Alvarez Rodríguez, J.M., Vafopoulos, M., Llorens, J. (2015). Enabling policy making processes by unifying and reconciling corporate names in public procurement data. The CORFU technique. *Computer Standards & Interfaces*, 41, pp. 28-38.

DOI: https://doi.org/10.1016/j.csi.2015.02.009

Enabling policy making processes by unifying and reconciling corporate names in public procurement data. The CORFU technique.

Jose María Alvarez-Rodríguez[a,*], Michalis Vafopoulos[b], Juan Llorens[a],

[a] *Carlos III University of Madrid*
Department of Computer Science
Avd. de la Universidad, 30, Leganés
Madrid, Spain
[b] *National Technical University of Athens*
9 Heroon Polytechneiou st. 15773, Zografou Campus,
Athens, Greece

Abstract

The present paper introduces a technique to deal with corporate names ambiguities and heterogeneities in the context of public procurement meta-data. Public bodies are currently facing a big challenge trying to improve both the performance and the transparency of administrative processes. The e-Government and Open Linked Data initiatives have emerged as efforts to tackle existing interoperability and integration issues among ICT-based systems but the creation of a real transparent environment requires much more than the simple publication of data and information in specific open formats; data and information quality is the next major step in the pubic sector. More specifically, in the e-Procurement domain, there is a vast amount of valuable meta-data that is already available via the Internet protocols and formats and can be used for the creation of new added-value services. Nevertheless, the simple extraction of statistics or creation of reports can imply extra tasks with regards to clean, prepare and reconcile data. On the other hand, transparency has become a major objective in public administrations and, in the case of public procurement, one of the most interesting services lies in tracking rewarded contracts (mainly type, location, and supplier). Although it seems a basic kind of reporting service the reality is that its generation can turn into a complex task due to a lack of standardization in supplier names or the use of different descriptors for the type of contract. In this paper, a stepwise method based on natural

language processing techniques and semantics to address the unification of corporate names is defined and implemented. After that, the technique is applied to the "PublicSpending.net" initiative to show how the unification of corporate names is a key activity to provide a visualization service that can serve policy-makers as a tool to detect and prevent new necessities in terms of products or services in a specific area. Furthermore, a research study to evaluate the precision and recall of the proposed technique, using as use case the public dataset of rewarded public contracts in Australia during the period 2004-2012, is also presented. Moreover, a robustness and refinement process is also outlined including company data from public contracts in United States, United Kingdom and the CrocTail project. Finally some discussion, conclusions and future work are also outlined.

Keywords: name disambiguation, natural language processing, public spending, semantics, linked data
68T30, 68T35, 68T50, 68W32, 68U15, 68U35

1. Introduction

Public bodies are continuously publishing procurement opportunities in which valuable meta-data is available. Depending on the stage of the process, new data pieces such as the supplier name that has been rewarded with the public contract arise. In this context, the extraction of statistics on how many contracts have been rewarded to the same company is a relevant indicator to evaluate the transparency of the whole process. Although companies that want to tender for a public contract must be officially registered and have an unique identification number, the reality is that in most of rewarded contracts the supplier is only identified by a name or a string literal typed by a civil-servant. In this sense, there is not usually a connection between the official company registry and the process of rewarding contracts implying different interoperability issues (Guijarro, 2009) such as naming problems and data inconsistencies that are spread to further stages hindering future activities such as reporting.

In the case of the type of contract and location, there are already standardized product scheme classifications (Rodríguez et al., 2012, 2014) such

*Corresponding author
Email addresses: josemaria.alvarez@uc3m.es (Jose María Alvarez-Rodríguez), vaf@aegean.gr (Michalis Vafopoulos), juan.llorens@uc3m.es (Juan Llorens)

as the Common Procurement Vocabulary (2003 and 2008), the Combined Nomenclature (2012), the Central Product Classification by the European Union, the International Standard Industrial Classification of All Economic Activities (Rev. 4) by the United Nations or the North American Industry Classification System (2007 and 2012) by the Government of United States that are currently used with different objectives such as statistics, tagging or information retrieval. Geo-located information can be also found in different common datasets and nomenclatures such as the Nomenclature of territorial units for statistics (NUTS) by the European Union, the Geonames dataset [1], the GeoLinkedData initiative (López-Pellicer et al., 2010; Husain et al., 2011) or the traditional list of countries and ISO-codes.

However, corporate, organization, firm, company or institution names (hereafter, these names will be used to refer to the same entity) and structure are not yet standardized at global scope and only some classifications of economic activities or company identifiers such as the TARIC database (On-line customs tariff database) can be found. Thus, the simple task of grouping contracts by a supplier is not a mere process of searching by the same literal. Technical issues such as hyphenation, use of abbreviations or acronyms and transliteration are common problems that must be addressed in order to provide a final corporate name. Existing works in the field of Name Entity Recognition (Nadeau and Sekine, 2007; Jung, 2012; Marrero et al., 2013) (NER) or name entity disambiguation (Sarmento et al., 2009; Klein et al., 2003; García et al., 2012) have already addressed these issues. Nevertheless, the problem that is being tackled in these approaches lies in the identification of organization names in a raw text while in the e-Procurement sector the string literal identifying a supplier is already known.

In the particular case of the Australian e-Procurement domain, the supplier name seems to be introduced by typing a string literal without any assistance or auto-complete method. Obviously, a variety of errors and variants for the same company, see Table 6 in the Appendix I, can be found: misspelling errors (Norvig, 2013; Lecture, 2013b), name and acronym mismatches (Yeates, 1999; Ratinov and Gudes, 2004) or context-aware data that is already known when the dataset is processed, e.g. country or year. Furthermore, it is also well-known that a large company can be divided into several divisions or departments but from a statistical point of view grouping data by a supplier name should take into account all rewarded contracts regardless the structure of the company.

[1] `http://www.geonames.org/`

On the other hand, the application of semantic technologies and the Linking Open Data initiative (hereafter LOD) (Berners-Lee, 2006; Heath and Ch. Bizer, 2011) in several fields like e-Government (e.g. the Open Government Data effort) tries to improve the knowledge about a specific area providing common data models and formats to share information and data between agents. More specifically, in the European e-Procurement context (European Commission, 2011), there is an increasing commitment to boost the use of electronic communications and transactions processing by government institutions and other public sector organizations in order to provide added-value services (Palacios et al., 2013) with special focus on SMEs (Small and Medium Enterprises).

In this context, the LOD initiative seeks for creating a public and open data repository in which one the principles of this initiative that lies in the unique identification or resources through URIs can become real. Thus, entity reconciliation techniques (Araujo et al., 2011; Maali et al., 2012) coming from the ontology mapping and alignment areas or algorithms based on Natural Language Processing (hereafter NLP) have been designed to link similar resources already available in different vocabularies, datasets or databases such as DBPedia or Freebase.

Nevertheless, the issue of unifying supplier names as a human would do faces new problems that have been tackled in other research works (Galvez and Moya-Anegón, 2006) to extract statistics of performance in bibliographic databases. The main objective is not just a mere reconciliation process to link to existing resources but to create an unique name or link (n string literals $\rightarrow$ 1 company $\rightarrow$ 1 URI). For instance, the string literals "Oracle" and "Oracle University" could be respectively aligned to the entity <Oracle_Corporation> and <Oracle_University> but the problem of grouping by an unique (*Big*) name, identifier or resource still remains. That is why, a context-aware method based on NLP techniques combined with semantics has been designed, customized and implemented trying to exploit the naming convention of a specific dataset with the aim of grouping n string literals $\rightarrow$ 1 company and, thus, easing the next natural process of entity reconciliation.

The remainder of this paper is structured as follows. Section 2 presents a literature review. Next section outlines main mismatches in corporate names and presents the CORFU approach to unify corporate names. Afterwards the possibilities of using public procurement data as policy-making tool and, more specifically the Public Spending initiative, are presented

as a client of the CORFU technique. Last section exposes and discusses the experimentation carried out to test the presented approach using as a dataset the rewarded contracts of Australia in the period 2004-2012. Finally the main outcomes of this work, conclusions and some open issues are also outlined.

2. Related Work

According to the previous section, some relevant works can be found and grouped by the topics covered in this paper:

- Natural Language Processing and Computational Linguistics. In these research areas common works dealing with the aforementioned data heterogeneities such as misspelling errors (Norvig, 2013; Lecture, 2013b) and name/acronym mismatches (Yeates, 1999; Ratinov and Gudes, 2004), in the lexical, syntactic and semantic levels can be found. These approaches can be applied to solve general problems and usually follow a traditional approach of text normalization, lexical analysis, pos-tagging word according to a grammar and semantic analysis to filter or provide some kind of service such as information/-knowledge extraction, reporting, sentiment analysis or opinion mining. Well-established APIs such as NLTK (Loper and Bird, 2002) for Python, Lingpipe (Bob Carpenter, 2012), OpenNLP (Lecture, 2013a) or Gate (Bontcheva et al., 2013) for Java, WEKA (Read et al., 2012) (a data mining library with NLP capabilities), the Apache Lucene and Solr search engines provide the proper building blocks to build natural-language based applications. Recent times have also seen how the analysis of social networks (Musial and Kazienko, 2013; Michalski et al., 2014) such as Twitter (Li et al., 2012; Gimpel et al., 2011), the extraction of clinical terms (Wang, 2009) for electronic health records, the creation of bibliometrics (Galvez and Moya-Anegón, 2006; Morillo et al., 2013; Mahmood et al., 2013), the identification of gene names (Krauthammer and Nenadic, 2004; Galvez and Moya-Anegón, 2012) or the suggestion of knowledge pieces (Palacios et al., 2014a) to name a few have tackled the problem of entity recognition and extraction from raw sources. Other supervised techniques (Nadeau, 2007) have also been used to train data mining-based algorithms with the aim of creating multi-label classifiers (Kajdanowicz and Kazienko, 2013).

- Semantic Web. More specifically, in the LOD initiative (Berners-Lee, 2006), the use of entity reconciliation techniques to uniquely identify resources is being currently explored. Thus, an entity reconciliation process can be briefly defined as the method for looking and mapping (Isele et al., 2010a, 2012) two different concepts or entities under a certain threshold. There are a lot of works presenting solutions about concept mapping, entity reconciliation, etc. most of them are focused on the previous NLP techniques (Maali et al., 2012; Araujo et al., 2011; Ngomo and Auer, 2011; Isele et al., 2010b) (if two concepts have similar literal descriptions then they should be similar) and others (ontology-based) that also exploit the semantic information (hierarchy, number and type of relations) to establish a potential mapping (if two concepts share similar properties and similar super classes then these concepts should be similar). Apart from that, there are also machine learning techniques to deal with these mismatches in descriptions using statistical approaches. Recent times, this process has been widely studied and applied to the field of linking entities in the LOD realm, for instance using the DBPedia (Mendes et al., 2011). Although, there is no way of automatically creating a mapping with a 100% of confidence (without human validation) a mapping under a certain percentage of confidence can be enough for most of user-based services such as visualization. However, in case of using these techniques as previous step of a reasoning or a formal verification process this ambiguity can lead us to infer incorrect facts and must be avoided without a previous human validation.

  On the other hand, the use of semantics is also being applied to model organizational structures. In this case the notion of *corporate* is presented in several vocabularies and ontologies as Dave Reynolds (Epimorphics Ltd.) reports [2]. Currently the main effort is focused in the designed of the Organizations Vocabulary (a W3C Recommendation) in which the structure and relationships of companies are being modeled. This proposal is especially relevant because of the next aspects:

  1. To unify existing models to provide a common specification.
  2. To apply semantic web technologies and the Linked Data approach to enrich and publish the relevant corporate information.
  3. To provide access to the information via standard protocols.

[2] `http://www.epimorphics.com/web/wiki/organization-ontology-survey`

4. To offer new services that can exploit this information to trace the evolution and behavior of the organization over time.

- Corporate Databases. Although corporate information such as identifier, name, economic activity, contact person, address or financial status is usually publicly available in the official government registries, the access to this valuable information can be tedious due to different formats, query languages, etc. That is why, other companies have emerged trying to index and exploit these public repositories; selling reporting services that contain an aggregated version of the corporate information. Taking as an example the Spanish realm, the Spanish Chambers of Commerce [3], Empresia.es [4] or Axesor.es [5] manage a database of companies and individual entrepreneurs. This situation can be also transposed to the international scope, for instance Forbes keeps a list of the most representative companies in different sectors. The underlying problems rely on the lack of unique identification, same company data in more than a source, name standardization, etc. and, as a consequence, difficulty of tracking company activity. In order to tackle these problems some initiatives applying the LOD principles such as the Orgpedia [6] in United States or "The Open Database Of The Corporate World" [7] have scrapped and published the information about companies creating a large database containing (76, 197, 263 of companies in July 2014) with high-valuable information like the company identifier.

  Apart from that, reconciliation services have also been provided but the problem of mapping (n string literals $\rightarrow$ 1 company $\rightarrow$ 1 URI, as a human would do and the previous section has presented) still remains. Finally, public web sites and major social networks such as Google Places, Google Maps, Foursquare, Linkedin Companies or Facebook provide APIs and information managed by the own companies that are expected to be specially relevant to enrich existing corporate data once a company is uniquely identified.

[3] http://www.camerdata.es/php/eng/fichero_empresas.php
[4] http://empresia.es
[5] http://www.axexor.es
[6] http://tw.rpi.edu/orgpedia/
[7] http://opencorporates.com/

3. The CORFU technique

According to (Galvez and Moya-Anegón, 2006; Morillo et al., 2013), institutional name variations can be classified into two different groups:

1. Non-acceptable variations (affect to the meaning) due to misspelling or translation errors.
2. Acceptable variations (do not affect to the meaning) that correspond to different syntax forms such as abbreviations, use of acronyms or contextual information like country, sub-organization, etc.

In order to address these potential variations the CORFU (Company, ORganization and Firm Unifier) approach seeks for providing a stepwise method to unify corporate names using NLP and semantic-based techniques as a previous step to perform an entity reconciliation process. The execution of CORFU comprises several common but customized steps in natural language processing applications such as:

1. Text normalization.
2. Filtering of stop-words.
3. Acronym expansion.
4. ...
5. Comparison and clusterization.
6. Linking to an existing information resource.

The CORFU unifier makes an intensive use of the Python NLTK API and other packages for querying REST services or string comparison. Finally and due to the fact that the corporate name can change in each step the initial raw name must be saved as well as contextual information such as dates, acronyms or locations. Thus, common contextual information can be added to create the final unified name.

1. Normalize raw text and remove duplicates. 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 the raw text (although some semantics can be lost, previous works and empirical tests show that this is the best approach); 3) remove duplicates and 4) lemmatize the corporate name. The implementation of this step to clean the corporate name has been performed using a combination of the aforementioned API and the Unix scripting tools AWK and SED. In this case, Figure 1 presents

```
rawname = filter(lambda x: x in string.letters or
                 x in string.whitespace, line)
...
def normalize(self, word):
  word = word.lower()
  word = self.lemmatizer.lemmatize(word)
  return word
```

Figure 1: Normalization and data cleansing using the Python NLTK API.

a snippet of code for cleaning the name and making a basic word normalization.

2. Filter the basic set of common stop-words in English. A common practice in NLP relies in the construction of stop-words sets that can filter some non-relevant words. Nevertheless, the use of this technique must consider two key-points:

   - There is a common set of stop-words for any language than can be often used as a filter.
   - Depending on the context, the set of stop-words should change to avoid filtering relevant words. In this particular case, a common and minimal set of stop-words in English provided by NLTK has been used.

   Thus, the normalized corporate name is transformed into a new set of words. Figure 2 presents the function for removing a set of words given a another set, it can also be applied to other stages that require filtering capabilities.

3. Filter the expanded set of most common words in the dataset. Taking into account the aforementioned step, this stage is based on the construction of a customized stop-words set for corporate names that is also expanded with Wordnet (ver. 3.0) synonyms with the aim of exploiting semantic relationships. In order to create this set, two strategies, as Figure 3 partially shows, have been followed:

   - Handmade creation of the stop-words set (accurate but very time-consuming).
   - Extract automatically the set of "most common words" from the working dataset and make a handmade validation (less accurate

```
from nltk.corpus import stopwords
self.stop_words_wn = Set(stopwords.words('english'))
...
def remove_set(self, set,  name):
  token_names= word_tokenize(name)
  filtered_token_list = [w for w in
token_names if not w in set ]
  cleaned_name = " ".join(["".join(filtered_token)
    for filtered_token in filtered_token_list])
  return cleaned_name

stop_unified_name = self.remove_set(self.stop_words_wn, name)
```

Figure 2: Filtering words with the Python NLTK API.

and time-consuming).

4. Dictionary-based expansion of common acronyms and filtering. A dictionary of common acronyms in corporate names such as "PTY", "LTD" or "PL" and their variants has been created in order to be able to extract, expand and filter acronyms.
5. Identification of contextual information and filtering. Corporate names can mainly contain nationalities or place names that, in most of cases, only add noise to the real corporate name. In this case, the use of external services such as Geonames, Google Places or Google Maps can ease the identification of these words and their filtering. In order to implement this functionality, the Geonames REST service has been selected due to its capabilities to align text to locations.
6. Spell checking (optional). This stage seeks for providing a method for fixing misspelling errors. It is based on the well-known speller of Peter Norvig (Norvig, 2013) that uses a train dataset for creating a classifier. Although the accuracy of this algorithm is pretty good for relevant words in corporate names, the empirical and unit tests with a working dataset have demonstrated that spell checking of non-relevant words is more efficient and accurate using a stop-words set/dictionary (this set has been built with words that are not in the set of "most common words", step 2, and exist in the Wordnet database). Furthermore, some spelling corrections are not completely adequate for corporate

```
from nltk.corpus import wordnet as wn
...
def create_syns_from_wn(self, word):
  syns = wn.synsets(word)
  lemmas = Set()
  for syn in syns:
    lemmas = lemmas | Set([lemma.name for lemma in syn.lemmas] )
  return lemmas

def expand_list_wn(self,  list):
  source = Set(list)
  expanded_set = Set()
  for word in source:
    expanded_set = expanded_set | self.create_syns_from_wn(word)
return expanded_set |  source

def list_most_used_words(companies, max):
  words = flatten(map(lambda company: company.rawname.split(),
                  companies))
  counter = collections.Counter(words)
  return [x[0].lower() for x in
  filter ( lambda x: x [1] > max,
   (itertools.islice(counter.most_common(), 0, MAX_WORDS)))]
```

Figure 3: Expanding a list of words with Wordnet syns and Counting "most used words" in a dataset.

names due to the fact that words could change and, therefore, a non-acceptable variant of the name could be accidentally included. That is why, this stage is marked as optional and must be configured and performed with extreme care.

7. Pos-tagging parts of speech according to a grammar and filtering the non-relevant ones. The objective of this stage lies in "classifying words into their parts of speech and labeling them accordingly is known as part-of-speech tagging" (Loper and Bird, 2002). In order to perform this task both a lemmatizer based on Wordnet and a grammar for corporate names ("NN"-nouns and "JJ"-adjectives connected with ar-

ticles and prepositions) have been designed, see Figure 4. Once words are correctly tagged, next step consists in filtering non-relevant categories in corporate names keeping nouns and adjectives, as an example Figure 4 also shows how to walk and filter nodes in the parsed tree.

```
self.lemmatizer = nltk.WordNetLemmatizer()
self.grammar = r"""
  NBAR: {<NN.*|JJ>*<NN.*>}
  NP: {<NBAR>}
      {<NBAR><IN><NBAR>}
    """
self.chunker = nltk.RegexpParser(self.grammar)

def leaves(self, tree):
  for subtree in tree.subtrees(filter = lambda t: t.node=='NP'):
    yield subtree.leaves()
```

Figure 4: Regular expression-based chunker in Python NLTK and Filtering words by the category "NP" (noun phrase) .

8. Cluster corporate names. This task is in charge of grouping names by similarity applying a string comparison function. Thus, if the clustering function is applied n times any name will be grouped by "the most probably/used name" according to a threshold generated by the comparison function. To do so, the CORFU technique has been configured to use the WRatio function to compare strings (available in the Levenshtein Python package) and a customized clustering function.
9. Validate and reconcile the generated corporate name via an existing reconcile service (optional). This last step has been included with the objective of linking the final corporate name with an existing information resource and adding new alternative labels. The OpenCorporates and DBPedia reconciliation services have been used in order to retrieve an URI for the new corporate name. As a consequence, the CORFU unifier is partially supporting one of the main principles of the LOD initiative such as unique identification.

4. Enabling policy-making process using public procurement data: The Public Spending Initiative.

In order to enable a policy-making process based on public procurement data different pieces of information should be considered: geographical information, data and time, amount, type of contract, payer and payee profiles (Senthil et al., 2014), etc. As the related work section has outlined, there are already approaches to deliver such information in a standard way easing the access to this valuable data and enabling a better way of exploiting information through the extraction of statistics. Thus, standardized geographical information can be found in some classifications such as NUTS (National Territorial Units) or the product scheme classifications that have been already unified (Rodríguez et al., 2014) under a common and shared data model (the Resource Description Framework-RDF).

In the context of this research work, corporate names or, more specifically, payers and payee names are being processed with the CORFU technique to deliver a common name that can be re-used in the "PublicSpending" initiative[8]. Once relevant public procurement data is unified in different datasets, network analysis techniques can be used to compare variables such as location, amount, time or type of contract in public spending data. As first step, public spending data is represented as a graph to create a payment network in which interesting relations among underlying agents can be easily captured, see Figure 5. This graph is created by the payments coming from payers (public institutions) to payees (mainly private organizations). Secondly, to interpret the graph every node is either the payer or the payee that are linked through a payment, which is characterized by its amount, time and type of contract. Moreover, measures of centrality, degree centrality (in-out degree, weighted degree) and measures relative to the rest of the network such as betweenness centrality are used to understand the public spending graph (Irani et al., 2014; Overbeek et al., 2012). Finally, as a detailed example, the process of promoting public spending data and unifying corporate names with the CORFU technique has been applied to the next geographical entities: United States, United Kingdom, Australia, Greece, the State of Alaska and Massachusetts and the city of Chicago, see also Table 1. Thus, a policy-maker is now able to graphically take the most of public spending data in a certain area and the purposed process also enables the possibility of preventing new necessities (products or services)

[8] `http://publicspending.net`

or ensuring transparency to name a few.

Region	**RDF Triples**	**Payers**	**Payees**	**Payments (€)**	**#Decisions**
United States	494,005	157	16,780	504,599,838,933	61,359
United Kingdom	613,3782	119	26,768	597,659,337,423	1,137,641
Australia	2,788,939	165	51,455	204,606,056,134	429,435
Greece	68,804,546	3904	204,406	48,095,068,098	2,303,360
State of Alaska	74,8319	22	28,333	16,127,397,046	139,821
State of Massachusetts	8,023,505	158	31,068	51,279,504,691	1,305,267
City of Chicago	63,6526	54	7,930	44,153,540,592	84,405
Total	87,629,622	4579	315,285	1,466,520,742,920	5,461,288

Table 1: Statistics obtained from "PublicSpending.net" after applying the CORFU technique.

- United States: there is one major node (payer) in the graph ("Department of Defense"), dispersing almost all (99%) the total budget (weight) of the graph. Obviously, defense has the lion's stake in sub prime awards. Furthermore, most of the money (92%) is received by CTA Inc., which is solely connected to this department (no connections to other nodes). The dispersion of the public budget is made through 42 agents to the contractors. There are either payer or payee nodes in the graph (no mixed mode both payer and payee-except Smithsonian Inst.), consequently there are no brokers in the network resulting the diameter to equal 1 and the modularity fairly low at 0.022. Due to the above characteristics, there are mainly corporates of the deference-military sector (only US companies are eligible to become vendors due to legal restrictions) coupled by some major global enterprises with less weighted degree as they do not awarded purely defense contracts.
- United Kingdom: is characterized by five major nodes (payers): health, family, education, business innovation and skills, local government

disperse 88% of the total budget. The major payees are local authorities or funds responsible for the proper exploitation of the funds received. There are also private companies receiving money for goods and services mainly information technology, telecommunications and consulting. The dispersion of the public budget is made through 26 agents (payers).

- Australia: there are two major nodes (payers) in the graph (Department of Defense and Defense Materiel Org.), dispersing almost half of the total budget (weight) of the graph. These two nodes are also the top out-degree nodes in the graph (22% of the payment links). This indicates that "Defense" is a major factor in the Australian economy sustaining a network of enterprises that selling goods and services suited for the defense needs of the state. The 35% of the budget is spent by institutions related to education, immigration, health and social security, taxation, public order and telecommunication. This reflects, in general, the priorities and major concerns of the Australian state. The dispersion of the public budget is made in a balanced way as there are no private enterprises receiving excessive amounts of money (except FMS and Central Office).

- Greece: there two major projects in Greece: 1) the subway in Athens and Thessaloniki and 2) Egnatia Runway in North Greece. The Ministry of Public Order was involved in significant construction works in regard to other ministries or regional authorities. Universities and research institutes are important hubs in the network maintaining a wide network of payees that offer a variety of services for them and are significant contributors to the dispersion of funds. This also applies to Regional Authorities in local scale. Pension Funds, Labor Office and other social security institutes have a significant amount of payments for services and are important agents in the network presented, independently of their actual spending for pensions or social security allowances.

- State of Alaska: there are distinct characteristics originating from the special conditions that apply to the region's low population, vast areas of natural resources and ecosystems, weather conditions, native (indigenous) population and distance from global markets. All the above result to a payment network where funds are allocated smoothly to local companies and authorities where health, education, environment,

natural resource management, transportation and construction have the lead.

- State of Massachusetts: the dispersion of the public budget is made in a balanced way through 157 payers to a network of local institutes and authorities. There are major global players as well but the amounts receiving are smaller due to the bigger amounts that are targeted to health, education and legal institutions and to local authorities. The graph diameter is 1 as there are payer/payee only nodes and modularity is 0.76 indicating the local structure of the payment network. It is worth noticing that there is great variety in the services offered, there are many companies present for every sector (competition) and the balanced value of the in degree indicates a mature market. Massachusetts is famous for its health and educational institutions and this fact is validated from the output data.

- City of Chicago is the only city examined (compared to countries or states) but the volume of data ranging from 1993 to 2013 offer a total amount for examination fairly comparable to a state or country. There are distinct characteristics originating from the fact that a city has different needs and priorities from a state/country and of course many resemblances to one (e.g. no need for defense/border safeguarding expenses). The dispersion of the public budget is made in a balanced way through 52 agents (payers) to a network of local companies and authorities and there are also major global players as well. The fact that the city is a transportation hub and resides to lake Michigan is pictured on the graph as major nodes both payers and payees are present and belong to transportation, water management and public utilities sector.

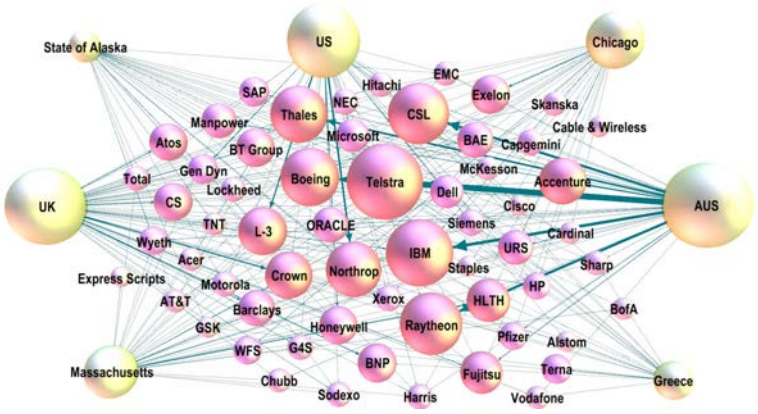

Figure 5: Public spending graph of the major vendors of United States, United Kingdom, Australia, Greece, Alaska, Massachusetts and Chicago.

5. Evaluation

5.1. Research design

Since the CORFU approach has been successfully designed and implemented [9], it is necessary to establish a method to assess quantitatively the quality of results. To do so, the following steps have been carried out:

1. Configure the CORFU technique, see Table 2.
2. Execute the algorithm taking as a parameter the file containing the whole dataset of company names.
3. Validate (manually) the dump of unified names.
4. Calculate measures of precision, see Eq. 1, recall, see Eq. 2, and F1 score (the harmonic mean of precision and recall), see Eq. 3, according to the values of tp (true positive), fp (false positive), tn (true negative) and fn (false negative).
   In particular, this evaluation considers the precision of the algorithm as "the number of supplier names that have been correctly unified under the same name" while recall is "the number of supplier names that have not been correctly classified under a proper name". More specifically, 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".

[9] `https://github.com/chemaar/corfu`

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

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

5.2. Sample dataset

As previous sections have introduced, there is an increasing interest and commitment in public bodies to create a real transparent public administration. In this sense, public administrations are continuously releasing relevant data in different domains such as tourism, health or public procurement with the aim of easing the implementation of new added-value services and improve their efficiency and transparency. In the particular case of public procurement, main and large administrations have already made publicly available the information with regards to public procurement processes. In this case of study, public procurement data coming from the Australia government are used to test and validate the CORFU unifier. More specifically, a dataset of supplier names in Australia in the period 2004-2012 containing 430188 full names and 77526 unique names has been selected. The experiment has been carried out executing the aforementioned steps in the whole dataset to finally generate a dump containing for every supplier the raw name and the unified name. On the other hand, the CORFU stepwise method has been customized to deal with the heterogeneities of this large dataset as Table 2 summarizes.

5.3. Results and Discussion

According to the results presented in Table 3, the precision and recall of the CORFU technique are consider acceptable for the whole dataset due to the fact that a 48% ($77526 - 40278 = 37248$) of the 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 precision is pretty good but the recall presents a low value because some corporate names were not unified under a proper name; some of the filters must therefore be improved in terms of accuracy.

In order to improve the results for relevant companies, the experiment has also been performed and evaluated for the first 100 companies in the Forbes list, actually 68 companies were found in the dataset. In this case, results show a better performance in terms of precision, 0.926, and recall,

Step	Name	Customization
1	Normalize raw text and remove duplicates	Default
2	Filter the basic set of common stop-words in English	Default
3	Filter the expanded set of most common words in the dataset	Two stop-words 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 the Geonames REST service
6	Spell checking (optional)	Train dataset of 128457 words provided by Peter Norvig's spell-checker (Norvig, 2013).
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 reconciliation service (optional)	Python client and Open Refine

Table 2: Customization of the CORFU technique for Australian supplier names.

0.926, and all these supplier names, 299 in the whole dataset, were unified by a common correct name. The explanation of this result can be found due to the fact that some of the parameters of the CORFU technique were specially selected for unifying these names because of their relevance in world economic activities.

On the other hand, it is important to emphasize that the last step of

Total number of companies	Unique names	CORFU unified names	% of correct 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

Table 3: Results of applying the CORFU approach to the Australian supplier names.

linking these names with existing web information resources using the reconciliation service of OpenCorporates or DBPedia in Open Refine can generate $37248 * 0.762 = 28383$ correct links (36.61%) instead of the initial 8% that was reached in the first mapping process (without name unification). Thus, the initial problem of linking (n string literals $\rightarrow$ 1 company $\rightarrow$ 1 URI) has been substantially improved.

Finally, the frequency distribution of supplier and number of appearances is depicted on Figure 6 with the objective of presenting how the cloud of points (appearances) that initially were only one per supplier has emerged due to the unification of names, for instance in the case of "Oracle" 75 apparitions can now be shown. On the other hand and due to the unique identification of supplier names, new RDF instances are generated, see Figure 7, and can be querying via SPARQL to make summary reports of the number of rewarding contracts by company, see Figure 8.

5.4. Robustness and Refinement

To illustrate the robustness of the presented approach to unify corporate names, a second experiment has been carried out as an extension of the previous one. This robustness experiment is necessary to ensure that results are creditable and it is based on similar studies that have been performed in the field of social network analysis (Dodds et al., 2011; Mitchell et al., 2013; Palacios et al., 2014b) when natural language processing techniques are used. In this case and due to the fact that human-validation is not completely possible, a test-campaign (23 tests) based on random walk techniques (Fouss et al., 2007) has been designed to measure again the precision and recall of the CORFU technique. Since the unification process generates a pair (corporate name, unified corporate name), it is possible to design a

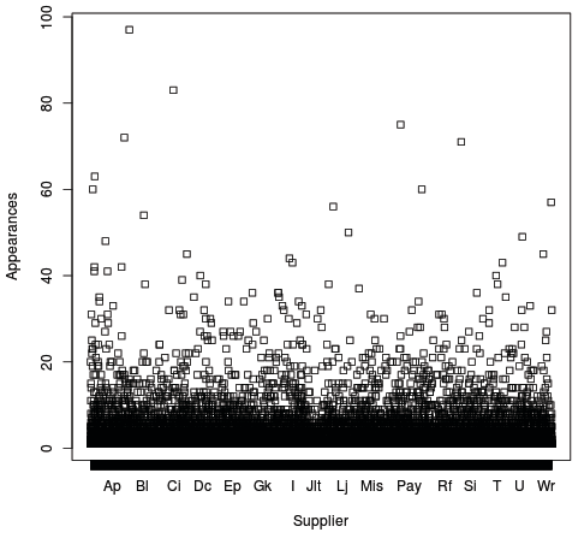

Figure 6: Full view of supplier and number of appearances in the Australian rewarded contracts dataset.

search process that taking as input a dataset of company names and a query (other corporate name), will match all relevant corporate names. Moreover, if we execute a query against a dataset which names have not been unified and compare the results to the ones generated using the same query against a dataset which names have been unified (expected results), we can extract metrics of precision, recall and the F1 score. In order to design this experiment, the following steps have been carried out creating a set of test cases:

1. Select a set of datasets of company names, $\mathcal{C} = \{\mathcal{C}^1, \mathcal{C}^2, ..., \mathcal{C}^k, ..., \mathcal{C}^n\}$, where every element $\mathcal{C}_k$ is a dataset of company names. In this case, three different groups of datasets, see Table 4, have been downloaded and processed to extract corporate names:
   - The "Public Spending Data" from the United States of America (USA) Spending Data portal[10]. In this case, the vendor names of every public contract between 2000 and 2015 have been selected as corporate names.

[10] `http://www.usaspending.gov/data`

```
:o1 a org:Organization;
        skos:prefLabel   ''Microsoft'';
        skos:altLabel    ''Microsoft Australia'',
        ''Microsoft Australia Pty Ltd'',
        ...;
        skos:closeMatch dbpedia-res:Microsoft;
        ...
.
```

Figure 7: Partial example of a RDF organization instance.

```
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)
```

Figure 8: Example of a SPARQL query for counting supplier names.

- The "Basic Company Data" from the United Kingdom (UK) data portal[11]. The company name has been selected as corporate name.
- The data dump of corporate information from "The CrocTail project"[12]. The company name has been also selected as corporate name.

2. For every company name dataset $\mathcal{C}^k \in \mathcal{C}$, apply the CORFU technique generating a new company name dataset $\mathcal{C}^k_{CORFU}$. The configuration of the technique has been the same as the presented in Table 2.
3. Create a common set of queries, $\mathcal{Q} = \{q_1, q_2, ..., q_k, ...q_n\}$. In the previous experiment, the list of the first 100 companies in the Forbes list was used to apply the CORFU technique. In this case, the list of

[11] http://data.gov.uk/dataset/basic-company-data
[12] http://croctail.corpwatch.org/

the world's biggest public companies (2000) in the Forbes web site[13] has been extracted and processed to create a set of queries for every company name in the list.

4. Design and implement a search process. To do so, a program on top of the Apache Lucene and Solr search engines has been implemented to index any dataset of corporate names and to provide a search engine. This engine has been configured using the standard filters and a RAM-stored index.
5. Run the search process taking as parameters every company name dataset in $\mathcal{C}^k_{CORFU}$ and the set of queries $\mathcal{Q}$ to generate for every test case a set of expected results.
6. Run the search process taking as parameters every company name dataset in $\mathcal{C}$ and the set of queries $\mathcal{Q}$ to generate for every test case a set of real results.
7. Extract measures of precision, recall and F1 score by comparing the expected and real results. A Python program has been implemented to automatically process all results generated by all test cases.

5.4.1. Results and Discussion

Table 5 shows the aggregated metrics of precision, recall and the F1 score after a total execution of 92000 queries, 2000 target corporate names * 23 datasets * 2 types of corporate names (unified and non-unified). According to the results, the average precision of the CORFU technique is closed to 0.5 for most of cases, see Figure 9. It also seems that the number of company names has a direct impact on the precision. The main reason of this behavior is due to the fact that as much company names are available as more contextual information can be extracted and, thus, corporate names can be easily unified. However, it is also clear that there is a decrease in the precision regarding the first experiment. Since the number of company names has been dramatically increased, it is possible that more variants for the same name have been also included implying the necessity of refining the CORFU technique to cover a broad scope of names. Furthermore, the contextual information of the experiment could be carefully revised to ensure higher precision values.

[13] `http://www.forbes.com/global2000/list/#page:20\_sort:0\_direction:asc\_search:\_filter:All\%20industries\_filter:All\%20countries\_filter:All\%20states`

Test	Datasource	Number of names	Number of unique names
t_1	Australia (2004-2012)	430188	77526
t_2	USA 2000	594427	538070
t_3	USA 2001	641841	574640
t_4	USA 2002	830364	733879
t_5	USA 2003	1183817	1096904
t_6	USA 2004	2001742	1725036
t_7	USA 2005	2921892	2593113
t_8	USA 2006	3795668	3337231
t_9	USA 2007	4110671	3708459
t_{10}	USA 2008	4504207	3979471
t_{11}	USA 2009	3495241	3010462
t_{12}	USA 2010	3535748	3107811
t_{13}	USA 2011	3390650	2999314
t_{14}	USA 2012	3109189	2778383
t_{15}	USA 2013	2492908	2293593
t_{16}	USA 2014	2206610	2077207
t_{17}	USA 2015	98952	98193
t_{18}	UK 1	849999	849756
t_{19}	UK 2	850000	849832
t_{20}	UK 3	850000	849836
t_{21}	UK 4	850000	849819
t_{22}	UK 5	101247	101228
t_{23}	CrocTail project	1370145	1364761
$\overline{T}$	All	44215506	39594524

Table 4: Summary of number of corporate names for every test.

Test	$\overline{Precision}$	$\overline{Recall}$	**F1 Score**
t_1	0.416	0.446	0.430
t_2	0.407	0.431	0.419
t_3	0.421	0.437	0.429
t_4	0.421	0.436	0.428
t_5	0.418	0.432	0.425
t_6	0.452	0.489	0.470
t_7	0.478	0.541	0.508
t_8	0.500	0.586	0.540
t_9	0.511	0.608	0.555
t_{10}	0.519	0.624	0.567
t_{11}	0.490	0.566	0.526
t_{12}	0.493	0.572	0.530
t_{13}	0.490	0.566	0.525
t_{14}	0.483	0.552	0.516
t_{15}	0.469	0.523	0.495
t_{16}	0.462	0.510	0.485
t_{17}	0.413	0.433	0.422
t_{18}	0.371	0.398	0.384
t_{19}	0.384	0.396	0.390
t_{20}	0.374	0.348	0.361
t_{21}	0.381	0.279	0.322
t_{22}	0.349	0.375	0.361
t_{23}	0.462	0.471	0.466
$\overline{T}$	0.387	0.378	0.381

Table 5: Average metrics of precision and recall of the test campaign.

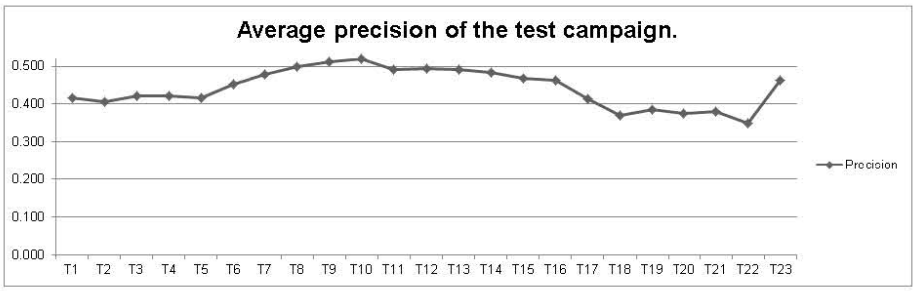

Figure 9: Average precision of the test campaign.

5.5. Research Limitations

Some key limitations of the presented work must be outlined. The first one relies on the sample size; our research study has been conducted in a closed world and, more specifically, using corporate names that have been extracted from a set of public sources. That is why, results in a broad or real scope could change, in terms of precision, since more complex names and contextual information could be found. For instance, we have evaluated the possibility of gathering corporate names from the public API of OpenCorporates/OpenLEIs or DBPedia but due to the restrictions on the use of the APIs we have preferred to download existing data dumps. However, the research methodology, the design of experiments and the creation of a kind of benchmark for testing the CORFU technique have been demonstrated to be representative and creditable.

On the other hand, we have automatically generated test cases from real data to avoid the necessity of human validation. In this case, we have focused on the creation and publication of set of datasets of corporate names for testing unification name processes due to the fact that the handmade creation of mappings between corporate names requires a great effort with a high probability of losing robustness (the same company can be named in a different way depending on the users and domain discourses). However, we consider that the precision and recall metrics are helpful to make a first estimation of the advantages of applying the CORFU technique to unify corporate names.

Building on the previous comment, we cannot either figure out the internal budget, methodologies, vocabularies, experience and background of specific sites to gather and create corporate information. We merely observe and re-use existing public and on-line knowledge sources to provide an accurate name unification process. Finally, we have also identified the

necessity of re-designing the CORFU technique to scale up and to support large datasets since the performance of the algorithm also decreases depending on the number of corporate names.

6. Conclusions and Future Work

A technique for unifying corporate names in the e-Procurement sector has been presented as a step towards the unique identification of organizations with the aim of accomplishing one of the most important LOD principles and easing the execution of reconciliation processes. The main conclusion of this work lies in the design of a stepwise method to prepare raw corporate names in a specific context, e.g. Australia supplier names, before performing a reconciliation process. Although the percentage of potential right links to existing datasets has been dramatically improved, it is clear that human-validation is also required to ensure the correct unification of names. As a consequence, the main application of CORFU can be found when reporting or tracking activity of organizations are required. However, this first effort has also implied, on the one hand, the validation of the stepwise method and, on the other hand, the creation of a sample dataset that can serve as input for more advanced algorithms based on machine learning techniques such as classifiers. Although the precision of the CORFU technique decreases when processing large datasets, it has been demonstrated to be creditable in a broad scope. From public administrations point of view, this technique also enables a greater transparency providing a simple way to unify corporate names and boosting the comparison of rewarded contracts.

Finally, further steps in this work consist in the extension of the stopwords sets for corporate names, a better acronym detection and expansion algorithm, other techniques to make string comparisons such as $n-grams$ (Sidorov et al., 2014) and the creation of a new final step to enhance the current implementation with a classifier that can automatically learn new classes of corporate names or automatically infer grammars for representing any type of corporate name. Furthermore, the technique must be reported to the international "Public Spending" initiative, as supporting tool, to be applied over other datasets to correlate and exploit public contracts meta-data.

7. Acknowledgements

This work is part of the "PublicSpending.net" effort carried out in cooperation with Giorgos Vafeiadis (Technical University of Athens), Giannis

Xidias (University of the Aegean) and Michalis Klonaras (OTE S.A.).

References

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

Berners-Lee, T. (2006). Linked Data.

Bob Carpenter, Mitzi Morris, B. B. (2012). *Text Processing with Java 6*, volume 1. LingPipe Publishing.

Bontcheva, K., Cunningham, H., Roberts, I., Roberts, A., Tablan, V., Aswani, N., and Gorrell, G. (2013). GATE Teamware: a web-based, collaborative text annotation framework. *Language Resources and Evaluation*, pages 1–23.

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., and Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PLoS ONE*, 6(12):e26752.

European Commission, D.-G. f. I. (2011). The eProcurement Map. a map of activities having an impact on the development of european interoperable eprocurement solutions. http://www.epractice.eu/en/library/5319079.

Fouss, F., Pirotte, A., Renders, J.-M., and Saerens, M. (2007). Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *Knowledge and Data Engineering, IEEE Transactions on*, 19(3):355–369.

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

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

García, N. F., Arias-Fisteus, J., Sánchez, L., and López, G. (2012). IdentityRank: Named entity disambiguation in the news domain. pages 9207–9221.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for Twitter: annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA. Association for Computational Linguistics.

Guijarro, L. (2009). Semantic interoperability in egovernment initiatives. *Computer Standards & Interfaces*, 31(1):174–180.

Heath, T. and Ch. Bizer, C. (2011). *Linked Data: Evolving the Web into a Global Data Space*, volume 1. Morgan & Claypool.

Husain, M. F., Al-Khateeb, T., Alam, M., and Khan, L. (2011). Ontology based policy interoperability in geo-spatial domain. *Computer Standards & Interfaces*, 33(3):214–219.

Irani, Z., Sharif, A. M., Kamal, M. M., and Love, P. E. D. (2014). Visualising a knowledge mapping of information systems investment evaluation. *Expert Syst. Appl.*, 41(1):105–125.

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

Isele, R., Jentzsch, A., and Bizer, C. (2010b). Silk server - adding missing links while consuming linked data. In *Proceedings of the First International Workshop on Consuming Linked Data, Shanghai, China, November 8, 2010*.

Isele, R., Jentzsch, A., and Bizer, C. (2012). Active Learning of Expressive Linkage Rules for the Web of Data. In *ICWE*, pages 411–418.

Jung, J. J. (2012). Online Named Entity Recognition Method for Microtexts in Social Networking Services: A Case Study of Twitter. *Expert Syst. Appl.*, 39(9):8066–8070.

Kajdanowicz, T. and Kazienko, P. (2013). Boosting-based Multi-label Classification. *J. UCS*, 19(4):502–520.

Klein, D., Smarr, J., Nguyen, H., and Manning, C. D. (2003). Named entity recognition with character-level models. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 180–183, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Lecture, S. N. L. P. (2013a). Apache OpenNLP Developer Documentation. http://opennlp.apache.org/documentation/manual/opennlp.html.

Lecture, S. N. L. P. (2013b). Spelling Correction and the Noisy Channel. The Spelling Correction Task. http://www.stanford.edu/class/cs124/lec/spelling.pdf.

Li, C., Weng, J., He, Q., Yao, Y., Datta, A., Sun, A., and Lee, B.-S. (2012). TwiNER: named entity recognition in targeted twitter stream. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 721–730, New York, NY, USA. ACM.

Loper, E. and Bird, S. (2002). NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 62–69. Somerset, NJ: Association for Computational Linguistics. http://arXiv.org/abs/cs/0205028.

López-Pellicer, F. J., Silva, M. J., Chaves, M. S., Zarazaga-Soria, F. J., and Muro-Medrano, P. R. (2010). Geo Linked Data. In *DEXA (1)*, pages 495–502.

Maali, F., Cyganiak, R., and Peristeras, V. (2012). Re-using Cool URIs: Entity Reconciliation Against LOD Hubs. In Bizer, C., Heath, T., Berners-Lee, T., and Hausenblas, M., editors, *LDOW*, CEUR Workshop Proceedings. CEUR-WS.org.

Mahmood, T., Jami, S. I., Shaikh, Z. A., and Mughal, M. H. (2013). Toward the modeling of data provenance in scientific publications. *Computer Standards & Interfaces*, 35(1):6–29.

Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., and Berbís, J. M. G. (2013). Named entity recognition: Fallacies, challenges and opportunities. *Computer Standards & Interfaces*, 35(5):482–489.

Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. (2011). DBpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, pages 1–8, New York, NY, USA. ACM.

Michalski, R., Kajdanowicz, T., Bródka, P., and Kazienko, P. (2014). Seed Selection for Spread of Influence in Social Networks: Temporal vs. Static Approach. *New Generation Comput.*, 32(3-4):213–235.

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., and Danforth, C. M. (2013). The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one*, 8(5):e64417.

Morillo, F., Aparicio, J., González-Albo, B., and Moreno, L. (2013). Towards the automation of address identification. *Scientometrics*, 94(1):207–224.

Musial, K. and Kazienko, P. (2013). Social networks on the Internet. *World Wide Web*, 16(1):31–72.

Nadeau, D. (2007). *Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision*. PhD thesis, School of Information Technology and Engineering, University of Ottawa, Ottawa, Canada.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Ngomo, A.-C. N. and Auer, S. (2011). LIMES: a time-efficient approach for large-scale link discovery on the web of data. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three*, pages 2312–2317. AAAI Press.

Norvig, P. (2013). How to Write a Spelling Corrector. http://norvig.com/spell-correct.html.

Overbeek, S., Janssen, M., and van Bommel, P. (2012). A standard language for service delivery: Enabling understanding among stakeholders. *Computer Standards & Interfaces*, 34(4):355–366.

Palacios, R. C., Casado-Lumbreras, C., Soto-Acosta, P., and Misra, S. (2014a). Providing knowledge recommendations: an approach for informal electronic mentoring. *Interactive Learning Environments*, 22(2):221–240.

Palacios, R. C., Cuadrado, J. L. L., Gonzalez-Carrasco, I., and Peñalvo, J. F. G. (2014b). SABUMO-dTest: Design and evaluation of an intelligent collaborative distributed testing framework. *Comput. Sci. Inf. Syst.*, 11(1):29–45.

Palacios, R. C., Messnarz, R., and Biró, M. (2013). Systems, software and services process improvement. *Computer Standards & Interfaces*, 36(1):1–2.

Ratinov, L. and Gudes, E. (2004). Abbreviation Expansion in Schema Matching and Web Integration. In *Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '04, pages 485–489, Washington, DC, USA. IEEE Computer Society.

Read, J., Bifet, A., Holmes, G., and Pfahringer, B. (2012). Scalable and efficient multi-label classification for evolving data streams. *Machine Learning*, 88(1-2):243–272.

Rodríguez, J. M. Á., Gayo, J. E. L., González, A. R., and de Pablos, P. O. (2014). 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*, 30:674–688.

Rodríguez, J. M. Á., Gayo, J. E. L., Silva, F. A. C., Alor-Hernández, G., Sánchez, C., and Luna, J. A. G. (2012). Towards a Pan-European E-Procurement Platform to Aggregate, Publish and Search Public Procurement Notices Powered by Linked Open Data: the Moldeas Approach. *International Journal of Software Engineering and Knowledge Engineering*, 22(3):365–384.

Sarmento, L., Kehlenbeck, A., Oliveira, E., and Ungar, L. (2009). An Approach to Web-Scale Named-Entity Disambiguation. In *Proceedings of the 6th International Conference on Machine Learning and Data Mining in Pattern Recognition*, MLDM '09, pages 689–703, Berlin, Heidelberg. Springer-Verlag.

Senthil, S., Srirangacharyulu, B., and Ramesh, A. (2014). A robust hybrid multi-criteria decision making methodology for contractor evaluation and selection in third-party reverse logistics. *Expert Syst. Appl.*, 41(1):50–58.

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., and Chanona-Hernández, L. (2014). Syntactic N-grams as machine learning features for natural language processing. *Expert Systems with Applications*, 41(3):853 – 860. Methods and Applications of Artificial and Computational Intelligence.

Wang, Y. (2009). 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. Association for Computational Linguistics.

Yeates, S. (1999). Automatic Extraction of Acronyms from Text. In *University of Waikato*, pages 117–124.

Appendix I

Raw Supplier Name	Target Supplier Name and URI
"Accenture" "Accenture Aust Holdings" "Accenture Aust Holdings" "Accenture Aust Holdings Pty Ltd" "Accenture Australia Holding P/L" "Accenture Australia Limited" ... "Accenture Australia Ltd"	"Accenture" http://live.dbpedia.org/resource/Accenture
"Microsoft Australia" "Microsoft Australia Pty Ltd" ... "Microsoft Enterprise Services"	"Microsoft" http://live.dbpedia.org/resource/Microsoft
"Oracle (Corp) Aust Pty Ltd" "Oracle Corp (Aust) Pty Ltd" "Oracle Corp Aust Pty Ltd" "Oracle Corp. Australia Pty.Ltd." "Oracle Corporate Aust Pty Ltd" "Oracle Corporation" "Oracle Risk Consultants" "ORACLE SYSTEMS (AUSTRALIA) PTY LTD" ... "Oracle University"	"Oracle" http://live.dbpedia.org/resource/Oracle_Corporation
"PRICEWATERHOUSECOOPERS(PWC)" "PricewaterhouseCoopers Securities Ltd" "PricewaterhouseCoopers Services LLP" "Pricewaterhousecoopers Services Pty Ltd" "PriceWaterhouseCoopers (T/A: PriceWaterhouseCoopers Legal)" ... "Pricewaterhouse (PWC)"	"PricewaterhouseCoopers" http://dbpedia.org/resource/PricewaterhouseCoopers
...	...

Table 6: Examples of supplier names in the Australian rewarded contracts dataset.