMAPSUMNIDA
GOZAIMASHITA
EFCHARISTO
JUSPAXAR
PALDIES
BOLZİN
MERCI
SPASSIBO
SNACHALHUYA
NUHUN
CHALTU
WABEEJA
MAITEKA
YUSPAGABTAM
HUI
GUI
HATUR
SPASIBO
DENKAUJA
NENACHALHYA
UNALCHÉESH
EKOJU
SIKOMO
MAKETAI
MINMONCHAR
DHANYABAAD
AHHIA
ATTO
MERSI
SAHCO
MERASTAWHY
GAEJTHO
LAH
TAVTAPUCH
MEDAWAGSE
AGUYJE
FAKAAUE
BAHKA

Acknowledgements...

- Dr. Michalis Vafopoulos
- Leader of the Public Spending initiative.
- E-mail: `vafo@me.com`
- WWW: `http://vafopoulos.org/`

The Public Spending team:

- Marios Meimaris
- Giannis Xidias
- Giorgos Vafeiadis
- Michalis Klonaras
- Panagiotis Kranidiotis
- Prodromos Tsiavos
- Georgia Lioliou
- WWW: http://publicspending.net

Roster...

- Dr. Jose María Alvarez-Rodríguez
- SEERC (until August, 2013) and Carlos III University of Madrid, Spain
- E-mail: josemaria.alvarez@uc3m.es
- WWW: http://www.josemalvarez.es

- Prof. Dr. Patricia Ordoñez De Pablos
- University of Oviedo, Spain
- E-mail: patriop@uniovi.es

- Prof. Dr. Jose Emilio Labra-Gayo
- University of Oviedo, Spain
- E-mail: labra@uniovi.es
- WWW: http://www.di.uniovi.es/~labra


J. M. Alvarez-Rodríguez, J. E. Labra-Gayo, A. Rodríguez-González, and P. O. D. Pablos.
Empowering the access to public procurement opportunities by means of linking controlled vocabularies. A case study of Product Scheme Classifications in the European e-Procurement sector.
Computers in Human Behavior, (0):–, 2013.


S. Araujo, J. Hidders, D. Schwabe, and A. P. De Vries.
SERIMI – Resource Description Similarity , RDF Instance Matching and Interlinking.
WebDB 2012, 2011.


J. Erickson.
TWC RPI's OrgPedia Technology Demonstrator, May 2013.
http://tw.rpi.edu/orgpedia/.


C. Galvez and F. Moya-Anegón.
The unification of institutional addresses applying parametrized finite-state graphs (P-FSG).
Scientometrics, 69(2):323–345, 2006.


C. Galvez and F. Moya-Anegón.
A Dictionary-Based Approach to Normalizing Gene Names in One Domain of Knowledge from the Biomedical Literature.
Journal of Documentation, 68(1):5–30, 2012.


R. Isele, A. Jentzsch, and C. Bizer.
Silk Server - Adding missing Links while consuming Linked Data.
In COLD, 2010.


M. Krauthammer and G. Nenadic.
Term identification in the biomedical literature.
J. of Biomedical Informatics, 37(6):512–526, Dec. 2004.


S. N. L. P. Lecture.

Spelling Correction and the Noisy Channel. The Spelling Correction Task, Mar. 2013.
http://www.stanford.edu/class/cs124/lec/spelling.pdf.

F. Maali, R. Cyganiak, and V. Peristeras.

Re-using Cool URIs: Entity Reconciliation Against LOD Hubs.
In C. Bizer, T. Heath, T. Berners-Lee, and M. Hausenblas, editors, LDOW, CEUR Workshop Proceedings. CEUR-WS.org, 2011.

P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer.

DBpedia spotlight: shedding light on the web of documents.
In Proc. of the 7th International Conference on Semantic Systems, I-Semantics '11, pages 1–8, New York, NY, USA, 2011. ACM.

M. M. Michail Vafopoulos and G. X. et al.

Publicspending. gr: interconnecting and visualizing Greek public expenditure following Linked Open Data directives, jul 2012.

G. Michalec and S. Bender-deMoll.

Browser and API for CorpWatch, May 2013.
http://croctail.corpwatch.org/.

P. Norvig.

How to Write a Spelling Corrector, Mar. 2013.
http://norvig.com/spell-correct.html.

C. Taggart and R. McKinnon.

The Open Database Of The Corporate World, May 2013.
http://opencorporates.com/.

M. Vafopoulos.

The Web economy: goods, users, models and policies. Foundations and Trends® in Web Science, volume 1.

Now Publishers Inc., 2012.

Y. Wang.

Annotating and recognising named entities in clinical notes.
In Proceedings of the ACL-IJCNLP 2009 Student Research Workshop, ACLstudent '09, pages 18–26, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

S. Yeates.

Automatic Extraction of Acronyms from Text.
In University of Waikato, pages 117–124, 1999.

Towards a stepwise method for unifying and reconciling corporate names in public contracts metadata. The CORFU technique.

Michalis Vafopoulos (speaker)
and
{Jose María Álvarez-Rodríguez, Patricia Ordoñez De Pablos and Jose Emilio Labra-Gayo}

MTSR 2013 | 7th Metadata and Semantics Research Conference
Track on Metadata and Semantics for Open Repositories, Research Information Systems and Data Infrastructures