Identifying References to Datasets in Publications

Katarina Boland[1], Dominique Ritze[2], Kai Eckert[2], and Brigitte Mathiak[1]

[1] GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany,
{katarina.boland, brigitte.mathiak}@gesis.org
[2] Mannheim University Library, Mannheim, Germany,
{dominique.ritze, eckert}@bib.uni-mannheim.de

Abstract. Research data and publications are usually stored in separate and structurally distinct information systems. Often, links between these resources are not explicitly available which complicates the search for previous research. In this paper, we propose a pattern induction method for the detection of study references in full texts. Since these references are not specified in a standardized way and may occur inside a variety of different contexts – i.e., captions, footnotes, or continuous text – our algorithm is required to induce very flexible patterns. To overcome the sparse distribution of training instances, we induce patterns iteratively using a bootstrapping approach. We show that our method achieves promising results for the automatic identification of data references and is a first step towards building an integrated information system.

Keywords: Digital Libraries, Information Extraction, Recognition of Dataset References, Iterative Pattern Induction, Bootstrapping

1 Introduction

In empirically oriented fields of research such as the social sciences, primary data from surveys, interviews and other studies lay the basis for publications and the continuing research process. Traditionally, primary data and publications are stored in separate systems. Libraries usually concentrate on publications while research institutions mainly focus on research data. Connections between these resources are usually not available. This leads to an unfavorable situation for researchers. They need to seek connections between publications and primary data and query structurally different information systems to gain access to further information.

A typical use case would be a researcher in the field of social sciences who investigates "the opinions of German citizens about the social state" and consults several publications concerned with that topic. These publications, however, may present differing results and draw divergent conclusions. Without knowing which research data they are based on, the results cannot directly be compared and interpreted. Unless the publications and corresponding studies are interlinked, the researcher has to read every document sentence by sentence to identify the

underlying research data. With existing links, however, the researcher would be able to recognize at first glance which publications base on the same data and which of the results can be compared directly.

Manually establishing links between research data and publications is a non-trivial and time-consuming task which does not fit to the demand of transparency in research. By generating these links automatically, we aim to increase transparency and thereby improve the traceability and reproducibility of research results. We identify connections by applying a bootstrapping algorithm based on pattern induction. Starting with a study name as seed, it generates patterns which detect other references to studies and primary data. These can again be used to generate new patterns. After several iterations, the algorithm returns a list of research data references for each publication. Identifying these links constitutes the first step towards our aim of creating an integrated retrieval system.

In this paper, we first discuss to what extent abstracts and bibliographic metadata for publications and studies can help to identify connections between the data types. We then analyze the specific characteristics of study references in publication full texts. Finally, we introduce and experiment with an iterative pattern induction method to recognize references to research data in publications and give an outlook on future applications of our work.

2 Related Work

2.1 Citation Mining

Several methods have been developed to automatically find connections between publications based on citations. Although finding citations and finding references to datasets appear to be very similar tasks, they differ in some important aspects. In contrast to study references, citations are usually listed in a bibliography which significantly simplifies the linking process. Systems like TIERL [1] do not consider any information except the bibliographies to detect citations. Other systems like ParsCit [2] additionally analyze the contexts of citation to apply machine learning methods. In [3], an approach for computing the similarity of publications based on the proximity of citations is introduced. However, for citations there are standardized specifications, albeit varying, that facilitate recognition and mining. This is not the case for references to research data.

2.2 Named Entity Recognition

A similar task to finding study references is named entity extraction. Within this field of research, texts are analyzed in order to find named entities, e.g. names of persons or cities. In this sense, study names can be seen as a specific group of named entities. Existing approaches try to extract named entities for example from speeches [4]. A training set is used to learn how to identify entities. In our case, such a training set is not available and its generation would be very time-consuming. Another technique [5] uses unsupervised learning in combination

with web search engines. This works quite well to extract facts from websites but studies are usually not frequently mentioned in the Web. Based on the specific characteristics of studies and the unavailability of a training set, it is very difficult to apply named entity extraction methods on our task.

2.3 Iterative Pattern Induction

Weakly supervised bootstrapping algorithms for automatic pattern induction have been applied in a variety of different fields. All these methods start with a small set of manually created patterns or training instances and iteratively expand their training sets by labeling subsets of the test data. One area of application is concerned with finding hyponyms [6], part-of relationships [7] or even more domain-oriented relations between concepts [8, 9]. Similar to our application, the creation of lexicons can be automated by learning patterns and extracting the appropriate noun phrases. For example, this is used to generate semantic lexicons with different categories [10] or to construct medical treatment lexicons [11]. In contrast to these approaches, prior candidates cannot be easily found in our application scenario. This is because study references often appear in captions or footnotes rather than in complete sentences. This makes syntactic preprocessing difficult. Furthermore, since the contexts of the unstandardized references are very heterogeneous, flexible patterns are needed that may capture mentions in contexts such as footnotes and captions as well as mentions in continuous text. The induction of patterns that, for example, only analyze the two preceding words of a predefined candidate may therefore not be sufficient for our task. Thus, we investigate the use of patterns that do not require the identification of prior candidates but instead enable the algorithm to automatically detect the boundaries of study names.

3 Approach

To find study references in publications, we investigated different types of data: bibliographic metadata, abstracts and full texts. The former two data types are often available even if full texts are not. We therefore started by analyzing whether they alone can be helpful for detecting references. For these preinvestigations, we used studies from da|ra[3], the registration agency for social science research data, and publications from SSOAR[4], the Social Science Open Access Repository. To find references to datasets, we extracted the study titles from the metadata and checked whether they can be found in the abstracts and full texts of publications. Unfortunately, this is rarely the case because the extracted study titles are usually very extensive, for example "ALLBUS/GGSS 1996 (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften/German General Social Survey 1996)." Such detailed names are seldomly referenced. Instead, it is more

[3] http://www.gesis.org/dara
[4] http://www.ssoar.info/

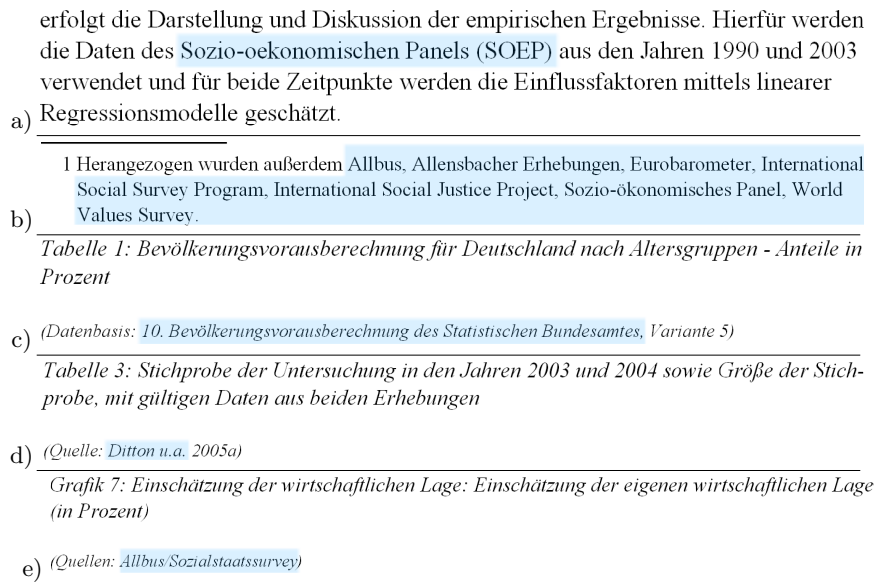

Fig. 1. Different referencing styles for datasets

common to use abbreviations like "ALLBUS 96." Without knowledge about abbreviations or synonyms, finding references is difficult. To check whether study references occur in abstracts at all, we analyzed the ALLBUS bibliography[5]. Only about 700 of 2000 publications which are listed in the bibliography mention "ALLBUS" or some longer version in combination with a year. Since ALLBUS is a very common study in Germany, we assume that other studies are referenced even more rarely.

Based on these preinvestigations, we decide to focus on full texts as they usually contain more references to datasets. Additionally, we refrain from searching for references based on lists of study names. In the following sections, we investigate the characteristics of references in full texts (Sect. 3.1) and finally introduce our pattern induction method (Sect. 3.2).

3.1 Characteristics of References to Research Data in Full Texts

Although citation standards have been proposed in the past [12], datasets are to date not referenced in a standardized way, cf. [13]. We therefore analyzed a random set of documents from SSOAR to gain insights on how datasets are actually referenced in full texts. Figure 1 demonstrates a few different reference styles we encountered. Study names are highlighted.

References to datasets are usually neither listed in a dedicated index nor included in the bibliography. They almost exclusively appear in the body of a

[5] http://www.gesis.org/fileadmin/upload/dienstleistung/daten/umfragedaten/allbus/Bibliographie/Biblio25_ris.txt

publication, except for some rare mentions in the abstract. Within the full texts, they can be found in captions of figures or tables (Figs. 1c, d and e), in footnotes (Fig. 1b) and in the continuous text (Fig. 1a).

As described before, the mention of study names themselves is not standardized. Studies are referenced using their proper title (Fig. 1b), an abbreviation (Fig. 1a), an alternative name or a different spelling (Figs. 1a and 1b). In some cases, authors cite the primary publication of a study instead of referencing the dataset itself (Fig. 1d). Figures 1b and e illustrate different ways of enumerating multiple datasets that were used.

Due to these various different possibilities to reference datasets, any method for their detection must be very flexible to recognize all types. Additionally, this variance makes manual creation of rules difficult. At the same time, it leads to the sparse data problem for machine learning approaches: as a large number of different reference styles exist, there is only a small number of mentions for each style.

3.2 Method

As our previous experiments with metadata of studies indicate, lists of study names are not helpful to reliably find references. Thus, we take a different approach: we try to identify contexts that typically include references to research data. Since we are not in possession of an annotated corpus to use for supervision and because examples for the different reference style are sparse, we apply an iterative bootstrapping method to overcome these problems.

The algorithm is depicted in Fig. 2. It starts with a study name as seed. This name should refer to an unambiguous dataset or study. It has to be mentioned frequently enough in the test corpus to allow the induction of patterns from its contexts. The selection of the seed is the only supervision our algorithm requires. As a first step, the text corpus is queried for the seed study name. For efficient search, we use a Lucene[6] index. Next, the contexts of all mentions are extracted. Based on this context set, the algorithm seeks to identify patterns that predict the existence of a study reference.

Construction of Patterns. Besides predicting occurrences of a reference, the patterns need to enable the algorithm to detect its boundaries. Patterns must therefore always include the words as well as brackets or punctuations which surround a dataset mention. For example, the whole sequence *(Quelle:Allbus)* would be extracted when the algorithm detects *Allbus* as a dataset mention. In order to avoid the generation of low-precision patterns, we require that each pattern must consist of at least one non-stop word based on a list of stop words consisting of determiners and prepositions. Additionally, at least one of the words must not consist of only punctuation or a single character. To achieve a higher recall, we normalize the words in the next step: year and percent specifications as well as numbers are substituted by placeholders.

[6] http://lucene.apache.org/

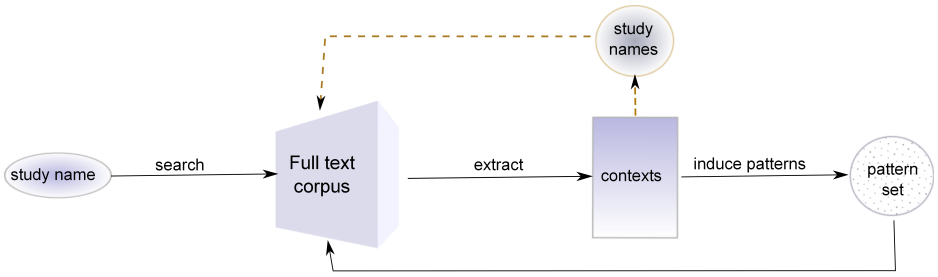

Fig. 2. Overview of the algorithm

For each context, the algorithm first tries to induce the most general pattern consisting of as less tokens as possible, i.e. <word1><study name><word2>. If it cannot find enough evidence for the validity of this patterns, it continues with more specialized patterns by expanding the range of surrounding tokens, e.g. <word1><word2><word3><study name><word4>. More specialized patterns have lower thresholds, but are only induced if the induction of more general patterns fails. For assessing pattern validity, we use a simple measure based on the relative frequency of matching contexts in the context set. We experimented with different thresholds and found the relative frequency of 0.25 (corresponding to 3 in 12 mentions) and a minimum number of occurrences of 2 to be optimal. Lower thresholds damage precision substantially. For more specialized patterns, the relative frequency threshold decreases by 0.025 for each additional word incorporated. We use a maximum number of 10 surrounding words.

Induction Strategies. In our experiments, we apply several strategies for pattern induction. At each iteration step, the algorithm can process each found seed and its contexts separately and afterwards merge the induced patterns (*separate*). Alternatively, the contexts of all found seeds can be merged first and the patterns induced based on this aggregated set (*merge*). These two strategies are depicted in Fig. 3. We will investigate the effects of the different strategies in Sect. 4.

Iterative Bootstrapping. In the next iteration step, the induced patterns are used to find new study names whose contexts are in turn retrieved from the corpus. At this step, ambiguous study names might be queried and wrong contexts might be added to the training set. If they do not resemble eacher other too much, the algorithm will not induce any patterns for those mentions.
The procedure is repeated until no new study names and patterns can be retrieved. Finally, a set of patterns, contexts and study names is returned.

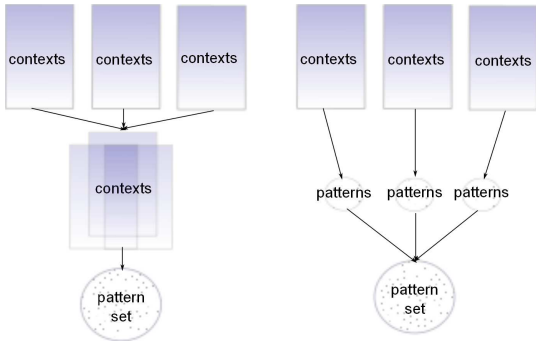

Fig. 3. Pattern induction strategies: merge (left) vs. separate (right)

4 Evaluation

4.1 Corpus and Preprocessing

For our experiments, we used publications contained in SSOAR. Due to the free availability, the transparency and replicability of our results is ensured. We selected all publications indexed with the keyword "empirisch-quantitativ" ("empirical-quantitative") which usually contain references to research data. To facilitate the manual verification of our results, we constrain our evaluation corpus to documents from the DGS corpus[7] which results in a total number of 259 documents. All considered documents are written in German. They are all available in PDF-format which requires several preprocessing steps. To extract the plain text of the PDF-files, we used the Python library PdfMiner[8]. Afterwards, we fixed the hyphenation, removed erroneous whitespaces and eliminated the bibliographies to avoid finding study names in citations. Even after these steps the documents still contain errors such as misclassified characters from OCR.

4.2 Experiments

We ran our algorithm with different seeds and pattern induction strategies to evaluate their influences. First, we compared the different induction strategies *separate* and *merge*. We hypothesized that strategy *separate* leads to higher recall but lower precision because the context sets are small and each context has a high influence on the resulting patterns. Therefore, even infrequent patterns might be found. At the same time, incorrectly classified seeds could cause the induction of wrong patterns. The more conservative strategy *merge* should prevent the induction of low-precision patterns because more evidence is needed for each pattern. In order to assess the influence of different seeds, we tested the seeds "ALLBUS," "SOEP," and "Eurobarometer." All of them are well known studies that are each referenced more than 20 times in our corpus.

[7] http://www.soziologie.de/
[8] http://www.unixuser.org/ euske/python/pdfminer/

4.3 Precision

High precision is an important criterion for the applicability of our approach: if a document containing an ambiguous study name which does not refer to the study in the particular context is linked to the dataset of the same name, this would hinder the information retrieval task for the user.

4.4 Recall

For lack of a gold standard, we generally apply the following approach to assess the recall of our method:

1. choose a study as reference study and use its contexts as reference set
2. run algorithm: if the reference study is used as seed, remove patterns induced from its contexts after first iteration
3. compare the algorithm's output to the reference set:
   how many of the mentions in the reference set were found by the algorithm?
4. do so for multiple reference studies to estimate the average recall

More precisely, we assess recall regarding three different factors:

Disambiguated Recall. Unambiguous titles can be queried and all contexts assumed to be referring to the study. For the recognition of ambiguous study names, only those contexts matching one of the induced patterns can safely be assumed to be a dataset reference. For the measurement of *disambiguated recall*, we therefore measure how many mentions of the reference study name the algorithm is able to find after having trained on several contexts. Note that the algorithm may find the reference study in one of its iterations and use its contexts for training.

Singleton Recall. Finding mentions for studies that have already been identified and whose contexts have been used for training is easier than identifying new studies in the first place, especially if they appear in dissimilar contexts. To get a meaningful assessment on the ratio of distinct study names that can be detected, we prevent the algorithm from processing the contexts of the reference study by instantly removing it from the seed set if it is found in any iteration. This measure corresponds to the recall anticipated for studies occurring only once in the corpus.

Identification of Alternative Names. The importance of finding abbreviations and synonyms was illuminated in the preinvestigations. Our algorithm should find alternative names as different study titles and thereby improve recall both for ambiguous and unambiguous studies. We examine whether the algorithm indeed manages to find abbreviations and synonyms in the final part of our evaluation.

Table 1. Comparison of different seeds and strategies

Seed	Strategy	Precision	Recall (mentions)	Recall (docs)	Mentions (total)
"ALLBUS"	merge	1.00	0.24	0.50	134
"ALLBUS"	separate	0.97	0.29	0.60	303
"SOEP"	separate	0.97	0.29	0.60	299
"Eurobarometer"	separate	0.97	0.29	0.60	306

In addition to these measures, we supply the total number of mentions for all studies that have been found. Together, this information permits a good approximation of the coverage of our method. To our knowledge, there has been no other work on this task so far. Therefore, we cannot compare our results to other approaches.

4.5 Results

Table 1 shows the results for different seeds and strategies. For assessment of recall, we used "ALLBUS" as reference study and *disambiguated recall* as measure. In total, the reference study has 34 mentions in 10 documents.

In sum, all strategies and seeds produce very high values for precision. The established links are therefore reliable enough to allow their integration into an information system. It is important to note that about 14% of all mentions (for the first configuration) in fact are references to literature. This is due to the fact that authors sometimes cite the primary publication instead of a study. In some cases, they also refer to information from secondary publications. Since these indirect references are nevertheless crucial to find, we count these citations as correct study references. For these mentions, a list of datasets and their primary publications is required to make the connection.

Recall for mentions does not exceed 0.29. Since we are interested in linking publications and datasets, it is only important to retrieve at least one reference to a dataset per document. The algorithm succeeds to do so in 60% of the cases when processing each context separately and in 50% of the cases when processing merged context sets. Consistent with our hypothesis, applying the more conservative strategy increases precision but hurts recall. As the more liberal strategy already performs well with regard to precision, this strategy appears to be favorable for the identification of study references.

For the total number of mentions retrieved for all studies, we can only provide absolute numbers as the total number of different studies contained in each document is unknown.

The choice of a seed has only minimal influence on the results. For each of them, the algorithm terminates after four iterations with very similar outcomes. There is a small number of diverging patterns which leads to slightly different precision values. The recall values do not differ which indicates that the diverging patterns are not relevant for finding the particular reference study.

Table 2. Different reference studies and measures for the measurement of recall

Study	Recall *singleton* (mentions)	Recall *singleton* (docs)	Recall *disamb.* (mentions)	Recall *disamb.* (docs)
ALLBUS	0.14	0.30	0.29	0.60
SOEP	0.01	0.04	0.42	0.65
Sozio-oekonomisches Panel	0.78	0.50	0.78	0.50
Wohlfahrtssurvey	0.42	0.57	0.42	0.57
Westdeutsche Lebensverlaufsstudie	0.30	0.50	0.78	1.00
EVS	0.22	0.25	0.22	0.25
ESS	0.23	0.50	0.31	0.50

For the previous experiments, recall has been measured for one particular study. However, we have no information about the relative difficulty to retrieve this study compared to others. To get more reliable information about the recall of our method, we therefore measured recall for a variety of different studies in our next experiments. Here, we used "ALLBUS" as seed and applied the more liberal induction strategy. Additionally, we compared the recall measures *singleton* and *disambiguated*. The results are listed in Table 2. Disregarding the outliers, *disambiguated* recall lies mostly between 0.3 and 0.4 for mentions and between 0.5 and 0.6 for documents. Since the recall values for various studies are similar, these numbers should constitute a good assessment of the algorithm's total recall. *Singleton* recall differs considerably from *disambiguated* recall for some reference studies ("ALLBUS," "SOEP") and remain unaltered for others (e.g. "Sozio-oekonomisches Panel" and "Wohlfahrtssurvey"). This result is caused by the fact that some studies like "ALLBUS" are frequently referenced by mentioning the full title followed by the abbreviation in brackets. In the *disambiguated* condition, the algorithm learns to use the full titles as a cue to identify the mention in brackets as a study mention. In the *singleton* condition, this information is missing and these mentions cannot be found. In addition, the exclusion of studies may also prevent the induction of more general patterns and thus lower recall for studies in general. This is the case for "SOEP" which is the most frequently referenced study in our test corpus and thus an important source for new patterns. The solution for this problem would be the usage of a larger corpus: in our test corpus, "SOEP" appears 81 times in only 23 different documents. We anticipate higher recall values for all conditions and recall values when applying the algorithm on larger corpora.

The absence of year or version specifications in some references helps the algorithm to expand its patterns and include these specifications, if present, into the study titles (e.g. it learns to identify "ALLBUS 2000" as a study title, not only "ALLBUS"). This sometimes produces duplicates which, however, can easily be removed automatically. Duplicates were not counted for precision and recall measurement.

As discussed before, recall for every single study is also determined by the number of different titles found. The algorithm succeeds in expanding its search to abbreviations, alternative titles and different versions of studies. For example, it learned that "SOEP," "Sozio-oekonomisches Panel," "SOEP-Ost," "SOEP Sondererhebungen," "SOEP-Zuwandererstichprobe," and "SOEP Pretest" constitute references to datasets – they all denote the same study series. In the preliminary examination, we highlighted the importance of knowing alternative study names.

Altogether, our algorithm receives highly precise results with acceptable recall. Note that references to unambiguous studies can be retrieved by searching the study name in the corpus instead of only using the unambiguously found mentions. The alternative titles and abbreviations can then be used to expand the search and improve recall. For these cases, the supplied recall values serve only as a lower bound.

5 Conclusions and Future Work

Connecting publications and datasets is a non-trivial task due to the absence of standards for dataset references. Although a wealth of information is available in bibliographic metadata and abstracts, these are generally not sufficient to establish connections between the data types. We therefore introduced an iterative pattern induction method to recognize dataset references in full texts. We showed that our approach achieves very useful results while requiring only minimal supervision (the manual selection of a non-ambiguous seed) and shallow features (e.g. no layout information is needed).

As part of our future work, we plan to investigate the use of pattern and instance ranking (see for example [11]) to enhance our frequency-based measure. Incorporating this additional evidence might allow the frequency thresholds to be decreased without damaging precision. Additionally, we are going to apply our algorithm on English documents and publications from other scientific fields to verify the language- and domain-independency of our approach. Furthermore, it would be interesting to investigate the different patterns emerging in different languages and domains and to thereby investigate how research data is referenced in different languages and scientific communities.

As a next step, we will match the acquired study references with metadata available in the DBK[9], a research data catalogue maintained by GESIS. We will use this information to integrate the links between research data and publications into the scientific information retrieval systems Sowiport[10] by GESIS and Primo[11] by Mannheim University Library. We will then be able to implement a first prototype of an integrated information system for both data types.

[9] www.gesis.org/en/services/research/data-catalogue
[10] http://www.gesis.org/sowiport
[11] http://www.bib.uni-mannheim.de/133.html

Acknowledgements. This work is funded by the DFG as part of the InFoLiS project (SU 647/2-1). We would like to thank Benjamin Zapilko and Christian Meilicke for their great support.

References

1. Afzal, M.T., Maurer, H., Balke, W.T., Kulathuramaiyer, N.: Rule based autonomous citation mining with tierl. Journal of Digital Information Management **8**(3) (2010) 196–204
2. Councill, I.G., Giles, C.L., Kan, M.Y.: Parscit: An open-source crf reference string parsing package. In: Proceedings of the Language Resources and Evaluation Conference, European Language Resources Association (2008)
3. Gipp, B., Beel, J.: Citation proximity analysis (cpa) - a new approach for identifying related work based on co-citation analysis. In: Proceedings of the 12th International Conference on Scientometrics and Informetrics. Volume 2. (2009) 571–575
4. Kubala, F., Schwartz, R., Stone, R., Weischedel, R.: Named entity extraction from speech. In: DARPA Workshop on Broadcast News Understanding Systems. (1998) 287–292
5. Etzioni, O., Cafarella, M., Downey, D., Popescu, A.M., Shaked, T., Soderland, S., Weld, D.S., Yates, A.: Unsupervised named-entity extraction from the web: An experimental study. Artificial Intelligence **165** (2005) 91 – 134
6. Hearst, M.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th Conference on Computational Linguistics, Stroudsburg, PA, USA, Association for Computational Linguistics (1992) 539–545
7. Berland, M., Charniak, E.: Finding parts in very large corpora. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, Stroudsburg, PA, USA, Association for Computational Linguistics (1999) 57–64
8. Pennacchiotti, M., Pantel, P.: A bootstrapping algorithm for automatically harvesting semantic relations. In: Proceedings of the Inference in Computational Semantics. (2006) 87–96
9. Meusel, R., Niepert, M., Eckert, K., Stuckenschmidt, H.: Thesaurus Extension using Web Search Engines. In: Proceedings of the 12th International Conference on Asia-Pacific Digital Libraries, ICADL 2010, Gold Coast, Australia, June 21-25, 2010, Springer (2010)
10. Thelen, M., Riloff, E.: A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In: Proceedings of the 2002 Conference on Empirical Methods in NLP, Stroudsburg, PA, USA, Association for Computational Linguistics (2002) 214–221
11. Xu, R., Morgan, A., Das, A.K., Garber, A.: Investigation of unsupervised pattern learning techniques for bootstrap construction of a medical treatment lexicon. In: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing. BioNLP '09, Stroudsburg, PA, USA, Association for Computational Linguistics (2009) 63–70
12. Altman, M., King, G.: A proposed standard for the scholarly citation of quantitative data. D-Lib Magazine **13**(3) (2007)
13. Green, T.: We need publishing standards for datasets and data tables. OECD Publishing White Paper (2009)