A Nominal Filter for Web Search Snippets: Using the Web to Identify Members of Latin America's Highly Qualified Diaspora

Jorge García-Flores and William Turner
LIMSI - CNRS,
B.P. 133, F-91403
Orsay Cedex, France
{*jgflores, turner*}*@limsi.fr*

***Abstract*—This paper presents efforts aimed at using Natural Language Engineering (NLE) techniques for evaluating the impact of talent mobility on the development of three Latin American countries: Argentina, Colombia and Uruguay. We explain the different steps of a research progam aimed at carrying out what we call a Mobility Trace Extraction Task. The first step enriches traditional person name disambiguation queries with Social Identity Markers (SIMs) extracted from the bibliographic records of the Web of Science database. The coherence of the snippets retrieved using this SIM-enriched Web search strategy is automatically verified by applying nominal filters based on a context-free grammar to eliminate those snippets which do not respect valid variations of personal names. Finally, the filtered results are ordered and presented to social scientists in a way which allows them to decide if they want to contact and interview a person or not. Our goal is to produce a computer supported infrastructure for doing this type of sociological research using data from the Web.**

***Keywords*-person disambiguation; semantic filtering; web people search; computer assisted sociology; highly skilled diaspora; mobility**

I. Introduction

Talent mobility in the global economy is attracting growing attention as a subject of social science research [1]. But what constitutes a population of talented people on the move? An answer to this question has generally been found using a variety of data sources. Population census are exhaustive in coverage, but are carried out too infrequently to provide a constantly updated picture of talent mobility. Demographic registers provide information on date of entry into a country, intended duration of stay but generally do not contain information on a migrant's level of education. Labour surveys and administrative sources provide useful but rather limited information on the role of highly skilled migrant populations in the economic activity of host countries. But when the focus of study is on the "brain drain" issue - that is, the idea that talent mobility is a serious problem affecting developing countries in their capacity for development - specific surveys designed to study the motivations, intentions and values of people on the move are required [2]. Our goal is to datamine the Web using natural language processing techniques in order to help sociologists design these specific surveys.

Many data sources exist on the Web that can be mined in order to identify a population of talented people on the move. However, a great deal still remains to be done in order to evaluate the quality of these sources. We are currently working with three types of data. The first are extractd from the reference database "Web of Science" on a weekly basis. They concern articles that have been published by Colombian, Uruguayan and Argentinian authors in the biotechnology field. In general, each article is co-authored by three scientists on an average. When one author has a Latin American address, we make two assumptions that we are trying to test. The first is that the Latin American author might be a "circular scientist": that is, a person who now lives back in her home country, but who has extensively trained abroad. The second is that one of her co-authors could well be a member of Latin America's scientific diaspora because scientists with a common cultural background are likely to publish together for affective reasons, even if one of those scientists resides at home and the other in a foreign country. So studying co-authorship seems like an appropriate strategy for detecting a stock of potentially talented people on the move. That said, this stock of names has to be verified in order to eliminate such things as spelling errors or homonymns.

A second data source are the great many documents on the Web containing biographical information. Online CVs, personal web pages, press articles, blogs, social networking sites and on-line academic databases are just some examples of where biographical information can be found. We call these documents "mobility traces" because they establish the fact that a university trained person left her country for either a short period ("is a circular scientist") or has elected to take up residence abroad permanently ("is a diaspora scientist").

Finally, a third data source is potentially the personal profiles posted on blogs or social networks but for the moment we are not using this data source because we do not feel that we know enough about the self-exposure dynamics underlying its production. For example, proof of a university training is a measure of talent in our approach, but it is not clear that people using Facebook consider their educational background as being fundamental to the

kinds of interactions they want to organize through that system. In the same way, we don't know how professionals manage their "immigrant status" when posting their profiles on a site like LinkedIn: Do they attempt to promote their origins in order to contact others or do they tend to consider that birthplace is irrelevant to their current expertise and not worth mentioning? Questions like these suggest that the informational value of personal profiles for determining membership in the talented mobility population is still very much under discussion by the social science community and, consequently, we have not yet addressed the issue of data-mining this data source [3].

Our approach to processing Web data sets has been very much influenced by work done in the Web People Search (WePS) task [4]. This task focuses on eliminating ambiguous name queries by developing and experimenting techniques for name disambiguation. However, our task is different: extracting mobility traces is not aimed at disambiguating homonyms; it aims at finding evidence that a person is a highly qualified diaspora member. The natural language processing problem we are faced with is that of combining a person's name with appropriate semantic, organizational and geographic attributes in order to correctly identify suitable mobility traces for that specific person. We call these attributes Social Identity Markers (SIM) and have identified four specific sets: name, places, topics and organizations [1] In theory, these SIM sets can be built and updated using any of the databases mentioned above but, for the moment, we've only used the Web of Science and a specific database provided to us by the colombian diaspora network, Redes Colombia, for that purpose (see Table I).

Finally, the last element to be noted in this general introduction to our approach, concerns the two step procedure we are using to datamine the Web for mobility traces of highly qualified diaspora members. During the first step, SIM-enriched queries are produced by using part of speech tagging, named entity extraction and heuristics tailored according to each SIM type. We extract people's names from the author section of WoS records; topics are extracted from the title section; organizations are extracted from the author affiliation section; and places are extracted from the author address section. In other words, we semantically enrich an invariable root (the author's name) by successively adding to that root noun phrases from a publication's title, organization names from the author's affiliation and place names from the author's address. The result of this SIM enriched approach is a name in context. Our hypothesis is that the probability of finding relevant mobility traces on the Web will be higher when queries are launched using a "name in context" than when using the name only.

During the second step, the snippet set obtained in response to SIM-enriched queries is filtered in order to eliminate those snippets that don't contain markers of talent mobility. In this paper, one of these filters is presented: the Nominal Filter, which serves to validate that a snippet corresponds to an individual's name. The filter eliminates name variations caused by abbreviations, the compression of a name to an initial, the expansion of an initial to a name and the possible inversion of first and last names (or their initials, or their expansions). These name variations are calculated by means of a context-free grammar which produces a set of regular expressions that are likely to be found in the snippet being filtered.

[1]A fifth type is under discussion: dates. But dates are still out of the scope of this study.

Table I
CORPORA

name	description
Reference Corpus	**23** bibliographical records of Argentinian biotechnologist manually extracted from *Pubmed* and *WoS* scientific databases with links to manually searched mobility traces.
Redes Colombia Corpus	**50** highly qualified members of Redes Colombia on-line diaspora community.
WoS Corpus	**50** bibliographical records from Latin American biotechnologist automatically extracted from WoS scientific publications database.

II. RELATED WORK

One way of looking at the Mobility Trace Extraction Task (MTT) would be to consider it as a name disambiguation task plus an information extraction task. Name disambiguation approaches send an ambiguous person's name query to a Web Search engine and classify the resulting documents according to the different homonyms for that name by means of clustering techniques [5] [6] [7]. The Web People Search (WePS) campaigns [8] take that approach. The second [9] and third [4] editions of the WePS campaign included an Attribute Extraction task [10], where personal information like date of birth, birthplace, occupation, and nationality had to be extracted from Web search results. The best system in the WePS 2010 campaign carried out the task by using a rule-based approach. The heuristics were custom-tailored to each attribute type and produced a precision reading of 0.22 and a recall of 0.24 as calculated against a gold standard [11].

However, the Mobility Trace Extraction Task (MTT) does not aim at classing homonyms. Instead, the goal is to find very specific moblity traces on the Web for a specific person. Therefore, we inverse the order of the WePS pipeline (name disambiguation + attribute extraction) by adding SIM attributes to the initial Web search query. We can do this because the starting point for the MTT is a "name in context", that is, a name which is located in a WOS bibliographical record as we said above. The starting point for the WePS task is an ambiguous name without

Table II
HEURISTICS FOR QUERY ENRICHMENT WITH SIM

$B \leftarrow$ bibliographical records from a scientific publications database
$Q \leftarrow \emptyset$ queries
$S \leftarrow \emptyset$ resulting snippets
for each $b \in B$ **do**
  $b.name \leftarrow$ authors_name(b)
  if $b.name$ contains an spanish first or last name **then**
    Q_{name} = expand_initials($b.name$)
    S += web_search($Q_{name}, english, spanish$)
  end if
  $b.geo \leftarrow$ affiliation's_city_and_country(b)
  for each $g \in b.geo$ **do**
    Q_{geo} = '$b.name$ + g'
    S += web_search($Q_{geo}, english, spanish$)
    g_{esp} translate($b.geo, spanish$)
    if $g \neq_{esp}$ **then**
      Q_{geo} = '$b.name$ + g_{esp}'
      S += web_search($Q_{geo}, spanish$)
    end if
  end for
  $b.topics \leftarrow$ publication's_title(b)
  for each $np \in$ noun_phrases($b.topics$) **do**
    Q_{topic} = '$b.name$ + np'
    S += web_search($Q_{topic}, english, spanish$)
  end for
  $b.orgs \leftarrow$ affiliation's_organizations(b)
  for each $o \in$ organizations(b) **do**
    $lang \in$ detect_language(o)
    Q_{org} = '$b.name$ + o'
    S += web_search($Q_{org}, english, spanish, lang$)
  end for
end for

context, that is, precisely the opposite to the MTT task. Consequently, while name disambiguation is an important feature of our approach, it only comes into play at the end of our pipeline and not at the beginning as in the case of the WePS task. Before disambiguating, we first have to establish that the snippet set produced by the SIM enriched query contains both valid variations of a person's name, and relevant markers of mobility. We will present the nominal filter used to control the coherence of name variations in the next section of this paper. A biographical filter for detecting relevant markers of mobility is currently under development and so will not be discussed in this paper.

III. ENRICHING WEB SEARCH QUERIES WITH SIMS

We have identified four different types of Social Identity Markers (SIMs):

- **Name**
- **Geography:** places where a person has lived or traveled or published.
- **Topics:** keywords describing the cognitive universe of a person's activity
- **Organizations:** institutions having some relationship with the person

As explained above, we currently extract SIM identifiers from the Web of Science database, but could potentially extract them from any data source on the Web. We tested this idea using the database provided to us by the colombian diaspora network, Redes Colombia. Extraction is done using heuristics tailored according to each SIM (see algorithm on Table II). [2]

Table III shows the performance of a SIM-query enrichment pipeline on the Reference Corpus (bibliographical records of 23 biotechnologist belonging to the highly qualified Argentinian diaspora, see Table I). Each of the four SIM sets serving to enrich a query contains a variable number of elements extracted from the WOS data source. The "cost" column in this table provides an indication of the average number of element combinations serving to produce queries when using a given SIM set. Thus, for the first line of the table, we see that using the simplest SIM enrichment strategy - querying a person's name with expanded initials (like in *James A* Joyce*) - we were able to identify 13 people or 57% of the corpus members with only 4 queries per person. When we combined the name and the geographical SIM sets we were able to identify 2 people more than with the name only SIM set (15 people) but at a higher query cost (5.2 queries on an average per person instead of 4). Of course, the highest costs are incurred when all SIM sets are combined. This is seen in the last line of the table. 20 of the 23 members (87%) of the corpus were found but at an average cost of 7.8 queries per person. This figure is higher than the 74% reported in previous work [16] and shows that our efforts to constantly improve our SIM enrichment pipelines have effectively produced better results.

Table III
SIM-ENRICHED QUERY GENERATION STRATEGIES COMPARED (CUMULATIVE ANALYSIS)

SIM type	**people found**			**cost**	
	distinct	*cumulative*	*%*	*queries*	*cumulative*
Name	13	13	57%	4.0	4
Geography	2	15	65%	5.2	9.2
Topics	3	18	78%	6.1	15.2
Organizations	2	20	87%	7.8	23.0
Total	**20**		**87%**		**23.0**

IV. NOMINAL FILTER FOR WEB SEARCH SNIPPETS

The goal of nominal filtering is to take out of a snippet set all those snippets which don't contain valid variations of a person's name. As we said, these variations are caused by abbreviations, name compression, initial expansion, or name inversion. To eliminate errors, we proceed in two steps: first we use a context-free grammar to parse the name following the name formation rules used in Spanish and English, and

[2]Common Spanish names come from data of the Spanish National Statistics Institute [12] and geo-demographic studies of common surnames in Spain and Mexico [13]. Spanish pos-tagging and named entities annotation are made using *Freeling* [14]; English ones using *NLTK* [15]. Language detection of organizations is done with Markov Chains. For Google queries, we used Peter Krumin's Python Library for Google Search (http://bit.ly/EUizu)

Table IV
CONTEXT-FREE GRAMMAR FOR PERSON'S NAME PARSING

```
PersonName -> FirstName LastName
FirstName  -> FirstName  Initial
FirstName  -> FirstName  CommonName
FirstName  -> Initial
FirstName  -> CommonName
FirstName  -> TypoName
FirstName  -> CommonName ParticleName
LastName -> MainLastName
LastName -> MainLastName CommonName
LastName -> MainLastName ParticleName
LastName -> MainLastName TypoName
LastName -> MainLastName Initial
MainLastName -> CommonName
MainLastName -> ParticleName
MainLastName -> TypoName
```

then we calculate regular expressions using the syntactic tree in order to control variations. The Grammar (see Table IV) has four terminal elements: 1) a common name, 2) a name with a particle (*Ana Ozores* ***de*** *Clarin*), 3) a name with a typographical link (*Ana Ozores-Clarin*) and 4) the initial. The syntax requires at least a name (or an initial) and a last name. It takes into account Spanish name formation rules, using father and mother last names, and ensures that the first last name can't be compressed as an initial. The first last name is also needed as a reference for name inversion.

A. Valid Nominal Variations

Table V describes valid operations on any branch of the Spanish name grammar tree. Operations are implemented as recursive algorithms that compress, expand or suppress elements from the tree [3]. Regular expressions are produced and then used to validate name variations found in the snippets. When any given regular expression is true, the snippet is considered as containing a valid name variation. A total of 27 possible variations for a name were identified during the acquisition process.

B. Quantitative Evaluation of the Nominal Filter

The goal of this experience was to calculate recall and precision of the nominal filtering process. Two training corpuses were used to acquire, test and correct the filtering rules: the Reference Corpus (23 Argentinian biotechnologists), and the Redes Colombia corpus (see Table I). The WoS Corpus was the testing dataset. Evaluation was made by hand by one human judge, who identified false positives and false negatives with the UNOPORUNO system, our on-line evaluation webapp. We analyzed the nominal filter's performance according to the SIM type on which the query enrichement strategy was based. The filtering rate is the highest for the snippets resulting from the SIM-type Name, and the lowest for the Organization type. The SIM type Topic got the worst F-measure because enriching the query with noun-phrases produced a lot of bibliographic references and, consequently, higher name density. Error analysis showed that the most common error was name inversion with an initial and no separating point.

[3]Implemented with NLTK formal grammars library [15]

C. Qualitative Evaluation of SIM-enriched Queries

The Nominal Filter evaluation was performed by a sociologist using the UNOPORUNO system. The system shows the resulting snippets of SIM-enriched queries (Fig. 1), and allows appraisal of the snippet set from the perspective of both the SIM-type and the nominal filter. In the future, UNOPORUNO will serve as both an evaluation tool and as an infrastructure to assist social scientist in organizing their qualitative research strategies. One improvement in the UNOPORUNO system was made by the sociologist who asked that snippets be classed according to whether they were identified or not in two or more SIM types. When results converge in this way, the sociologist felt that the time taken to identify relevant mobility traces would be reduced. The initial Redes Colombia file contained 333 names of which 50 were selected for testing by the social scientist. The initial file was provided to us by our Colombian Social Science colleagues and contained a list of names that they had manually collected of Colombians living abroad. Our goal was to use the UNOPORUNO system to do two things. First, we wanted to verify that the people identified by the Colombian social scientists as potential candidates for interviewing were in fact members of the Colombian diaspora. For that we needed to find a document on the Web materializing a "mobility trace" which we defined earlier in the paper. Our second goal was to evaluate user satisfaction because when snippets are returned in response to a web search request, their number raises the needle in the haystack problem. If it takes too much time to find the evidence a user is looking for – in other words, if she has to click on a great many snippets in order to consult a large number of Web pages only to find that that the documents retrieved don't contain valid mobility traces – the user will tend to get discouraged. For this reason, we wanted to evaluate the time taken by a social scientist to verify that a person is actually a Diaspora member when using the UNOPORUNO

Table V
VALID OPERATIONS ON A NOMINAL TREE

Operation	**Description**
n: name	$n \rightarrow N$
a: surname	$a \rightarrow A$
C: compression	$CnLa$(Noe Lopez) →N\.?Lopez
E: expansion	$EnLa$(Eva M Perez) →Eva M$[a-z]$+Perez
L: literal	$LnLa$(Noe Lopez) →Noe Lopez
X: extra element	$LnXLa$(Eva M Perez) →Eva M$[A-Z][a-z]$+Perez
V: inversion	$VCnLa$(Eva M Perez) →Perez, $?+E[\.]?[-]?M \setminus .?$
SI:supress initial	$SInSIa$(Noe J Lopez F) →Noe Lopez

system. A scale ranging from less than 2 minutes to more that 15 minutes was used in the test, the idea being that user satisfaction will be highest when time of access to relevant evidence is fastest. We found that the 5 point scale we had designed for evaluation purposes wasn't necessary. For 45 out of the 50 people identified by the Colombian social scientists, access to relevant Diaspora evidence was obtained in less than 2 minutes. For the other five, the sociologist spent over 15 minutes opening and closing Web pages and gave up without a definite result. All in all, then, it took just a little over 2 hours for a social scientist to obtain evidence needed to validate her decision to interview 41 members of the Colombian diaspora.

V. Discussion and perspectives

In this section we raise two questions for discussion. First, how should we evaluate the Mobility Trace Extraction Task (MTT). The clustering measures used in the first WePS campaign don't apply in MTT, but the more traditional recall and precision measures used in the later WePS-3 Attribute Extraction Task [4] are useful for measuring the performance of the nominal filter. However, these measures only capture a part of the story in MTT. We showed that in order to have a more complete picture, we need to proceed by SIM type, measuring for each of the four sets the number of queries, the processing time and the number of people found per set. The reason is that a tradeoff has to be found between a "deep" extraction strategy and a more "shallow" extraction strategy which is less costly in terms of queries and processing time. And this is probably the most serious drawback of our approach: it requires using a Web search engine and so we depend upon how fast this engine is able to process our SIM-enriched queries. Second point: homonimy. Having only an ambiguous person's name query as input, WePS and name disambiguation approaches deal with homonimy by clustering web search documents. However, the MTT has a richer input. The semantic information generated through the use of SIM's can be used to eliminate the ambiguity associated with targeting a specific person. An individual lives in a specific area, is interested in specific subjects and works in a specific organization. Our hypothesis is that these three variables could be suffisant to identify a person without ambiguity. Furthermore, the filters we are developing should ensure coherence between a SIM-enriched query strategy and the retrieved snippet dataset (or eventually the full document dataset that the snippets reference). Our goal in this paper was to show how we have started out on our project with the Nominal Filter. However we know that geographic, biographic and semantic coherence filters will have to be developped. This will be the next phase of our program.

Table VI
Nominal filter performance on the WoS Corpus (50 persons)

SIM type	snippets	filtering%	recall	precision	F
Name	1230	69%	91%	86%	89%
Geography	2066	49%	93%	89%	91%
Topics	1975	37%	86%	81%	83%
Organizations	3447	30%	87%	87%	87%
All heuristics	8718	41%	86%	85%	85%

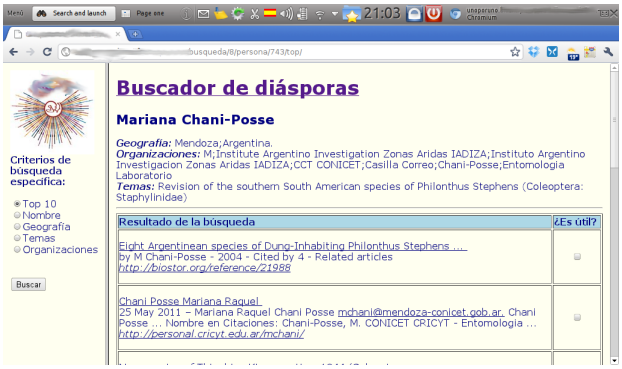

Figure 1. UNOPORUNO application searching highly qualified Latin American diaspora

VI. Conclusion

Our goal in this paper was to explain the different steps of a research progam aimed at carrying out a Mobility Trace Extraction Task. First, we showed how POS-tagging and NE techniques can be used to extract Social Identity Markers (SIMs) from the records of the Web of Science (WOS) database. Second, we showed that by using SIMs we can launch "name in context" Web search queries using noun phrases to identify the topics a person is interested in; place names to locate her geographically; and names of organisations to locate her institutionally. Third, we showed that by adopting this SIM enriched approach we were able to improve by 30% the number of mobility traces found by our method as compared to what is found using traditional name-only Web search queries. And, finally, we showed that the coherence of a snippet set produced by SIM-enriched queries can be controled using a nominal filter to eliminate name variations. The nominal filter was constructed using context-free grammar, obtaining an average F-measure of 85%. Future research will aim at building geographic and biographic filters to complete the nominal filter presented in this paper, and to work on other sources than the Web of Science database such as patent databases and the personal profiles found on social network sites.

References

[1] R. Barrere, L. Luchilo, and J. Raffo, "Highly skilled labour and international mobility in south america," in *OECD Science, Technology and Industry Working Papers 2004/10*. OECD Publishing, 2004.

[2] OCDE, *International Mobility of the Highly Skilled*. OCDE, 2002.

[3] T. Stenger and A. Coutant, "Ces réseaux numériques dits sociaux," *Revue Hermes*, no. 59, 2011.

[4] J. Artiles, A. Borthwick, J. Gonzalo, S. Sekine, and E. Amigó, "WePS-3 evaluation campaign: Overview of the web people search clustering and attribute extraction task," in *Conference on Multilingual and Multimodal Information Access Evaluation (CLEF)*, 2010.

[5] M. B. Fleischman and E. Hovy, "Multi-document person name resolution," in *ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL), REFERENCE RESOLUTION WORKSHOP*, 2004, pp. 66–82.

[6] D. Bollegala, Y. Matsuo, and M. Ishizuka, "Extracting key phrases to disambiguate personal name queries in web search," in *Proceedings of the Workshop on How Can Computational Linguistics Improve Information Retrieval?*, ser. CLIIR '06. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 17–24. [Online]. Available: http://portal.acm.org/citation.cfm?id=1629808.1629812

[7] D. Rao, N. Garera, and D. Yarowsky, "JHU1 : An Unsupervised Approach to Person Name Disambiguation using Web Snippets," in *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Prague, Czech Republic: Association for Computational Linguistics, 2007, pp. 199–202. [Online]. Available: http://www.aclweb.org/anthology/W/W07/W07-2042

[8] J. Artiles, J. Gonzalo, and S. Sekine, "The semeval-2007 weps evaluation: Establishing a benchmark for the web people search task," in *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. ACL, 2007. [Online]. Available: http://www.aclweb.org/anthology/W/W07/W07-2012

[9] ——, "WePS 2 evaluation campaign: overview of the web people search clustering task," in *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, 2009. [Online]. Available: http://nlp.uned.es/weps/weps2/papers/weps2-clustering-task-description.pdf

[10] S. Sekine and J. Artiles, "Weps2 attribute extraction task," in *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, 2009.

[11] I. T. Nagy and R. Farkas, "Person attribute extraction from the textual parts of web pages," in *In Third Web People Search Evaluation Forum (WePS-3), CLEF 2010*.

[12] I. N. de Estadística, "Análisis y estudios demográficos. nombres y apellidos más frecuentes en españa," Online database), 2010, spain. [Online]. Available: http://www.ine.es/daco/daco42/nombyapel/nombyapel.htm

[13] P. Mateos, P. Longley, and R. Webber, "El analisis geodemografico de apellidos en mexico," *Papeles de Población*, no. 65, pp. 73–103, 2010.

[14] L. Padró, M. Collado, S. Reese, M. Lloberes, and I. Castellón, "Freeling 2.1: Five years of open-source language processing tools," in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, Eds. Valletta, Malta: European Language Resources Association (ELRA), may 2010.

[15] S. Bird, E. Klein, and E. Loper, *Natural Language Processing*, 8 2008. [Online]. Available: http://www.nltk.org/book-1

[16] J. Garcia-Flores, M. Deckert, W. Turner, and M. Hurault-Plantet, "The semantics of involvement: mining the web to locate biomedical expertise," in *Proc. 4th Symposium on Semantic Mining in BioMedicine (SMBM'10), Cambridge, UK, October 25-26*. Cambridge, UK: (in press), 2010.