*Moim Rodzicom*

POZNAN UNIVERSITY OF ECONOMICS
MACQUARIE UNIVERSITY SYDNEY

AGATA FILIPOWSKA

# Spatial Indexing for Creating Company Profiles

PhD Dissertation

Supervisor: dr hab. Witold Abramowicz, prof. nadzw. UEP
Cotutelle Supervisor at Macquarie University Sydney: prof. Leszek Maciaszek

Poznań 2009

## *Table of Contents*

## List of figures

## List of tables

## *List of abbreviations*

| | |
|---|---|
| ACE | *Automatic Content Extraction* |
| ADM | *Administrative regions* |
| ADR | *Addresses with zip codes and building numbers* |
| CIT | *City* |
| CMN | *Commune* |
| CNT | *County* |
| CPT | *Cost Per Thousand* |
| CRY | *Country* |
| CTR | *Click Through Ratio* |
| DAML+OIL | *DARPA Agent Markup Language + Ontology Inference Layer* |
| DB | *Database* |
| DNS | *Domain Name System* |
| DOM | *Document object model* |
| ED | *Euclidean Distance Measure* |
| GIR | *Geographic Information Retrieval* |
| GIS | *Geographic Information Systems* |
| GKB | *Geographic Knowledge Base* |
| GPE | *Geographical and political entity* |
| GPS | *Global Positioning System* |
| HD | *Hierarchical Distance Measure* |
| HTML | *HyperText Markup Language* |
| HTTP | *Hypertext Transfer Protocol* |
| ICANN | *Internet Corporation for Assigned Names and Numbers* |
| IE | *Information Extraction* |
| IF | *Information Filtering* |
| IP | *Internet Address* |
| IR | *Information Retrieval* |
| LAN | *Land forms (e.g. continent names, geographical regions)* |
| MEP | *Maximum Entropy Principle* |
| MUC | *Message Understanding Conference* |
| NE | *Named Entity* |
| NLP | *Natural Language Processing* |
| OWL | *Web Ontology Language* |
| PR | *Public Relations* |
| PRO | *Province* |
| RDF | *Resource Description Framework* |
| RSS | *Rich Site Summary* |
| SoV | *Share of Voice* |
| SPIRIT | *Spatial Search Engine* |
| SProUT | *Shallow Processing with Unification of Typed feature structures* |
| STP | *Shallow Text Processing* |
| TDL | *Type Description Language* |
| TF*IDF | *Term Frequency*Inverse Document Frequency* |
| TGN | *Thesaurus of Geographic Names* |
| TSD | *Total Spatial Distance* |
| UML | *Unified Modelling Language* |
| URL | *Uniform Resource Locator* |
| WAT | *Water bodies* |
| WSML | *Web Service Modeling Language* |
| XML | *Extensible Markup Language* |
| XTDL | *Extended Type Description Language* |
| ZON | *Zones* |

# Introduction

*Everything is related to everything else,*
*but near things are more related than distant things.*
*(Tobler, 1970)*

## Motivation

Currently an increasing number of initiatives in the domain of information retrieval concern providing users with more relevant and contextualized information. Information provided as a response to a query should not only be relevant but also as concise as possible. Geographic information retrieval systems (Bucher et al., 2005, Gey et al., 2006) as well as local search services such as Google Maps[1], Yahoo! local[2] or Zumi[3] are now of paramount importance in practice and research. (Schockaert and Cock, 2007) report that the geographical search engines suffer from two important limitations. Firstly, their knowledge base is built usually based on structured information available to the system in a given point in time and because of that it is rarely updated. Such knowledge doesn't go far beyond what one may find in the Yellow Pages. The second issue concerns the fact that only very simple queries are supported by the system.

The importance of the Internet as a major information system for companies dealing with public relations (PR) has been growing instantly (Wright and Hinson, 2008). Not only portals of all leading news agencies provide news, but also support Web 2.0 activities such as discussion forums and blogs. A vast amount of precious information is also published by organizations and users on their private web pages. Internet resources are therefore perceived as the most important sources of information for PR agencies when establishing, maintaining or changing the company image (Seitel, 2006). However, because tools supporting PR experts in the process of acquiring and processing information from the Web are still far from being perfect, the Internet is not used at a bigger scale (Kaczmarek-Śliwińska, 2006).

Assuming that a PR agency monitors hundreds of traditional sources when preparing its daily reports, the Internet offers many more resources that can no longer be dealt with manually by

[1] http://local.google.com

[2] http://local.yahoo.com

[3] http://www.zumi.pl/

people employed in PR agencies (Wojcik, 2001). (Pouliquen et al., 2004) mentioned in 2004 the need for semantic annotation of web sources in order to process the information automatically, but the Semantic Web has not yet fulfilled its promise (Berners-Lee et al., 2001).

Another issue underlined by the PR experts is the need of imposing structure on content acquired. Currently, in order to answer a question "what is the image of a given company?" one needs to gather a vast amount of data and then rework it in order to achieve a coherent view enhanced with graphical elements (e.g. charts).

One of dimensions of analysis that is carried out by PR experts is the geographical dimension. The United Nations also recognize the importance of location in order to "improve knowledge and decision-making by extending the traditional role of maps to support the rapid integration, analysis and modeling of information critical to achieve improved operational readiness and responsiveness" (Group). But the Internet doesn't deliver the geographically-oriented information in any explicit manner and even the search engines and catalogs do not provide users with any such structured view on this information as one may find when using Yellow Pages (Himmelstein, 2005).

Nowadays, a lot of research is carried out regarding the geographical description of documents based on information extracted from their contents (Kimler, 2004, Leidner, 2007a). This trend originated from the series of Message Understanding Conferences,[4] which aimed at choosing the best mechanisms of information extraction from documents obtained from different sources. If one analyses documents from a single corpus[5], it can easily be noticed that the key to understanding of the text is to properly annotate proper names it mentions. (Gey, 2000) says that about 30% of content-bearing words are proper names. (Friburger and Maurel, 2002) provided statistics for a model corpus of documents saying that 10% of the words in newspapers are proper names out of which 43,9% are locations. Moreover, when taking into account the analysis of the websites, approximately 4,5% of web pages contain a recognizable US zip code, 8,5% contain recognizable phone number, and 9,5% contain at least one of these (however, no details on the population investigated are provided) (McCurley, 2001). According to (Delboni et al., 2005) at least 20% of the web pages include one or more easily recognizable and unambiguous geographic identifiers. The

[4] http://www-nlpir.nist.gov/related_projects/muc/

[5] Set of texts manually annotated with named entities (LEIDNER, J. L. (2004). Towards a reference corpus for automatic toponym resolution evaluation. *Workshop on Geographic Information Retrieval, SIGIR.*)

geographical annotation of web resources is especially important. This may be observed in changing requirements and expectations of consumers of information, namely that:

- locality of information: information from a number of web pages is interesting only for local communities and most of the pages explicitly manifest their geographic context, e.g. by including phone numbers or addresses (Buyukkokten et al., 1999, McCurley, 2001). Studies based on the Excite query log show that approximately 18% of all queries contain some kind of geographic information (Sanderson and Kohler, 2004). Similar studies carried out by (Jones et al., 2008) indicated that 12,7% of queries contained a place name. The research outcomes published in (Spink, Jansen et al. 2002) show that 19,7% of their random sample of 2,453 unique queries contained "people, places or things";
- query mechanism: interest of users to provide spatial queries increases (McCurley, 2001, Vestavik, 2004). Search engines still ignore the geographic scope of the Web pages and do not consider natural language spatial relations such as "close to", "inside", "in front of", "5 minutes from" and all of their terminological or semantic variations to be equivalent (Delboni et al., 2005). Such a need is also underlined by (Bilhaut et al., 2003), who after an in-depth analysis state that documents should not only be matched against the query, but also both need to be semantically analyzed. Use of the geographical information may lead to reducing and reordering the result set in order to match the user's interests (Markowetz, Brinkho et al. 2004);
- presentation: change in approach to presentation of the geographically important documents, e.g. users are given an interface containing a traditional text search form combined with a map enabling them to zoom in on areas of the world that are of interest, and the results of textual queries are plotted on the map (Rauch et al., 2003);
- new application possibilities: extensive development of localization services (via mobile phones, GPS) and the need to find local information while traveling on business or for leisure, plus wide and free access to mapping services (e.g. Google Maps, Street View).

However, not only search engines community underlines a need for investments in this domain. A recent strategic assessment by the U.S. government identified bio-technology, geo-spatial technology and nano-computing (the triad "bio, geo, nano" for short) as the three key technology growth areas in the first decades of the 21st century. This directly highlights the importance of this area of research (Leidner, 2007b).

One of the specific features of geographical information is its non-relativity. If we provide a group of people able to read maps with coordinates of a place, they will all point to the same location. Extracting geographical data from documents enables us also to provide such a non-relative description of the documents and to provide user with a different method of browsing (Arikawa et al., 2000). (Hill 2000) proposed to use place names to identify point locations for navigation purposes. Users can enter a short name identifier for, say McDonald's restaurant, and find McDonald's locations nearest to their current position. Moreover, such information may be then further exploited by Geographical Information Systems (GIS).

Current approaches to extraction of information from documents provide precision and recall scores over 90% (it was so even more than 10 years ago (Agency, 1996)). However, the challenge now is not only to recognize a fragment of text as named entity, but also to recognize entities on a semantic level and provide all details related to them (Maynard et al., 2003).

The information extraction for Polish is not a well explored field. Moreover, the task is more demanding for Polish than for e.g. the English language because of such phenomena as free order syntax and inflection that are discussed in the next chapters.

The dissertation firstly discusses methods and algorithms used in public relations and geographical information retrieval. Then it proposes an ontology-based indexing mechanism addressing the above described limitations and requirements emerging from the public relations domain. This mechanism is to provide an acceptable level of precision and recall when processing texts written in Polish. It is to be further compared with a second developed spatial indexing method, namely a source-based document indexing method. The source-based document indexing is supposed to provide an initial geographical index of a document that may be further used e.g. while disambiguating named entities using the ontology-based spatial indexing method.

## *Main aim and thesis of the dissertation*

The dissertation focuses on the issue of spatial (geographical) document indexing. It is to propose an original spatial indexing method. This method is to process free text documents written in the Polish language and filtered from the Internet. An effect of this processing are document indexes (surrogates) that are to be further used in the PR campaign.

In particular, two methods for geographical indexing of documents and their sources will be worked out. These mechanisms emerge from the state of the art achievements in the field of geographic information retrieval.

Therefore the main research hypothesis of this dissertation was formulated as follows: t*he introduction of a semantic-based accurate and precise geographical indexing mechanism will provide functionalities needed for creation of a new kind of a search engine that may be used within the entry and output phases of the PR process.*

Moreover, the following research goals are defined:

- structuring of public relation analysts' requirements towards a search engine to be used in the entry and output phases of the PR process,
- development of a corpus of news articles written in Polish, which can be used e.g. during an evaluation of the spatial indexing mechanism,
- development of a geographical ontology and a gazetteer for specification of artifacts emerging from geographical resources for Poland, that may be processed automatically,
- definition of the spatial indexing methods allowing for producing accurate surrogates of documents.

Addressing the research goals requires investigation of the following issues:

- reviewing the current approaches to analysing documents with regard to their geographical dimension,
- analysing and structuring requirements towards a spatial indexing mechanism emerging from requirements on the PR search engine,
- development of annotation guidelines for preparation of a corpus of news articles written in Polish,
- definition and development of a document-based spatial indexing mechanism,
- definition and development of a source-based document indexing mechanism.

To summarise, the following research questions resulting from the research goals, are discussed within this dissertation:

- What are the requirements of public relation analysts working on PR campaigns towards the search engine to support the process?
- What requirements for a spatial indexing mechanism emerge from the requirements on the PR search engine?
- How the spatial indexing method should be developed in order to deal with challenges emerging from the Polish language?

- How to link existing gazetteers with ontologies to exploit advantages of ontologies within an extraction process?
- What features of a source should be taken into account while developing the source-based document indexing mechanism?

The first proposed spatial indexing method is to apply the information extraction techniques (including various heuristics) to produce a document index. This index is used by a second method while training and validation of the source-based document indexing approach.

The research work was carried out with regard to design science paradigm discussed in the next section. According to this methodology the following workplan is defined. Firstly, the analysis of the specifics of the PR domain is carried out. Then the analysis of information systems and computer science fields from the point of view of this dissertation follows. Further, methods for spatial indexing are proposed. These methods are then tested against the corpora of documents and conclusions are drawn.

## *Research methodology applied*

Research in the Information Systems discipline is characterized by two paradigms: behavioural science and design science. The behavioural science tries to develop theories that explain or predict behaviour or rather analyse effects of application of information systems on individuals and organisations. The design-science paradigm focuses on extending boundaries of human and organizational capabilities by creation of new artefacts (Hevner et al., 2004). These artefacts are to solve an existing organisational problem, and therefore the design-science originates from the engineering disciplines.

This dissertation provides new, original mechanisms for the spatial document indexing. These methods are to support PR analysts in the PR process. Following the second, design-science paradigm, the dissertation delivers artefacts to efficiently solve the described PR problem.

Figure 1 presents the conceptual framework for understanding, executing as well as evaluation research in the domain of information systems (Hevner et al., 2004). The environment defines the problem space for the domain addressed. Information systems (IS) environment consists of people, organizations and technologies (Silver et al., 1995). These are closely related to goals, tasks, problems and opportunities organizations face. Business needs are evaluated by people based on the context of organizational structure, culture, technological infrastructure and recent developments, etc. These factors define problem perceived by a researcher.

Information systems design research based on the defined needs focuses on building and evaluating theories and models designed to meet these needs. Evaluation helps to identify weaknesses in the theory artefacts and suggest possible refinements of the proposed model.

The knowledge base provides foundational theories, frameworks, instruments, methods used in the build phase of the research conducted. This also incorporates computational methods that are used to evaluate the quality and effectiveness of artefacts (including also empirical methods).

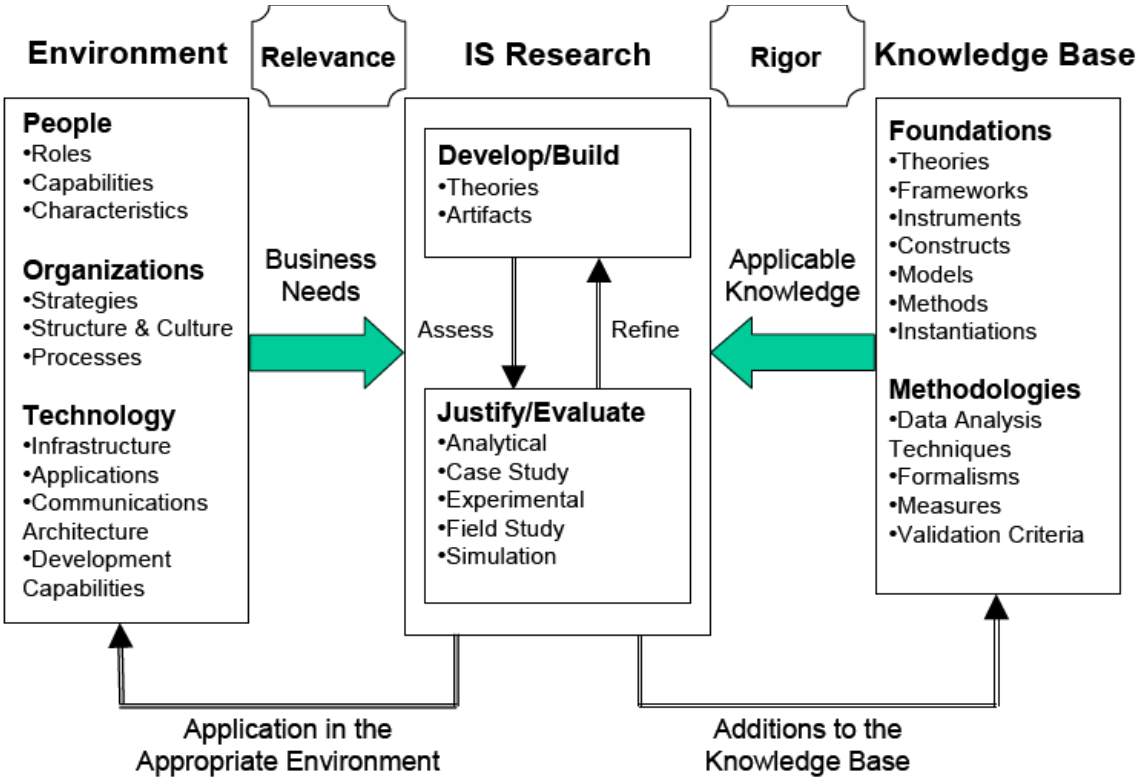



**Figure 1. Information Systems Research Framework. Source: (Hevner et al., 2004)**

The approach described by (Hevner et al., 2004) provides also a classification of results of a design-science approach. They are divided into:

- constructs i.e. vocabulary and symbols used while defining problems and solutions,
- models, which use constructs to represent a problem and a solution space,
- methods solving the defined problem,
- instantiations proving that models and methods may be implemented in a working system.

The artefacts resulting from this thesis are as follows:

- constructs: chapters 1, 2 and 3 provide a comprehensive overview of the terminology used within this dissertation. Additionally, chapter 4 presents type hierarchy, geographical ontology and gazetteer that are developed and used in this research.
- models: chapter 4 introduces both spatial indexing methods providing also a high-level description of a framework that may incorporate them,
- methods: chapter 4 provides an in-depth overview of both spatial indexing methods including heuristics and information extraction rules used,
- instantiations: chapters 5 and 6 provide evaluation and validation of proposed approaches.

To summarise, this dissertation follows the Information Systems Framework and methodology as described by (Hevner et al., 2004). Therefore, three parts of the dissertation were distinguished. Firstly, the research environment is described. The research problem originates from the domains of information retrieval and information extraction. Therefore, an overview of these domains is presented. Then, conceptual models of both spatial indexing methods are created taking into account remarks of PR analysts. Finally, both methods are evaluated on the corpus of annotated news filtered out from RSS feeds.

## *Structure of the dissertation*

This dissertation as indicated in the previous section is divided into three parts.

Part I includes chapters 1, 2 and 3. The chapter following the Introduction discusses the notion of information and its role in the public relations domain. It elaborates on the public relations process and the requirements it imposes on tools that aim at supporting its phases.

Chapter 2 presents technologies that are used when developing the method of the spatial indexing. It introduces the fundamental knowledge from the area of information extraction and information retrieval with the focus put on the ontologically-supported information retrieval.

The issues of geoparsing and geocoding addressed by this dissertation are discussed in chapter 3. This chapter was structured to present the importance of analysis of both documents and the sources that the documents were retrieved from. We believe that information on sources is especially useful when disambiguating the geographical name mentions and producing initial document indexes.

Part II includes chapter 4, where two methods for spatial indexing are described. These methods combine experience from different fields extending it with ontologies as well as applying it to the Polish language.

Part III provides evaluation, validation and application examples of the developed methods. Chapter 5 presents the evaluation of the two proposed indexing methods. Experiments to prove the usefulness of the method are based on the developed corpus of documents retrieved from RSS channels. The accuracy of the proposed mechanisms is evaluated and discussed. The annotation guidelines are introduced as they were also developed for the needs of this research.

Chapter 6 presents the application of a spatial indexing method in the public relations domain. It also compares the functionalities that may be offered after implementation of the spatial indexing method with the functionalities of currently existing tools.

The dissertation concludes with a summary explaining the results achieved as well as the benefits for the public relations process. In this final chapter also some future work directions are presented and shortly elaborated.

# Chapter 1.
# The role of the information in the public relations process

*The most important thing in communication*
*is to hear what isn't being said.*

*Peter F. Drucker*

## 1.1. Introduction

In 2001 the Nobel Prize in Economics was awarded to George Akerlof (Akerlof, 1970), Andrew Spence (Arrow, 1963) and Joseph Stiglitz (Stiglitz, 2000) "for their analyses of markets with asymmetric information"[6], thereby underlining the importance of information in economics research. Effective use of this strategic resource by companies may lead to a market success (Jarosz-Palach, 2005). Information, in comparison to "traditional resources", such as labour and capital, is a very specific resource – expensive, when it comes to the cost of its creation, and cheap in case of a reproduction (Kelly, 1997, Shapiro and Varian, 1999). Moreover, companies are interested in gaining an efficient access to already existing one.

The Internet is currently the largest existing source of information. The volume of information resources stored in the Internet is estimated for about 281 exabytes of data (Gantz et al., 2008).

The main advantage of the Internet, or the drawback depending on the viewpoint, is its public availability. Therefore, the statement that an access to the Internet *doesn't guarantee achieving a success by a company, but without it this success will be extremely hard to achieve* (Szyfter, 2005) seems to be justified. The confirmation of this fact may be found in the recent research showing that the Internet contributes to the reduction of information asymmetry (Levitt and Dubner, 2005). However, the Internet is not a perfect resource in terms of quality of information and information overload e.g. (Berghel, 1997, Edmunds and Morris, 2000, Grise and Gallupe, 2000). The assumption that business entities are aware of all possible alternatives and have an ability to assess all possible outcomes is not valid anymore (Balcerowicz, 1997). The problem of information overload is widely discussed and addressed in the Information Systems domain. Some solutions to this problem may be found e.g. in the field of information filtering and retrieval (Abramowicz, 2003) as well as data extraction and aggregation (Kowalkiewicz et al., 2006a).

[6] http://nobelprize.org/nobel_prizes/economics/laureates/2001/index.html

One of domains dealing with a large amount of data and information, that actually benefits from the potential of the Internet, is public relations (PR) (Kaczmarek-Śliwińska, 2005a). PR experts have identified a number of sources in the Internet, that may be useful when carrying out a public relation process (sometimes referred also as a PR campaign) (Kaczmarek-Śliwińska, 2006). However, the problem of the information overload and the need of eliminating the unnecessary information is still valid. A solution that would lead to restricting the amount of information that is used in the PR process, would substantially influence the efficiency of analyses carried out.

## 1.2. Information processing

The research in the fields of economics and public relations clearly indicates that adequate information is crucial in order to take the most appropriate business decisions. Before describing the role and types of information used within the PR process, it would be worth to shortly discuss the definition of information (Shapiro and Varian, 1999, Shannon, 1993, Floridi, 2005). Following the Cambridge Dictionary of Philosophy (Audi, 2001), information is *an objective (mind independent) entity. It can be generated or carried by messages (words, sentences) or by other products of cognizers (interpreters). Information can be encoded and transmitted, but the information would exist independently of its encoding or transmission*. Other suggestion incorporates the notion of data into the definition of information: *data is the raw material that is processed and refined to generate information* (Silver and Silver, 1989). From the point of view of information economics, information is perceived as *a kind of resource, that enables extending our knowledge about us and the surrounding world* (Kisielnicki and Sroka, 2001).

The definitions such as those presented above turn our attention to the specifics of information. Information is a resource being interpretation of messages that enables enhancing (obtaining) knowledge about object's surroundings. These definitions however, need to be extended by the business dimension of information. Following a definition of economical information provided by Oleński, it is *information that relates to complex economical objects (subjects) or economical categories (a company or consortium of companies, domain, region, or country economy), and its features are metrics of economical categories applied to these objects* (Oleński, 2001).

For the purpose of this dissertation, information is defined as a resource that enables building knowledge on a company's environment from the public relations' point of view.

Information, especially when transmitted via network such as e.g. the Internet, may be distorted. This may be caused by people involved in the process of transmission. Distortion or loss of information on the way between its sender and receiver is defined as information noise (Shannon, 1948). (Sznajder, 1993) differentiates three types of information noise:

- Physical – caused by the illegibility of information: takes place when receiving a message or missing parts of information.
- Semantic – related to wrong use of words or symbols (especially important when taking into account globalization – not all the words or symbols are uniform).
- Disinformation – informing receivers about fictitious facts.

In order to reduce the information noise, feedback (checking efficiency of transmission) in communication between a message sender and a receiver should take place.

The figure below presents the communication system as drawn by (Shannon, 1948). The central part of the system where the noise has its source may be substituted by the Internet and search engines. The noise can be then reduced e.g. by using proper and accurate information filters.

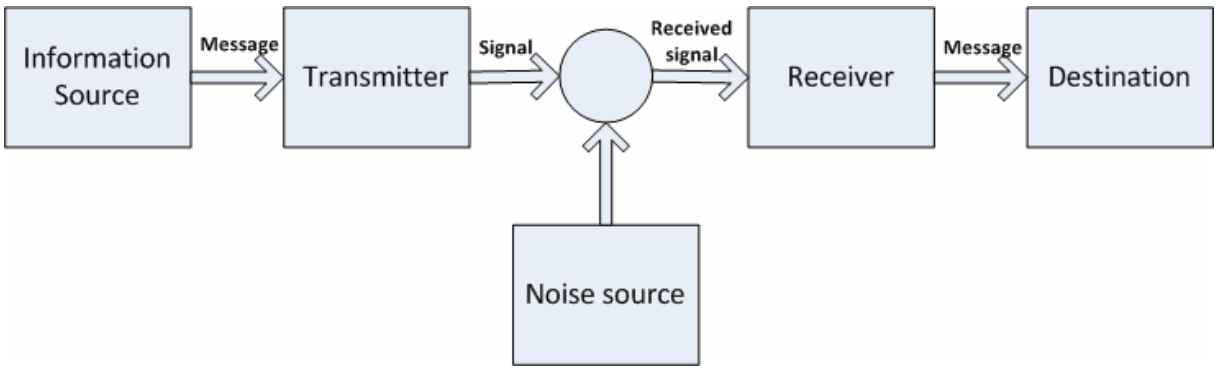



**Figure 2. General communication system. Source: (Shannon, 1948)**

Information found in the Internet may be categorized in many ways. Usually, categorization depends on its aim as well as on type of information. One of classifications of business information was proposed by (Januszko, 2001). According to Januszko business information may be categorized as follows:

- Information about a company – its profile, contact data, category of products or services, benchmarks;
- Information about products and services of a company – categories and prices, parameters, method of payment and delivery;
- Information about cooperative needs of the company – bids, offers, subcontracting;
- Information on fairs and exhibitions;

- Information about business meetings;
- Information on provided services including detailed documentation on exchange of goods with other countries e.g. certificates and similar documents;
- Information about the financial support – loans and other sources of financing;
- Information about law;
- Information about organizations supporting economical activities;
- Information about countries from the point of view of their economy and demographic information;
- Bibliographic information – outcomes of market search, statistical analyses and reports, press articles;
- Information on references to documents on economy;
- Information on people.

This research focuses on geographic dimension of mentioned information categories with regard to a company profile that is presented further in the dissertation.

## *1.3. Public relations*

Public relations is one of domains that widely utilize information. Public relations according to the mostly used definition by Public Relation Society of America *"helps an organization and its publics adapt mutually to each other". (Wojcik, 2001)* defines PR as an *"art and social science that analyses trends, forecasts their consequences, consults company management and implements planned programs aiming at satisfying needs of both organizations and public communities"*. Other definitions present PR as e.g. *"a set of management, supervisory, and technical functions that foster an organization's ability to strategically listen to, appreciate, and respond to those persons whose mutually beneficial relationships with the organization are necessary, if it is to achieve its missions and values"* (Heath, 2004). Practitioners define PR in a more general way e.g.: *"PR makes organizations more effective by building relationships with strategic publics"* (Grunig and Huang, 2000) or *"PR, simply defined as doing the right thing – of performing – and communicating the substance of that performance"* (Seitel, 2006).

Following these definitions it is worth mentioning that PR activities should be systematic and should implement certain plan that aims at establishment of mutual understanding between organization and its surroundings ((Crisford, 1974) after (Doktorowicz, 1999)). Similarly the "perfect PR" is perceived by James E. Grunig (who defined four different models of

communication) as steps to perfect PR that may be implemented within a company (Wojcik, 2005). He states that public relations activities mature in the following four stages:

1. press agency, when a PR Department acts as a press agency and offers various press releases on the company to be used by media;
2. public information broadcasting, when a company employs people (usually journalists and reporters) to work for a PR Department. New employees should be familiar with the way media works and how information should be prepared in order to be used. Activities they perform are targeted at informing people. There is also no clear plan concerning establishment of public relations in the organization.
3. asymmetric PR, when an organization uses research methods to uncover and understand human motivation and uses this knowledge in its PR activities. Activities performed are carried out according to a previously defined plan.
4. symmetric PR aiming at assuring communication in order to establish mutual understanding between an organization and its surroundings (Wojcik, 2005).

Public relations is often compared to marketing and treated as an add-in to marketing activities (Kotler, 1999). (Sznajder, 1993) diminishes the role of public relations and describes it as one (along with selling, complementary promotion and sponsoring) approach to company promotion. He also states that PR activities aim at dissemination of journalistic information about company's activities.

Public relations activities differ because of the diversity of application domains, the number of goals, etc. One of the diversity reasons is related to historical issues. Ryszard Ławniczak defines the notion of transitional PR. This notion underlines the need for different approach to public relations (techniques and methods used) in former communist countries undergoing economical transitions (i.e. in Central and Eastern Europe) (Ławniczak, 2003b, Ławniczak, 2003a, Ławniczak and Kaczmarek-Śliwińska, 2004, Kaczmarek-Śliwińska, 2005b). In these countries PR should deal not only with activities focused on building images of companies. While performing "typical" PR activities, activities targeted at a mental change of people being introduced to a market economy and having different attitude toward promotion should be held.

To summarize, public relations deal with activities that promote a company. However, less "aggressively" than traditionally understood advertisements. PR activities concern data gathering and processing, consulting, creation of communication rules with diverse communities (media, scientific institutes, cultural institutions, government, etc.), planning in

case of crisis situations, using various means to create company image that is reliable, friendly and attractive for customers (McKeone, 1995).

However, in all types of PR activities PR analyst needs precise information. One of examples of such activities carried out by a PR expert is a PR process.

### 1.3.1. Public relation process – maintenance of a company image

According to (Wojcik, 2001) public relations activities deal with "implementation of plans targeted at fulfilling needs (interests) of both: companies and public communities". Activities to be performed when implementing such a plan are called PR processes or PR campaigns. They concern development of description of a company's or country's image and undertaking actions that may correct or maintain it.

In practice not every organization commits effort for establishing and maintaining contacts with its environment and therefore, PR programs differ from each other. According to (Wojcik, 2001) there are two groups of organizations that may be distinguished, taking into account their approach to PR activities, namely active and defensive. Defensive strategy deals only with defending against attacks targeted at organizations and solving crisis and conflict situations. In turn, the active strategy predicts emerging conflicts, tries to minimize threats for company image and conducts research on perception of company's image within groups from company's surroundings. Although different, these two approaches can coexist in a company depending on the target group defined for the PR activities.

Besides the communication strategy, companies run continuous monitoring of their environment in order to introduce changes to communication strategies or PR programs (Austin and Pinkleton, 2006). Monitoring is defined as „a constant, systematic, professional observation of a company environment. It deals with searching and processing information about changes in the environment, learning trends (and so called weak signals) and reporting the monitoring results to management. The information obtained is needed for building corporate strategy and amending PR plans or communication strategies (Wojcik, 2005)." A traditional way of carrying out the monitoring activities involves regular browsing of news releases mentioning an organization or its competition, in domain, regional and multiregional magazines, newspapers, TV and radio programs, observing the statements of supporters, opponents, competitors, or representatives of the research institutes. However, in the Internet era most of this information can be found more efficiently. According to (Szyfter, 2005) monitoring deals mostly with regular browsing of the Web with a focus on news portals, e-magazines and e-newspapers (or electronic versions of "traditional" ones), corporate portals

and discussion forums. (Seitel, 2003) highlights also the need for monitoring websites that may negatively affect the company's image, that are maintained by a company's competitors or unsatisfied customers. Recently also blogs are highlighted as an important information source (Kaczmarek-Śliwińska, 2006, Wright and Hinson, 2008, Weber, 2007).

Data acquired during the monitoring process is used to control company's image and establish the starting point for the PR campaign (process). To summarize, the aim of the PR campaign is to develop authentic and accurate, at the same time positive, company's image via the continuously performed PR activities (Adamus-Matuszyńska, 1999). In order to achieve this aim the PR process can divided into a number of steps (Nitsch, 1975, Wojcik, 2001):

1. Evaluation of Entry Situation.
2. Analysis and Interpretation.
3. Establishing Goals.
4. Planning.
5. Communication.
6. Measuring Efficiency.

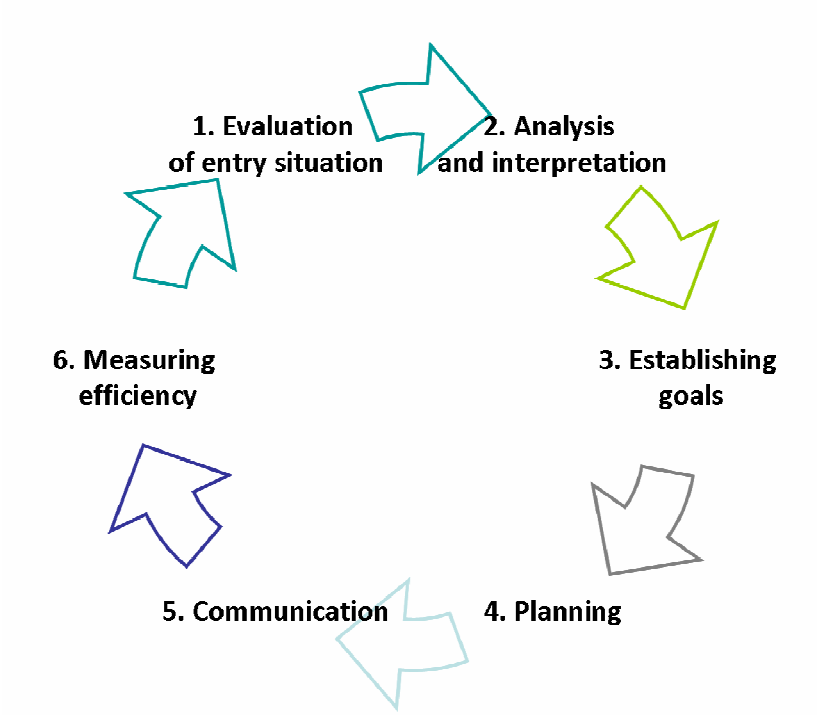



**Figure 3. Public relations process. Source: (Nitsch, 1975, Wojcik, 2001)**

Similar approach is presented by (Parsons, 2004), who defines four steps of public relations process:

1. Gathering and analysis of available data (research).

2. Establishment of plan for the PR process.
3. Implementation of the plan.
4. Measuring efficiency of the PR process.

The analysis of stages of the public relations process leads to an important observation – stages 1., 2. and 6. according to (Nitsch, 1975, Wojcik, 2001) or stages 1. and 2. according to (Parsons, 2004) are strongly related to collecting, processing and analysing information from different sources. These phases are often referred in the literature as research phases (Stacks 2002): „Research is essential to any public relations activity or campaign. As evidenced in many public relations models, research is the beginning of a process that seeks to bring about a specific objective. Hendrix's ROPE (Research, Objectives, Program, Evaluation), Marston's RACE (Research, Action, Communication, Evaluation), and Cultip, Centerm and Broom's four-step process (Defining PR Problems, Planning and Programming, Taking Action and Communicating, Evaluating the Program) models posit that any serious public relations activity must begin with a research step."

Research phases focus mainly on acquiring information that is to be used within the subsequent phases. They are described in detail in the section below.

**Evaluation of entry situation**

This stage of the PR process starts when a decision on the establishment of a PR campaign is taken. It focuses on the current corporate image and in the literature is referred to as an "information audit" focusing on building the current company image and confronting it with the reality (Adamus-Matuszyńska, 1999, Wojcik, 2001, Seitel, 2006, Smith, 2004).

This phase starts with collecting information about the organization (its internal and external surroundings). The efficient use of data (Wojcik, 2001) gathered at the stage of evaluation of entry situation in the subsequent steps of the PR campaign depends on its accuracy, timeliness and certainty.

There is a variety of information sources that can be utilized in PR campaigns in order to establish a common view on the organization and its surroundings and create a profile of a company. The following information sources can be distinguished as the most important ones (Wojcik, 2001, Smith, 2004):

- Shareholders,
- Minutes from executives' meetings,
- Data from yearly budgets, statistics, literature,

- Correspondence from members of an organization,
- Correspondence and phone calls with other organizations,
- Press releases / newspaper articles containing information about the organization and related events,
- Statements of advocates, opponents, competitors, representatives of research institutes, etc.
- Regional and multiregional newspapers, journals, brochures, radio and TV programs.

Reports (analyses) currently available in the company or received from external parties can be also included in the list presented above (Treadwell and Treadwell, 2005). From the point of view of information technologies and in the context of this dissertation the most important ones are those that are available in the digital format i.e. meeting minutes, data from yearly budget, statistics, literature, press releases, regional and multiregional newspapers, journals, and brochures.

Recently, PR practitioners started adopting changes in communication channels and models, caused by development of the Internet. The Web 2.0 mechanisms for collaboration, Wikis, blogs impose a need to analyse not only what newspapers write about the company, but also what people think (Wright and Hinson, 2008, Reich and Solomon, 2008). There are, however differences between traditional Internet sources and Web 2.0. These differences concern inter alia (Abramowicz et al., 2007a):

- dynamics – unparalleled ratio of changes, together with decreasing time span between the event and revealing the information about it, make Web documents a perfect source for tracking reactions. However, due to high dynamics, one must be careful not to miss the initial reactions to the content published. It is hard to monitor all dynamic sources, though there are tools that make it a less tedious task e.g. RSS feeds;
- reader's involvement – the introduction of forums or blogs leads to unprecedented reader's involvement in the creation of content. Currently, there exist systems that support content creation by visiting Internet users (most successful story of such involvement is Wikipedia; there were also experiments with online editing conducted by major press portals e.g. http://www.gazeta.pl). This enables capturing not only "official" messages prepared by trained reporters, but also spontaneous reactions of eye witnesses – a measure unavailable to such extent before;
- networking effect – it is currently possible, thanks to the search engines, to find "similar" documents that cover the same information, but possibly with a different

information flavour. This broadens the spectrum of views that can be analysed in order to undertake appropriate PR actions.

- data coverage – lots of data that nowadays just awaits to be gathered, was very hard to obtain in the past. Soon there will be means to monitor online the shift of prices of competitive products or track customer complaints and therefore shorten the reaction time.

From the public relations' perspective the most important resources available in the Internet are websites of the traditional media, websites of opinion-building communities, and consumer organisations (Wright and Hinson, 2008, Kaczmarek-Śliwińska, 2006, Reich and Solomon, 2008). The significance of blogs, forums, discussion lists and groups have also increased over the last years (KnowledgeStorm, 2006, Wright and Hinson, 2008). (Wright and Hinson, 2008) report on increasing importance of blogs and other social media for PR practitioners based on their study among 328 PR practitioners coming from North America (57%), Europe (20%), Asia (10%), Australia (10%), Africa (9%) and South America (1%). 61% of respondents of this study believe that the emergence of blogs and social media changed the way their organizations communicate with surrounding.

According to the outcomes of the research published in (Kaczmarek-Śliwińska, 2006) (see Table 1) regarding the Polish PR market, the Internet sources most frequently monitored by Polish PR agencies are:

- portals, vortals, information portals, thematic pages,
- discussion list and groups,
- Internet portals of traditional newspapers and magazines.

**Table 1. Importance of different information sources in monitoring of the Internet. Source: (Kaczmarek-Śliwińska, 2006)**

| Statement | very important | important | neutral | small importance | not important |
|---|---|---|---|---|---|
| Monitoring of portals, vortals, information portals, thematic pages | 66,7% | 33,3% | 0% | 0% | 0% |
| Monitoring of information published at portals of the newspapers and magazines (online media) | 41,7% | 50% | 0% | 8,3% | 0% |

| | | | | | |
|---|---|---|---|---|---|
| Monitoring of information published at radios' websites | 0% | 83,3% | 0% | 16,7% | 0% |
| Monitoring of information published on TV portals | 0% | 75% | 8,3% | 16,7% | 0% |
| Discussion lists monitoring | 41,7% | 25% | 8,3% | 25% | 0% |
| Discussion groups monitoring | 50% | 16,7% | 8,3% | 25% | 0% |
| Chats' monitoring | 8,3% | 33,3% | 16,7% | 33,3% | 8,3% |
| Forums monitoring | 33,3% | 16,7% | 8,3% | 33,3% | 8,3% |
| Blogs monitoring | 16,7% | 25% | 16,7% | 41,7% | 0% |
| Monitoring of resources indexed by search engines taking into account textual data | 16,7% | 41,7% | 25% | 8,3% | 8,3% |
| Monitoring of resources indexed by search engines taking into account graphics | 0% | 16,7% | 58,3% | 8,3% | 16,7% |

**Analysis and interpretation**

The second step of the PR process focuses on analysis and interpretation of data collected in its first stage. This phase aims at creation of a company image. Then, an analysis and definition of reasons of differences between desired and existing company image follows. All these activities establish a foundation for the third step of the PR process.

The next steps cover definition of goals of the PR process, planning of activities with regard to budget and methods of implementation, carrying out tasks according to a schedule and communicating in order to ensure proper accomplishment of the plan. If needed, some amendments to the plan are introduced. However, these activities, establishing the core of the PR campaign, are less interesting in terms of the IT support as most of them are performed by PR analysts and therefore are out of scope of this dissertation.

**Measuring efficiency**

The last stage of the PR campaign deals with measuring efficiency of activities that were undertaken to change the company image. According to (Wojcik, 2001) the subjects to control are: PR process concept, methods of implementation including establishing of efficient communication with media, problems and their sources, outcomes of the PR process (efficiency, accuracy, budget spent). The efficiency analysis is based on data and information

acquired from internal and external surroundings of an organization (like in the 1st stage of the process). There is a number of methods of control of the PR process: observation, analysis of information from different departments dealing with organization-external surroundings, analysis of correspondence with business partners and free telephone lines, correspondence checking the effects of the activities performed, questionnaires, interviews with target groups, interviews with experts, DAGMAR[7] method, method of Walter K. Lindenmann, press clipping, letters to editorial office (comments on web pages), press releases and measuring the existence of the information in media (Wojcik, 2005). The most important issue connected with the last point is a character and a length of information, its perception, and place of its publication. The importance of a press release and its influence on the company image depends on the quality of the source, page number and placement of information on a page.

Regarding Internet sources, interesting approach to the estimation of efficiency of the PR process was proposed by Walter Lindenmann (Kaczmarek-Śliwińska, 2005a, Lindenmann, 2003), who defined three methods for efficiency estimation:

1. **Measuring OUTPUTS.** This method deals with measuring of media resonance through the estimation of the number and the evaluation of the importance of press releases, press conferences, events, materials provided to media, etc. The typical metrics that can be distinguished are e.g.:

- number of website hits,
- SoV – the ratio of company's banners compared to all banners that appear on the website,
- CPT – cost of PR campaign compared to the overall number of recipients of the message,
- CTR (Click Through Ratio) – the number of clicks on the promotional element compared to the number of its displays.

[7] Defining Advertising Goals for Measured Advertising Results – a method described by Cooley, which deals with measuring of the advertising effects in comparison to aims defined – getting people acquainted with the subject of the advertisement, its utility and benefits, and making people act according to advertisement's subject (COLLEY, R. H. (1961). *Defining Advertising Goals for Measured Advertising Results,* New York, Association of National Advertisers, Inc.).

Most of PR agencies present in their offer not only types of advertisements offered, but also some ratios regarding sources where the advertisements are published. Such comparison of sources depending on their value for the PR process is presented in the table below, however, a similar comparison may be also found at e.g. http://www.arbomedia.pl/.

**Table 2. Metrics for Internet Business Sources. Source: http:// www.bizonmedia.pl**

| | | 1 month | | 1 week | | |
|---|---|---|---|---|---|---|
| | | users (real users) = UU / month | displays = PV /month | users (real users) = UU / week | displays = PV / week | Source of data |
| 1 | ABC.com.pl | 75 458 | 270 007 | 23 495 | 64 233 | BBelements 2009.05 |
| 2 | Banki.pl | 7 792 | 11 492 | 1 823 | 2 609 | BBelements 2009.05 |
| 3 | Bankier.pl | 358 364 | 2 734 283 | 121 599 | 642 155 | BBelements 2009.05 |
| 4 | Biznes-firma.pl | 4 937 | 8 078 | 1 086 | 1 778 | BBelements 2009.05 |
| 5 | BiznesNet.pl | 2 085 | 5 148 | 414 | 981 | BBelements 2009.05 |
| 6 | BiznesPrawo.pl | 909 | 2 196 | 208 | 501 | BBelements 2009.05 |
| 7 | BizPoland.pl | 683 | 2 088 | 182 | 530 | BBelements 2009.05 |
| 8 | biznespolska.pl | | | | | |
| 9 | Comperia.pl | 15 422 | 40 868 | 4 593 | 9 490 | BBelements 2009.05 |
| 10 | e-podatnik.pl | 1 370 | 4 213 | 358 | 1 028 | BBelements 2009.05 |
| 11 | ePorady24.pl | 2 820 | 5 401 | 730 | 1 311 | BBelements 2009.05 |
| 12 | e-prawnik.pl | 37 599 | 102 993 | 9 642 | 25 235 | BBelements 2009.05 |
| 13 | eGospodarka.pl | 41 268 | 105 495 | 13 164 | 25 064 | BBelements 2009.05 |
| 14 | epr.pl | 47 055 | 109 517 | 12 200 | 25 366 | BBelements 2009.05 |
| 15 | eurobankier.pl | 13 831 | 222 414 | 5 179 | 51 324 | BBelements 2009.05 |
| 16 | FinanseOsobiste.pl | 2 271 | 4 676 | 718 | 1 282 | BBelements 2009.05 |
| 17 | Forsal.pl | 73 645 | 386 341 | 20 431 | 88 199 | BBelements 2009.05 |
| 18 | franchising.pl | 22 913 | 66 321 | 6 355 | 15 547 | BBelements 2009.05 |
| 19 | GazetaPodatkowa.pl | 29 779 | 134 301 | 9 426 | 33 035 | BBelements 2009.05 |
| 20 | GazetaPodatnika.pl | 1 281 | 5 013 | 336 | 1 160 | BBelements 2009.05 |
| 21 | GazetaPrawna.pl | 172 799 | 903 751 | 57 116 | 218 384 | BBelements 2009.05 |
| 22 | gieldowe.pl | | | | | |
| 23 | gofin.pl | 102 141 | 545 233 | 41 255 | 134 126 | BBelements 2009.05 |
| 24 | interaktywnie.com | | | | | |
| 25 | GoldenLine.pl | 793 918 | 9 974 054 | 220 636 | 2 313 172 | Google Analytics 2009.02-03 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 26 | ITBiznes.pl | 2 750 | 7 846 | 1 166 | 2 834 | BBelements 2009.05 |
| 27 | Karieramanagera.pl | 1 825 | 6 716 | 644 | 1 911 | BBelements 2009.05 |
| 28 | lex.com.pl | 70 468 | 204 827 | 20 227 | 48 503 | BBelements 2009.05 |
| 29 | Manager.Money.pl | 104 111 | 142 408 | 32 491 | 43 047 | BBelements 2009.05 |
| 30 | marketingprzykawie.pl | 6 062 | 16 016 | 1 779 | 3 915 | BBelements 2009.05 |
| 31 | media2.pl | 65 348 | 535 503 | 20 353 | 119 877 | BBelements 2009.05 |
| 32 | mediafm.net | 3 621 | 12 440 | 1 078 | 2 864 | BBelements 2009.05 |
| 33 | medialine.pl | 3 361 | 16 861 | 928 | 3 863 | BBelements 2009.05 |
| 34 | mediamikser.pl | | | | | |
| 35 | medialink.pl | 6 614 | 19 301 | 3 185 | 6 371 | BBelements 2009.05 |
| 36 | MediaRun.pl | 17 414 | 58 102 | 5 538 | 13 497 | BBelements 2009.05 |
| 37 | MojePrawo.pl | | | | | |
| 38 | Msp.Money.pl | 123 434 | 197 731 | 39 191 | 60 939 | BBelements 2009.05 |
| 39 | Mybank.pl | 5 410 | 17 679 | 1 679 | 4 151 | BBelements 2009.05 |
| 40 | NoweBiuro.pl | 12 941 | 36 604 | 3 952 | 11 625 | BBelements 2009.05 |
| 41 | NowoczesnaFirma.pl | 78 786 | 238 211 | 21 373 | 55 917 | BBelements 2009.05 |
| 42 | parkiet.com | 182 844 | 1 999 773 | 63 453 | 472 285 | BBelements 2009.05 |
| 43 | pit.pl | 40 964 | 97 938 | 9 379 | 18 823 | BBelements 2009.05 |
| 44 | podatki.pl | 31 492 | 82 802 | 8 639 | 19 709 | BBelements 2009.05 |
| 45 | poradaprawna.pl | 3 378 | 7 719 | 855 | 1 825 | BBelements 2009.05 |
| 46 | prawo-pracy.pl | 8 284 | 16 874 | 2 216 | 4 231 | BBelements 2009.05 |
| 47 | proto.pl | 14 740 | 63 765 | 5 024 | 14 522 | BBelements 2009.05 |
| 48 | PRnews.pl | 9 630 | 46 506 | 3 185 | 10 623 | BBelements 2009.05 |
| 49 | przepisnabiznes.pl | 4 574 | 8 278 | 1 118 | 1 982 | BBelements 2009.05 |
| 50 | PSZ.pl | 11 064 | 58 842 | 3 327 | 13 500 | BBelements 2009.05 |
| 51 | rzeczpospolita.pl | 752 970 | 8 344 624 | 245 675 | 1 981 690 | BBelements 2009.05 |
| 52 | salon24.pl | 103 167 | 1 789 408 | 33 539 | 473 856 | BBelements 2009.05 |
| 53 | skarbiec.biz | 33 958 | 42 153 | 8 322 | 10 398 | BBelements 2009.05 |
| 54 | sparing.pl | 566 | 1 471 | 154 | 395 | BBelements 2009.05 |
| 55 | spedycje.pl | 7 089 | 50 834 | 1 768 | 11 562 | BBelements 2009.05 |
| 56 | stooq.pl | 257 150 | 7 693 762 | 89 799 | 1 666 419 | BBelements 2009.05 |
| 57 | szkolenia.com | | | | | |
| 58 | Twoja-Firma.pl | 6 484 | 16 351 | 1 613 | 3 842 | BBelements 2009.05 |
| 59 | vat.pl | 16 181 | 31 648 | 4 123 | 7 196 | BBelements 2009.05 |
| 60 | warsawvoice.pl | | | | | |
| 61 | wirtualnemedia.pl | 74 811 | 339 388 | 25 964 | 80 500 | BBelements 2009.05 |
| 62 | zyciewarszawy.pl | 108 120 | 744 298 | 37 531 | 230 369 | BBelements 2009.05 |

The important issue to notice is that statistics concern mainly a number of visitors. They do not include an analysis of distribution of locations of website visitors or links the users were redirected from. The importance of a source is estimated mainly based on the number of website hits.

2. **Measuring OUTTAKES** – this approach may be applied only for PR programs that have been running for some time. This approach takes into account the scope and accuracy of understanding of messages being disseminated to the public.

3. **Measuring OUTCOMES** – this method takes into account changes in target groups after the PR process was carried out. The evaluation of outcomes is closely related to measuring the influence of activities undertaken on the increase in importance of a company on the market.

Similar evaluation models were also proposed by (Cutlip et al., 2005, Macnamara, 1999). Main difference between them concerns the fact that Lindenmann doesn't evaluate PR strategy, choice of PR channels and clarity of messages being disseminated. These models, however, are quite complicated. Therefore, companies define also their own success metrics to evaluate efficiency of the PR process.

### 1.3.2. Company image

The aim of an evaluation of entry situation as well as of the control phase involves acquiring data and information regarding relations between organization and its environment, image of the organization, and previous communication activities (Wojcik, 2001). Worth emphasizing is the notion of a company image. According to (Boorstin, 1963) a company image is a pseudo ideal desired by companies (especially in the age of Internet economy (Oliver, 2005)), that should be artificially created, compact, reliable, vital and enigmatic. (Cenker, 2000) and (Budzynski, 2002) define a company image as a perception of an organization by people from its environment – direct (e.g. as customers) or indirect (e.g. as market participants). Also in (Wojcik, 2005) one may find an analogy to the previously quoted definition of the company image brought in after (Grunig, 2001), according to which an image is the picture that one or more communities have about the person, organization or institution.

To summarize, an image concerns perception of organization possessed by different parties and therefore one can distinguish the following types of images (Budzynski, 2002):

- real image (foreign) – an image of the company possessed by parties that contact the organization,
- mirror image (own) – an image of the company possessed by its employees,

- desired image – a target image that company would like to achieve – the way it should be perceived by its environment,
- optimum image – compromise between the previous images that is achievable in current conditions.

Image being something existing in people's minds doesn't have any formal representation. However, people build it based on information read in newspapers, heard from other people, etc. Image may be formalized in a form of a company profile.

## *1.4. Company profile*

A company profile is usually defined as a set of information describing the company, helping people, organizations or institutions to build the perception (image) of the company (Abramowicz, 2008). However, in the PR domain a profile is usually understood differently. (McKeone, 1995) defines a profile similarly to an image saying that it is the way of perception of the organization and foundation for creation of brand and product reputation. According to the PR literature, notion of a profile is disjunctive from the notion of a company image. A profile of the company is defined as a technique supporting measurement and evaluation of the company image that deals with ranges usually containing seven levels of adjectives describing opposing states (e.g. small/big, wide/narrow) used to express the perception of an organization (Altkorn, 2004). An example of such a profile, called semantic profile, is presented in the Figure 4.

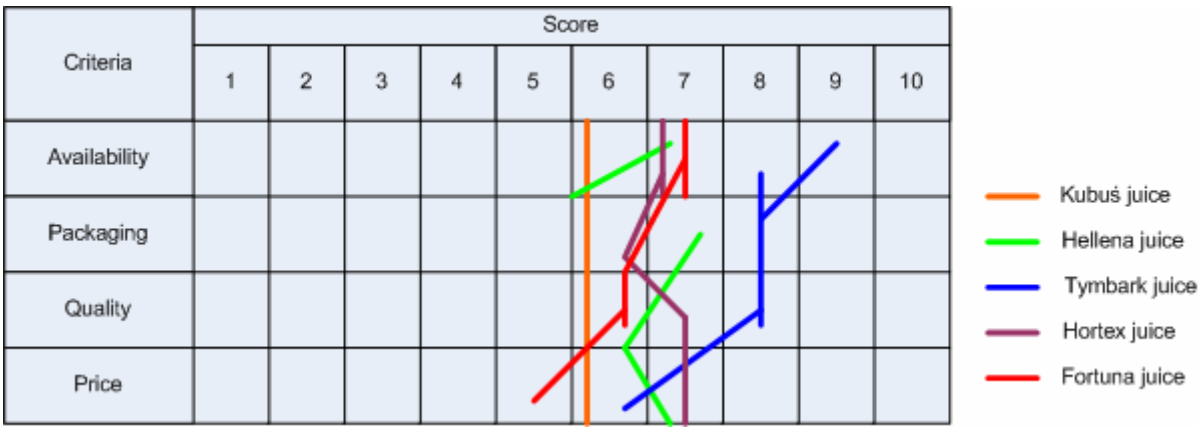



**Figure 4. Semantic profile. Source: (Altkorn, 2004)**

On the other hand, in information systems domain a profile is defined as a method of expressing the information needs of a certain entity (Belkin and Croft, 1992, Baeza-Yates and Ribeiro-Neto, 1999, Abramowicz, 2008). Information needs concern information, thanks to

which one can perform work or research, as it is recognized by its receiver (Line, 1969, Mizzaro, 1998).

Approach taken by (Grunig et al., 1992) connects these two perspectives on the organization profile. Grunig states that a profile depicts internal state and capabilities of a company (for specified categories, so similarly to the semantic profile), at the same time enabling interactive analysis aiming at creation of links between a company profile and the company-external world (so describing categories using data and information on a certain enterprise).

A similar approach to a definition of profile one may find in the field of information extraction. According to (Nado and Huffman, 1997) a profile (of certain object) is a list of features that may be found about a certain object in the collection of documents that are presented as search outcomes in the web search engine. An exemplary profile as defined by (Nado and Huffman, 1997) is presented in the Figure 5:

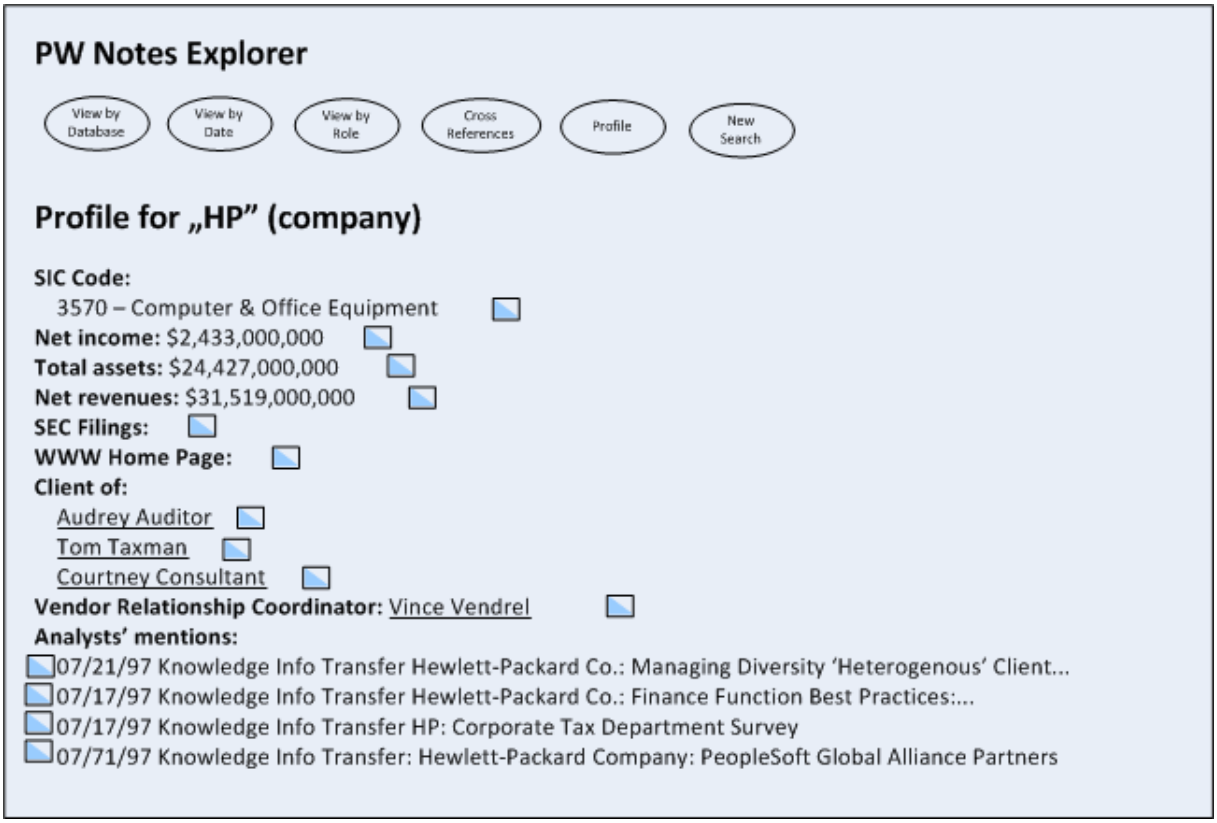



**Figure 5. Company Profile. Source: (Nado and Huffman, 1997)**

Enhancing this definition, (Srihari et al., 2003) define profiles as information-rich objects describing an entity as well as related entities and events. Such profiles can be built based on information acquired from different documents utilizing information extraction techniques (e.g. solving coreferences).

### 1.4.1. Methods of supplying a company profile with information

Nowadays, PR companies prepare reports on the existing company image mainly manually. Automating this process is not a trivial task. A PR Analyst would like to receive 100% complete and accurate data.

In IT, especially in the field of information retrieval, completeness and accuracy are measured using precision and recall (Węcel, 2002, Rijsbergen, 1979, Baeza-Yates and Ribeiro-Neto, 1999). Applying these measures to the field of public relation is not unintended. If monitoring concerns *„constant, systematic, professional observation of company environment, searching for and processing of information about its changes, learning development tendencies (and so called weak signals) and reporting them to company's management, as information needed for building corporate strategy and amending PR plans and/or communication strategies in response to the changes in the environment*" (Wojcik, 2005), then one can here notice features of information filtering and benefit from its techniques and evaluation methods in PR. According to (Loeb and Terry, 1992) information filtering deals with monitoring of a stream of documents taking into account user information needs (described in a profile) in order to acquire relevant information (in this case for creating the corporate strategy and aligning it to changes in expectations and needs of the organization surroundings). After receiving such information, a PR analyst can undertake specific actions in order to meet targets defined for the PR process.

Among the technologies that can be used to support a PR expert, besides information filtering and retrieval are also methods for content and information aggregation and integration (Kaczmarek, 2006, Kowalkiewicz et al., 2006b, Abramowicz et al., 2007b). Sections below present a short overview of these techniques and methods.

**Information filtering and retrieval**

As it was already mentioned, information filtering (IF) is a constant process that enables acquiring documents that match user preferences. In opposite to it, information retrieval (IR) is about finding documents that are relevant to a user query in a collection of documents. Two main differences between filtering and retrieval are as follows:

- IR is performed on a fixed collection of documents, while IF techniques are able to process stream of (new) documents.
- In IR user needs are constantly changing, whereas in IF they are relatively stable.

However, there is one common idea in both processes. Both approaches need properly described documents as well as user needs. These descriptions are called indexes. Document indexing is providing each of documents with a machine readable representation of document's content (Baeza-Yates and Ribeiro-Neto, 1999). However, dealing with different types of content in the document (time, geography, proper names) actually means applying different methods in order to create document indexes.

Context is defined as an *"environmental information that is a part of an application's operating environment and that can be sensed by the application*[8]*"* or as information characterizing situation of a person, place or object, that cooperates with a user or application (Dey and Abowd, 2000). There are several types of contexts described in the literature: computing context, physical context and user context (Schilit, 1995), human factors and physical environment contexts (Schmidt et al., 1999). Based on these ideas the following types of context were differentiated in our research (Abramowicz, 2008):

- cognitive context,
- time context,
- geographical context,
- proper names context.

Applying IR&F techniques for PR leads to development of search techniques and interfaces fulfilling needs of the public relation analysts.

**Information integration tools**

Information filtering and retrieval approaches perform the best when applied to unstructured data, however, the Internet is also about data stored in the databases i.e. deep Web (He et al., 2004, Kabra et al., 2005). The main problem with accessing this kind of data is that the sources are dispersed, they are not directly searchable and the process of obtaining data may be tedious. Therefore, these sources should be tackled using information integration technologies.

A typical data source on the Web exposes some forms to ask queries. These forms are very often of limited functionality. Moreover, when we get the results, they come in the form of a Web page, what is not very convenient for manipulating data or comparing it between different sources. The purpose of information integration tools is to overcome these issues as well as to ease the processes of obtaining and manipulating the data. The aim is therefore to

[8] http://www.cs.cmu.edu/~anind/context.html

provide a single interface for posing queries (or even allow to automatically query) and gather data in unified, ready for reuse form.

In order to achieve this aim, information integration techniques have several steps that need to be performed before data sources can be used (Kaczmarek, 2006):

- source description – each of the sources to be accessed needs to be introduced to the system, showing its structure, access means and special features;
- integration – after all sources are introduced, they have to be integrated to be seen under single view – this step allows later on to ask a single query that will be forwarded to all the sources without the need to query each of them separately;
- query translation – this activity occurs each time the query is posted – it has to be automatically translated for each source;
- extraction – data is gathered from Web pages delivered by the source. It is crucial for this step that it is exact and robust – no garbage is gathered and all what is required is extracted;
- consolidation of results – the final step includes merging results from all the sources and presenting them to the user.

With the advent of such tools it is possible to monitor any Web enabled data source, track changes or periodically obtain relevant information necessary to guide PR activities (Abramowicz et al., 2007b, Flejter and Kaczmarek, 2007).

**Information aggregation tools**

Other tools that may support a PR expert are information aggregation tools. In data and information integration systems it is usual to integrate the extraction results. This includes identifying data that refer to the same real world entities, normalizing it, equalizing the measures, formatting etc. and removing redundancies. However, this shouldn't be done when aggregating content. The data integration tries to find corresponding entity descriptions and remove duplicate information, as opposite content aggregation only brings the information together. Content is subjective in nature and therefore it may be important to retain this subjectivity and preserve the meta-information about the source. Moreover, it is essential for the overall assessment of the aggregated content if human users are able to read the unchanged excerpts of news or financial information about different enterprises.

Therefore, information aggregation focuses on presentation of possibly large amounts of information. It is sensible to present the information obtained from Web sources as a Web page (or portal), but other technologies can be also applied (like RSS aggregators).

Information aggregation is an area that has not been very well investigated so far. One and the most significant of the few scientific publications is the research of (Stonebraker and Hellerstein, 2001). They define the term content aggregation to refer to the integration of operational information across enterprises. The authors understand content as semi-structured and unstructured information. Stonebraker and Hellerstein describe a Cohera Content Integration System, helpful in content integration, which consists of the following components: Cohera Connect, providing browsing and Document Object Model (DOM) based wrapper generation functionality, Cohera Workbench for content syndication, and Cohera Integrate for distributed query processing. Another example of such information integration tools is myPortal (Kowalkiewicz et al., 2006b) that provides user with the possibility to define a portal aligned to his needs.

## 1.5. Conclusions

The Internet is one of the most important information sources for the public relations domain (Seitel, 2003). However, the expansion of the Internet[9] is also followed by a growing amount of information that is to be managed by an organization (Oliver, 2005). Dealing with the information overload necessitates the use of diverse tools to address different sources of information, especially that the bigger the number of sources utilized, the better (Berghel, 1997, Oliver, 2005).

PR analysts are not always aware of changes in communication channels and models caused by an increasing role of the Internet. Differences between traditional documents and electronic ones should lead to different approaches when dealing with these documents within the PR campaign. PR agencies are mainly interested in automated clip tracking, extranets and automation of the PR campaign, often forgetting about data and text mining, content syndication, improving efficiency of search engines, etc. (Holtz, 2002).

Nevertheless, PR analysts see the need of introducing tools that will help them in the process of acquiring and initial processing of data and information (Oliver, 2005). However, tools offered are still suffering from limitations, e.g. regarding efficiency (Nado and Huffman, 1997, Srihari et al., 2003). Furthermore, there exist a strong need of alignment of PR process

[9] http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/

to the needs of specific communities (Oliver, 2005), e.g. regarding the geographical distribution.

These reasons provide motivation for developing a method of geographical indexing, such as described and validated within next chapters of this dissertation.

# Chapter 2.
# Technologies for spatial indexing

*Any sufficiently advanced technology*
*is indistinguishable from magic.*

*Arthur C. Clarke*

## 2.1. Introduction

This chapter provides an overview of information extraction (IE) and information retrieval (IR) domains. It provides foundations for methods that are introduced in the next chapter as well as for spatial indexing mechanisms being the main contribution of this dissertation.

(Turmo et al., 2006) noticed that the most of texts produced for humans consist of unrestricted natural language and lack explicit structure. Extracting their contents e.g. for the needs of further computer processing involves detailed linguistic knowledge. Information extraction provides techniques and methods for extraction of desired content from such free-text documents (Appelt and Israel, 1999, Strzalkowski, 1999).

Information retrieval techniques deal with selecting documents satisfying restrictions of a query provided by a user (Baeza-Yates and Ribeiro-Neto, 1999). This selection is usually based on matching the query against already available document indexes (Abramowicz, 2008). The role of information extraction techniques is considered to be marginal in IR, as the latter usually delivers its results based on the set of keywords (Turmo et al., 2006). However, content extracted from text may be used in IR to provide extensive indexes of documents what may lead to obtaining more accurate responses (Appelt and Israel, 1999, Turmo et al., 2006). Such an approach is also followed in this dissertation.

This chapter is structured as follows. Firstly, it discusses approaches used in information extraction. Then, the spatial information retrieval is elaborated highlighting the role of ontologies both in the processes of document indexing as well as query matching. Finally, some concluding remarks for the spatial indexing model, proposed in this dissertation, are made.

## 2.2. Information extraction techniques

The aim of the information extraction (IE) is to pull out all pieces of information from an unstructured text. These pieces are linked to a predefined set of related concepts (template)

defined as an extraction scenario (Peshkin and Pfeffer, 2003, Appelt and Israel, 1999, Turmo et al., 2006). Obtained information is then persistently stored, e.g. in a database, for further processing.

(Appelt and Israel, 1999, Dimitrov, 2002, Turmo et al., 2006) identify the following characteristics of an IE process:

- addressing only unstructured sources (e.g. news articles, legal documents),
- identification of information that is important only for the previously specified problem and a defined template,
- extraction of information to a predefined format (applying structure to an unstructured content),
- domain-dependence – most of the IE systems are designed for a specified domain in order to achieve high precision and recall measures.

Basically, there are two approaches used in IE, namely knowledge-engineering and machine-learning (Kaiser and Miksch, 2005, Appelt and Israel, 1999).

Machine-learning approaches use the manually annotated sample texts to automatically discover knowledge about a domain that is used when extracting new information from documents (Frantzi et al., 2000). Therefore, a crucial point in the machine-learning approach is the development (annotation) of corpus of domain-relevant texts. It's also important to note, that the more documents in the corpus, the better performance of IE tools with regard to accuracy measures. (Appelt and Israel, 1999) report that for a corpus of 30.000 words F-measure[10] was 81%, whereas for 1,2 mln words – it was estimated for 91%. This approach to information extraction is further addressed by the following work done in our research group (Abramowicz and Wiśniewski, 2008, Wiśniewski, 2009).

In contrast, knowledge-engineering approach involves domain experts familiar with requirements and specifics of the domain of discourse. They are responsible for development of rules used for the extraction of the desired information. In this case skills of knowledge engineers play a major role in the level of performance that is achieved by the IE system (Appelt and Israel, 1999).

The development of the IE systems was sped up by the launch of the Message Understanding Conference Series[11] (MUC) taking place from 1987 until 1998. Their goal was to evaluate IE systems developed by different research groups by means of comparing their performance

[10] One of the measures of quality of information extraction, explained later in this section.

[11] http://www.itl.nist.gov/iad/894.02/related_projects/muc/index.html

against free text documents restricted to a given domain. Each year different domain was selected for a conference and all participants were provided with an extraction scenario as well as a set of training documents. The evaluations of reports provided by participants were carried by Beth Sundheim from the US Naval Ocean Systems Center (NOSC) and continued with DARPA funding under the TIPSTER Program for support by Dr. Nancy Chinchor of Science Applications International Corporation (SAIC).

MUC, being the first initiative in the field of information extraction, actually spawned other initiatives such as e.g. ACE[12], LRE Programme, TIDES funded by DARPA, European Commission or national governments. Additionally, MUC provided a corpus of documents for development of the machine learning approaches.

Moreover, for the needs of the MUC evaluation, the IR measures of precision (P) and recall (R) were redefined:

**Precision** $$P = \frac{COR}{COR + INC + SPUR}$$

**Recall** $$R = \frac{COR}{COR + INC + MISS}$$

where *COR* stands for correctly extracted slots, *INC* for incorrectly extracted slots, *SPUR* for spuriously extracted slots, and *MISS* for missing slots.

There is a well-known trade-off between precision and recall. One may improve the precision by sacrificing the recall, and vice versa. Therefore, a combined measure was proposed to alleviate this phenomenon, namely F-measure defined as a harmonic mean of precision and recall.

**F-Measure** $$F = \frac{2 \times P \times R}{P + R}$$

Currently in ACE[13] different evaluation methods are applied in order to show not only precision of identification of single named entities in text, but also precision of discovering relations between them. These measures, moreover, may be adapted to various evaluation needs (Maynard et al., 2003). (Lavelli et al., 2004) provide discussion on a number of approaches to evaluation of different IE methods.

[12] http://projects.ldc.upenn.edu/ace/

[13] http://www.nist.gov/speech/tests/ace/index.htm

Though detailed techniques for evaluation and scoring are out of the scope of this dissertation, it is worth to mention the interannotator agreement. It proves that the task of annotating or extracting information from text is difficult even for humans. For various aspects of IE tasks, the interannotator agreement is usually in the range of 60-80% (Appelt and Israel, 1999). Similar measures are observed even when the annotation is done using predefined vocabulary (ontologies) (Passonneau et al., 2006). Similar results, around 60%, apply to systems presented at MUC (Appelt and Israel, 1999). This may be caused by:

- nature of the texts – as (Turmo et al., 2006) mention texts are written for humans not bearing in mind their automatic processing;
- complexity and variety of the kinds of information sought;
- appropriateness of output chosen to the information requirements of the task.

The following sections elaborate on extraction methods used in IE and their characteristics as well as present remarks that influence the model of the spatial indexing to be proposed.

### 2.2.1. Named entity recognition

The aim of the named entity (NE) task (in ACE program defined also as Entity Detection and Tracking) is to identify proper names (locations, persons), dates, times, monetary amounts and percentages in text (Bikel et al., 1999). Proper names are especially important for IE systems, because usually one wants to extract events, properties, and relations about a particular object (Appelt and Israel, 1999). Moreover, based on research of (Marsh and Perzanowski, 1998) proper names constitute about 70-80% of all named entities; date and time expressions account for 10-20% and the numerical values are less than 10%.

The NE extraction task was defined for the purposes of the MUC that also assumed that mechanisms developed should be of practical use, largely domain independent and perform with a high accuracy (Grishman and Sundheim, 1996).

ACE, in addition to NE MUC task, defines also five subtasks dealing with detection of (Maynard et al., 2003):

- entities: person, organization, location, facility and geo-political entity (GPE);
- entity attributes;
- entity mentions;
- mention roles;
- mention extents.

In order to identify named entities, a free-text document has to be analysed following the classical IE approach (Turmo et al., 2006, Maynard et al., 2003):

- document preprocessing phase – text segmentation (dividing text into a set of text zones, then segmenting text into appropriate text units), selection of relevant text fragments, tokenization (obtaining lexical units), morphological analysis as well as NE recognition and classification, disambiguation of name mentions, stemming, lemmatizing, etc.;
- syntactic parsing – using full or partial parsing approaches to identification of name mentions. The full parsing approaches were adopted from the NLP techniques for the needs of information extraction. However, at MUC-3 they were outperformed by the selective concept extraction proposed by (Lehnert et al., 1992). The reason why nowadays the partial parsing approaches are more popular is that all NLP techniques are computationally expensive, require a lot of time and computing resources, produce ambiguous results and broad-coverage grammars are difficult to maintain.
- semantic interpretation of extracted content – dealing with analysis of discourse in order to link semantic interpretations among sentences using e.g. anaphora resolution. IE systems as dealing with filling our predefined extraction templates, have also to resolve issues such as ellipsis, coreference, etc.
- generation of output template to translate the final interpretations into the desired format.

Typical modules of an IE system built following the approach described above are presented in the Figure 6.

The architecture presented is quite mature and widespread and most of the systems take advantage from the proposed approach. However, complexity of systems i.e. the number of modules they need for proper functioning and achieving the best possible results, depends on the following factors (Appelt and Israel, 1999):

- language of the text - e.g. Polish in comparison to English requires morphological analysis, Chinese needs also word segmentation processing;
- text properties – very long texts may require additional IR support to identify parts for IE processing. A separate treatment is also needed for texts including large amount of multimedia and tables.
- definition of extraction tasks – for simple tasks only the general components are involved.

- genre – informal texts, transcripts, emails may include mistakes, misspellings, ungrammatical constructs – such texts need different techniques to deal with.

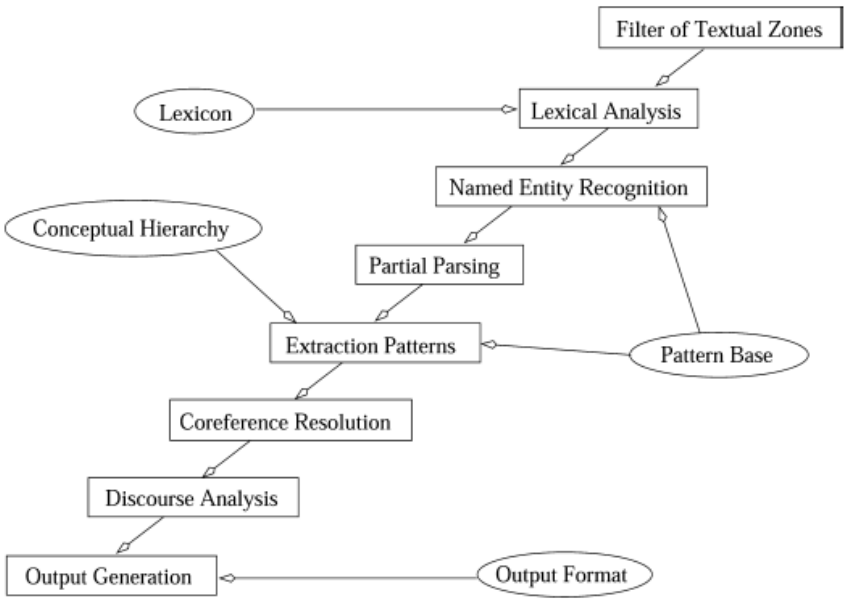



**Figure 6. Architecture of an IE system.**
**Source: (Turmo et al., 2006, Yangarber and Grishman, 1997)**

(Dimitrov, 2002) analysed results of information extraction of different tools from various conferences and noticed the quality of the information extraction results achieved in the NE task is usually very high - the best F-Measure achieved in the MUC-7 competition was 93% while humans achieved 98%.

As for the plethora of available tools including inter alia GATE, Ellogon, BRIEFS, it was not easy to choose one that would be used for the needs of this dissertation. After an analysis of functionalities provided, we decided to choose SProUT. SProUT (Shallow Processing with Unification of Typed feature structures) is a platform for the development of multilingual shallow text processing systems consisting of several linguistic processing resources, which can be coupled in a flexible way as well as provide the environment for development and testing of integrated grammars (Drozdzynski et al., 2004). The reasons for our choice were threefold. Firstly, SProUT offers similar functionalities to other available extraction tools, but it is fully integrated with a morphological analyser for the Polish language (Morfeusz). This analyser is incorporated with a module for development of extraction grammars. Secondly, SProUT is available as a web service. This enables combining it with other modules that were

developed for the needs of this research. Moreover, close cooperation between DFKI[14] and Poznan University of Economics also influenced our choice.

### 2.2.2. Extraction patterns

Extraction patterns (extraction rules, case frames) are used to extract information relevant to a particular extraction task (Muslea, 1999). As it was already mentioned, these patterns may be generated automatically, using learning algorithms or manually by a knowledge engineer. Applying them for text analysis means, that a text fragment that matches the defined pattern is memorized and the information it contains is extracted using the extraction rule, to fill in the template.

There are two approaches to manually designing the domain-relevant patterns, namely molecular and atomic approach (Kaiser and Miksch, 2005).

The molecular approach involves matching most of the arguments (or all if possible) to a single pattern. In order to achieve this, a number of small rules that capture the most common situations from the domain is developed. Then, the set of rules is extended with rules including patterns less frequently occurring in the corpus. This leads to a decreased precision of the original set of rules and at the same time to an increased recall.

The atomic approach builds a model that extracts arguments for an event and fills in the template structure based on "intelligent" guesses rather than syntactic relationships. This approach also often results in a very high recall as it assumes domain-relevant events for any recognized entity and in low precision. It is often used when entries in the domain have easily determined types and when templates are structured so that there are only a few slots that may be filled in by entities of a given type.

In SProUT Unification-Based Grammars are used for rule-based information extraction (Krieger et al., 2004). They provide declarative representation of linguistic knowledge which is defined using a high-level representation language. Thus, the parser that takes them as an input behaves like an inference engine. The representation is using the feature structure allowing for integration of different linguistic description levels, spanning phonology, syntax and semantics. Feature structure in SProUT may be perceived as a collection of feature-value pairs, where feature expresses functional property and the value of feature may be an atom or another feature structure (allowing recursive embedding).

[14] http://www.dfki.de

The grammar rules are defined using XTDL (eXtended Type Description Language) - a formalism combining the typed feature structures and regular expressions. XTDL was built on top of TDL (Type Description Language) and enables definition of rules as well as establishment of a type hierarchy of linguistic entities. A rule in XTDL is defined as a recognition pattern written as a regular expression. Regular expressions over feature structures describe sequential successions of linguistic signs. XTDL provides standard operators: disjunction, Kleene star, Kleene plus and optional existence represented by |, *, +, and ? respectively; $\{n\}$ following an expression denotes the n-fold repetition, whereas $\{m,n\}$ repeats at least m times and at most n times (Krieger et al., 2004).

The example of a rule written in XTDL is presented in Listing 1. Some more examples of developed rules along with explanations are presented in Chapter 4.

**Listing 1. Developed XTDL rule for extraction of squares.**

```
pl_place :>
(
((morph & [STEM "plac", CSTART #cs]) | (morph & [STEM "Plac", CSTART #cs]))
  |
 (((token & [SURFACE "Pl", CSTART #cs]) | (token & [SURFACE "pl", CSTART
#cs]) | (token & [SURFACE "PL", CSTART #cs]))  token & [TYPE dot] ))

(@seek(pl_geo_all_names) &  [SURFACE #nazwa_do_adresu, CEND #ce])
(@seek(pl_numer_domu) & #numer_domu)?

-> ne-location-postal & [LOCTYPE postal_address, STREET #nazwa_do_adresu,
STREET_NUMBER #numer_domu, LOCNAME #cala_nazwa, NCSTART #cs, NCEND #ce],
where #cala_nazwa=ConcWithBlanks("Plac", #nazwa_do_adresu, #numer_domu).
```

### 2.2.3. Type hierarchy

Tagging of Named Entities is an important functionality of most of the IE applications, particularly when developing extraction patterns or annotating texts. However, various sources define only a limited number of Named Entity Types.

The first developed entity set had only 7 types, namely organization, location, person, date, time, money and percent expressions (Grishman and Sundheim, 1996). Target application for this entity set was extraction of information on business activities. Unfortunately, this hierarchy proved to be not extensive enough for specific applications, and had to be extended by interested entities to be useful. Another formalization of named entities was provided by

(ACE, 2009). In this formalization two new entities were added, namely GPE (geographical and political entity) and facility. This however was still not enough to cover general information extraction tasks as usually more specific descriptions were needed. This motivated the development of the most extensive hierarchy proposed up to now. (Sekine et al., 2002) proposed a named entity hierarchy consisting of about 150 types. Its development was based on analysis of corpus of newspaper articles, types used in previous systems and definitions and thesaurus such as WordNet and Roget. Authors proposed division of all named entities into three major classes; name, time and numerical expressions. Each of them is then further refined, e.g. name is divided into person, organization, location, facility, product, event, natural_object, title, unit, vocation, disease, god, id_number, colour and name_other. Detailed hierarchy is depicted in Figure 7.

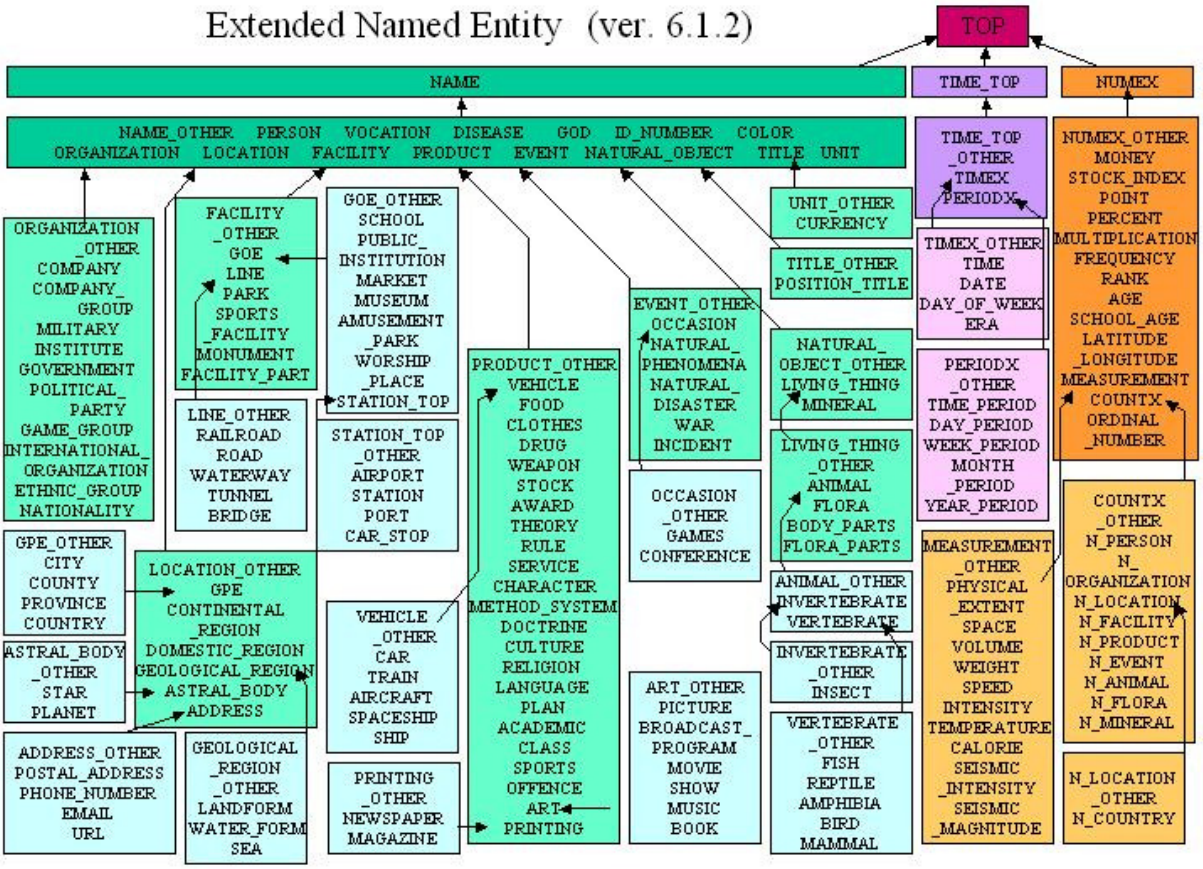



**Figure 7. Extended Named Entity Hierarchy. Source: http://nlp.cs.nyu.edu/ene/**

From the point of view of this dissertation the most important part of this hierarchy concerns locations. They were divided into seven groups: location_other, GPE, continental_region, domestic_region, geological_region, astral body and address. Detailed description of all categories is available at: (Sekine, 2003).

There are also commercial efforts to develop the hierarchy of named entities, but they are all tailored to the purpose of extraction that is to be performed, e.g. (Brunstein, 2002).
All these efforts influenced our type hierarchy that is presented in Chapter 4 of this dissertation.

### 2.2.4. Anaphora resolution

Another issue that is currently widely addressed in the IE domain is anaphora resolution, also known as coreference resolution. The term *anaphora* denotes *"the phenomenon of referring to an entity, already mentioned in text, by different ways - most often with the help of a pronoun or a different name* (Dimitrov, 2002)*"*. Resolving anaphora is important especially when dealing with semantic information extraction, which aim is not only to identify name mentions in text, but also to discover relations between them. Moreover, such resolution is important for assuring a correct understanding of text because without finding the proper antecedent the meaning and role of the anaphor in text cannot be realized.
The following types of coreferences were identified and are addressed by various researchers in the field, e.g.(Bagga, 1998, Denber, 1998, Dimitrov, 2002, Kaiser and Miksch, 2005):

- pronominal coreference – using pronouns such as he, she, my, your, myself instead of the name of the entity we refer to. According to (Mitkov, 1998) the pronominal reference concerning the personal pronouns is the most important type of anaphora and resolving it significantly increases the performance of an IE system;
- name-alias coreference (or proper names coreference) – as addressed by (Piskorski, 2002) – names and their variants should be perceived as referring to the same entity;
- appositions – typically used to provide some additional information on the entity, e.g. "Poznan, one of the biggest cities in Poland";
- predicate nominal (subject complement) – completes reference to a subject of a clause, e.g. Koszalin is the name of the city and county in the Zachodniopomorskie Region in Poland. If in the same text one refers to a county, then this mention should be substituted by the name of the county (disambiguation issue).
- identical set or types – hypernym/hyponym relation – if in one text different names are assigned to the same group of cities, people, etc.
- function-value coreference – when in one sentence appears a word that is a function for other phrase appearing in the same sentence, e.g. this year the total amount of revenue is at the level of 20%;

- definite description coreference – enables reasoning on the relation between entities, e.g. Apple Computer being a product of Apple.

As our research deals only with geographical location names – some types of coreferences are out of scope in this work. Further in the thesis we focus on resolving pronominal and predicate nominal coreferences.

## 2.3. Spatial information retrieval

This dissertation builds on top of information extraction techniques and provides a mechanism for improving precision and recall of spatial information retrieval. As it was already mentioned information retrieval (IR) is concerned with retrieving documents that "are likely to be relevant to the user's information needs as expressed by his request" (Rijsbergen et al., 1998). However, documents subject to information retrieval contain information that may be viewed from different aspects. Therefore, different IR approaches were developed. Next section presents concepts and resources used in geographical information retrieval (GIR).

### 2.3.1. Characteristics of geographic information retrieval

Geographic information retrieval may be defined as an approach to dealing with queries that can be formalized as a triple <*what, rel, where*>, where *what* specifies the non-geographic object, *where* specifies the georeferenced term, and *rel* relates the geographic and nongeographic terms (Fu et al., 2005). (Larson, 1996) defines the GIR as an applied research area combining aspects from various fields such as databases, human-computer interaction, geographic information systems, and information retrieval. Furthermore, he underlines that in order to perform the GIR tasks effectively, indexing, searching, retrieving, and browsing techniques must be carefully developed taking into account the specifics of the geographical information.

GIR is currently dealing with issues such as identification of geographic context in plain text (Woodruff and Plaunt, 1994), resolution of ambiguities in place names (Pouliquen et al., 2006, Leidner, 2007a), building spatial indexes for documents (Fu et al., 2003) as well as the use of semantic technologies (Andrade and Silva, 2006, Martins et al., 2006).

(Martins et al., 2005c) list the following challenges for the GIR:

- development of the geographical ontologies,
- handling geographical references in texts,
- assigning geographical scopes to documents,

- providing ranking of documents according to the geographic criteria,
- building interfaces for GIR.

This work addresses the first three challenges.

The whole process of the GIR resembles the classical information retrieval. Firstly, all documents need to be properly indexed. Then, after a query is submitted, system searches for relevant documents. Typical query may contain terms and operators (i.e. disjunction, conjunction and filters). Searching involves the use of the previously created indexes and computing a similarity score between a query and each document. However, in case of GIR the problem of computing similarity is more complex. Simple keyword matching in classical IR neglects underlying spatial relationships and doesn't support complex spatial queries. To solve the problem, a representation that takes into account both textual and spatial features of web pages need to be developed (Zhou et al., 2005). Then, estimating a spatial relevance of document to a query may be simplified to the problem of computing similarity between two geographic locations, one specified for a query and the other mentioned in the document (Andrade and Silva, 2006).

Consequently, a ranking of documents based on their relevance to a submitted query is produced. Sometimes, a cosine measure from a classic IR is applied for estimation of relevance between a query and indexed documents (Abramowicz et al., 2002). It is a cosine angle between the vectors that represent the document and the query (Baeza-Yates and Ribeiro-Neto, 1999).

$$sim(d_j, q) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}$$, where:

$d_j$ represents a document, $q$ stands for a query and $w$ is a weight of a term within a document or a query.

For web information retrieval when establishing ranking also link-based ranking scores are computed (Jin and Dumais, 2001, Martins et al., 2005b).

Geographical IR developed a number of geo-specific approaches for ranking documents. (Martins et al., 2005c) propose to use measures similar to georeferencing techniques where a spatial relevance of a location with respect to a queried region is inversely proportional to the Euclidean distance between them.

Other approaches, e.g. (Beard and Sharma, 1997), (Hill, 1990), propose to use extent of overlap between geographic regions when calculating the spatial relevance. We can also calculate a topological distance between places based on adjacency, connectivity or hierarchical containment. The problem of measuring the similarity in hierarchic structures has been studied in detail by (Li et al., 2003b).

The SPIRIT project proposed geographic query operators that use different metrics for ranking, e.g. "near" operator triggers Euclidean distance estimation, "east of" operator - angular distance (Vaid et al., 2005).

(Jones et al., 2001) propose to use the potential criteria for assessing the geographical similarity: distance in map or geographical coordinate space between query and candidate, travel time between query and candidate, number of intervening places, spatial inclusion of the candidate within the query place, containment of the query place by the candidate, containment of candidate within or overlap of candidate with, regions that contain or overlap the query place, boundary connectivity between query and candidate.

Interestingly, the authors prove that the most effective way is the use of geographical hierarchies in combination with Euclidean distances between places. Worth underlining is the fact that handling the concept@location queries is usually done based on two dimensions, namely content and geography. Nevertheless, some authors extend these dimensions and introduce time, e.g. (Bassara, 2009, Abramowicz et al., 2002).

The earliest approaches to GIR were based on use of simple gazetteers where all place names were associated with a map-grid or geographical coordinate. Therefore, the information processing was similar to the one taking place in typical GIS (Jones et al., 2001). The first attempt to introduce relationships between places was made in Thesaurus of Geographic Names (TGN) (Harpring, 1997). TGN stored not only coordinates and alternative versions of place names, but also relationships of administrative areas with physical features. However, since then relatively little progress has been made when it comes to representation of geographical concepts for the purposes of information retrieval (Jones et al., 2001). Current approaches pay a lot of attention to benefiting from IE efforts and reduce the boundary between IR and IE (Martins et al., 2005c).

### 2.3.2. Representation of locations

Typically, locations in information systems are represented as keywords or two-dimensional spatial objects (Zhou et al., 2005).

Textual keywords include, besides name mentions, also numerical data such as postal codes, telephone numbers (McCurley, 2001). However, using keywords it is not possible to define a shape of place. (Zhou et al., 2005) claim that because of that it is impossible to define relationships between places. In our work, we show that ontologies may provide a remedy for this problem.

Two-dimensional spatial objects may be represented using either vector model or raster model. With regard to a raster model, the precision of this representation depends on the size of grid cells. In vector model, point locations are represented as points, whereas regions are polygons or minimum bounding rectangles. Representation using polygons is more accurate, but minimum bounding rectangles (being a simple approximation of region shape using only two diagonal points to represent a shape) enable decreasing the storage cost and computation overload (Zhou et al., 2005, Markowetz et al., 2005, Ma and Tanaka, 2004).

### 2.3.3. Ontologies to support information retrieval

Ontology is a "a catalogue of everything that makes up that world, how it's put together, and how it works" (Sowa, 1984). According to one of the most frequently quoted definitions ontology is "a specification of a conceptualization" (Gruber, 1993) or "a logical theory which gives an explicit, partial account of a conceptualization" (Guarino and Giaretta, 1995). As pointed out by (Hepp, 2007) there are three aspects that usually influence the understanding of the notion of ontology, namely:

- truth vs. consensus – ontology as a detailed model of the reality vs. a consensus on this model reached by a group of people,
- formal logic vs. other modalities – for some of the ontology researchers formal logic as means to express the semantic account is the core characteristic of the ontology. However, it is arguable if formal logic is the only or even the most appropriate modality for specifying the semantics of a conceptual element in an ontology (Hepp, 2007),
- specification vs. conceptual system – if it is an abstraction over domain of interest (entities and their relationships) or explicit specification of this abstraction using one of the available formalisms, e.g. WSML, OWL.

(Smith and Mark, 2001) define ontology as "a neutral and computationally traceable description or theory of a given domain which can be accepted and reused by all information gatherers in that domain".

According to (Maedche and Staab, 2001) every ontology comprises the following elements:

- a set of strings describing lexical entries for concepts and relations,
- a set of concepts,
- a taxonomy of concepts with multiple inheritance (hierarchy),
- a set of non-taxonomic relations – described by their domain and range restrictions,
- a hierarchy of relations,
- a set of relations between concepts and their lexical definitions,
- a set of axioms describing additional constraints on the ontology in order to make implicit facts explicit.

(Tomai and Kavouras, 2004) summarizing the work of (Maedche and Staab, 2001) claim that ontology, especially geographical ontology, consists of four main elements: concepts, lexicon, relations and axioms. These elements as well as relations between them are presented in Figure 8.

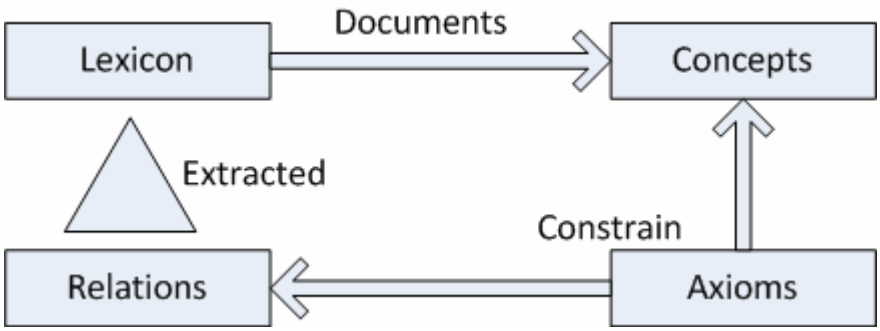



**Figure 8. Relations between ontology elements. Source: (Tomai and Kavouras, 2004)**

Concepts are an integral part of any ontology. They are denotations for real-world objects. The knowledge about ontology concepts is usually expressed by a lexicon providing documentation of concepts constituting the ontology. Relations that may be distinguished involve semantic relations between concepts, properties defined for concepts or relations among relations. Semantic relations are relationships between concepts such as hypernym/hyponym (is-a, kind-of relation), meronym/holonym (has-part, is-a-part of). Axioms are constraints imposed on concepts or on relations among concepts, properties and their values (Tomai and Kavouras, 2004).

(Smith and Welty, 2001) propose other explanation of the Gruber's definition. This specification is depicted in Figure 9. It presents different levels of knowledge representation starting from the one of the lowest complexity. The simplest ontology is a catalogue in which each product type has a unique code. Such a catalogue may be perceived as an ontology of things that company sells. More advanced information systems provide natural language texts

and allow for string matching. The next level are glossaries providing natural language descriptions of terms. Thesaurus in addition to definition of terms provides relations between them (synonyms, antonyms, etc.). Taxonomies are used for example in object-oriented systems and in frame-based systems. They provide not only relations between concepts but also enable inheritance of properties from more general classes to more specific ones. Additionally, they provide restrictions (e.g. cardinality) on relations between objects. Finally, the most expressive ontologies use axioms defined in first order, higher order or modal logic. An important issue to note is that authors do not distinguish between ontology and knowledge base.

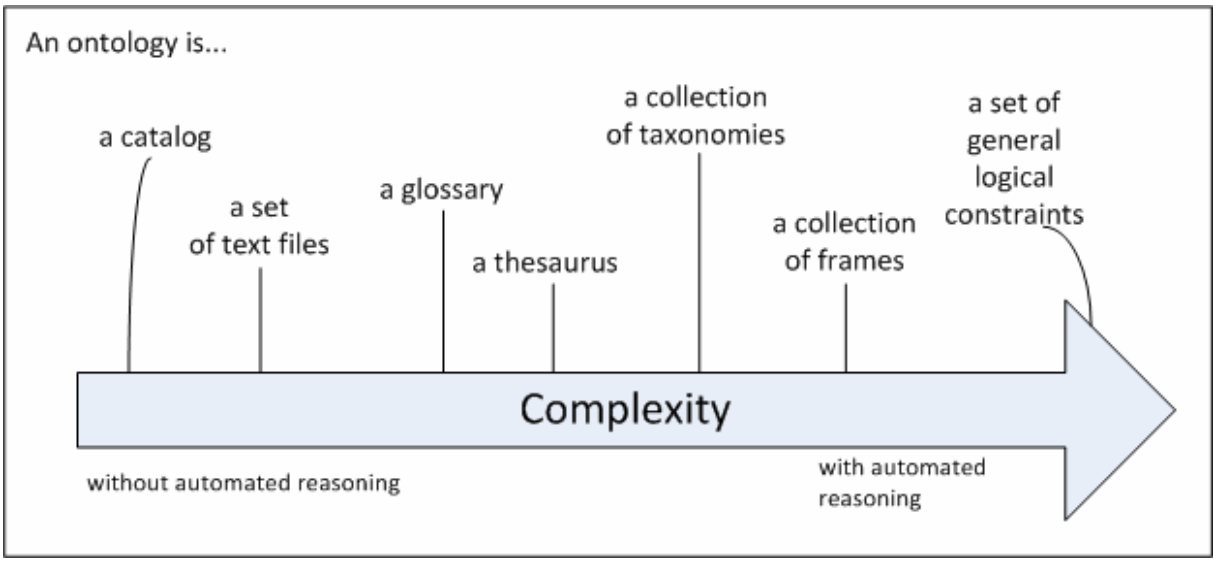



**Figure 9. Different levels of expressiveness of ontologies.**

**Source: (Welty et al., 1999)**

From this dissertation's point of view, an ontology is understood as a model of geographical domain that enables storing and reasoning on geographical data and information. The geographical ontology follows design approach defined by (Maedche and Staab, 2001). Also the geographical knowledge base is provided. The ontology is specified using WSML[15] that was chosen from available ontology specification languages (RDF[16], DAML+OIL[17], OWL[18]).

[15] http://www.wsmo.org/wsml/

[16] http://www.w3.org/RDF/

[17] http://www.w3.org/TR/daml+oil-reference

[18] http://www.w3.org/TR/owl-features/

### 2.3.4. Geographical ontology

As mentioned in passing, geographical ontology is a conceptualization of a geographical domain. This conceptualization should support the measurement of spatial similarity between geographical places (Jones et al., 2001). The spatial similarity is understood as matching named places with other named places that are equivalent or similar in geographical location. (Fu et al., 2003) similarly define the geographical ontology saying that such ontology should encode names of places and the terms to describe spatial relationships. For the needs of proper query answering, the ontology should also describe the footprint of a place, i.e. its geometric representation.

The creation of geographical ontology therefore may be done in the following steps (Cai, 2007):

- definition and selection of the aspects of the world to be modelled,
- specification of important entities and their properties,
- recognition of instance existence,
- demarcation of the entity instances on the ground, i.e. creation of their geographic footprint.

However, representing geographic object by its footprint is not an easy task as the geographical object's boundaries may be crisp or fuzzy. Moreover, they may be changeable. (Smith, 1995) divides geographical objects into two types: *bona fide* and *fiat*. *Bona fide* boundaries are river banks, coastlines, etc. These boundaries exist independently of humans and are influenced only by nature. *Fiat* boundaries are dependant on human decisions, laws or political decrees. Sometimes *fiat* boundaries may be exactly the same as *bona fide* boundaries, e.g. when a border between two countries lies on a river. Because of these problems many authors e.g. (Jones et al., 2001) admit that for many practical purposes detailed geometric data are not necessary. They propose to maintain a parsimonious spatial model of geographical space instead of detailed GIS data on spatial relationships of topology and proximity.

As an ontology enables to store relations between objects, detailed GIS data seem to be an enhancement that may be used for disambiguation of geographical name mentions.

In information extraction the most important resources when extracting information about places from documents are gazetteers. Some authors, e.g. (Martins et al., 2005c), assume that there is not much difference between gazetteers and geographical ontologies as they both store information on places. The reasons for introducing the geographical ontology instead of a gazetteer, reported in previous research were enumerated by (Fu et al., 2003). Gazetteers

usually do not include support for storing spatial relationships. Some gazetteers, e.g. TGN, encode hierarchical relationships, however do not support adjacency or overlap. This issue is also unresolved in the SProUT gazetteer (that is used in this research as well), which provides support for simple hierarchies, but unfortunately it does not support descriptions of more complex relations or storing additional data on locations.

Moreover, there is a need to define relations between different geographic feature types. (Fu et al., 2003) mention, that the Netherlands uses the term province for what in other countries may be referred to as canton, region, state, etc.

Other issue concerns support for encoding geographical features of different types in the same manner. All geographic places in gazetteers are described using the same properties and ignoring the specifics of a named entity described.

Moreover, existing gazetteers vary in many dimensions (e.g. scope, completeness, correctness, granularity, balance and richness). They also lack standardization formats, contents and service interfaces (Martins et al., 2005c). Therefore, sharing data between these resources is not a trivial task. Ontologies propose to overcome these problems using the Semantic Web standards (Egenhofer, 2002).

Similarly to gazetteers, geographical ontologies are not fully accessible and applicable. Their content depends on the requirements that were elicited before the ontology development.

Before presenting details of few geographical ontology development methodologies, it is worth to mention that they follow two distinct approaches (Cai, 2007):

- philosophical approach – targeted at identification of only top-level categories of geographic domains independent from any particular application, e.g. (Smith and Mark, 2001, Mark et al., 1999);
- knowledge engineering approaches – aimed at design of application-specific ontology for developed system. Such ontologies provide concept categories and relations for a specific area of application and therefore are difficult to apply in other solutions.

(Smith and Mark, 2001, Mark et al., 1999) carried out a series of experiments to find how non-experts conceptualize geographic domain. Based on asking general questions such as “something that could be portrayed on a map”, ”a kind of geographic feature”, “something geographic”, “a geographic concept”, etc. In the final classification they decided not to include terms mentioned with statistical frequency less than 10% of responses. The developed geographical ontology consists of 35 concepts concerning mainly physical geography (river, lake, ocean, sea, mountain). The ontology comprises also concepts such as city, continent, building, county, etc.

(Jones et al., 2001) propose an ontology created for their OASIS system (Ontologically-Augmented Spatial Information System). OASIS was designed to store cultural information about archaeological findings and historical buildings that have been classified according to AAT (Art and Architecture Thesaurus) and geographically referenced using data from Thesaurus of Geographic Names. A place concept in OASIS is implemented as a Geographical Concept. The figure below presents the way the Place concept was modelled within the proposed ontology.

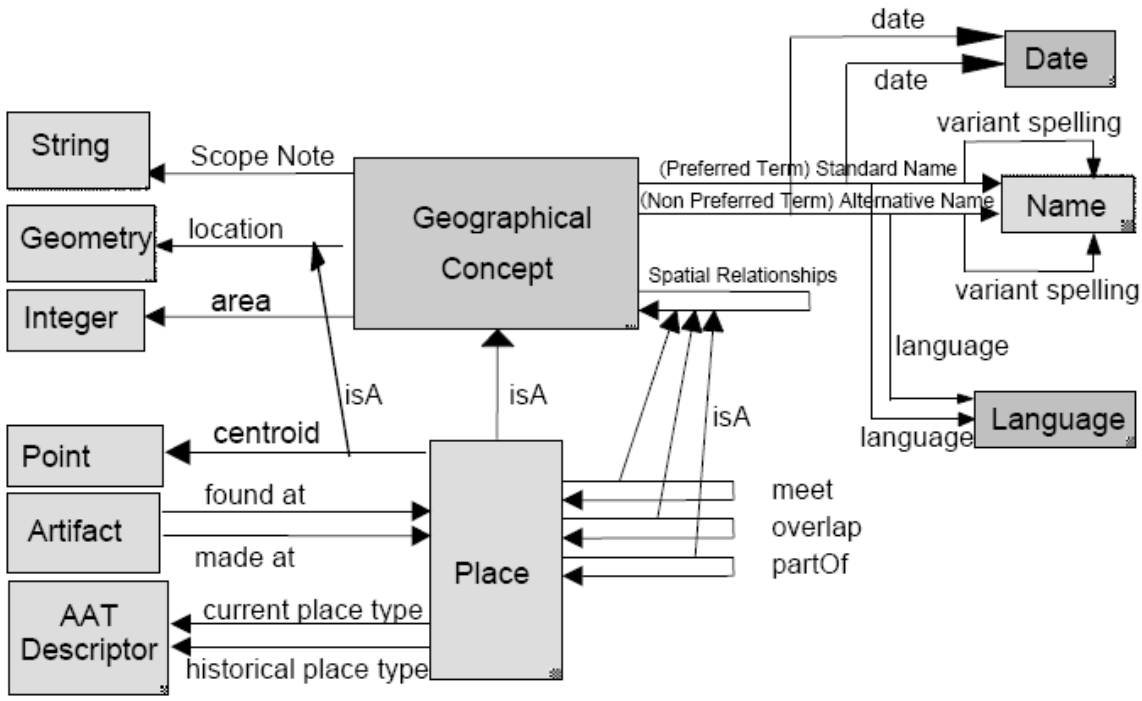



**Figure 10. Place as a type of geographical concept within the OASIS system. Source: (Jones et al., 2001)**

Geographical concept has a standard name and alternative names. These names may have many variants of spelling, a date of origin and language. A scope note provides explanation of the concept. Geographic location may be associated with a location defined by a geometry object, a measurement of the area and spatial relationships. In the Place concept a location relation is specialized to a centroid relation having defined latitude and longitude coordinates. (Fu et al., 2003) proposed a structure of such an ontology based on their experience of development of various gazetteers as well as taking into account requirements for the geographical information retrieval. According to them the spatial information retrieval requires that ontology is enhanced with geometric footprints associated with a place that may take the form of points, polylines and polygons and for each place more than one geometric

footprint is supported. The ontology designed by them stores additionally information on classification of places, such as counties and districts. For disambiguation purposes this ontology stores also containment relations between places.

The ontology they designed was used in the SPIRIT project and up till now is one of the most important efforts in the domain of geographic ontologies. This ontology consists of three components, namely: geographical feature type ontology, geographical feature ontology and spatial relationship ontology.

Geographical feature type ontology encodes various types of the geographic feature instances. The second component, the geographical feature ontology, stores the concrete feature instances in a given geographic space. Authors decided not to store too detailed information about features because of complexity issues it won't be used either in the indexing or in searching.

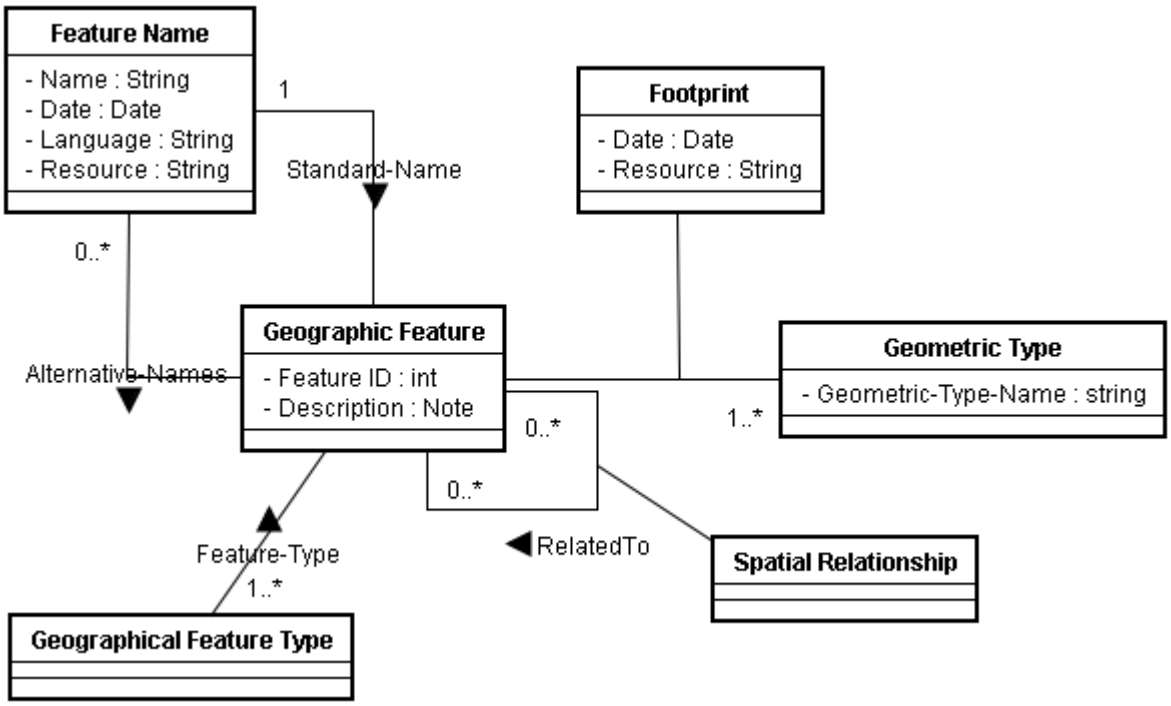



**Figure 11. Schema of Geographical Feature Ontology.**
**Source: (Fu et al., 2003)**

The schema of the ontology is presented in Figure 11 and extended in Figure 12 and Figure 13. Each geographical feature has its own unique ID. Interesting issues concern the geometric type, i.e. footprint of the geographical feature. The geographical object in the SPIRIT ontology may be represented as a point, a polyline (composed of two or more coordinate

points) or a polygon (representing the boundary of a geographic feature). Geographical feature may have spatial relationships with other features that include part-of relationship, contains (being the inverse of part-of), adjacent-to (features that share the boundary with the geographical feature given) and overlapping.

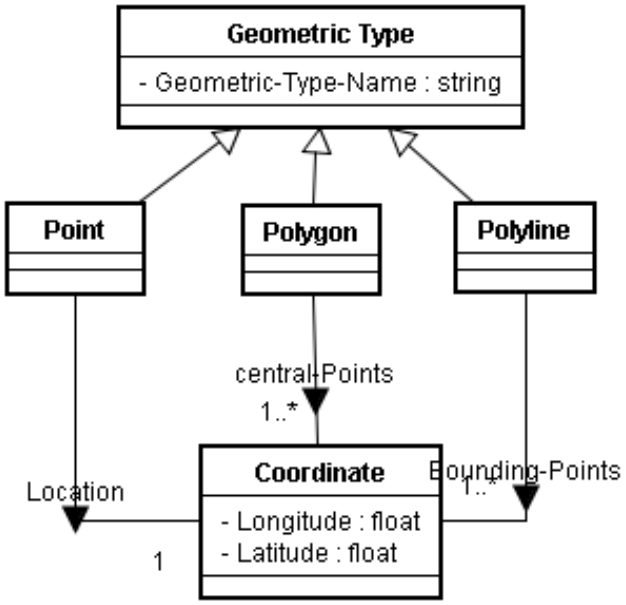



**Figure 12. Geometric Feature Types. Source: (Fu et al., 2003)**

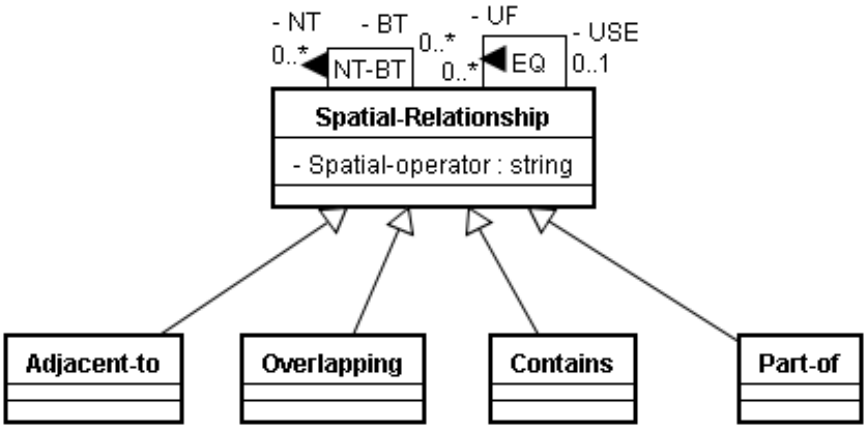



**Figure 13. Spatial Relationship in SPIRIT Ontology. Source: (Fu et al., 2003)**

All ontologies developed must be populated with content. In case of the SPIRIT ontology, the content was provided using resources from SABE (EuroGeographics, 2004) - a digital map dataset representing the geometry of administrative boundaries for Europe and TGN (Getty,

2004) and providing a structured vocabulary with information on places with a global scope (Fu et al., 2005).

Recently, (Chaves et al., 2007) presented a Geographic Knowledge Base (GKB), environment for integrating geographic data and generating ontologies in OWL format. Firstly, they defined GKB metamodel supporting representations of multiple information domains related to geography, such as administrative and physical domains.

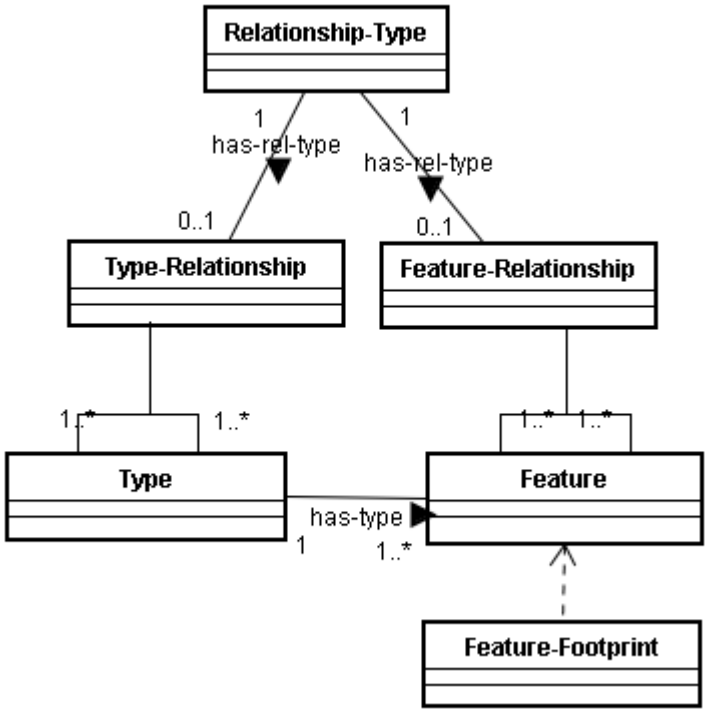



**Figure 14. GKB base class metamodel. Source: (Chaves et al., 2007)**

The class Feature is associated with the class Type that enables specification that e.g. Wisła is of type river. The class Type-Relationship defines relationships between types. The class Relationship-Type stores other relationships including geographical ones. Features may be specialized by their footprints capturing coordinates using the Feature-Footprint concept. This base model is then further extended by authors. Having defined the model, they used data from Portuguese Ministries, Governmental Institutes and Wikipedia (about 23000 features) to feed the GKB, creating the first public geographic ontology of Portugal Geo-Net-PT01[19].

[19] Ontology is available after sending a request from http://poloxldb.linguateca.pt/index.php?l=geonetpt&f=en

### 2.3.5. Spatial document indexing

In information retrieval an index is required to organize documents and limit retrieved set of data after processing a user's query (Larson, 1996). (Martins et al., 2005b) define document indexing as providing representation of data for the purposes of increasing efficiency in the information retrieval; usually this is done by splitting text into tokens that will become keywords in queries. According to other definition, document indexing in the collection means transforming document text to its surrogate for use by an IR system (Lalmas and Ruthven, 1998). The authors claim that indexing relies on selecting good document descriptors, such as keywords and terms, in order to represent the content of the documents being indexed. A good descriptor should help to describe the information content of the document and at the same time differentiate it from other documents in the collection (this is called a discriminatory power of the descriptor) (Abramowicz, 2008).

In GIR, depending on different representations of locations, various index structures are proposed (Zhou et al., 2005). Place names can be organized in a form of a flat list or hierarchy tree when representing administrative relations between them. Such a hierarchy structure is especially useful when identifying the geographical scope of web pages (Wang et al., 2005b). Representations of spatial objects are usually in the form of R-tree, quad-tree or grid-files (Baeza-Yates and Ribeiro-Neto, 1999, Zhou et al., 2005). For searching, these indexes are usually combined in order not only to achieve high recall and precision, but also to assure efficiency of the process. (Markowetz et al., 2005) propose to perform a conventional search and only the geographic footprints of pages are compared. (Ma and Tanaka, 2004) use the spatial index in order to increase effectiveness of the retrieval process by building dynamically R-tree index on the Google results. (Lee et al., 2000) integrate these two (textual and spatial) approaches using Oracle database to handle the user queries. This approach however is not efficient enough to be used in the Internet.

(Vaid and Jones, 2004) summarize that from the plethora of available indexing methods two seem to be frequently applied in various search engines, namely inverted index and spatial access methods (SAM). Summary of their work is presented in the table below.

**Table 3. Index Options. Source: (Vaid and Jones, 2004)**

| Index Type | Description |
|---|---|
| A. Pure Textual (PT) | This index type consists of a lexicon and inverted list. The major difference is that when answering spatial queries the |

| | |
|---|---|
| | query undergoes a spatial term enrichment process by interacting with the ontological component. |
| B. Text Followed by Spatial (TS) | In this case a two-stage index is constructed. First, a lexicon for the text collection is created. Then to every inverted list a spatial index is created containing a certain number of cells. |
| C. Spatial Followed by Text (ST) | In this case firstly the document space is divided into a certain number of cells. Then for each cell a textual index is created. |
| D. Spatio-Textual Index (SP) | Similarly to type B and C, the document space is divided into cells and a lexicon created for each cell. However, the index in this case is constructed by combining the id of the spatial cell with text terms identified for documents contained in this cell. It may be easily noticed that this index has similar spatio-textual characteristics to the TS and SP types. The reason for distinguishing it as a separate type of index is due to its implementation, which is much easier when compared to the TS and ST types. |

The most popular indexing technique to create and maintain textual indexes is the inverted index (Martins et al., 2005b). It consists of a set of inverted lists, one for each word (or index term) that may occur in the document. The inverted list for a term is a sorted list of positions, or hits, where the term appears in the collection. A hit consists of a document identifier and the position of the term within it, often containing additional information useful for ranking (e.g. HTML mark-up).

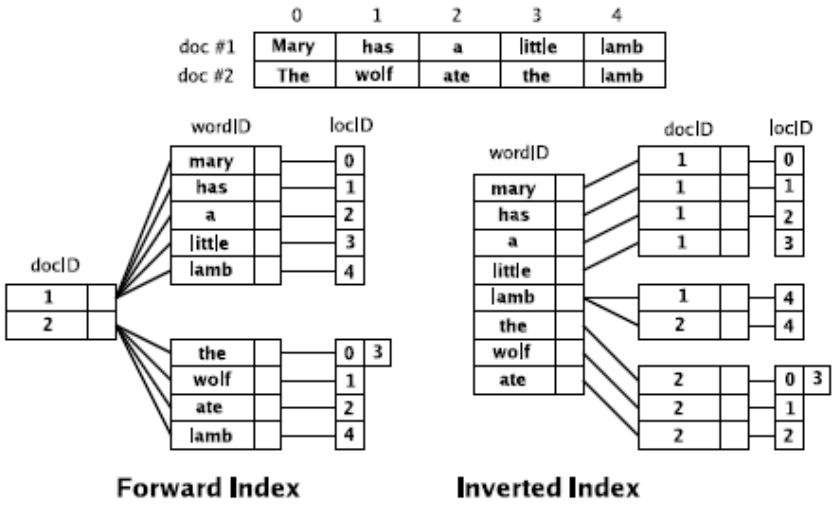



**Figure 15. Indexes for text retrieval. Source: (Martins et al., 2005b)**

Such an index may be derived in a number of ways (Larson, 1996):

1. Simple extraction (extracting keywords from the text)
2. Inferential extraction (mapping from text words to thesaurus terms)
3. Intellectual analysis and assignment of index terms.

These methods can be applied in combination. (Larson, 1996) provides an example of a GIR system that uses intellectual analysis for assignment of bounding box coordinates to an aerial photograph and inferential indexing for assignment of coordinates for places mentioned in a text.

## 2.4. Conclusions

The delivery of information in IR systems heavily depends on the retrieval model applied. What they provide as an output is ranked based on some measure of "goodness", so that the "best" matches receive the highest ranks.

The most important in the IR process is the representation of the document that is to be compared with a user query. Currently a lot of research is being done in the area of combination of textual indexes with spatial indexes, however disregarding the importance of creation of better textual geographical indexes.

Recent research in information retrieval explores the potential of applying ontologies. They are accepted as panacea for most of the geospatial semantic problems but nobody validated thoroughly this assumption. Even the basic set of concepts, categories and entities is still discussed and no agreement has been reached (Cai, 2007). Some work, however, was done in order to produce shared conceptualization of the geographical domain (Smith and Mark, 2001, Mark et al., 1999). This research benefits from the work already done and it provides relevant ontology for Poland. It is also worth to note that by application of ontologies in the post-processing phase of information extraction, we are able to eliminate deficiencies of gazetteers as discussed in this section.

(Martins et al., 2005c, Purves et al., 2007) in their work identified challenges for the geographical information retrieval. Among others, they distinguished:

- development of geographical ontologies,
- handling geographical references in texts,
- assigning geographical scopes to documents – geotagging the document,
- assigning geographical scopes to web resources.

Our research addresses these issues and aims at proposing a method for the spatial indexing, i.e. providing textual geographical description of documents retrieved from Internet sources. The intended outcome is to provide an ontological representation of a single document using the geographical ontology developed. It is important to note that our research builds on the spatial indexing techniques described in this chapter (mainly molecular, rule-based approach to shallow text processing), but it focuses on the georeferencing techniques to be described in the next chapter.

# Chapter 3.
# Extracting geographical information from documents

*Geography has made us neighbours. History has made us friends.*
*Economics has made us partners, and necessity has made us allies.*
*Those whom God has so joined together, let no man put asunder.*

*John F. Kennedy*

## 3.1. Introduction

The objective of this chapter is to provide an overview of methods, used for extraction of geographical name mentions from free text Web documents. This chapter presents approaches that may be used while creating indexes of documents for the purpose of GIR. Therefore, methods used for analysis of Internet sources as well as approaches for discovering geographical place name mentions within documents are presented. The analysis takes also into account the fact that two types of geographical references are associated with each web page, i.e. a source geography (the origin of the page, the location of a server, address of its author) and a target geography determined by the web page contents (Amitay et al., 2004).
This chapter is structured as follows: firstly it elaborates on geoparsing and geocoding as well as categories of geographic data and information. Then, it provides an overview of methods for determination of web sites' scopes followed by the description of approaches for analysis of documents. It also discusses common problems with geoparsing and geocoding and presents some solutions to these problems described in the literature. Finally, remarks regarding the spatial indexing models that are to be proposed are provided.

## 3.2. Geoparsing and geocoding – approaches for identification and grounding of place name mentions

Identification of a geospatial context of a web page, including the place of creation of information, locations it mentions as well as the target audience of the information, is of the utmost importance when improving precision and recall of the geosearch (Markowetz et al., 2004). The process of extracting geographical name mentions from documents and grounding them to real world entities is named differently by different authors, e.g. geoparsing,

geocoding, grounding, georeferencing. The most of approaches to identification of the geographical name mentions from the free-text documents (usually web pages) differentiate two phases of this process, namely: the first dealing with identification of place name mentions in text, and the second trying to bind these name mentions to real world entities; the phases, however, are not named consistently among authors.

(Densham and Reid, 2003, McCurley, 2001, Bilhaut et al., 2003) use the term of geoparsing for the first phase, while naming the process of tagging of the candidate (assigning geographic coordinates) as geocoding. (Scharl, 2007) differentiates two types of geocoding stating that only automated approach can be named as geocoding, while the not automated one should be called geotagging.

(Leidner et al., 2003, Amitay et al., 2004) define the process of determination which place name is meant by a particular occurrence in text as grounding, but the same process is called by Leidner the toponym resolution when talking about the issue from the perspective of information extraction (Leidner, 2007b).

From the point of view of information retrieval this catalogue may be extended with the notions of geo-referencing meaning "automatically index and retrieve a document according to the geographic locations discussed, displayed, or otherwise associated with its content" (Larson, 1996), and indexing - generating a document footprint determining the location of the document and possibly various others (Fu et al., 2003).

This diversity in terminology is caused by the variety of approaches and domains, dealing with the notion of geography in the free text documents.

This dissertation names the process of identification and grounding of geographical references as spatial (geographical) indexing, similarly to (Fu et al., 2003).

## 3.3. Geographically-referenced data and information

The focal point of interest of all above approaches is geographical (spatial) data and information. Therefore, before presenting details of these approaches, some explanation about different types of spatial data and information is warranted.

Two types of geographical information may be distinguished, namely coordinates and geo-references (Arikawa et al., 2000). Each type of the information has its own characteristic that is elaborated further in this section, however it is worth to emphasize that a lot of effort is put in matching these two representations, i.e. to align georeferences with their coordinates. Coordinates, are perceived to be the best kind of spatial data (when compared to georeferences), because of their (Larson, 1996):

1. persistency – they are independent of changes e.g. political or boundary;
2. simplicity of use and visualization – they may be simply connected to spatial browsing interfaces and GIS or GPS data;
3. consistency – providing framework for geographic information retrieval (GIR) applications and spatial queries.

(Michigan, 2007) states that besides georeferences there are other means of describing the location e.g. *"coordinates (latitude/longitude), coordinates whose relationship to the earth is not known (such as some survey data), linear referencing (such as street addresses and mile markers), more general land referencing (such as tier/range/section in the Public Land Survey System) and indirect positional references (such as distance and direction descriptions from known markers in the field - [30' west of the ‹old oak tree›])"*. Therefore, we need standards to collect and maintain this data in a unified way.

(Arikawa et al., 2000) present classification of the geographic data (see Table 4). As it was already mentioned, they differentiate two types of spatial data, namely georeferences and coordinates, recognizing different levels of their structuring. This dissertation deals with the georeferenced unstructured data and transforming it into the fully-structured geographic data.

**Table 4. Categories of spatial data. Source: (Arikawa et al., 2000)**

| | **Geo-referenced Data** | **Geographic Data** |
|---|---|---|
| **Fully-structured** | Non–spatial data extended with address data, e.g. database table of clients and their addresses | Data stored in GIS databases; Spatial DBs |
| **Semi-structured** | XML documents including data on addresses | XML documents including coordinates of places |
| **Non-structured** | Texts including addresses, e.g. on websites | Web pages where one may find spatial coordinates |

For a classification of the fully-structured geographic data please see Table 5.

**Table 5. Types of fully structured geographic data. Source: (Hill, 2000)**

| **Type of Representation** | **Description** |
|---|---|
| Point | Single pair of latitude and longitude coordinates |
| Bounding box | Double pair of coordinates representing the maximum and minimum of latitude and longitude extent |
| (Poly)Line | Set of points that do not enclose a space |
| Polygon | Set of points that do enclose a space |
| Grid representation | Grid references to a location according to an identified grid referencing scheme |

Furthermore, several types of the geo-referenced non-structured data may be distinguished. They are usually referred to as spatial information because of providing abstraction over coordinates. (Bilhaut et al., 2003) provide classification of spatial expressions that one may find in free text documents, based on the corpus of documents on educational system in France. They consider three kinds of spatial expressions:

1. expressions containing only [zone] part and denoting the geographical area – e.g. Paris, north of France. These expressions contain "named geographical entity" on top of which some "geographical operations" are performed – "north of", "south of", etc.;
2. expressions in the form of [quantification] + [type] + [zone], where quantification refers to a determiner (all, some, most, the, of, etc.), e.g. fifteen districts of north of France. They distinguished four types of quantification: absolute (fifteen), relative (32 per cent, half of the towns), exhaustive (all, the), exhaustive-negative (no, not any);
3. [quantification] + [type] – where zone is not specified implicitly, e.g. some seaboard towns (zone is dependent on the context).

Some authors consider also other types of geographic data such as telephones and postal codes as the geo-referenced data (McCurley, 2001).

## 3.4. Geographical importance of sources

All web resources are built bearing in mind their target audience. (Ding et al., 2000) introduce a concept of a geographical scope of a web resource as "*the geographical area that the creator of the resource intends to reach"*. They also define measures, namely spread and power of a web resource, to estimate what is the audience (the location covered) of the web site as well as its importance. This information in comparison to information on a targeted audience and envisioned importance may provide important and useful data for the public relations analysts as already elaborated in chapter 1.

Some of the existing approaches to estimation of scopes of web pages propose to analyse distribution of logs, however this data is not publicly available. (Ding et al., 2000) proposed an approach based on the publicly accessible data that provides the foundation for other methods. They assumed that if there exists a web resource *w* that is targeted at audience in location *l,* then a significant fraction of *l*'s web pages contain links to *w,* and references to it are distributed smoothly across the location *l*. Measures proposed by (Ding et al., 2000) and another approach to their evaluation proposed by (Asadi et al., 2006) are elaborated further in this section.

(Asadi et al., 2006) differentiate three types of locations that may be associated with a web resource, namely:

- host's location (physical location) – as every website is stored on one or more servers which are located somewhere in the world, the psychical location of a resource or *L(w)* is defined as the location of the server which is hosting the resource *w:* $L_{(w)}$ *:* location of *w*'s hosting services
- content location – locations mentioned in a web page's content, including geographic features mentioned in text, footnotes, contact addresses, metatags, headings, etc. In this case, the location of a webpage is the geographical location of its objects or entities: $L_{(w)}$: location of geographic entities in *w*.
- target location – location of the intended audience of the website: $L_{(w)}$: location of people that use *w*.

In the next section the issues related to the target location, namely measuring popularity and uniformity of web resources, are presented in detail. Approaches dealing with analysing location of the information host and addressing the association of content with a location are addressed later in the chapter.

### 3.4.1. Measuring popularity and uniformity of web resources

Intuitively, web pages from location *l* targeted by a web page *w* should express interest in contents of *w* by including links to it. Therefore, the power of a specific resource *w* may be then estimated as follows:

$$Power(w,l) = \frac{Links(w,l)}{Pages(l)}$$

where *Links(w, l)* stand for the number of pages in location *l* containing links to web resource *w*, and *Pages(l)* provide the total number of web pages in location *l*.

(Asadi et al., 2006) provide an alternative to this measure and estimate it based on the data extracted from logs obtained from Internet providers. They define power as popularity of web resource, i.e. the number of people accessing the resource *w* from *l* in comparison to all people accessing this resource.

$$Power(w,l) = \frac{Visitors(w,l)}{Vistors(w)}$$

If in location *l* there is an interest in a web resource *w*, then its power is greater than 0. However, (Ding et al., 2000) propose also to check the smoothness of distribution of links to resource *w* from location *l* in all sublocations of *l*, e.g. for a newspaper informing on specific region most of the links may come from a city situated in this region. Spread is high if power of a web resource *w* for all *l*'s sublocations is high.

Spread, based on the Vector Space Model from IR, may be estimated as cosine of the angle between vectors of pages and links. The formula for spread is presented below.

$$Spread(w,l) = Pages \otimes Links = \frac{\sum_{i=1}^{n} p_i \times l_i}{\sqrt{\sum_{i=1}^{n} p_i^2} \cdot \sqrt{\sum_{i=1}^{n} l_i^2}}$$

(Asadi et al., 2006) calculate the spread measure differently - as the distribution of people interested in the resource *w* coming from sublocations of *l* i.e. $l_i$ (by *l* in the formula below they mean the number of visitors from the location *l*).

$$Spread(w,l) = \frac{\sum_{i=1}^{i=n} l_i}{\sqrt{\sum_{i=1}^{i=n} l_i^2}}$$

These measures however do not take into account the number of people inhabiting a certain location or its area. They also extend the approach proposed by (Ding et al., 2000) and define the notion of stability. Stability means, that a location is eligible as target, if it remains in the scope of a webpage over time. In order to ensure the freshness of a webpage, the extracted target locations can receive different time-dependent weights. The formula for defining the stability is as follows:

$$Stability(l,w,T) = \frac{\sum_{n=t_0}^{n=t_n} \frac{1}{2^n} l}{\sum_{n=t_0}^{n=t_n} \frac{1}{2^n} L}$$

where $t_i$ is a subdivision of time *T*, *l* is a location selected as a temporal target of *w* in $t_i$, and *L* refers to all temporal targets selected in $t_i$.

The approach suggested by (Asadi et al., 2006) is, however, inconsistent in terms of notation (sometimes by *l* they mean location, sometimes the number of visitors). Also, the approach lacks normalization. The question that appears is why not to use standard deviation of probability distribution in this case..

(Wang et al., 2005a) using information from logs propose a slightly different approach. They assign scope of web resource calculating this as weight of location *l* in *w* as follows:

$$weight(w,l) = \begin{cases} \alpha_1 Userfreq(w,l) + (1-\alpha_1) Contloc(w,l) & i = 0 \\ \alpha_2 \sum_{j=1}^{n} Srvloc_{i-1}(w_j,l) + (1-\alpha_2) Srvloc_{i-1}(w,l) & i > 0 \end{cases}$$

where:

- *Userfreq(w,l)* is the frequency of *w*'s accesses by all users within a location *l*;
- *Contloc(w,l)* equals to 0 or 1 meaning that l is contained (or not) in the content location of *w*;
- $w_j$ is the web resource that has links to *w*, ($1 \leq j \leq n$, where *n* is the number of all the web resources that have links to *w*);
- $Srvloc_{i-1}(w,l)$ indicates if *l* is hierarchically contained in the intermediate serving location of *w* after the *(i-1)* iteration;
- $\alpha_1$ and $\alpha_2$ are the weights of user access frequency and serving location of previous iteration.

As it may be easily seen from the formula above they propose to constantly refine scopes of web resources based on the new information obtained from the server logs.

(Zhang et al., 2008) modified the approach suggested by (Ding et al., 2000). They define the notion of the web site serving area differentiating it from the scope of a web resource. The basis for this approach is that if a web resource has a local geographical scope it will probably contain mainly the location names or other named entities covered by its geographical scope. The geographical scope of a web resource can be used as an approximation of its serving area, but the serving area describes user interest and not a geographical coverage of the web resource. The authors provide the an example of http://www.newzealand.com that has a geographical scope of New Zealand, but is of interest of global users. (Zhang et al., 2008) estimate the serving area of a website as follows:

$$weight(w.l) = \frac{Click(w,l) / Population(l)}{Click(w, Parent(l) / Population(Parent(l))}$$

where:

- *Click(w,l)* is the number of clicks opening a web resource *w* by representatives of location *l*,
- *Population(l)* is the population of location *l*,
- *Parent(l)* is the parent location of *l* in the administrative hierarchy.

This metric is estimated based on analysis of logs, so at first they map all users' IPs to locations (unfortunately the exact method is not defined). Then, they map these locations to geographical hierarchy and estimate spread and weight measures in order to prune some nodes if the values estimated do not exceed threshold given. The second more interesting approach proposed by them relates to analysis of query logs for the web page. They extract all query terms for a given web page and based on them they build a query document enabling them to compute the serving area of the website. The second algorithm proved to be more efficient than the first one with both precision and recall of above 85% (tests were carried out on the MSN search engine) (Zhang et al., 2008).

### 3.4.2. Other approaches to evaluation of importance of web sources

Two previous approaches estimated a scope and importance of a website by analysing the number of links pointing to it or a number of users that visited a website. There exist also a number of approaches trying to estimate the geographical scope of web resources differently.

(McCurley, 2001) proposes to derive context for server sites for which he uses information from Whois database[20] for domain registrations associating server IPs with their owners. This approach was earlier proposed by (Buyukkokten et al., 1999), who implemented a three-step approach involving:

1. mapping IPs of web sites to the phone numbers of all Class A and B domains administrators based on the RS.internic.net database;
2. creating database (using information from http://www.zipinfo.com) of cities with assigned area codes what enabled them to assign IPs to cities (that had their zip codes assigned);
3. and, as the final step, using zip code mapper (also from http://www.zipinfo.com) to download zip codes and corresponding coordinates (latitude and longitude) what enabled to assign each IP to the exact location.

(McCurley, 2001) proposed also to study "contact" pages appearing within commercial websites that provide information on location of the entity that established the website.

Additional source of information provides also analysis of the network IP addresses either by obtaining data from Internet providers' catalogues or tracing data packages on the web. Most of sites are connected to the Internet using one or two individual connections. Using tools such as nslookup, traceroute, whois it is possible to discover the last hop (an important server)

[20] http://www.who.is/

before reaching the server we are interested to locate (Buyukkokten et al., 1999, McCurley, 2001). However, this approach is not always useful – especially when providers rent some space on the server on which different sites may be published. Similar method is used also by the Gtrace tool (Periakaruppan and Nemeth, 1999). However, according to (McCurley, 2001) data collected in this way is of very limited accuracy.

Other approaches suggest incorporating the location information into the DNS system. SRI International[21] submitted a request to ICANN to incorporate into the DNS system new .geo top level domain that would be further broken down by region depending on latitude and longitude, but ICANN rejected this idea. There are also other methods that were proposed regarding the information that may be obtained from DNS, for example defining LOC, GPOS or TXT resource records for locations in the Domain Name System (Farrell et al., 1994, Davis et al., 1996). However, they do not work in practice, since only few sites utilize them.

Another approach suggested by (McCurley, 2001) is to identify the language of the website in order to identify its geographical context (the analysis performed exceeds the scope of the linguistic analysis and involves meta tags in HTML pages, HTTP response headers, etc.). Such an approach is not always correct, but when combined with other methods it can guarantee good precision.

(Buyukkokten et al., 1999) proposed an approach that was also utilized by Google's PageRank – the more pages link to a given website, the more important it is. However, as they were interested also in defining the scope of resources, they analysed locations of servers linking to these resources. Based on this approach they were able to confirm within their study that the New York Times has a more global reputation than the San Francisco Chronicle.

(Markowetz et al., 2004) introduced a concept of a *web page locality* to distinguish between sites of local and global importance. They propose to first assign pages to their geographical locations and then recalculate their geographical contexts measuring a distance between locations that were assigned to the linking pages and a given web resource. They also propose to distinguish between inbound and outbound locality, taking only in- or outgoing links into account. The authors admitted, however, that for most applications their approach would be an overkill.

(Martins et al., 2005a) divided the procedure of assigning the geographical scope of web resource into three steps. In the first step, all Geographical Named Entities are identified and

[21] http://www.ai.sri.com/dotgeo/

weights are assigned. In the second step, weights for all entities recognized in document are divided between all linking documents associated with it in the Web graph. This value is then assigned to the same entity in all linking documents. This propagation procedure applies only to pages being one hyperlink away from the source. Also heuristics are applied at this stage. For example, documents available on the same site are considered more likely to relate to the same geographical concept and therefore receive an extra credit in the weight propagation process.

### 3.4.3. PageRank algorithm

PageRank algorithm was developed by the originators of Google: Sergey Brin and Lawrence Page and is used as the Google's ranking method (however a number of modifications and extensions have been introduced since then). Brin and Page developed a mechanism similar to quantifying the academic citation literature metrics. They extended it by normalizing the number of links from a web page and used it to estimation of importance of web pages in the Internet. Simply, PageRank is the probability that a random user will visit a page starting from the given webpage and only following the available links. It is defined as follows:

**Definition 1. PageRank. Source: (Brin and Page, 1998)**
*We assume page A has pages T1...Tn which point to it (i.e. provide citations to A). The parameter d is a damping factor, reflecting the probability that at any step the person will continue following the link structure, which can be set between 0 and 1 (usually set to 0.85). Also C(A) is defined as the number of links going out of page A. The PageRank of a page A is given as follows:*
$PR(A) = (1-d)/N + d\,(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$

PageRank is calculated in an iterative way and corresponds to the principal eigenvector of the normalized link matrix of the web. It is worth noting that if the webpage is a sink (it does not contain any outbound links) then it is assumed to link out to all other pages in the collection and its PageRank is divided equally between all pages in the collection (Green, 2005). PageRank is a very efficient algorithm – its authors claim than a PageRank for 26 million web pages can be computed in a few hours on a medium size workstation. Since 1998 PageRank algorithm has been significantly modified. However, its current version is not available to the public, it is assumed that while estimating the importance of the resource for a given topic, it takes into account over 200 different features of a document.

### 3.4.4. HITS algorithm

Another algorithm dealing with computing the importance of the web pages in the Internet is HITS that is based on the idea of hubs and authorities for the WWW search topics (Kleinberg, 1999). After specifying a query to a search engine by a business user HITS applies two steps:

1. first step concerns searching for web pages containing the query terms the user is searching for – the full text search techniques are applied. The set of retrieved documents is used as a root set of pages (e.g. first 200 of results retrieved by a typical search engine e.g. AltaVista are taken into account).
2. In the second step – algorithm determines numerical estimates of hub and authority weights by an iterative procedure in order to provide users with a complete list of search results with proper weights assigned.

The problem is, that it is possible that not all of the authoritative web pages are in this set. Nevertheless, the assumption is that some of the pages describing the given topic link to them. Therefore, the root set of documents is expanded by including all pages that are linked to by pages in the root set and all pages that link to a page in the root set (up to a designated size cut-off). This makes set of documents grow to about 1000-5000 documents among which is a large number of pages that one would subjectively view as authoritative for the search topic. Then, all links between pages within the same domain from the sub graph induced by the base set are deleted. In the next step, weights for good hubs and authorities are calculated similarly to the simplified Page Rank algorithm. The problem with this algorithm is that after the initial keyword based query it disregards the query and works only on links.

### 3.4.5. CLEVER project

CLEVER, being the extension of HITS, deals with its weakness, namely with ignoring the textual content of pages after assembling the root set of documents. This is especially important for websites discussing different topics extended by various links, when HITS doesn't perform well (Chakrabarti, 1999). Therefore, CLEVER extends HITS by creating weights and assigning them to each link based on the query terms. Heuristics applied in order to compute these weights are as follows:

1. anchor text analysis,
2. division of large hub pages containing the number of links into mini-hubs or pagelets.

Moreover, the authors proposed not to delete links between pages within the same domain but solve the problem by decreasing their weights.

To summarize, the main difference between CLEVER/HITS and Google's PageRank is in the scope of outcome: while PageRank focuses mainly on authoritative web pages, the former two approaches also try to provide good hubs (pages providing links to various resources from the field).

## 3.5. Analysis of documents

From the point of view of this dissertation, more important than estimation of geographical importance of sources, is the analysis of documents' content. To represent a document one may define a document index. However, the index cannot just contain a the set of geographic names appearing in the document as for example the phrase "north of Paris" doesn't have the same meaning as "Paris" (Bilhaut et al., 2003) and needs to be resolved.

Moreover, not only geographical names appearing in the text are important, but also (McCurley, 2001) addresses and postal codes that may be resolved by using an external dataset associating addresses (especially zip numbers) with their coordinates as well as telephone numbers bearing also location-important information.

In the literature two types of approaches for extraction of geographical name mentions may be found, namely intra and interdocument approaches. Intradocument approaches cover all methods that enable extracting name mentions without using any of external resources containing geographical entities. Interdocument approaches use external resources – mainly gazetteers. These two approaches are elaborated in next sections.

### 3.5.1. Intradocument approaches for extraction of geographical information

The aim of the intradocument approach is extraction of geographical named entities without use of external resources. These approaches use techniques originating from artificial intelligence, e.g. neural networks, Hidden Markov Models, maximum entropy models as well as linguistic ones, e.g. grammars or part-of-speech tagging (Sang and Meulder, 2003).

The main issue when applying these techniques is that they should be developed in relation to a collection of documents they are to analyse. Before they are applied they must be tested, and improved if needed. All machine-learning approaches must be trained on the subset of documents from the collection and only then tested. An important issue may be also oversampling. If the collection of documents is rather similar and there is only a small subset of documents being entirely different than others, a sample for training and testing should be

constructed differently, i.e. not by representing equal proportion of each of document types in training and testing sets but overrepresenting the rare types of documents.

Other group of approaches concern linguistic analysis of documents by using grammars or part-of-speech tagging. These approaches, due to their importance for this dissertation, are elaborated in detail in the next chapter.

According to (Amitay et al., 2004), techniques based on machine-learning algorithms may perform better than interdocument approaches as they may spot also geographical name mentions from outside the available list of entities. The drawback is, however, that they end up with much more complicated algorithms. (Amitay et al., 2004) claim that "the most of the published algorithms do not employ machine learning, but are rather based on various NLP heuristics."

### 3.5.2. Gazetteer-based approaches

(Leidner, Sinclair, Weber 2003) define gazetteers as "large lists of geographic entities, usually enriched with further information, such as their class (e.g. town, river, dam), their size, and their location (i.e. with respect to some relative or absolute coordinate system such as longitude and latitude)." Gazetteers are also defined as geospatial dictionaries of geographic names (Hill, 2000). They enable to find a place name mention in the text and then translate it into coordinates. Therefore, essential information that should be stored in gazetteers on geographical location is name, footprint (spatial location) and type (Hill, 2000).

According to (Fu et al., 2003), contemporary gazetteers share a number of limitations that actually hamper their use in geographic information retrieval. Firstly, they do not encode spatial relationships apart from using region hierarchies (Amitay et al., 2004), (Harping, 1997). Secondly, generic relations between objects are also not stored. Moreover, they usually do not include all significant details needed e.g. previously used names. They also do not include geographic footprints for fuzzy defined places e.g. north of Paris, eastern Poland, etc. (Souza et al., 2005) add also to the drawbacks of gazetteers the lack of city landmarks, well-known locations used by citizens as reference points.

Researchers try to eliminate these deficiencies. According to (Kimler, 2004), more advanced gazetteers include information concerning importance of a place that may be useful in disambiguation of a place name. (Pouliquen et al., 2004) introduce *exonyms* to gazetteer – names used in certain language to describe places situated in other country (name quoted after (Piton and Maurel, 2001)).

An example of such an expanded gazetteer is GeoXwalk (Densham and Reid, 2003) that stores not only information on locations but also their detailed geometry. This gives the ability to derive relationships between features using geometric computation what leads to provision of more accurate results than can be ascertained by simple lookups based on hierarchical thesauri methods as in traditional gazetteers.

(Mikheev et al., 1999) proved that gazetteer-based techniques lead to significantly higher precision and recall values than intradocument approaches. Their research showed that precision/recall for gazetteer based approaches were about 90%, whereas for intradocument approaches they accounted to 46% precision and 59% recall.

(Krupka and Hausman, 1998) show that performance of an information extraction system does not change much when decreasing the number of entities in gazetteer from 25000 to 9000, however one can dramatically increase its role adding only 42 entries. Their experiment shows that there is no need to develop extensive gazetteer, but only gazetteer that contained the most important places.

(Mikheev et al., 1999) proved also that gazetteer doesn't have to contain all geographical name mentions. Small gazetteers containing only most important places for most of text collections work nearly the same - 91% vs. 92% recall. Experiment they held shows that dealing with geographical named entities is different than dealing with persons and organizations. While system working on gazetteers for geographic named entities presented quite good results, this was not the case for organizations and persons. Results of their research is presented in Table 6. Learned lists column presents results for a gazetteer built based on the MUC training set and applied to the rest of collection, common list depicts gazetteer built using publicly available resources (e.g. CIA Fact Book, financial web sites), while the third column shows results achieved using the combination of these two gazetteers.

**Table 6. NE recognition with single list lookup. Source (Mikheev, Moens et al. 1999)**

| Category | Learned lists | | Common lists | | Combined lists | |
|---|---|---|---|---|---|---|
| | **Recall** | **Precision** | **Recall** | **Precision** | **Recall** | **Precision** |
| Organization | 49% | 75% | 3% | 51% | 50% | 72% |
| Person | 26% | 92% | 31% | 81% | 47% | 85% |
| Location | 76% | 93% | 74% | 94% | 86% | 90% |

Analysis of documents using gazetteer approach is language-independent. When the language of the discourse changes – the only thing that should be changed is gazetteer (and when multi-

language gazetteer is applied nothing has to be changed) – and the efficiency should not be affected (Kimler, 2004).

Main problem with gazetteers concerns the fact that they only help with extraction of name mentions they store. Moreover, development and maintenance of such lists are very time consuming. There are approaches to automatic population of ontology and adding time stamps to entities, however, they are neither extensively tested nor widely used. Therefore, it seems that the best approach would be to combine intra and interdocument approaches.

There are a number of gazetteers existing that are used in research e.g. (Leidner et al., 2003, Kimler, 2004, Pouliquen et al., 2006). The simplest gazetteer is actually a section within atlases that can be used to look up a geographic name and find page and grid reference, where it is shown (Fu et al., 2003).

Examples of other more advanced gazetteers are as follows:

- GEONET22,
- KNAB23,
- UN-LOCODE24,
- Alexandria gazetteer (Smith et al., 1996)25,
- GIS WWW Resource List26,
- Gazetteer of Planetary Nomenclature27,
- Getty Thesaurus of Geographic Names28,
- NIMA GEOnet Names29,
- Place Name Servers30,
- U.S. Census Gazetteer31.

One of the most frequently used gazetteers is the Getty TGN (Thesaurus of Geographic Names) that was developed by the Getty Research Institute. It structures geographical

[22] http://earth-info.nga.mil/gns/html/

[23] http://www.eki.ee/knab/knab.html

[24] http://www.unece.org/cefact/locode/service/main.htm

[25] http://www.alexandria.ucsb.edu/gazetteer

[26] http://www.geo.ed.ac.uk/home/giswww.html

[27] http://wwwflag.wr.usgs.gov/USGSFlag/Space/nomen/nomen.html

[28] http://shiva.pub.getty.edu/tgn browser/

[29] http://164.214.2.59/gns/html/index.html

[30] http://www.asu.edu/lib/hayden/govdocs/maps/geogname.htm

[31] http://www.census.gov/cgi-bin/gazetteer/

thesaurus containing over 1 million of geographic feature names as well as other information on places. Each TGN record describes a place, its alternative names, hierarchical positions of the place in the physical and administrative world, coordinates. Figure 16 presents an exemplary record from the TGN gazetteer.

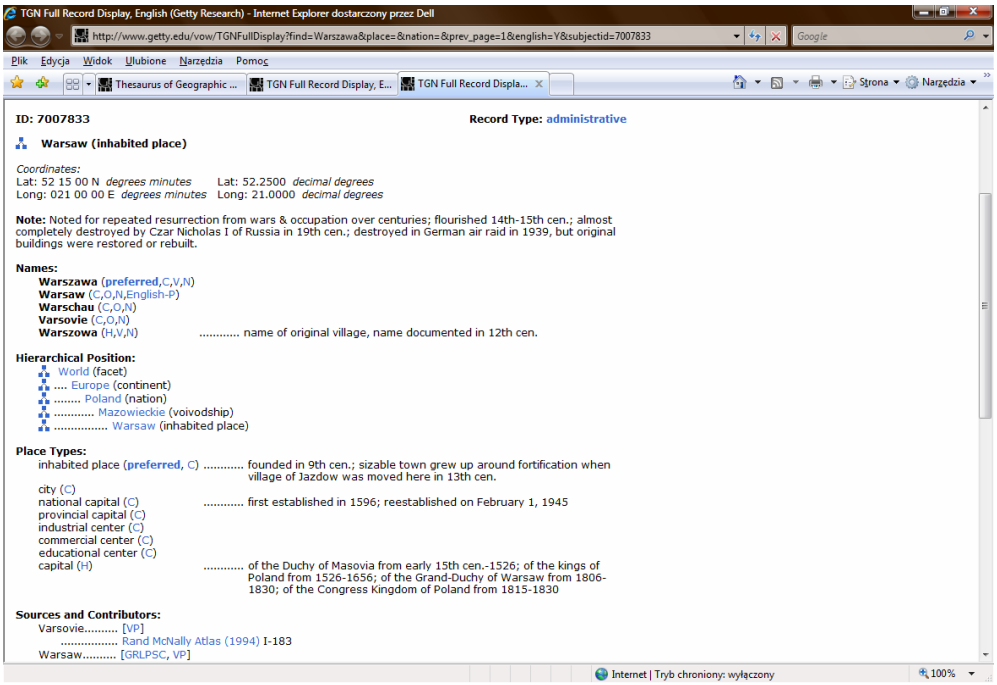



**Figure 16. TGN Record. Source: http://www.getty.edu**

## 3.6. Challenges for geoparsing and geocoding

The most common problem that geoparsing and geocoding have to tackle is the ambiguity of geographical place names. (Amitay et al., 2004) provide statistics for web pages according to which 37% of the potential geographic-name mentions have several possible geographic meanings. The average number of possible meanings for a single name mention is roughly 2. (Smith and Crane, 2001) report that in their corpus (of historical texts) 92% of all occurring names is ambiguous.

The abovementioned ambiguities concern both geo/geo as well as geo/non-geo ambiguities. (Amitay et al., 2004) define geo/non-geo ambiguity as the case when a place name has another, non-geographic meaning, e.g Friday (Texas, US), Mobile (Alabama, US) or Reading (England), Black (Montana, US). Some of the common English, French or German words are also place names. A list of such exemplary place names is presented in Table 7.

**Table 7. 10 out of 30 the most common English, French and German words being also names of places. Source: (Kimler, 2004)**

| English | | French | | German | |
|---|---|---|---|---|---|
| And | Ireland | De | Burkina Faso | Die | France |
| To | Ghana | Du | Ghana | Den | Etiopia |
| Be | India | Un | Russia | Zu | Zaire |
| By | Sweden | Une | Columbia | Ist | Hungary |
| Are | Nigeria | Est | The Netherlands | Im | Russia |
| His | France | Il | Iran | Dem | Cameroon |
| But | Afghanistan | Au | Austria | Als | Denmark |
| Had | Oman | Par | Great Britain | Auch | France |
| She | India | Sur | Oman | An | Mexico |
| We | Zaire | Pas | Turkey | Aus | Namibia |

The georeferencing algorithms have also to deal with place names being homonymic with people's names e.g. Victoria (Australia), Annan (UK), Jackson (USA), Ford (Ireland), George (RPA), Blair (Malawi) (Pouliquen et al., 2004, Kimler, 2004).

However, one of the biggest difficulties concerns dealing with disambiguation between several places with the same name, namely solving the geo/geo ambiguity, e.g. Victoria (being one of 200 different locations), Athens (capital of Greece and town in USA), Jerusalem (in USA there are 18 cities of this name). (Amitay et al., 2004) report that almost every major city in the Old World has its counterpart in the New World, e.g. London, Paris, Vienna, Berlin, Moscow, Cairo, Rome and Jerusalem. (Pouliquen et al., 2004) states that there are 244 places named Aleksandrovka, 199 named San Antonio or Santa Rosa, 144 San Franciscos and 18 Londons (Pouliquen et al., 2006).

In Gazetteer for Poland the most often occurring place is Zalesie (63) but here also a problem of ambiguity between different types of geographical entities emerges (e.g. city – commune ambiguity) (Abramowicz et al., 2006).

Moreover, there are also places having various names in different or even in the same language. (Pouliquen et al., 2004) recalls example of Saint Petersburg (being also Saint Pétersbourg, Sankt-Peterburg, Leningrad, Petrograd, etc.). However, in their publications authors seem to disregard the possibility of place names changing with the flow of history.

In order to improve georeferencing various disambiguation techniques were developed, e.g. taking into account at least two more parameters: the hierarchy of places and the geographical distance between two places (Pouliquen et al., 2004), maximum weight spanning trees (Li et al., 2003a), utilizing geocontext of the place (Ignat et al., 2003) and (Pouliquen et al., 2004). They are elaborated in detail in section 3.7.1.

## 3.7. Disambiguation of geographic name mentions

In extraction of geographic information two kinds of ambiguities exist, namely geo/non-geo and geo/geo and therefore different methods addressing these ambiguities were developed.

### 3.7.1. Disambiguation heuristics for geo/geo ambiguities

(Buscaldi and Rosso, 2008) divide the disambiguation of toponyms into three groups:

- map-based – methods using representation of places on the map, e.g. proximity of places heuristics,
- knowledge-based exploiting external knowledge sources such as gazetteers, Wikipedia or ontologies, e.g. superordinate mention heuristics e.g. relative importance of place heuristics,
- data-driven or supervised utilizing standard machine learning techniques.

The most important and most frequently described heuristics are shortly introduced in this section.

**Relative importance of a place**

This heuristic states that a place name that appears in text is likely to stand for an alternative with a high importance (capitals, major cities) (Kimler, 2004). Therefore, weights for named entities have to be introduced to a gazetteer. Kimler divides all places into classes, and each alternative is given a weight between 5 (small villages) and 80 (countries, administrative units, counties; capitals, major cities). The weight assignment is authoritative based on data acquired from different gazetteers and author's experience. More objective approach is presented by (Rauch et al., 2003) who use population numbers for assigning the weight for a place in the MetaCarta gazetteer.

**Context-based triggering**

(Kimler, 2004) claims that most of news articles have a geographical focus and defines this focus as a geo-context of text. Then based on a defined context the author disambiguates the

geographical name mentions. For all places being the geographical context of the text analysed, he increases their weight in the gazetteer. In his work he differentiates three indicators for the inference of a geo-context:

1. **place of publishing** – defined as a region (area) that is covered by articles published. Author assigns this context manually for each source that articles will be retrieved from. Sometimes the context can be created based on subtitles of the news' sources e.g. East Anglican Daily Times – The morning newspaper for Suffolk and Essex, or categories found on the web pages. However, the problem with this approach is that when no context can be set, it is just left empty. Moreover, the authors seem to forget that when it comes to global news agencies publishing news from all over the world where the context of the source cannot be precisely defined[32].
2. **place of writing** – is based on the assumption that for many articles published, a country they concern, is mentioned at the beginning of the text e.g. „Washington (CNN)…", „(Lagos) July 1, 2004…". (Kimler, 2004) analyses first 50 characters of the text (for each word a query in the gazetteer is performed) and if a name from the gazetteer is spotted, it is added to geo-context of a text. The problem of this approach is that it is not universal for newspapers written in other languages.
3. **shallow parsing of text** - where geo-context is set based on the most important places that appear in it. The text is parsed in order to find at least three references to one place. Alternatively, the country is set as a geo-context of the text, if references to it make up at least 50% of all references in the text.

**Comparison of a location to other places in the text (also called superordinate mention) heuristics**

Even if a country (commune, region) name is not in the geographical context of a news article, it may appear in the text. Using the gazetteer as the hierarchy of places it is possible to assign a city mention to the level (of more) above in the hierarchy and based on this solve the ambiguity (Pouliquen et al., 2004, Kimler, 2004).

(Martins et al., 2005a) in their ontology-based approach built a geographical ontology (being actually a hierarchy of places) based on the information from the GKB. When extracting a name mention from a document they activate each node in the ontology with the weight

[32] It would be worth to divide such a source into a set of sub-sources and for each of the sub-sources setting the context would be valid.

associated to the named entity spotted in the document. These values are then propagated across the ontological relationships between the entities, using inference methods from probabilistic graphical models. Their aim is to set up the geographical context for the web page, and therefore they are interested in assigning only the highest weighted entity as the scope (index) of the document.

**Proximity of places (geometric minimality and distance to unambiguous text neighbours)**

As geography is concerned with continuous space the number of approaches to disambiguation dealing with proximity of places (X near Y) were proposed.

1. (Kimler, 2004) proposes a heuristic based on the assumption of a news "epicentre". He states that news texts often have a place of occurrence, an "epicentre". To disambiguate other locations in the text, he calculates a distance between these locations and the "epicentre". He assumes that it's more likely that a place name refers to an alternative which is closer to that epicentre than to ones that are further away.
2. Spatial minimality heuristic proposed by (Leidner et al., 2003) assumes that when there are two or more name mentions in a discourse, the smallest region that incorporates all places mentioned is the one that gives them their interpretation. Figure below describes this situation in detail. Let's assume that mention of place A is ambiguous and refers either to A' or A". But our text refers also to F, E, G, so we assume that our A is actually A' because it leads to a smaller spatial context.

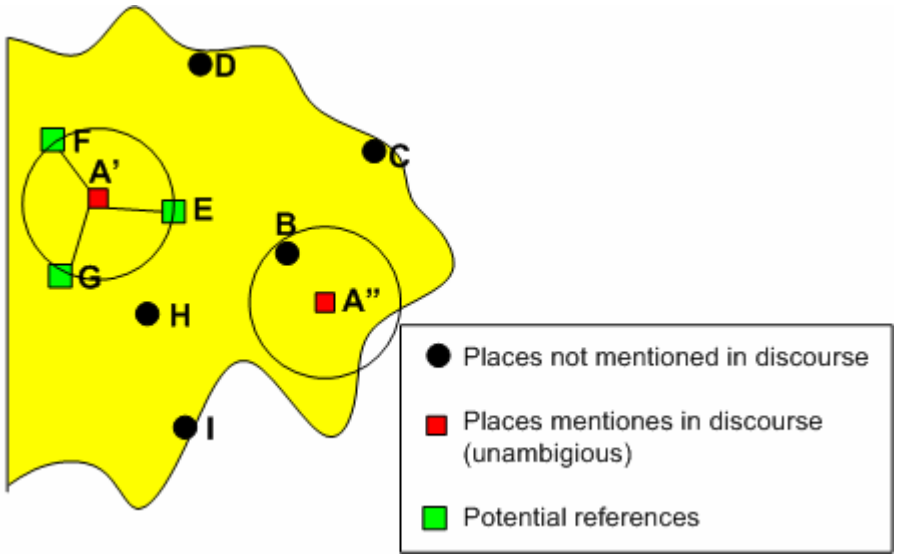



**Figure 17. Spatial minimality heuristic. Source: (Leidner, Sinclair et al.)**

3. (Pouliquen et al., 2004, Kimler, 2004) propose also disambiguation heuristics based on calculating an average distance between references appearing in text. They apply a measure for calculating distance between two locations on earth based on the work of (Sinnott, 1984).

**One referent per discourse**

This heuristic assumes that a place name mentioned in a discourse refers to the same location in the whole discourse unless there are other indicators that suggest a different understanding (Gale et al., 1992, Leidner et al., 2003, Gardent and Weber, 2001). This approach proved to provide a very good precision (nearly 1)(Leidner, 2007b).

**Additional information appearing in text**

(Wang et al., 2005a) for disambiguation use zip codes appearing within text. Similar approach is also suggested by (McCurley, 2001). (Rauch et al., 2003, Hauptmann and Olligschlaeger, 1999) for disambiguation use additional information stored in gazetteers.

**"Contained-in" qualifier**

This heuristic for disambiguation uses information appearing just after the phrase to be disambiguated e.g. London, UK or Warsaw (Poland). This approach provides precision of about 93% when disambiguating name mentions.

In order to achieve good results the above mentioned approaches are frequently combined.

(Li et al., 2002, Li et al., 2003a) present an algorithm consisting of five steps. First, only words in the text which are also listed in the gazetteer are filtered out. Then, NLP techniques are applied to exclude the non-geo terms (based on prefixes like e.g. "Mr.", "Mrs")). Next, the "single sense per discourse" principle is applied. After that one of the meanings of each ambiguous place name is selected using the heuristic minimizing the overall distances between all mentions. Names being unresolved after this step are assigned a default sense (based on the weights from gazetteer). The authors report 93,8% precision on news and travel guide data.

(Smith and Crane, 2001) changed the order of the first two phases – only potential geographical name mentions are being looked up in the gazetteer. They also introduce new approach to disambiguation – they minimize distance not between all places appearing in the article, but only between unambiguous and ambiguous ones.

(Li et al., 2006) propose to deal with implicit name mentions. When a text mentions e.g. North America, all implicit locations (based on the concept subsumption hierarchy) should be also added to the list.

(Amitay et al., 2004) propose a multi step approach to solve ambiguities they implemented for the Web-a-Where system. The procedure starts with spotting all place name candidates appearing in text. In this step abbreviations, such as IN (Indiana, India) or AT (Austria) are not spotted because they may also be a common English prepositions. In the second step disambiguation follows. Firstly, all tokens that can uniquely qualify a name mention are used e.g. if Chicago is followed by IL, then meaning of Chicago, IL is assigned with a confidence of 0,95-1,00. Then, for each reference the place with the biggest population is assigned with confidence of 0,5. For all spots of the same name the same meaning is assigned, with confidence in range of 0,8-0,9 depending whether the initially recognized meaning was the same as the current one (the application of a "single sense per discourse heuristic"). Finally, the disambiguation context for all of the remaining spots is being sought based on the already resolved place names. All of the places identified within this phase are given confidence of 0,65-0,75 depending on whether the initially recognized meaning was the same as the current one. In this approach however, the authors do not take advantage of applying information on the source of a document they analyse.

### 3.7.2. Solving geo/non-geo ambiguities

For disambiguation of entity types (Rauch et al., 2003) propose a confidence measure based on surrounding words e.g. community, college after a geo name, or city of…, mayor of.. (Amitay et al., 2004) propose to analyse ambiguous place names. For names that appeared more than 100 times and in the most of cases were not capitalized as a name should be, e.g. "Asbestos" (Quebec) and "Humble" (Texas), they include them in a stop list (the gazetteer's non-geo section). Moreover, (Amitay et al., 2004) exclude from further analysis names mentioned much more frequently than their population would suggest. "Grove" (Spain) and "Atlantic" (Iowa) were encountered in the corpora they analysed as capitalized most of the time, but their high frequency did not match their small populations (10,976 and 7,474, respectively). The authors seem to disregard the meaning of Atlantic Ocean, however the approach suggested seems to be working as desired.

(Pouliquen et al., 2004) propose to manually create lists of persons that frequently appear in media (sometimes also using automatic detection of person names) and using this list

similarly to Amitay's non-geo section of the gazetteer. However, they propose not to exclude these words definitely, but decrease their weights by a certain factor.

## 3.8. Conclusions

There exists a number of research prototypes attempting to estimate geographical locations of web documents by describing sources they come from as well as their content, e.g. systems such as Columbia GeoSearch (Gravano et al., 2003), Geotags GeoSearch[33], Kokono Search (Yokoji et al., 2001), SPIRIT (Purves et al., 2007).

However, a number of problems remains unsolved and the accuracy of search using these search engines doesn't exceed much the traditional keyword-based search or even performance of Yellow Pages-based systems. Combining linguistic analysis with the use of ontologies providing means for modelling fuzzy relations between places such as "near", "close to" could improve the process both from the point of view of document indexing as well as documents retrieval. The ontology-based approach would enable to store historical geographical information („political changes in the world moving faster than geological changes, and borders, country and region names, even the existence of political entities may change at any time" (Larson, 1996)).

The geographical index of a document should take into account not only document itself but also a source the document originates from. The spatial indexing should take into account such issues as:

- disambiguation of geographic name mentions;
- taking into account references to landmarks that are indirect geographical references , e.g. Buckingham Palace or Wawel;
- addressing anaphoric expressions (coreferences) such as "these regions" in order to improve precision and recall of indexing mechanisms.

Moreover, although a number of these methods were already implemented, none of them was tested on a collection of texts written in the Polish language.

[33] http://geotags.com/

# Chapter 4. Spatial indexing methods

*If I have ever made any valuable discoveries,*
*it has been owing more to patient attention,*
*than to any other talent.*

*Isaac Newton*

## 4.1. Introduction

This chapter provides a conceptual model of the spatial indexing method addressing goals defined within this dissertation. This method aims at enabling complex and precise description of geographical content of documents. The method is presented in the context of GIR as well as its role for PR experts is also underlined. This method is then further compared with a second simple, source-based indexing method developed as a response to requirements of efficient pre-indexing for the needs of information filtering.

This chapter is structured as follows. Based on the chapter 1, presenting motivation for research from the public relations point of view as well as chapters providing analysis of foundations from the domains of information retrieval, information extraction as well as geographical referencing, requirements for development of spatial indexing method are presented. They are followed by description of challenges emerging from the Polish language, for which the method is developed.

Then resources used for extraction i.e. gazetteer and ontology that were developed and are used by the method are shortly described. Next, a definition of a source for the needs of this dissertation is provided and second geo-indexing method is introduced and described in detail. The chapter concludes with a summary.

### 4.1.1. Requirements for the spatial indexing method

Definition of requirements is crucial when developing an information system (Maciaszek, 2007). The requirements should be defined using techniques such as writing use cases with customers, requirements workshops that include both developers and customers and demonstration of achieved results after iterations to gain feedback (Larman, 2004). Requirements may be also defined by extensive studies on business documents, however they need to be agreed with customers (Larman, 2004).

The requirements defined for this dissertation are twofold. On one hand they emerge from the public relations domain showing the need of support of PR experts in their work and enabling

practical application of results. On the other hand they focus on providing advantage over the previously described state of the art of the current spatial indexing techniques.

The public relations experts deal with constant monitoring of Internet sources in order to find information influencing a company image as soon as it appears. These sources are often predefined as not all sources existing in the Web are of the same quality and importance. PR experts are especially interested in sources such as portals of newspapers and other media, portals of regional institutions, forums, etc. as these sources change often and the need to react to messages published there appears. Moreover, PR analysts need to browse messages acquired in the process of source monitoring according to different criteria. An interesting parameter concerns geographical distribution of information acquired. Current search engines as well as tools used by PR experts do not enable them to perform a search according to different types of criteria (they have to provide one coherent query and it's up to the system how the query is resolved) and re-ranking of documents based on these types. These needs constitute requirements for the mechanism of spatial indexing from the PR point of view.

An extensive set of technical requirements for the geographical information retrieval system was presented by (Purves et al., 2007). Based on a set of mock-ups together with scenarios developed with use case partners, semi-structured interviews to collect information about user's views on functionalities of the system as well as a study on existing web-based systems, they defined functionalities and characteristics for the SPIRIT project. An interesting, similar overview of requirements for GIR was also presented by (Martins et al., 2005c).

Taking into account these requirements, the analysis of literature from the PR domain as well as workshops organized with the Department of Economic Journalism and Public Relations at Poznan University of Economics, the following requirements were defined for the public relations search engine:

- viewing results of query on the map, where points of the map should be linked to relevant web documents,
- ability to query for geographic areas which boundaries are not precisely defined but are usually identified by a name,
- ranking of documents depending on their spatial and thematic relevance for a query,
- ability to change ranking of documents based on preferences regarding criteria specified (time, geography, object),
- handling of different names for the same location,

- support for defining high quality sources that should be always presented on top of ranking (for internal search engines),
- improvement of quality of information retrieved in terms of recall and precision.

Based on these requirements, functionalities of a PR search engine were defined and are presented in the Use Case diagram below (Figure 18).

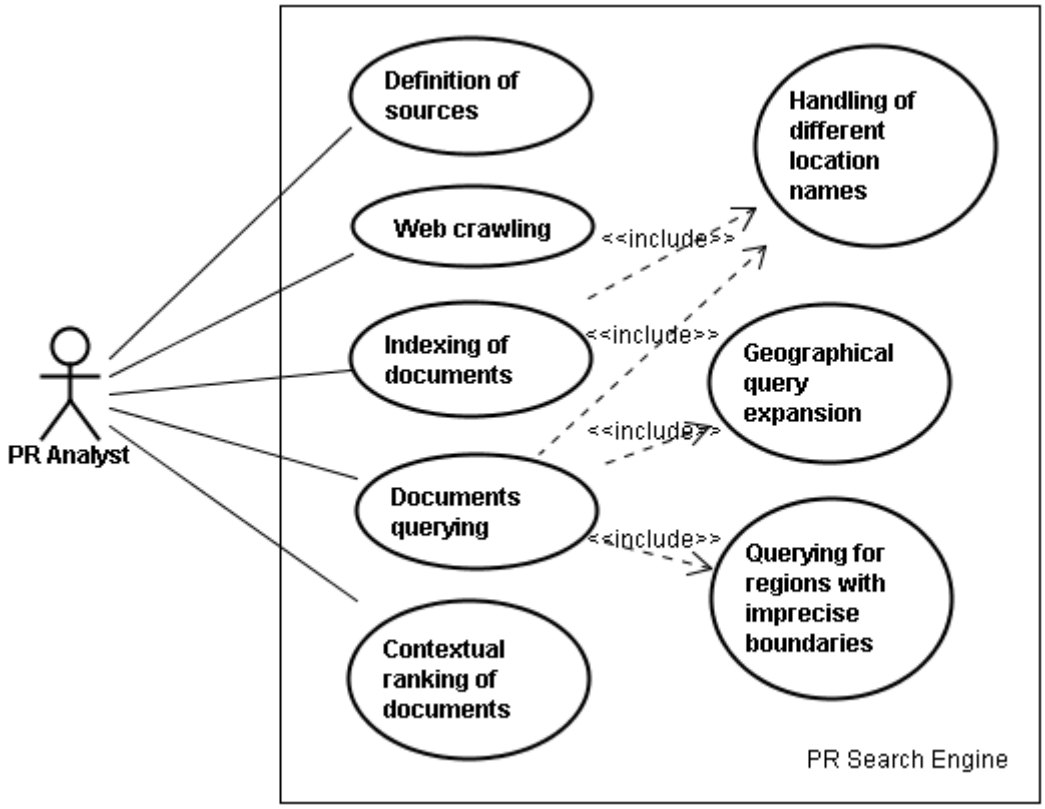



**Figure 18. PR Search Engine Functionalities. UML Use Case Diagram**

These requirements show that users are interested in extending a "typical search" by imposing different categories on content retrieved. They also would like to extend a search engine interface by provision of map where documents would be seen as a small icons. The challenge is in provision of a document index that would provide not only keyword such as a place name e.g. Gdansk, Poznan, but also coordinates of this place. Then it could be easily linked to a map using e.g. Google Maps[34] or Zumi[35].

[34] http://maps.google.com/

[35] http://www.zumi.pl/

Therefore, in this dissertation we focus on requirements towards the spatial indexing mechanism that would enable implementation of a new kind of spatial search engine. The following table presents the requirements gathered.

**Table 8. Requirements for the Spatial Indexing Mechanism**

| | |
|---|---|
| **Functional** | • delivery of an ontology-based spatial indexing mechanism for news articles retrieved from monitored Web sources, which includes:<br>o mechanism for extraction of geographical named entities for the Polish language,<br>o heuristics used for extraction of named entities taking into account specifics of the Polish language,<br>o disambiguation of place names (including geo/geo and geo/non-geo disambiguation),<br>o identification of geographical context of web resources. |
| **Nonfunctional** | • index should be the most precise document surrogate<br>• development of IE resources for Polish (gazetteer; type hierarchy taking into account specific of administrative division of Poland; extraction rules, corpus of documents) for mechanism evaluation<br>• development of geographical ontology for Poland |

Regarding the spatial indexing mechanism, there is only one functional requirement: to provide a spatial indexing mechanism for documents retrieved from the Internet. However, this requirement decomposes into a few more specific ones dealing with all components and methods that need to be developed in order to assure producing of a correct and detailed geographical index of a document. Each of these requirements is addressed in details in this chapter of the dissertation.

### 4.1.2. Context of the spatial indexing

The previous section presented requirements for the public relation search engine emphasizing a need for a proper spatial indexing mechanism. We claim that in order to provide a search engine for the public relations domain it is essential to provide spatial indexing mechanism that may extend or exchange currently existing indexing mechanisms

within search engines. Therefore, this section shortly presents an architecture of a search engine on the example of the first version of Google (the last publicly available description) showing place for introduction of a spatial indexing mechanism. Then the SPIRIT search engine (Purves et al., 2007) is also shortly elaborated.

The first phase of operation of every search engine is web crawling i.e. obtaining documents from the Web. In Google (see Figure 19) the web crawling is performed by distributed crawlers using URLs provided by the URL Server. Then all retrieved web pages are sent to the Store Server which compresses them and stores in a Repository. The indexing of documents is performed by the Indexer and the Sorter. The Indexer reads the Repository, uncompresses documents and parses them. At this step a document is converted into a set of word occurrences called hits. Hit represent a word, its position in a document, font size as well as its capitalization. Then Indexer distributes hits into a set of Barrels creating a forward index. Indexer also retrieves URLs from documents and stores the information in an anchor file. The URLresolver based on the information stored in an anchor file produces absolute URLs and a database of links, which is used to compute PageRank for all documents.

The Sorter takes the Barrels and generates the inverted index for documents. The Sorter also generates a list of wordIDs and offsets into the inverted index. Then, a program called DumpLexicon takes this list and lexicon produced by the indexer and based on them generates a new lexicon that is to be used by the Searcher. The searcher uses this lexicon, inverted index and PageRanks to answer user queries (Brin and Page, 1998).

This description of a search engine shows in general how a search engine works. Precision and recall depends here mainly on a quality of indexing. Our mechanism could therefore extend the existing approach by providing additional information while building index as well as constructing the lexicon, especially that is aims at producing index being a taxonomy of geographical entities over a set of documents.

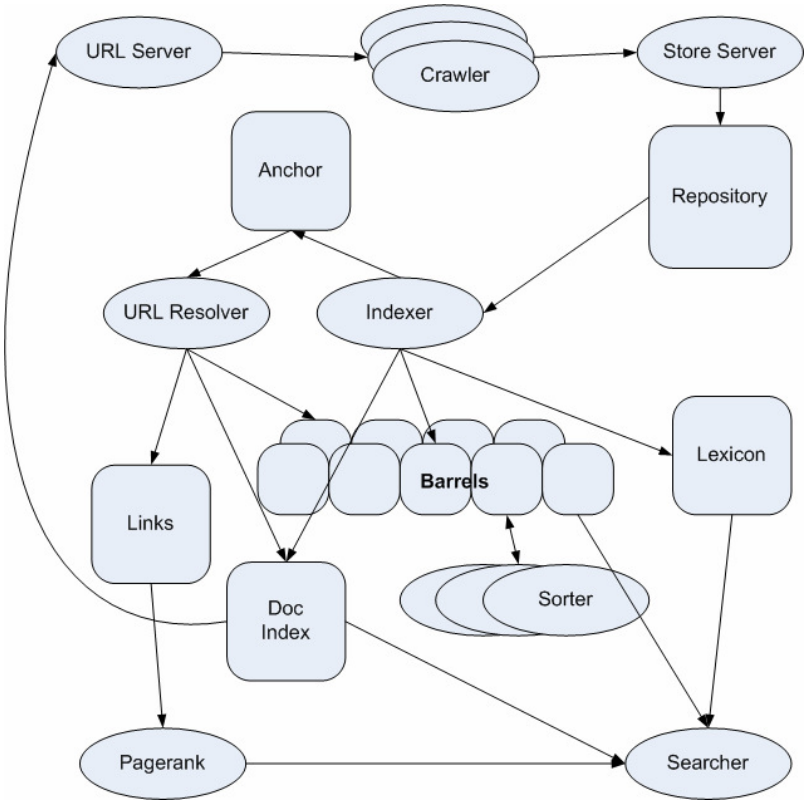



**Figure 19. Google Architecture. Source: (Brin and Page, 1998)**

In SPIRIT, the architecture of a search engine is similar to the one presented above. However, as the aim of SPIRIT is to deal with geographical queries few new components were introduced and should be described. SPIRIT operates on geotagged web resources extracted from a collection of 94 million web pages. Geographical tags added to these documents represent geographical places which are then associated with their geographical footprint by the metadata component. This is used both in generation of a spatial index and while producing a spatially-aware relevance ranking. Spatial indexing extends text indexing by associating documents with one or more cells of a subdivision of geographic space (Purves et al., 2007). The search engine uses indexes in order to answer user queries. The retrieved documents are firstly scored only based on textual matching. Then the geography relevance-ranking component provides methods for ranking of retrieved web resources using different types of spatial relevance scores. Geographical ontology is here used as a repository of knowledge on place names and their relationships. This data is used to recognize presence of names in web resources and to ground them to geographic locations. It is also used to provide geographical footprint of a query submitted by a user.

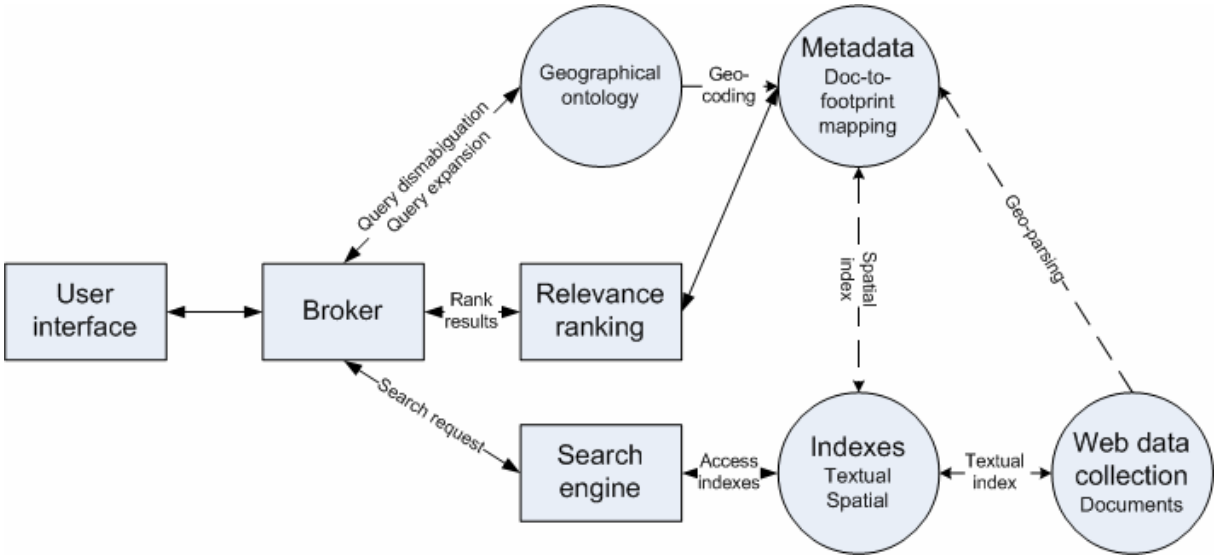



**Figure 20. SPIRIT architecture including run-time and pre-processing components and links. Source: (Purves et al., 2007)**

From our point of view the most important component of this architecture is the one dealing with geotagging of documents. In SPIRIT for geotagging a simple gazetteer lookup was utilised. It applies context rules, list of proper names and commonly occurring terms to filter out non-geo ambiguities. The accuracy of this mechanism was around 72% while 25% of false positives for all annotations were found (Clough, 2005). The grounding mechanism achieved the accuracy of about 89%.

Our work is to deliver a method for geotagging and geocoding that could be considered as an alternative in such a geographical search engine for the Polish language.

### 4.1.3. Extraction of information from documents in Polish - challenges

Information extraction to achieve its goals needs to deal with syntactic processing of various languages differing from one another. The spatial indexing method developed as part of this research deals with addressing the Polish language.

Polish belongs to a family of Slavonic languages, and its linguistic phenomena make it harder to process than Germanic or Romance languages. Therefore, although being a relatively large language, linguistic resources and tools supporting extraction are still in an early phase of development (Przepiorkowski, 2007, Piasecki, 2007).

(Collins et al., 1999, Przepiorkowski, 2007) identified the following characteristics of Slavonic languages that make them difficult for automatic processing, namely:

- rich nominal inflection – Polish nouns apart from inflecting for number (singular and plural), inflect also for 7 cases. Moreover, besides Morfeusz (Wolinski, 2006)

containing very few proper names, morphological treatment is not available for Polish proper names and it is not possible to lemmatize proper names, either.

- free word order – that enables specification of information in many different ways,
- idiosyncratic inflection of Slavonic proper names - (Piskorski et al., 2008, Piskorski, 2002) discuss that many Polish names have the same base forms that common names. This constitutes the problem not only in recognizing these names but also these names may have different gender values and different inflection.
- unstable inflection of some foreign names – this problem however being governed by strict prescriptive rules of the Polish language, native speakers (even journalists) produce many variants of these names in written texts. Moreover, inflection of foreign names depends also on their pronunciation connected with their origin.
- high degree of trans- and intraparadigmatic syncretisms – rich inflection (2 numbers, 7 cases, 5 genders) makes the size of the tagset very large. Such detailed tagsets make achieving accuracy in syntactic processing difficult, if not impossible. Moreover, rich inflection leads also to a high number of syncretisms.
- quirkiness of numeral phrases – in Polish the case of the noun depends on the numeral and position of the numeral phrase in a sentence. Secondly, the numeral phrase in Polish does not agree with the verb – as the verb occurs in the default 3$^{rd}$ person singular neuter form. Moreover, sometimes numeral phrase triggers atypical agreements with predicate adjective, which phrase is difficult to learn from the corpus of documents as being rare.

Other issue concerns coreferencing. In Polish one should deal not only with resolving name aliases and pronouns and noun phrase lemmatization, but also has to deal with zero-anaphora (e.g., in Polish it is common not to include a subject in a sentence as this information can be derived from a specific verb form) (Abramowicz et al., 2006).

As it may be easily noticed development of information extraction methods for Polish causes a number challenges for indexing method dealing with analysis of unstructured text documents. These challenges are addressed while development of spatial indexing mechanism within this dissertation.

## 4.2. Resources to support the spatial indexing methods

### 4.2.1. Gazetteer

Gazetteer together with an ontology and its knowledge base constitute the main resources used in this research. Therefore, a significant number of new entries related to locations were added to the existing SProUT gazetteer. For a part of the entries, morphological variants have been produced and tailored to the modified hierarchy of named-entities. This was done in a semi-automatic manner by application of general inflection patterns and manual correction of erroneous entries. Since extensive gazetteers are indispensable, a new compression technique for storing gazetteers has been developed, which is described in detail in (Piskorski, 2005b).

Currently, our gazetteer contains a hierarchical view of the world, divided (in our current implementation) into continents, countries, states (for some countries), and cities. This hierarchy besides providing names for places, also associates each geographic entity (i.e., place) with a canonical taxonomy node. It is an abstraction over the geographical ontology as it stores only the place name mentions (without attributes), what is enough for extraction of information from documents. Hence, for disambiguation and spatial indexing the geographical ontology is used. It was proved by (Martins et al., 2005c) that the use of "flat" gazetteer "can suffice for tasks involving the recognition of geographical references, but other Geo-IR tasks require additional information".

Table 9 summarizes the gazetteer resources. The type of entries for which morphological variants were created are marked with an asterisk. While most of the entries are from Polish, some English names have been incorporated as well, since they appear in Polish texts frequently. Further, location entries are typically enriched with an additional attribute that reflects the administrative division of the country.

**Table 9. Gazetteer resources**

| Type | PL | ENG |
|---|---|---|
| city* | 155810 | 1006 |
| commune | 2489 | - |
| county | 375 | - |
| province* | 16 | - |
| country* | 1763 | - |
| region* | 488 | 52 |
| river | 283 | 43 |

| | | |
|---|---|---|
| sea | 69 | - |
| lake | 48 | - |
| zone | 23 | - |
| facility | 68 | - |
| given-name | 1796 | 16714 |
| surname | - | 13376 |
| postion | 530 | - |
| facility | 68 | - |
| company | 262 | 91 |
| org-government | 60 | 83 |
| org-education | 21 | 1276 |
| org-recreation | 12 | - |
| org-other | 119 | - |
| other | 402 | - |
| **Total** | **164634** | **19265** |

According to (Li et al., 2002) in Tipster Gazetteer a location entry has on an average 1,39 senses and around 19% of the location entries have at least one meaning. In Polish over 12,9% of city name variants are morphologically and semantically ambiguous. For instance, there are about 70 cities and villages named *Zalesie*. Another complicacy is caused by the name convention for counties. 21,1% of county names are also city names (urban counties), where each of these cities is also a capital of land county whose name is an adjectival form of a city. The most common ambiguity types in our gazetteer are summarized in the table below.

**Table 10. Gazetteer ambiguities**

| Ambiguity type | Frequency |
|---|---|
| city-commune | 2082 |
| city-given-name | 204 |
| county-city | 61 |
| county-commune | 56 |
| county-city-commune | 55 |
| city-river | 44 |
| city-country | 33 |

| | |
|---|---|
| city-region | 21 |
| company-city | 19 |
| company-city-commune | 13 |

Below we present exemplary records from our gazetteer.

The following records present definition of cities. It may be easily noticed that also a taxonomy based on a administrative division of Poland is provided in the gazetteer.

```
Koszalin | GTYPE:gaz_city | G_GMINA:Koszalin miasto | G_POWIAT:Koszalin |
G_WOJEWODZTWO:zachodniopomorskie | G_CONCEPT:Koszalin miasto
Sławno | GTYPE:gaz_city | G_GMINA:Sławno | G_POWIAT:sławeński |
G_WOJEWODZTWO:zachodniopomorskie | G_CONCEPT:Sławno
```

Below descriptions of communes and counties follow.

```
Wronki | GTYPE:gaz_gmina | G_POWIAT:szamotulski |
G_WOJEWODZTWO:wielkopolskie | G_CONCEPT:Wronki
Wróblew | GTYPE:gaz_gmina | G_POWIAT:sieradzki | G_WOJEWODZTWO:łódzkie |
G_CONCEPT:Wróblew
```

```
lubartowski | GTYPE:gaz_powiat | G_WOJEWODZTWO:lubelskie |
G_CONCEPT:lubartowski
łukowski | GTYPE:gaz_powiat | G_WOJEWODZTWO:lubelskie | G_CONCEPT:łukowski
zamojski | GTYPE:gaz_powiat | G_WOJEWODZTWO:lubelskie | G_CONCEPT:zamojski
```

Also countries with their inflection are included.

```
Afganistan | GTYPE:gaz_country | G_CASE:acc_nom | G_CONCEPT:Afganistan
Afganistanu | GTYPE:gaz_country | G_CASE:gen | G_CONCEPT:Afganistan
Afganistanowi | GTYPE:gaz_country | G_CASE:dat | G_CONCEPT:Afganistan
Afganistanem | GTYPE:gaz_country | G_CASE:ins | G_CONCEPT:Afganistan
```

### 4.2.2. Type hierarchy

For the purpose of our work we also developed a fine-grained named-entity (NE) hierarchy (Kabra et al., 2005), which is a blend of results of previous endeavours in this area, including the work on NE taxonomies presented in (Sekine et al., 2002; Chinchor, 1998; Doddington et al., 2004). This NE hierarchy was then utilized to define the annotation guidelines and ontology developed within the scope of this work.

Clearly, location is the most structured category within the NE hierarchy developed. It groups together entities, which are relevant for geo-indexing and can be mapped onto geographical coordinates (e.g. facilities, water bodies, land forms, etc.). Particularly, administrative subcategory conforms to the geo-political division of Poland, which is organized into provinces, counties, and communes. The category product is motivated by the fact that product names often include valuable clues such as brand and company names, which can be utilized for inferring organization names and implicit location names. The category person groups named mentions of persons that are identified only via their first and/or second names or people named after a country.

Each main category is subdivided into eventually non-disjoint subtypes. An excerpt of the instantiated entity hierarchy is given in Figure 21.

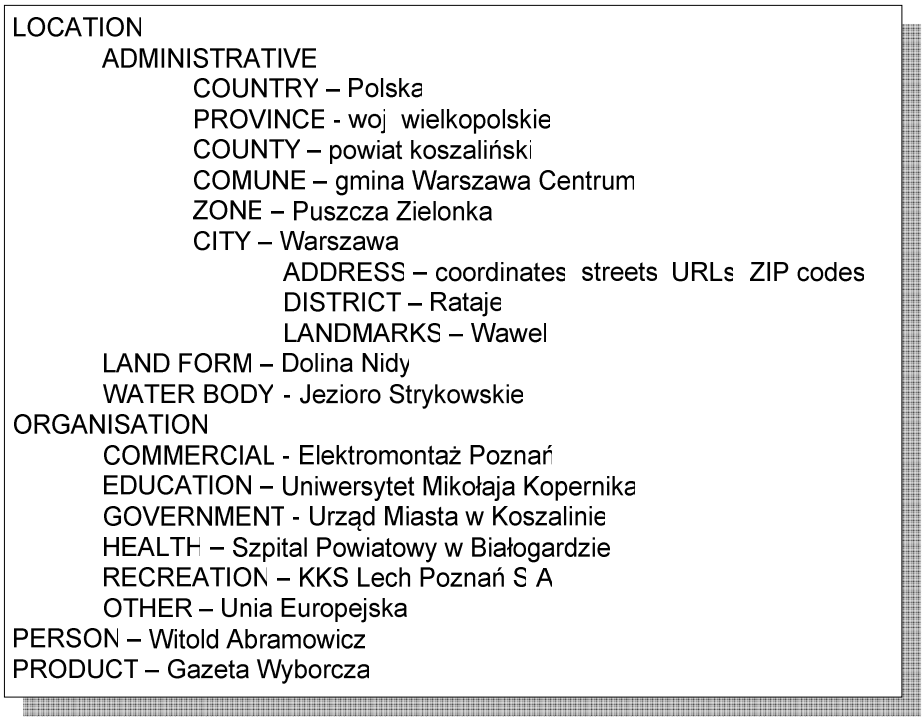



**Figure 21. Named entity hierarchy with corresponding examples**

This hierarchy is represented in SProUT (please see figure below) and used when developing rules and extracting information from documents.

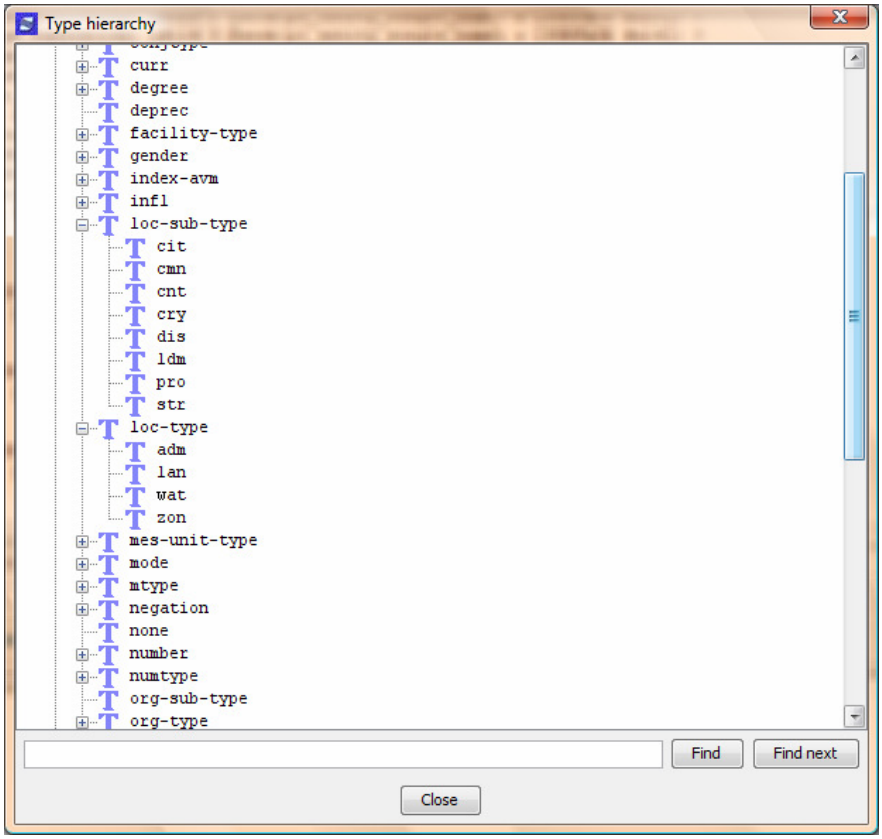



**Figure 22. Visualization of the type hierarchy in SProUT**

### 4.2.3. Geographical ontology

Geographical ontology provides an extension over the gazetteer and the type hierarchy described in the previous section. In order to provide an efficient spatial indexing method, we decided to take advantage from the additional information that may be associated with spatial objects. This information includes spatial footprints, time stamps, number of inhabitants, etc. as well as relations between objects. It was already described that the most appropriate way to store this kind of information is within an ontology. Designing ontology we decided to follow the application-specific approach, however the ontology developed may easily be applied in other solutions as well.

When defining requirements for our ontology, we took into account the work of (Fu et al., 2003). The requirements defined for this research with regard to the geographical ontology are as follows:

- functional:
  - providing information on place names enabling for recognition of presence of a place name in a document,

  - providing alternative place name given in a document (as one place may be described using few different names because of historical issues, language, etc.),
  - providing place names that are in relation to other place names specified (e.g. cities situated in the same commune),
  - specifying a type of the place spotted in a document (e.g. city, country, landmark),
  - associating geographical footprints to place names,
  - defining a time stamp of a place (as administrative division of the country may evolve in time, it is important to store information when an instance of a concept is defined);
- nonfunctional (quality requirements):
  - consistency – all defined concepts should be on the same level of granularity, the ontology should be consistent with other geographical resources including gazetteer as well as with type hierarchy defined within SProUT,
  - operational – available in an formalism for which scalable repositories, reasoning support, APIs, and tools are available,
  - reusability – an ontology should be applicable also for other tools and mechanisms,
  - clarity – readable, well structured and easy interpretable.

Taking into account all these requirements, we developed a geographical ontology presented in Figure 23. The ontology was developed in WSML-Flight[36]. A description of the ontology follows.

The ontology was developed using the WSMO Studio[37]. For the reader's confidence and as it is not possible to visualize ontology in WSMO Studio including all relations and properties, we present this ontology using an UML class diagram, where UML classes correspond to WSML concepts, UML associations to object properties, UML inheritance to is-a relations and UML attributes to properties of concepts. Such an approach for ontology visualization was proposed for OWL ontologies by (Brockmans et al., 2004).

The proposed geographical ontology (see Figure 24) consists of two structures: location structure providing taxonomy of geographic entities and artifacts defining classes needed for definition of geographical concepts.

[36] http://www.wsmo.org/wsml/

[37] http://www.wsmostudio.org/

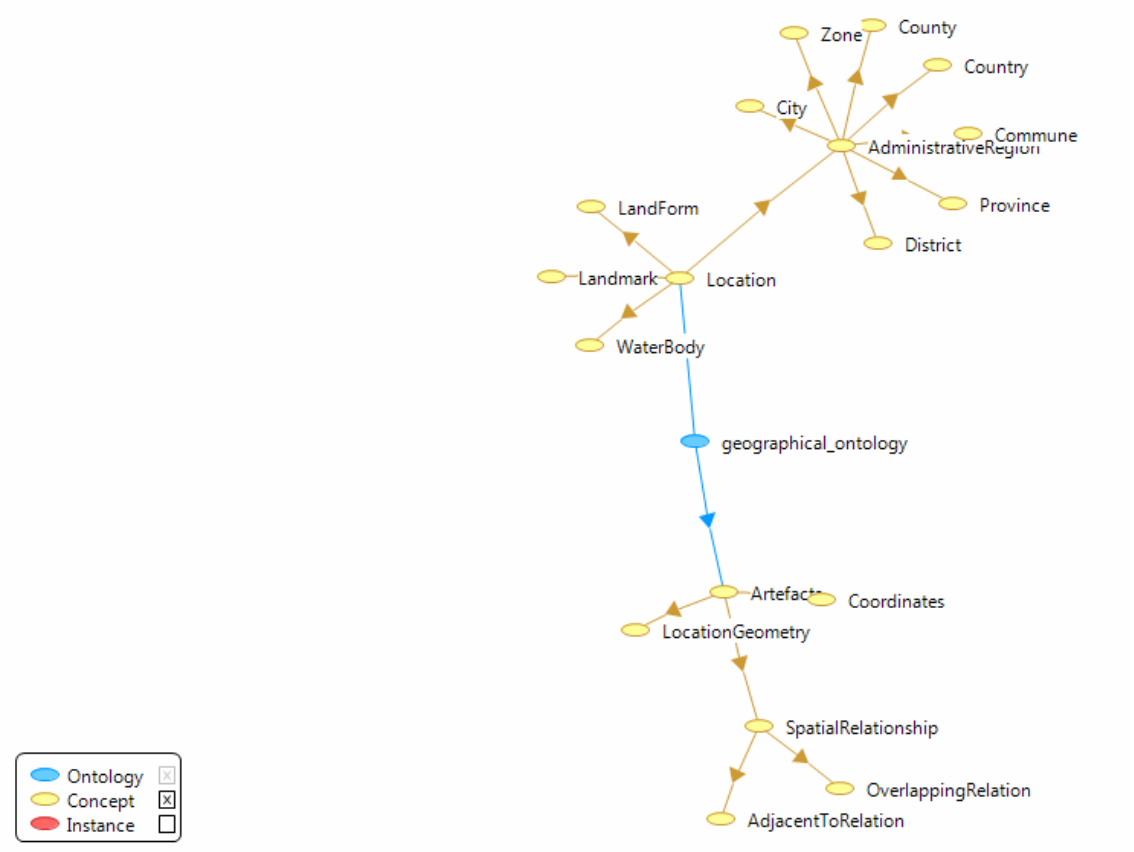



**Figure 23. Geographical ontology visualization (developed using WSMO Studio)**

Location tree presents a taxonomy of geographical entities similarly to the type hierarchy proposed in the previous section. This enables for definition of relations between different instances of cities, counties or provinces. Location has attributes enabling for cataloguing variant names of concept instance, area of the entity, its spatial footprint and associating a respective time stamp. Moreover, it may also have an overlap or adjacent relation with other instances.

The structure of artefacts includes a definition of a location name, location geometry (coordinates) and coordinates.

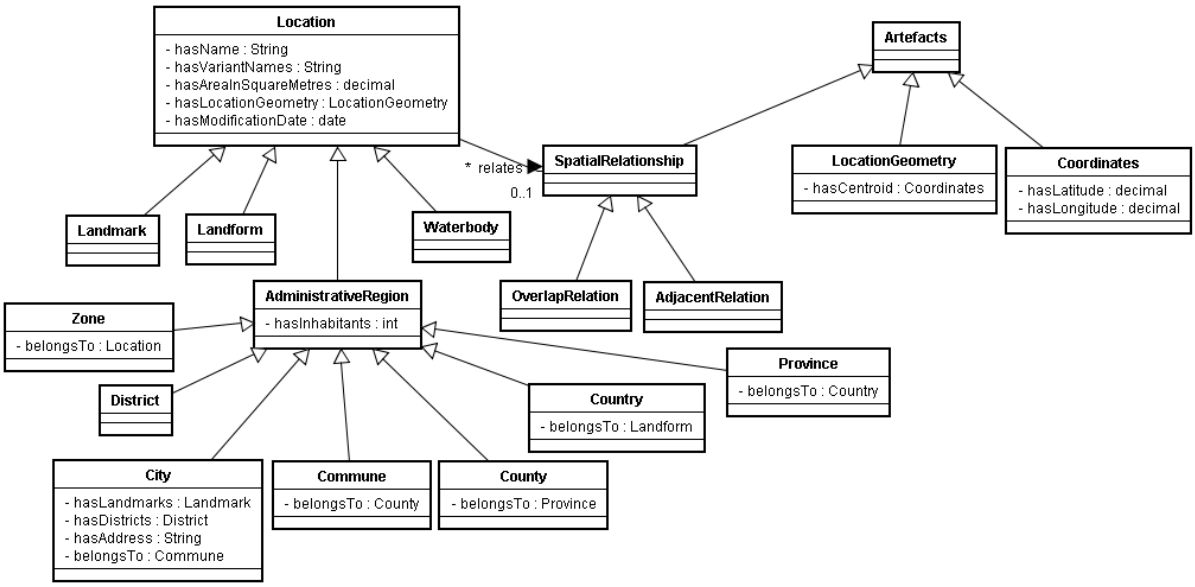



**Figure 24. Geographical ontology. UML Class Diagram**

Definition of a location concept from the ontology is presented in the listing below. Annex 2 presents geographical ontology in detail.

```
concept Location
hasName ofType (1) LocationName
hasVariantNames ofType (0 1) LocationName
hasDescription ofType (0 1) _string
hasAreaInSquareMetres ofType (0 1) _float
hasGeometry ofType LocationGeometry
hasModificationDate ofType _date
```

Another issue concerns a definition of non-administrative relations between places e.g. near Poznan, north of Warszawa, etc. For such relations (Fu et al., 2003) propose two paths that may be followed. One of them is to include such regions in the geographic ontology specifying their boundaries. This approach however would make the ontology unmanageable and it is not possible to predict every possible type of proximity. Another proposal concerns having an interactive dialogue with the user when such a phrase appears in the query (and disambiguate it based on his answers). However, there is also a third possibility to solve this issue i.e. to explicitly say that near means the circle around a place of a diameter of 50km, etc., what can be done using coordinates stored in the ontology.

Measuring the ontological distance within an ontology hierarchy may also be implemented using the measure defined by (Jones et al., 2001). They define the Hierarchical Distance Measure (HD) between a query place *q* and a candidate place *c* as follows:

$$HD(q,c) = \sum_{x \in \{q.PartOf - c.PartOf\}} \frac{\alpha}{L_x} + \sum_{y \in \{q.PartOf - c.PartOf\}} \frac{\beta}{L_y} + \sum_{z \in \{q,c\}} \frac{\gamma}{L_z}$$

The $L_x$, $L_y$, $L_z$ values represent the hierarchical levels of the individual places within their respective hierarchies. The set of places *x* includes these distinctive super-parts of the query term that belong to the query but not to the candidate. Set of places *y* include the distinctive super-parts of the candidate, that are not included in the query. The *z* set of places includes terms from the query and the candidate terms. The sets of terms *q.PartOf* and *c.PartOf* refer to the transitive closure of the super-parts of *q* and *c* in the part-of hierarchy. $\alpha$, $\beta$ and $\gamma$ are weights that enable making candidates that are sub-parts of the query more (or less if required) similar to the query than are candidates that are super-parts.

(Jones et al., 2001) also define a total spatial distance measure defined as follows:

$$TSD(q,c) = w_e ED(q,c) + w_h HD(q,c)$$

where $w_e$ and $w_h$ are weights of *ED* and *HD*, and *ED* is the Euclidean Distance Measure between two places.

In this research we benefit from ontological relations between places when estimating weights for geographic references included in the document index. We use method based on the TF*IDF approach (Baeza-Yates and Ribeiro-Neto, 1999) extending it to work with the ontology-based index.

The geographical ontology developed was populated using data from Główny Urząd Statystyczny (GUS)[38] and Wikipedia[39]. GUS provides the taxonomy and the main statistical data for places in Poland. This data was further detailed using data from Wikipedia infoboxes that are relatively easy to parse, but do not always contain precise and updated information.

The exemplary city description that may be found in Wikipedia is presented in Figure 25.

[38] Central Statistical Office for Poland, http://www.stat.gov.pl

[39] http://pl.wikipedia.org/wiki/Strona_g%C5%82%C3%B3wna

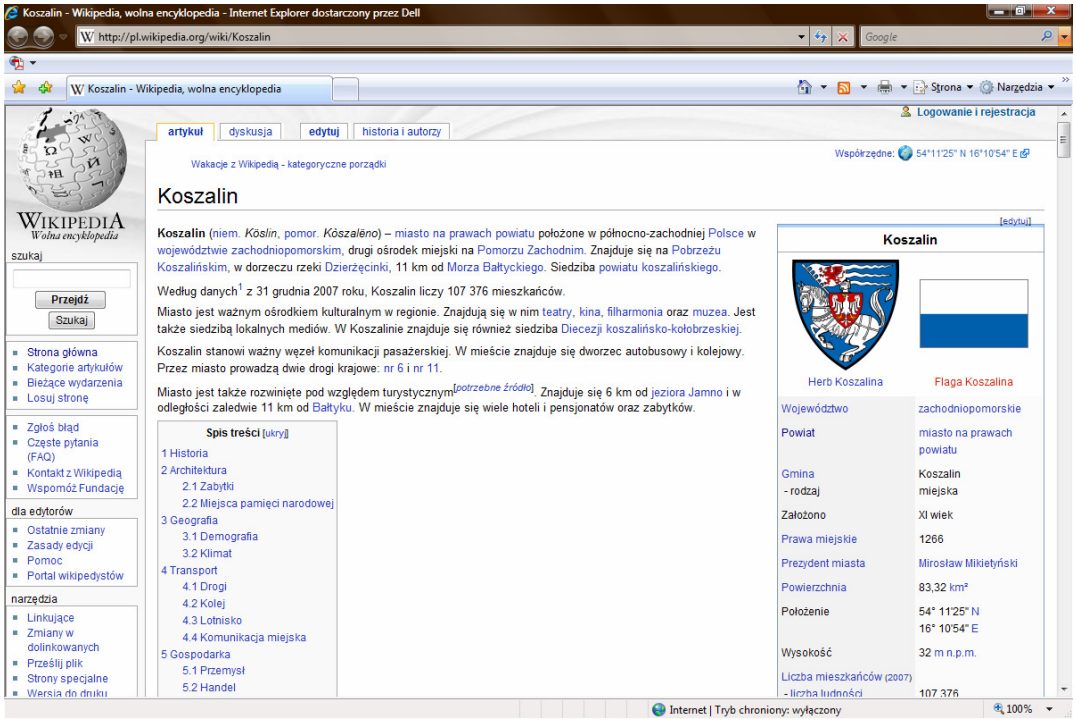



**Figure 25. Information on a given city in Wikipedia.**

**Source: http://pl.wikipedia.org/wiki/Koszalin**

The geographical ontology developed is to be validated against the following set of competency questions. The list presented below includes the most important competency questions from the point of view of requirements defined for the ontology. These competency questions are to be further formalized for the needs of automated ontology validation.

1. What is a name of a given location (e.g. city or county)?
2. What is an alternative name of a location?
3. What cities are situated in a given commune?
4. What communes belong to a given province?
5. What is a type of the place spotted in a document?
6. What is the geographical footprint of the location?
7. What is the time stamp and language variant of a location?
8. What is the area of a given commune?
9. What is the name of the landmark?
10. What is the name of the city the landmark is situated in?

Additionally, the ontology was validated by instantiation of all concepts in the ontology and using it for spatial document indexing, where the results achieved were compared with manually created document indexes.

## 4.3. Definition of a source

Public relations analysts while monitoring Web sources try to find information relevant for a company being subject to an ongoing PR process. They are interested mainly in sources that frequently update their content. Among these sources are web information portals, websites of newspapers and other media as well as blogs. These sources, usually provide also information about all updates syndicating content to interested parties in the RSS format.

The RSS format is an XML-based format for syndication and subscription of information (Young Geun et al., 2008). RSS stands for Rich Site Summary or Real Simple Syndication (Richardson, 2005). However, RSS is not a single format. The first RSS 0.9 was designed by Netscape in 1999. It was envisioned as a format for building portals of headlines to mainstream news sites (Pilgrim, 2002). However, it was too complex for goals in comparison to its goals, therefore its simpler version, namely 0.91 was proposed. This version was picked by UserLand Software that intended to utilize this format while developing webblogging products and other web-based software. At the same time, a RSS-DEV Working Group, based on RSS 0.9 and RDF developed RSS 1.0 specification. UserLand however, didn't accept this format and evolved 0.91 version producing versions 0.92, 0.93, 0.94 and finally 2.0. These standards, so RSS 1.0 and 2.0, are two versions currently available for content syndication technologies (Pilgrim, 2002, Young Geun et al., 2008).

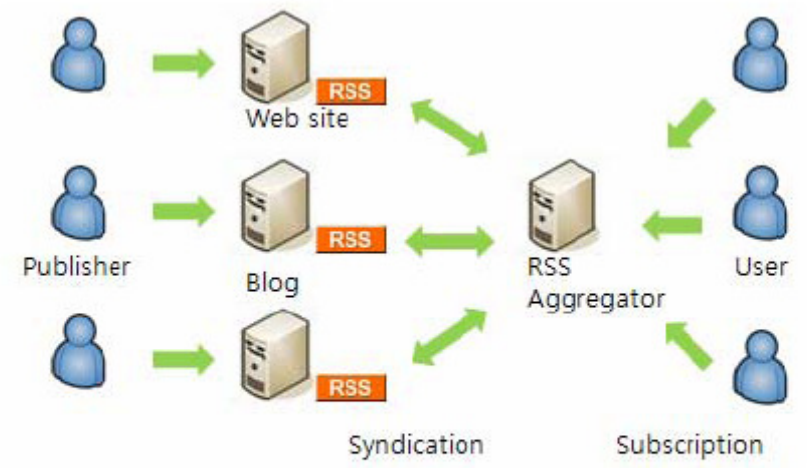



**Figure 26. RSS feed syndication and subscription.**
**Source: (Young Geun et al., 2008)**

RSS is operated based on a pull-based protocol. Publishers generate posting into the feed. Subscribers register feed addresses to an RSS feed aggregator that may aggregate data coming from various RSS feeds. Then the aggregator aggregates the postings and may show them to the subscriber.

As it may be easily noticed, RSS matches perfectly the idea of source monitoring in public relations. PR expert points to sources that he would like to focus on (represented as RSS feeds). Then, constantly updated content (a stream of documents) is presented to him in an aggregated manner. He doesn't have to visit these information portals few times a day, instead he may find all content in one tool using one interface.

Taking advantage from this fact, a source of information for the need of this research is defined as a stream of URLs, namely RSS feed, linking to documents on similar topics. We also assume that the same URL may be syndicated by few different feeds.

Because we deal with indexing of documents written in the Polish language, sources that are taken into account provide news only in Polish.

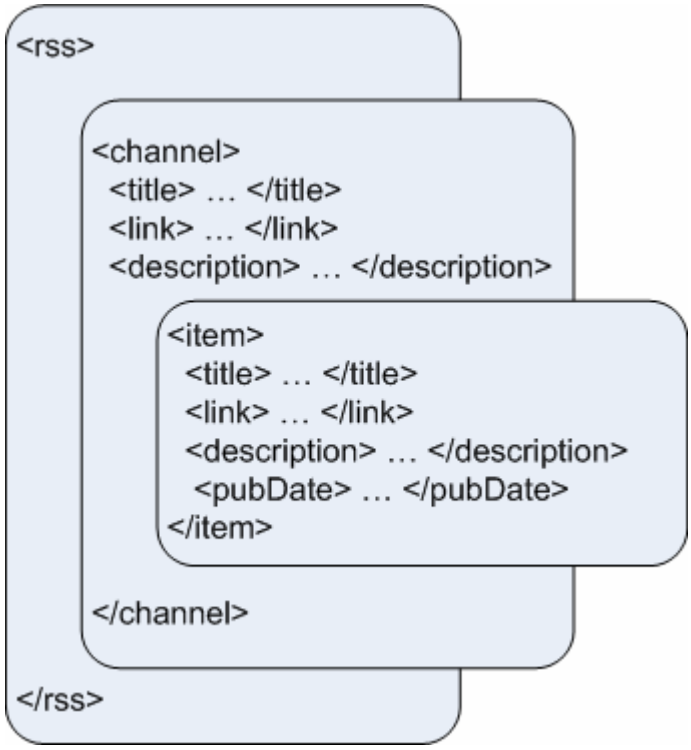



**Figure 27. RSS 2.0 Structure. Source: (Young Geun et al., 2008)**

It is also worth to note that in this research we are interested only in RSS feeds developed according to the RSS 2.0 standard. The structure of such message is presented in Figure 27. <channel> provides metadata on RSS feeds (title, link and description). A channel may contain a number of <item>s representing individual articles that may be found at websites of different newspapers. The structure of the feed is simple, what eases automation of the content extraction process.

## 4.4. Document-based spatial indexing method

The spatial indexing method that is described in this section assumes that an organization has already filtered out (e.g. from RSS feed) potentially interesting documents and they are available in the corporate repository. The method of acquisition of documents for the purpose of this research is described in section 5.2.

The indexing mechanism for each document retrieved, calls SProUT that utilizing its morphological resources, grammars and the gazetteer, extracts geographical name mentions from the document. It also disambiguates types of geographical references based on the context e.g. in Poland there are 65 counties (called in Polish *powiaty grodzkie*) of the same name as a city and commune (that overlap). This concerns the biggest Polish cities such as Poznań, Wrocław or Warszawa. In this case, if for a name mention no additional context information appears, it is assumed that author of the document meant a city.

In the second step all geographical ambiguous name mentions are disambiguated using heuristics and taking advantage of the developed ontology. The disambiguation procedure is presented in detail in section 4.4.2. The disambiguation concludes with a set of unambiguous geographical name mentions referring to only one place in domain of discourse. Moreover, in this step subtaxonomies of geographic ontology are built. This concerns assignment of addresses, districts and landmarks to a specific city and a city to a commune, etc.

The whole process ends with assignment of weights based on the number of references appearing in a document. After this step propagation across the geographical ontology is performed. This concludes the process of spatial indexing and index in the form of an inverted file may be stored in a database ready for comparison with a user's query.

This procedure is depicted in Figure 28.

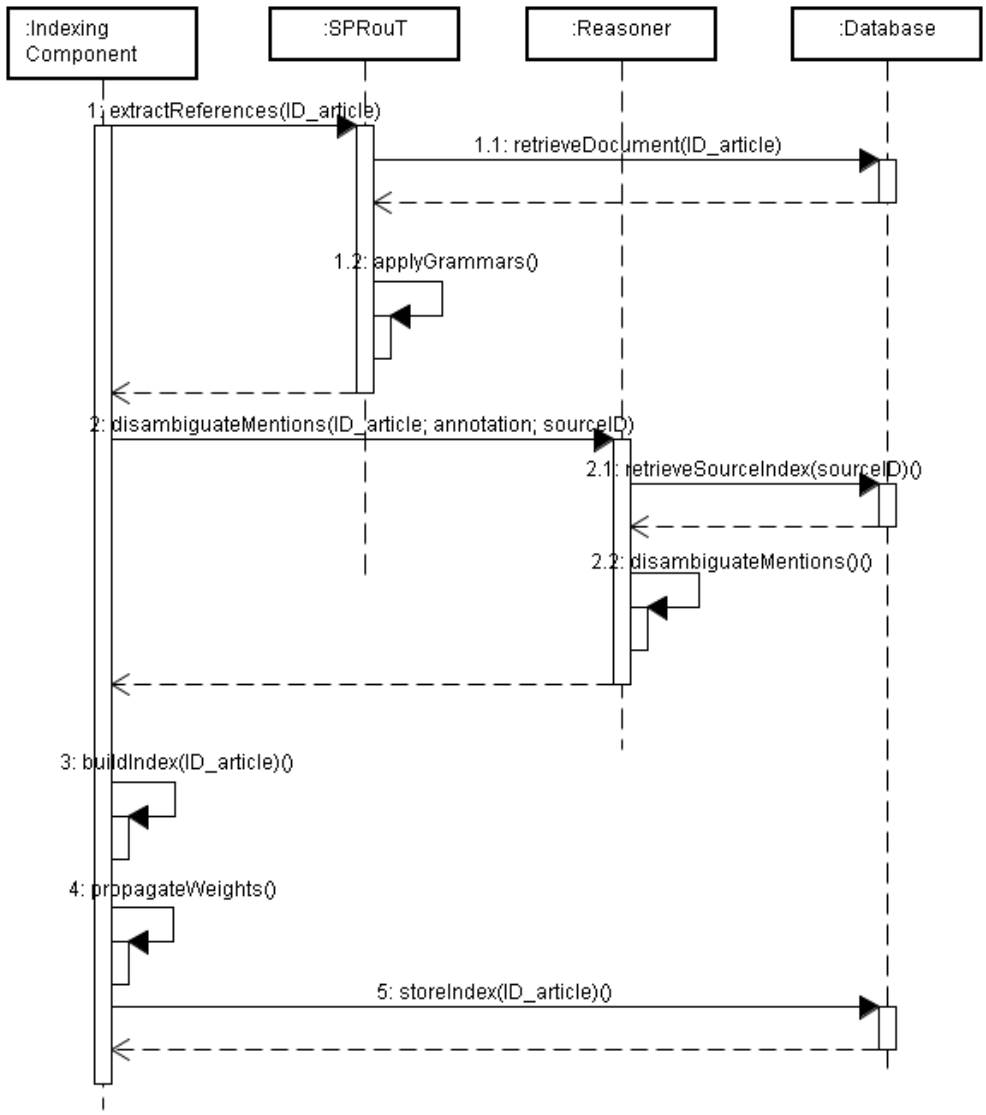



**Figure 28. Document-based spatial indexing method. UML Sequence diagram**

Components involved in the spatial indexing of documents are as follows:

- indexing component - responsible for: creation of a document index based on unambiguous name mentions retrieved by the SProUT and the reasoning component, propagation of weights based on the number of different place name mentions, storage of documents' indexes in a database,
- SProUT – described in section 2.2.1., where the geographical names' extraction rules are applied and types of geographical name mentions are disambiguated,
- reasoning component – taking advantage of the geographical ontology and heuristics developed, it performs disambiguation of geographical name mentions and builds the geographical subtaxonomies for cities and their sub-types extracted,

- database – stores articles, their indexes as well as indexes of sources, they were filtered from.

The functionalities of these components are described in detail in the following sections.

### 4.4.1. Extraction of named entities from free text documents

In free text documents one may distinguish the following types of geographical references (Filipowska and Węcel, 2004):

- absolute place reference – where a place in a document is mentioned explicitly e.g. Warszawa, Wrocław. This kind of geographic reference within a document may be catalogued as a triple "{place name, place type, coordinates}".
- relative place reference:
    - strong relative reference – where geographical concepts to which the reference points to, is resolved using some additional information e.g. on neighbourhood of a place e.g. 5km from Poznan, 500m from the Koszalin Town Hall;
    - weak relative references (in Information Extraction defined as coreferences), which may be resolved based on additional information appearing in text e.g. "in our city" ("w naszym mieście").

In this dissertation we focus on absolute place references as well as resolution of typical weak relative references. It is also possible to extend the proposed mechanism to deal with some strong relative references as the developed ontology stores information on the city coordinates. It is important to note however, that in case of such references as "north of Poznan" – additional information on what is meant by "north of…" should be provided (whether it is 5, 10 or 50 km north of Poznan).

For the purpose of extraction of name entities from documents, the XTDL grammar was developed. The rules it contains are based on the extensive study of documents written in Polish and are an effect of annotation workshops held at the Department of Information Systems as a part of the enIRaF project[40].

The current grammar contains 56 rules. They may be divided into two groups: supporting rules and rules for extraction of geographical name mentions. Supporting rules deal with:

- extraction of persons' and organization names (including historical names),
- extraction of noun phrases of different types (adjective-noun, noun-adjective, noun-noun, preposition-noun-adjective, preposition-noun),

[40] http://eniraf.kie.ue.poznan.pl

- extraction of adjective phrases,
- extraction of dates.

All these name mentions are extracted only as a part of a geographic name e.g. street or avenue, as they are initiated only by geographic extraction rules and do not perform as standalone rules (this is indicated by the ":/" sign after the name of the rule). An exemplary supporting rule is presented in the listing below.

```
;; extraction of nominal phrases e.g. Nad Wierzbakiem

pl_phrase_PN :/
(
morph & [POS prep, SURFACE #surface1, INFL infl_prep & [CASE_PREP #case]]
morph & [POS noun, SURFACE #surface2, INFL infl_noun & [CASE_NOUN #case]]
)
-> #geo_fraza_PN, where #geo_fraza_PN=ConcWithBlanks(#surface1, #surface2),
Capitalized(#surface1), Capitalized(#surface2).
```

This rule extract phrases of type "preposition + noun". If the noun was inflected, the output phrase returns only the nominal form. In the output the preposition and noun are merged and presented as a one phrase.

The second part of the grammar concerns rules for extraction of geographical name mentions. These rules were developed for each type from the type hierarchy mentioned in Section 4.2.2. Therefore, they involve rules for extraction of:

- countries,
- references to administrative concepts – such as provinces, counties, communes, cities,
- zone mentions such as Kaszuby, Pomorze, Mazury,
- references to city districts and landmarks (monuments, castles, etc.),
- street and avenue names, crossroads, roundabouts and related streets, etc.,
- references to water bodies – rivers, sees, lakes,
- references to continents and land forms (e.g. Kilimandżaro, Afryka).

In Section 3 of this dissertation it was mentioned that gazetteer proved to be an important resource for extraction of geographical name mentions. However, because of inflection of Polish geographical names and the fact that our gazetteer doesn't include all inflected forms the contextual rules were also developed. Such an approach also enables geo/non-geo disambiguation that is explained further in the section. The current set contains 36 rules for

extraction of geographical name mentions. They were developed in a way enabling disambiguation of mention types.

An exemplary rule for extraction of water bodies i.e. oceans, sees, lakes, rivers, streams is presented in a listing below. This rule calls two supporting rules `pl_org_name_innerwords` and `pl_entity_single_name` and extracts geographical name mentions preceded by a preposition (`pl_org_name_innerwords` rule).

```
pl_geo_wat :> (
@seek(pl_entity_single_name) & [SURFACE #n0, CSTART #cs]
(@seek(pl_org_name_innerwords)   &#i1   ?   @seek(pl_entity_single_name)   &
[SURFACE #n1]) ?
@seek(pl_org_name_innerwords) &#i2 ?
morph & [STEM "ocean", SURFACE #inne] |
morph & [STEM "morze", SURFACE #inner] |
morph & [STEM "jezioro", SURFACE #inner]|
morph & [STEM "rzeka", SURFACE #inner] |
morph & [STEM "strumień", SURFACE #inner]
@seek(pl_org_name_innerwords)   &#i3   ?   @seek(pl_entity_single_name)   &
[SURFACE #n3, CEND #ce]
) -> ne-location & [
        LOCTYPE wat,
        LOCNAME #n,
        NCSTART #cs,
        NCEND #ce],
        where #n = ConcWithBlanks
(#n0,#i1,#n1,#i2,#i3,#n3).
```

All rules were developed and compiled using the XTDL formalism in SProUT. The figure below presents the SProUT interface. On the left side the rules navigator is presented. The right side presents the way rules are developed. There is no editor available, so the rules need to be written by hand. The grammar developed is presented in Annex 1 of this dissertation.

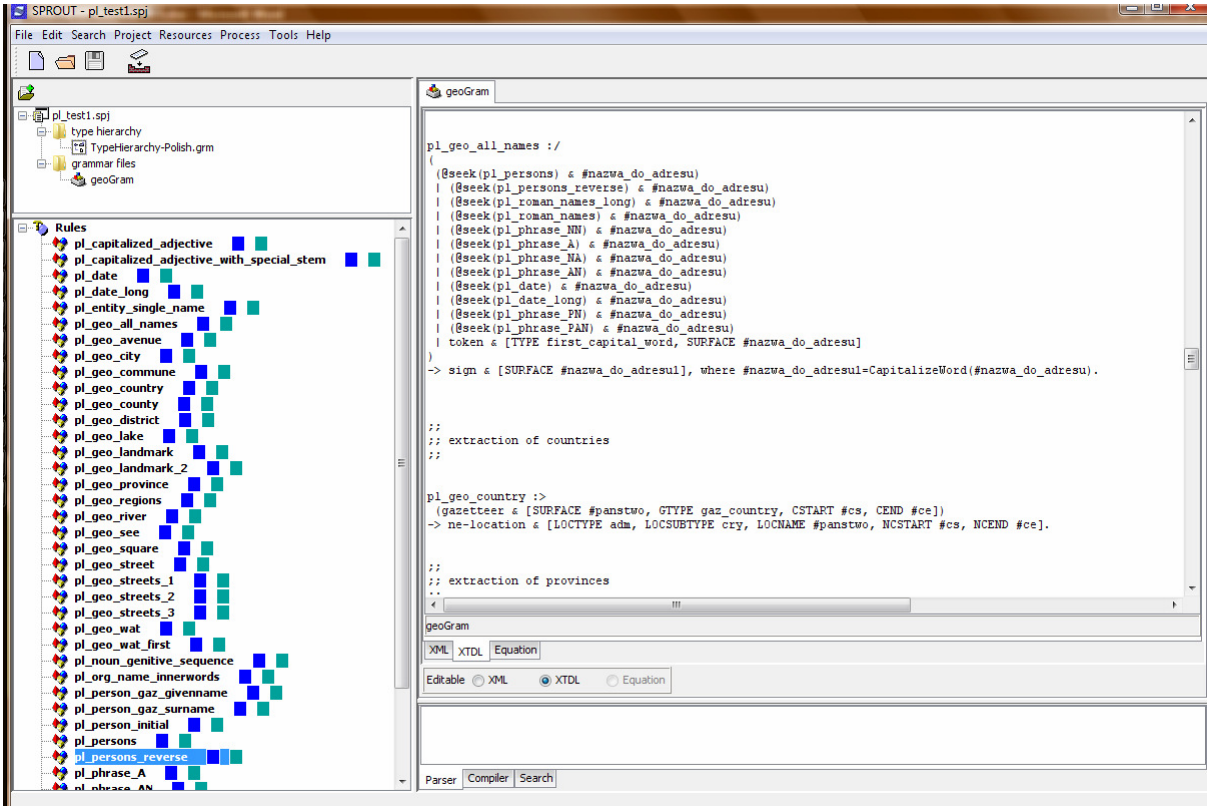



**Figure 29. SProUT interface for development of rules**

It is worth to note that the more rules, the more conflicts. The current grammar is an effect of a careful evaluation carried out in this research. Final evaluation outcomes are presented in Chapter 5 of this dissertation.

### 4.4.2. Disambiguation methods

Once extracted, the references need to be disambiguated. As discussed, there are two types of ambiguities: geo/non-geo ambiguity and ambiguity between different geographical entities. Because of applying context grammars, instead of simply spotting in texts names of places included in the gazetteer, we rarely spot geo/non-geo ambiguity and therefore, this kind of heuristics is left out of scope of this research.

Instead, we decided to prefilter geographical references based on a stop list prepared for the Polish language. This list includes frequently used nouns being also names of small places, that in our corpus appear over ten times more often that ten biggest Polish cities.

We assume one sense per discourse disambiguation, what means that if in a document a place of a given name appears several times, we assume that it refers to only one geographic entity. Our disambiguation approach is shortly presented in Figure 30.

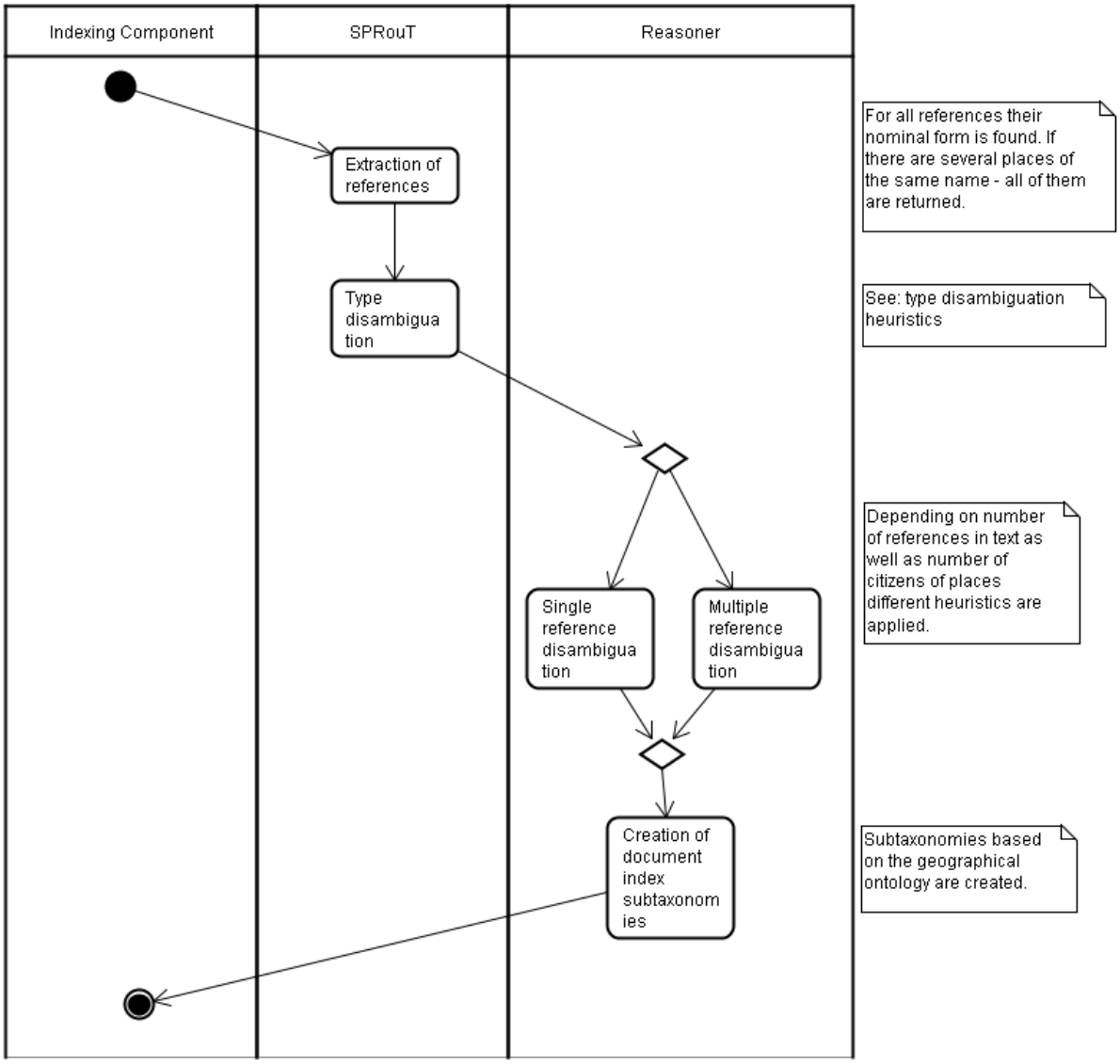



**Figure 30. Disambiguation procedure. UML Activity Diagram**

The geographical references received from SProUT are disambiguated to the level of type. It means that there are no geo-type/geo-type ambiguities and the remaining ambiguities concern only choosing from one of the potential cities or communes, etc. More comments on this follows in the remainder of this section.

In the next step heuristics disambiguating place name mentions follow. They depend on size of candidates (number of inhabitants), source the documents were retrieved from, distance between places, etc. The whole process finishes with development of document index built using the geographical ontology developed. More details on each heuristics utilized follows.

**Geographical type disambiguation**

One of the issues regarding the geographical ontology of Poland is that there are entities of the same name at the same time being of different types. So, a problem to resolve is not only that there are two cities called Białystok, but also that Białystok is also a name of a county and a commune. This situation is common for 65 biggest Polish cities, that besides being cities are also communes and are granted also the rights of counties. Similar situation concerns also city-commune ambiguities (for the number of all type-type ambiguities in gazetteer for Poland please refer to Table 10). Therefore, an approach suggested by many authors that a simple gazetteer lookup provides satisfactory results, could not be applied in our case. After analysis of documents from our corpus we concluded that, if author means the county or a commune, he explicitly mentions so. In all other cases he means city when referring to a place name. Examples of this are as follows:

- „W naszym województwie lepsze okazały się tylko trzy gminy wiejskie – Choceń, Lubiewo i Golub-Dobrzyń…" (In our province, better performed only three rural communes – Choceń, Lubiewo and Golub-Dobrzyń). All three names refer also to city names (document ID: 00000000040580[41]).
- „…w ostatnich dniach podpalili fragmenty lasów w gminie Wizna i Jedwabne…" (…in the last days they set fire to forests in Wizna and Jedwabne communes). These communes are also cities (document ID: 00000000013741).

Similar situation concerns disambiguation between river and city names e.g. in Poland Wisła is a name of the biggest Polish river and also a name of a popular skiing resort.

Taking this into account we developed contextual rules for extraction of name mentions. An example of such a rule is presented in listing below. We extract all counties mentions that are preceded by "pow." or inflected word "powiat" and afterwards mention the county name that is looked up in the gazetteer.

```
pl_geo_county :>
 ((token & [SURFACE "pow", CSTART #cs] token & [TYPE dot] ) (gazetteer &
[SURFACE #powiat, GTYPE gaz_powiat, CEND #ce]))|
 ((morph & [STEM "powiat", CSTART #cs]) (gazetteer & [SURFACE #powiat,
GTYPE gaz_powiat, CEND #ce]))
-> ne-location & [LOCTYPE adm, LOCSUBTYPE cnt, LOCNAME #powiat, NCSTART
#cs, NCEND #ce].
```

[41] All document IDs refer to IDs of documents from the annotated corpus of documents.

Such an approach was applied for extraction of all provinces (that are referred to by adjective pointing to a region), counties, communes and waterbodies.

**Extraction of references to districts**

District heuristics was developed based on the assumption that, if there is a city, that has more than 100 000 inhabitants, then if after it's name appears a word (or phrase) written from a capital letter, then this word (or phrase) refers to a city district. The corpus of documents developed delivers several examples confirming this fact e.g. Warszawa Praga (a district Praga in Warszawa), Warszawa Śródmieście (Warszawa city centre), Gdańsk Wrzeszcz (Wrzeszcz district in Gdańsk).

Moreover, districts that are of a unique name such as Niebuszewo (Szczecin), Gumieńce (Szczecin), Wrzeszcz (Gdańsk) were included in gazetteer and are searched using a typical approach for a gazetteer lookup.

The most important heuristics concern disambiguation of mentions pointing to few real-world counterparts of the same type. They are presented in subsequent paragraphs.

**Source-based disambiguation of name mentions**

Each source of documents is usually assigned different characteristics. Some of them relate to topics addressed by this source, other concern quality of information, type of source, etc. (Kaczmarek, 2006). We assume that a source may also have its geographic index i.e. index developed after analysis of a representative set of documents retrieved from this source.

A source index is understood as a vector of places that assigns each place a probability that this place name appears in a new document in this source. This is based on the assumption that regional newspapers and portals write mainly about their neighbourhood and do not refer often to cities from outside the addressed area.

The source for the need of this research is defined as: "a stream of documents syndicated from a one known source, namely RSS". The sources we monitored were RSS feeds retrieved from the main pages of websites catalogued in DMOZ[42] in category *Top/World/Polska/Regionalne/Europa/Polska/**. For each RSS feed, for one month we have been retrieving all documents appearing in the source. Then using our extraction grammars

[42] http://www.dmoz.org/

and heuristics that are elaborated in this section disregarding only the source-based heuristics, we developed index of the source as the average of document indexes. That was our initial source index, that using a propagation method (described in section 4.4.3) was then aggregated to the level of counties. In the end the source index consists of counties assigning each county a probability that a new article appearing in the source refers to a specific point (city, commune, city district) in this county.

**Heuristics for disambiguation of single name mentions**

There are documents that contain only a single geographic reference. When disambiguating such a reference in a first step we check, if a name mention refers to a city with a number of inhabitants exceeding 100.000 people (this number was validated in our research). Otherwise, we check an index of a source a document was retrieved from. The source index is defined on a county level and contains a probability of inclusion of a city from this county in a document that is retrieved. If a city bearing the reference name is found in the geographical ontology as a city belonging to this county, then a city name is disambiguated. If a city name is not found in the county given, then the reference is disambiguated as a city of this name with the highest number of inhabitants in Poland.

**Multiple reference heuristics**

In most cases a document contains more than one geographic reference (or contains no references at all). Documents bearing no references are left out of scope of this research.

Focusing, however on documents with multiple references it is worth to note, that most of them refer to the biggest Polish cities. Out of 1122 references to location of type LOC-ADM, 518 referenced cities being the biggest Polish cities.

However, even the biggest Polish cities have its smaller counterparts. There are 2 places called Poznań, 5 cities called Warszawa, 2 cities called Olsztyn, etc. To deal with these ambiguities the heuristics depicted in Figure 31 was defined.

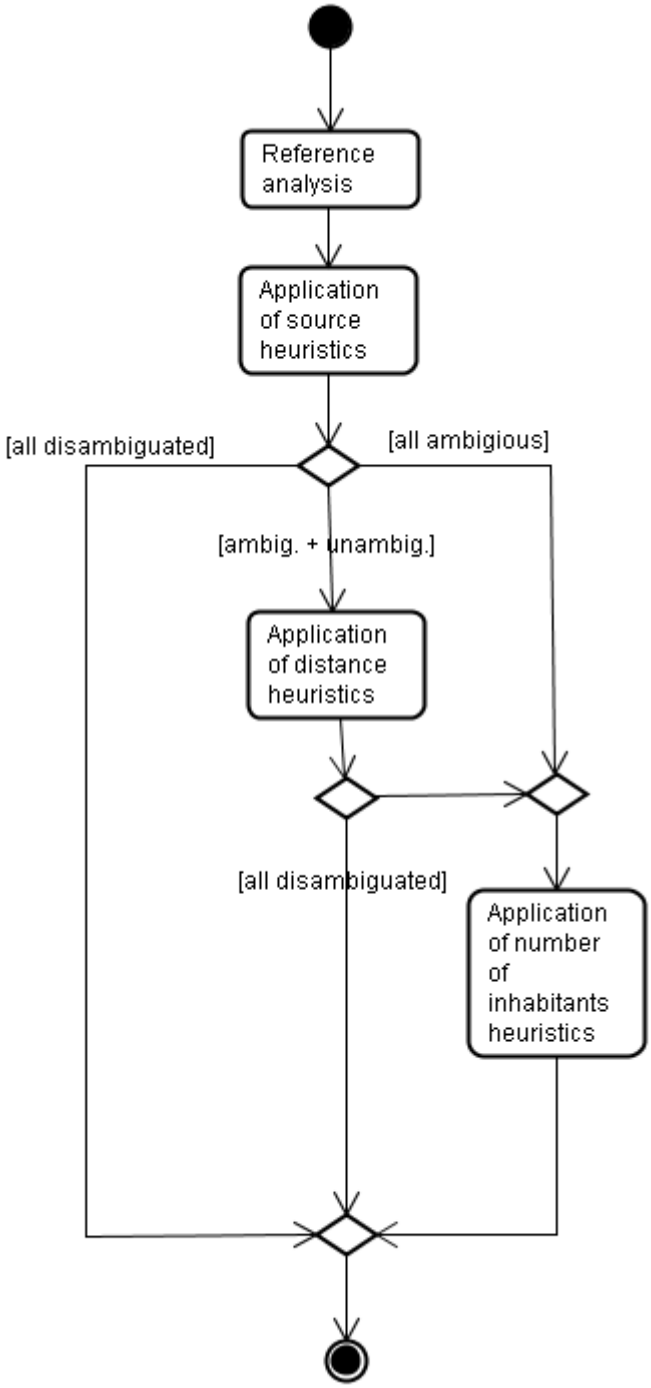



**Figure 31. Multiple reference heuristics. UML Activity Diagram**

In the first step we analyse all references from a given document. There may be a case that all references in text are ambiguous but it also may happen that some of them are unambiguous and may be used while disambiguating. In the next step we apply the source heuristics as we monitor mainly local resources. It means that for each source we defined the most probable[43]

[43] Or a few equally probable.

county the document concerns. If a city of such name is found in this county, it is disambiguated. In case there are two or more cities of the same name, the bigger one based on the number of inhabitants is chosen. After this step we may have three different situations:

- all references are disambiguated,
- there are ambiguous and unambiguous references in a given document,
- there are only ambiguous references.

In the third situation we apply heuristics based on the number of inhabitants – the biggest city bearing a reference name is chosen. This is also applied for all references that couldn't be disambiguated in other way.

In the second situation we have ambiguous and unambiguous geographical references. In this case we apply distance heuristics. Based on the unambiguous references we calculate a centre of polygon described by points (coordinates) of unambiguous places. Then we calculate the distance between the centre and the most distant point. This is the radius of the circle we draw. In case there is only one unambiguous city we draw a circle having radius of 80km. If the radius is longer than 80km – then we draw circles (of 80km radius) around each unambiguous location. All ambiguous places that have its alternatives in this circle are disambiguated. All other mentions are disambiguated based on the number of inhabitants heuristics.

In order to calculate the centre of polygon we use the following formula, where x and y stand for coordinates of vertices and A is area of the polygon.

$$A = \frac{1}{2}\sum_{i=0}^{n-1} (x_i y_{i+1} - x_{i+1} y_i)$$

$$\bar{x} = \frac{1}{6A}\sum_{i=0}^{n-1} (x_i + x_{i+1})(x_i y_{i+1} - x_{i+1} y_i)$$

$$\bar{y} = \frac{1}{6A}\sum_{i=0}^{n-1} (y_i + y_{i+1})(x_i y_{i+1} - x_{i+1} y_i)$$

The figure below presents a document index for one of documents from our corpus. This article is retrieved from Nowa Trybuna Opolska (http://www.nto.pl) and its geographical index concerns mainly powiat opolski (opolski county). This means that even if there are 4 cities named Opole in Poland, the capital of the Opolskie province would be disambiguated based on the source index. If there is only one reference disambiguated, then a circle having radius of 80km is drawn. All the references would be exact in the circle drawn.

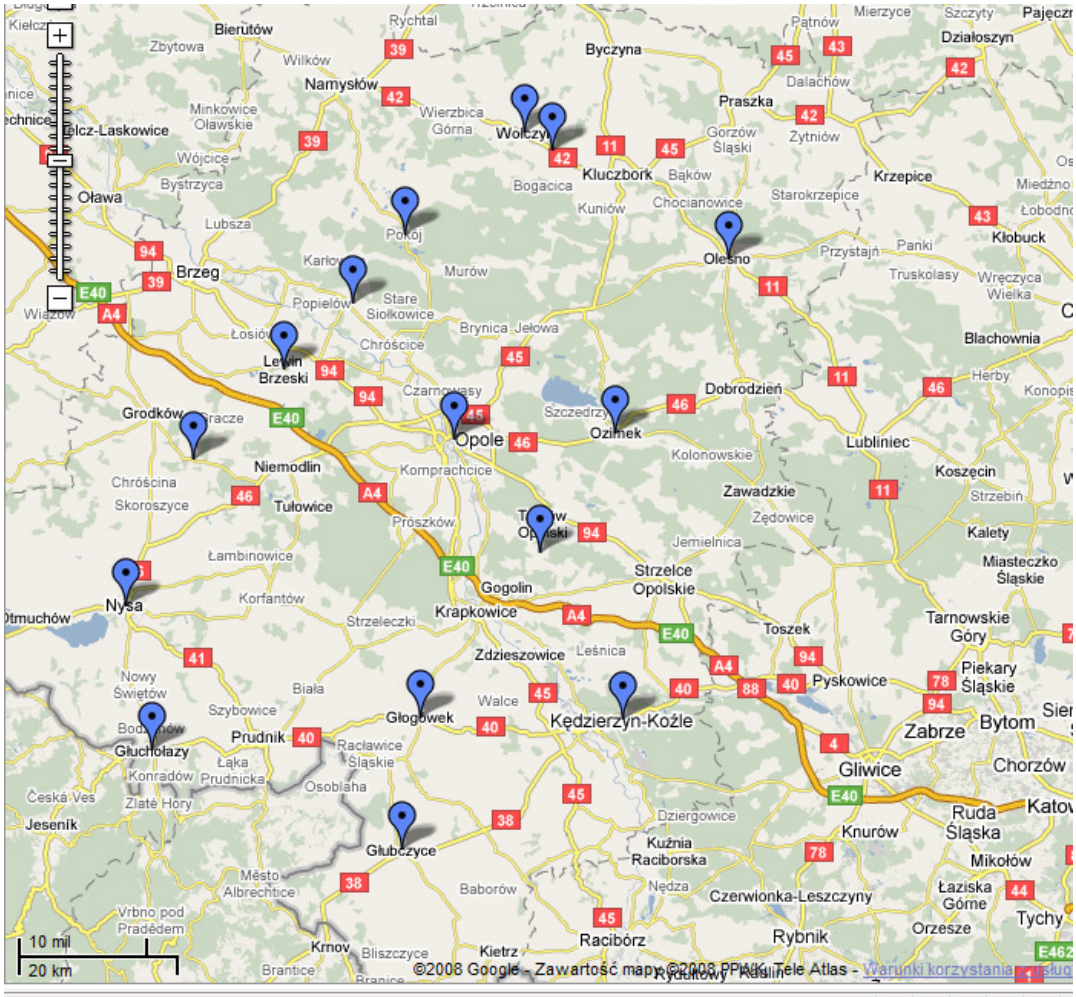



**Figure 32. Index of the document no. 00000000018932**

This approach helps also to disambiguate such mentions as in fragments:

- "…przejeżdżałem przez bliską mym uszom miejscowość Lisowice między Opolem a Częstochową…" (..I was driving through a familiar to me city Lisowice between Opole and Częstochowa…). Document ID: 00000000014067.
- „Do miejscowości Karnity koło Miłomłyna na zgrupowanie sportowe wyjechała…" (…for a sport camp to Karnity situated near Miłomłyn…). Document ID: 00000000028577.

**Heuristics for streets, landmarks and city districts**

Authors in documents quite often refer to streets, landmarks (characteristic points of the city) or city districts. Therefore, except from heuristics dedicated for districts also a heuristics assigning streets, landmarks and districts to appropriate cities is proposed. Some of them are also unambiguous landmarks such as *Jasna Góra* or *Wawel* or districts such as *Wrzeszcz* or

*Zaspa* that may be associated with a specific city. As landmarks and districts are included in our geographical ontology, in such a case, the city is straightforwardly associated with them. Streets' names and some of districts and landmarks however, duplicate in multiple cities and need to be associated only with the exact place in space. Another argument supporting this fact would be that developing a gazetteer for street names would be resource expensive.
After analysis of numerous references the following heuristics was developed.
We extract all mentions that are streets, landmarks and districts using SProUT rules. Exemplary rules are presented in listings below.

**Rule for extraction of landmarks**

```
pl_geo_landmark_2 :> (
morph & [STEM "hotel", SURFACE #inner, CSTART #cs] |
morph & [STEM "pomnik", SURFACE #inner, CSTART #cs] |
morph & [STEM "stadion", SURFACE #inner, CSTART #cs] |
morph & [STEM "teatr", SURFACE #inner, CSTART #cs] |
morph & [STEM "kino", SURFACE #inner, CSTART #cs] |
morph & [STEM "park", SURFACE #inner, CSTART #cs] |
morph & [STEM "twierdza", SURFACE #inner, CSTART #cs] |
morph & [STEM "pałac", SURFACE #inner, CSTART #cs]
@seek(pl_org_name_innerwords)   &#i1   ?   @seek(pl_entity_single_name)   &
[SURFACE #n1, CEND #ce]
) -> ne-location & [
        LOCTYPE adm,
        LOCSUBTYPE ldm,
        LOCNAME #n,
        NCSTART #cs,
        NCEND #ce],
        where #n = ConcWithBlanks(#inner,#i1,#n1).
```

**Rule for extraction of districts**

```
pl_geo_district :>
(
  ((morph & [STEM "osiedle", CSTART #cs]) |(morph & [STEM "Osiedle", CSTART
#cs]))
  |
  (((token & [SURFACE "Os", CSTART #cs]) | (token & [SURFACE "os", CSTART
#cs]) | (token & [SURFACE "OS", CSTART #cs]))
    token & [TYPE dot]))
```

```
( @seek(pl_geo_all_names) &  [SURFACE #nazwa_do_adresu, CEND #ce])
->   ne-location-postal   &   [LOCTYPE   adm,   LOCSUBTYPE   dis,   STREET
#nazwa_do_adresu, LOCNAME #cala_nazwa, NCSTART #cs, NCEND #ce],
where #cala_nazwa=ConcWithBlanks("Osiedle", #nazwa_do_adresu).
```

**Rule for extraction of streets**

```
pl_geo_street :>
(
  morph & [STEM "ulica", CSTART #cs]  |
  ( ( token & [SURFACE "Ul", CSTART #cs] | token & [SURFACE "ul", CSTART
#cs] | token & [SURFACE "UL", CSTART #cs] ) (token & [TYPE dot] | token &
[TYPE comma]))
 @seek(pl_geo_all_names) &  [SURFACE #nazwa_do_adresu, CEND #ce]
)
->   ne-location-postal   &   [LOCTYPE   adm,   LOCSUBTYPE   str,   STREET
#nazwa_do_adresu, LOCNAME #cala_nazwa, NCSTART #cs, NCEND #ce],
where #cala_nazwa=ConcWithBlanks("ul.", #nazwa_do_adresu).
```

Assignment of these place name mentions to an appropriate footprint takes place when all city names are disambiguated. As in Polish some addresses are in the following format: “ul. Ludowa 11, 64-920 Piła (address in Piła, document ID: 00000000007135), we apply the following procedure:

1. We look for a city name in the distance of 50 words before and 10 after the reference. The closer one is perceived as the target city. If there is a name of the county or commune, we disregard it and search further for a city name.
2. In case no reference to a city is found, we search for the most popular reference to a city that appears in this document. If there is such a city, we assign the reference to this city (e.g. in document ID: 00000000014067).
3. In other case, if there is no city name in the document or there are few cities with an equal number of references, we choose a city that precedes reference to a street or landmark. Otherwise, we leave the reference unresolved.

**Ontology-based reference merging**

Some of the geographical places in documents may be referred to using the historical or common names as e.g. (examples mentioned here are based on the developed corpus of documents):

- Polska (Poland) is sometimes referred to also as Rzeczpospolita Polska (and its abbreviation RP) or using a previous formal name Polska Rzeczpospolita Ludowa (and its abbreviation PRL);
- China is sometimes referred using its formal name Chińska Republika Ludowa (People's Republic of China);
- some cities in Poland have also its common names used in writing W-wa for Warsaw, K-lin for Koszalin, or Wrocek for Wrocław.

Such place name alternatives are stored in the geographical ontology and are grounded while extracting information using SProUT.

**Solving indirect references**

We also solve weak relative references of type: "naszej gminy" (our commune - document ID: 00000000001612), "naszego miasta" (our city - document ID: 00000000006970), "naszej gminie" (our commune - document ID: 00000000012836). We decided not to resolve references only by a word „miasto" (city) as it may be used in various contexts, out of which only a few mean a reference to a particular city.

This heuristics is similar to the heuristics for streets, districts and landmarks. If in a document a following phrase is spotted: "nasz(a) gmina/miasto/powiat/województwo" (our commune/city/county/province), first it is searched for all references to a place of such type preceding the phrase (to the beginning of a document). The first one reference found is set as a resolution of this phrase. If such a reference is not found, then it is resolved based on the source index – it is grounded as the most frequently city referred by this source. If there is no source index, the phrase is left without resolution.

**Other heuristics**

The city name may be also disambiguated based on the name of the city mayor based on phrases such as "Zapraszam na bezpieczne kąpielisko do Lewina - mówi Leszek Paukła, wiceburmistrz miasta…" (I kindly invite you to visit the swimming place in Lewin – said Leszek Paukła, vicemayor of the city). Document ID: 00000000018932. However, in our case we decided to leave such heuristics out of scope of our work.

**Other issues**

Currently, we do not support disambiguation of statements such as "małych miast Polski wschodniej (small cities of Eastern Poland)", unless an expert disambiguates the notions of a small city (based on the area, number of inhabitants, etc.) and Eastern Poland.

We understand European Union as in definition provided at European Union website[44], where the EU is defined as "a unique economic and political partnership between 27 democratic European countries". Therefore, we understand EU as an organization and not as a zone (type of geographical location).

We do not support analysis of imprecise references such as "a Jan Kowalski wojewodą dużego województwa, ot, choćby wielkopolskiego". We also do not address the issue of resolving spelling and orthographic errors in names of places (from a small letter, without Polish letters, etc.) e.g. wroclaw, poznan.

For all references disambiguated using the ontology developed, we prepare a tree-like structure being the index of a document. This index is a vector of ontology instances where an importance of an instance for a document is calculated using propagation mechanism described in the next section.

### 4.4.3. Spatial document indexing propagation mechanism

The aim of the propagation mechanism is to assign each instance of the geographical ontology a weight showing the importance of a given location for an analysed document taking into account relations between places. This importance is calculated based on the initial document index (disambiguated references along with their number of occurences). The approach applied is a modification of the TF*IDF scheme (Baeza-Yates and Ribeiro-Neto, 1999) that describes an importance of word to a document from a given corpus. Finally, all indexed documents are represented by a vector of place names and associated weights for the needs of further storage and processing.

Let's assume we extracted 10 geographical references from a particular document: 1 to Lubelskie Province (woj. lubelskie), 3 to Goraj (a commune in the lubelskie province), 3 to Andrzejówka (a small city in Biłgoraj commune), one to Cicha street in Biłgoraj and 2 to koszaliński county. Based on this information and using knowledge base of the ontology we develop an initial document index presented in Figure 33.

[44] http://europa.eu/abc/panorama/index_en.htm

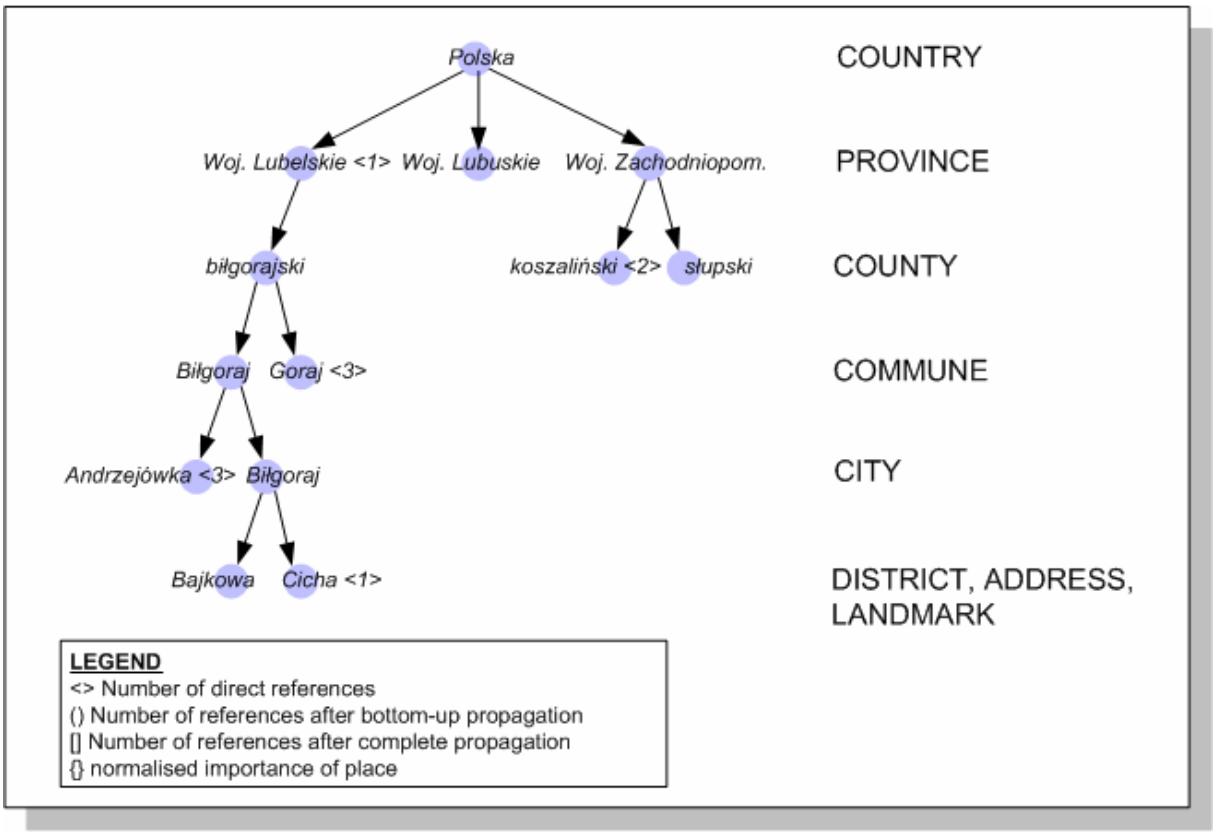



**Figure 33. Excerpt of instances of geographical ontology after extraction of named entities from text**

Then we initiate our propagation procedure. Both propagations: bottom-up and top-down may be performed in parallel as they are independent from the results of one another. However, for a clarity purpose we demonstrate them sequentially.

We assume that a reference to an instance of e.g. commune may be perceived also as an indirect reference to a county or province this instance refers to using *belongsTo* attribute. Based on the *belongsTo* relation a tree-like structure is built that represents a hierarchical structure between ontology instances.

After the tree-like structure is built, we perform a bottom-up propagation of references. This involves adding the number of references from the lower level nodes to the upper level nodes. As it may be noticed after this propagation the number of references in text for Lubelskie province increased to 8 (which is the total number of references to geographical entities from this region).

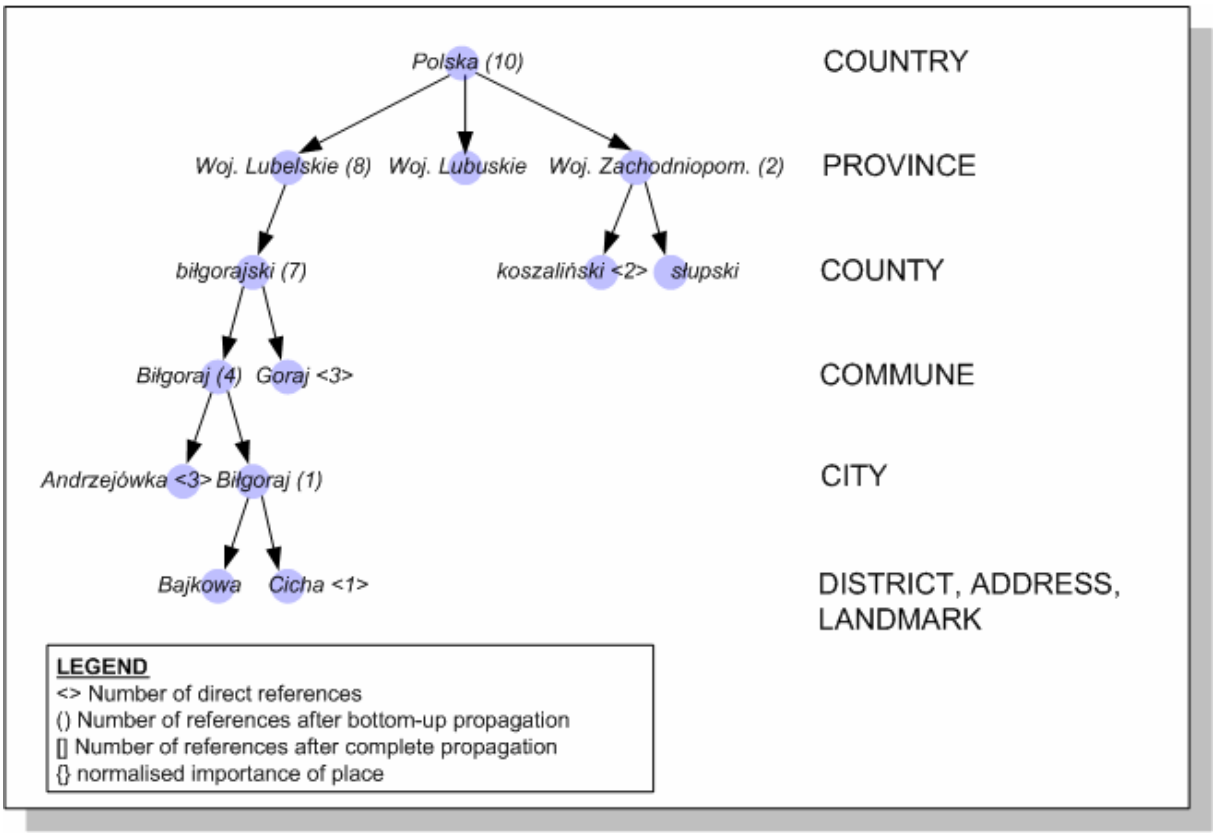



**Figure 34. Excerpt of a tree-like structure of ontology instances after the bottom-up propagation**

If we denote *R(c)* as a number of direct references to a node *c*, then we may define *B(c)* - number of references after bottom-up propagation as:

$$B(c) = \begin{cases} \sum_{n=1..\infty} R(s_n^c), l = 0 \\ \sum_{n=1..\infty} B(s_n^c) + \sum_{n=1..\infty} R(s_n^c), l = 1..4 \end{cases}$$

where *l* indicates level of a tree based on administrative division (*l=0* indicates cities, *l=1* communes, etc.) and *s* subconcepts of concept *c*.

After performing the bottom-up propagation, we proceed with a top-down propagation. This involves dividing the number of direct references to a node by a number of its subnodes e.g. if there is a direct reference to the Lubelskie province, it means that it also concerns all counties, communes and cities situated in this region. Lubelskie province consists of 24 counties, biłgorajski county consists of 13 communes and there is 89 cities in Biłgoraj commune. Based on this information (that may be easily retrieved using knowledge base of the geographical ontology we have) we calculated new weights for places in the document index.

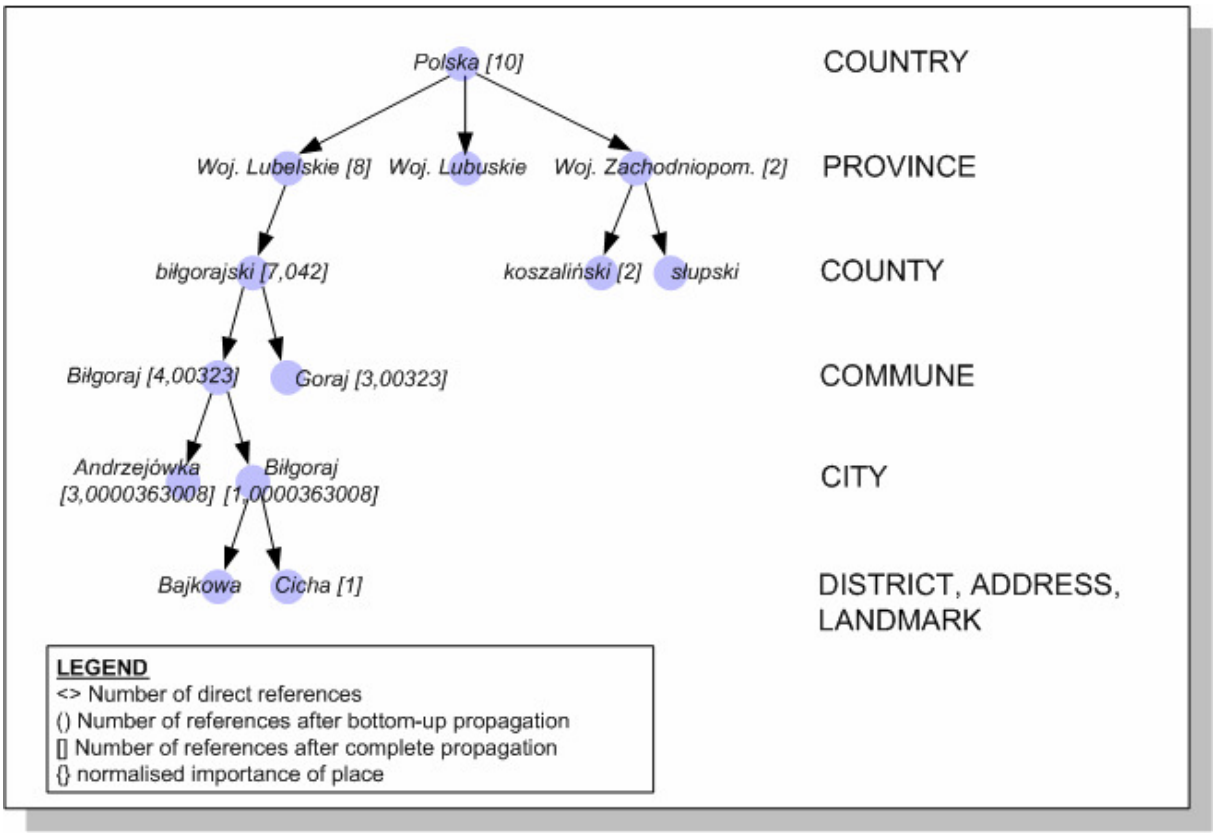



**Figure 35. Excerpt of a tree-like structure of ontology instances after the top-down and the bottom-up propagations**

The top-down propagation T(s) may be defined as follows:

$$T(s_n^c) = \begin{cases} \frac{1}{n} R(c), & l = 4 \\ \frac{1}{n}\left(R(c) + T(c)\right), & l = 0..3 \end{cases}$$

where *s* is a subnode of a node *c* and *R(c)* is the number of direct references to *c*.

The last step concerns normalization of weights. Firstly, S(c) is defined as a sum of all references (direct and propagated).

$$S(c) = T(c) + B(c) + R(c)$$

Then, S(c) is normalized according to a formula:

$$r(c_i) = \frac{s(c_i)}{\sum_{n=1..\infty} s(c_n)}$$

After this step the sum of weights for each type of geographic named entities sums up to 1. In the figure below, however each reference type may not sum up to 1 (because of clarity issues some nodes that were also assigned weights in the process of the top-down propagation

including counties in woj. Lubelskie, communes in biłgorajski county as well as cities in Biłgoraj commune are not visualized).

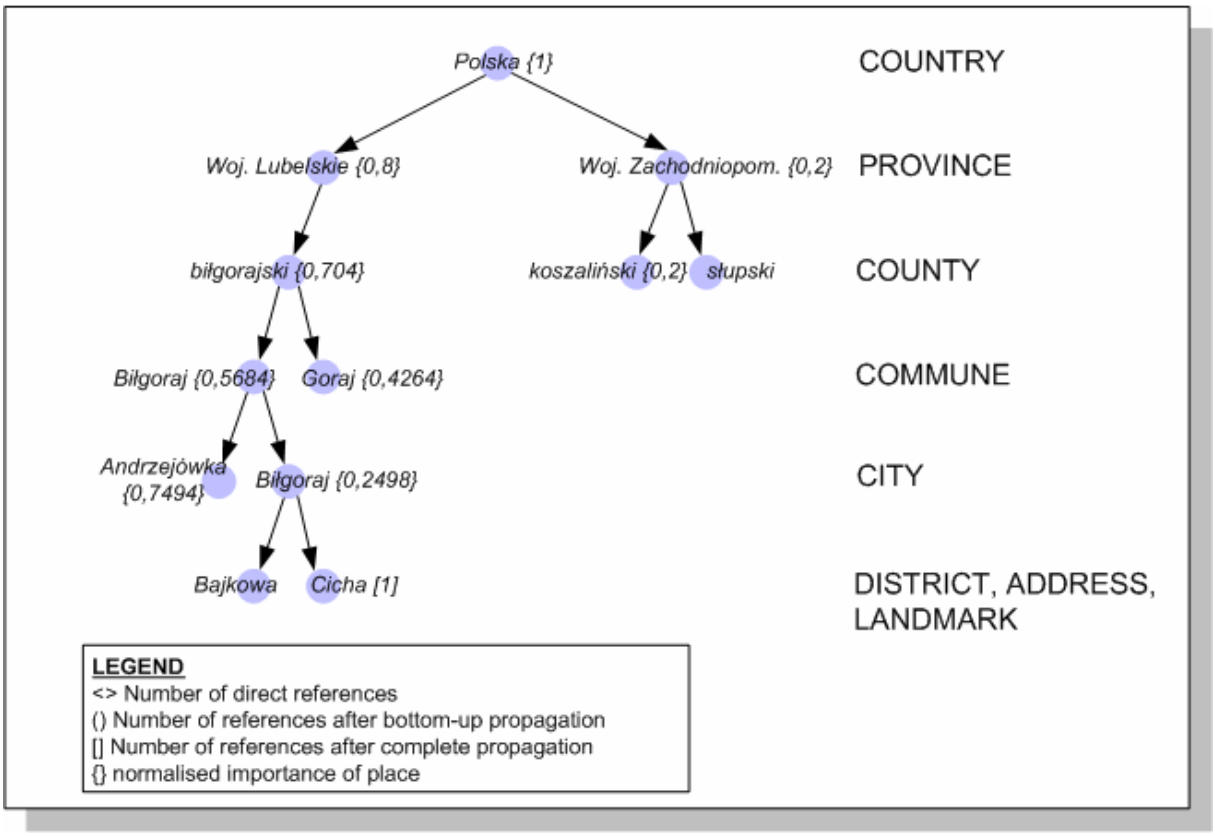



**Figure 36. Excerpt of the instances of geographical ontology after normalized weights for geographical NE were created**

Summarizing the example, the propagation procedure may be described as follows:

1. Assign to ontology instances a number of direct references to them in the text.
2. Build a tree-like structure of ontology instances.
3. Then perform a bottom-up propagation of references. Count all references of subnodes of a given node, and set it as a new number of references for this node (increase the number of references for a node by the total number of references in all of its direct subnodes).
4. Then perform top-down propagation. The number of initial references to an upper node should be divided by the number of its direct subnodes, and their number of references should be changed accordingly.
5. Normalize weights for the each type of geographic named entities.

This step concludes with development of the geographical index of document. The index should be then stored, optimized (if needed) and used in the process of Geographical Information Retrieval.

## 4.5. Development of a source-based spatial document index

Georeferencing or spatial indexing of documents while disambiguating sometimes refers to source-based heuristics. These heuristics use geographical information on a source the documents were retrieved from, namely a source index. Source index is a vector of places that assigns each place a probability that a place name appears in each new document published in this source. This is based on the assumption that news published in regional newspapers and portals are biased towards the newspaper's neighbourhood and rarely refer to small cities from outside of the addressed area (and these cities are usually the name mentions to be disambiguated).

This section describes how we build the source index being a kind of a reference place for a document. The research question is as follows: may the source-based document index substitute a document-based index produced by the previously described spatial indexing method.

### 4.5.1. Assumptions for the source-based spatial document indexing method

While developing this method, we focus on sources that are RSS channels providing URLs that give access to news documents in HTML format. We assume that:

- news accessible from single RSS feed are thematically related,
- majority of RSS feeds address a specific geographic area(s),
- single URL may be present in multiple RSS feeds,
- presence of URL in a specific set of RSS feeds may be related to geographical range of news they point to,
- URL addresses follow some patterns that may be related to geographical range of news they point to.

### 4.5.2. Overview of the source-based spatial document indexing method

Source-based spatial document indexing method takes advantage of features of an URL address pointing to a specific web resource. We assumed that that these features may enable us to produce a source-based document index that may be used e.g. for disambiguation of

named entities within a document or even substitution of a traditionally developed document index.

We create a source-based spatial document index using a stochastic model predicting a geographical scope of a given document based on features of its source. This model is built applying the Maximum Entropy Principle (MEP). We decided to apply the MEP because of the following reasons (Ratnaparkhi, 1997):

- the MEP model assumes that features used may correlate with one another,
- as a result of MEP algorithm we do not receive a classification function but a probability distribution; if a fact is unknown, then probability is distributed evenly among all unknown facts,
- neural network that would probably also fulfil these needs would be extremely difficult to built and apply because of 50000 entry features and only 300 possible outputs.

The ME principle states that the correct distribution of probability *p(a,b)* is that, which maximizes the entropy, subject to constraints representing facts already known (Jaynes, 1957). In other words, ME algorithm models all facts that are known, assumes nothing about facts that are unknown and distributes the probability evenly among the unknown facts thereby maximizing the entropy (Berger et al., 1996).

To apply the MEP, firstly it is needed to determine a set of statistics that captures the behaviour of a random process. In our case that is to determine a set of features that influence the geographic index of document. Then, these facts must be correlated to create an accurate model of the process predicting the most probable document index.

First phase also focuses on determining information sources important from the point of view of the domain being modelled. These sources are usually chosen based on a large training set using a feature function *f(x,y)*. The value of this function is 1 when there is a relation between a feature and an occurring event and 0 otherwise. An important characteristics of all feature functions is that they may interrelate with each other.

To choose from a set of all feature functions, only these that describe a logical relation between a feature and an event, a large set of training data is needed. Therefore, only functions fulfilling the condition about the expected value of *f* with respect to the empirical distribution $\tilde{p}(x, y)$. The expected value is denoted as:

$$\tilde{p}(f) = \sum_{x,y} \tilde{p}(x, y) f(x, y)$$

When it is discovered that a statistics is useful, its importance is acknowledged by constraining the expected value that the model assigns to the feature function *f*. The expected value of *f* with respect to the model $p(x|y)$ is:

$p(f) = \sum_{x,y} \tilde{p}(x,y)\, p(x|y)\, f(x,y)$, where $\tilde{p}(x)$ is the empirical distribution of x in the training sample.

This value is constrained to the expected value of *f* in the training sample: $p(f) = \tilde{p}(x)$.

Taking into account all constraints, from the set *P* of all probabilities $p(x|y)$ we distinguish a subset C defined as:

$$C \equiv \{ p \in P \mid p(f_i) = p(\tilde{f}_i) \ for \ i \in \{1,2,..,n\} \}$$

Among the models $p \in C$, the maximum entropy principle imposes to choose the most uniform distribution of probability. This function *H(p)* called entropy is denoted as:

$$H(p) \equiv -\sum_{x,y} \tilde{p}(x)\, p(y|x) \log p(y \mid x)$$

The entropy is bounded from below by 0 (in case of a model with no uncertainty) and from above by $\log |Y|$ (the entropy of the uniform distribution over all possible |*Y*| values of *y*.

All above, enables us to define the Maximum Entropy Principle (after (Berger et al., 1996, Ratnaparkhi, 1997)):

**Definition 1.**

"To select a model from a set *C* of allowed probability distributions, choose the model $p_* \in C$ with maximum entropy *H(p)*:

$$p_* = \arg\max_{p \in C} H(p)."$$

The spatial indexing procedure with application of the MEP model is described below and presented in Figure 37. Firstly, PR expert identifies sources he would like to monitor (this can be also done automatically based on a catalogue of web pages filtered out according to previously specified criteria). It is also possible to create an RSS feed for the website that does not provide RSS feed using such portals as: http://www.feedity.com.

Then, RSS feeds for all sources are retrieved and their monitoring is initiated.

URLs linking to documents of interest are extracted from the received feeds and documents they point to are retrieved. URLs and these documents constitute training and evaluation sets for the source-based document indexing method.

The analysis of URLs and documents retrieved from these URLs is performed in order to extract possible URL features. We divided these URL features into three groups:

- basic features – encompassing such characteristics as protocol, domain, port, domains including geographical names, IP address, etc.,
- tokens – such as appearance of sequence of tokens (or a token on a given position in an address), city name, parameter such as a number on a given position in after a domain name, parameters – appearance of a specific parameter with a text or numerical value.
- source features – belonging to specific RSS feed is also treated as binary feature function.

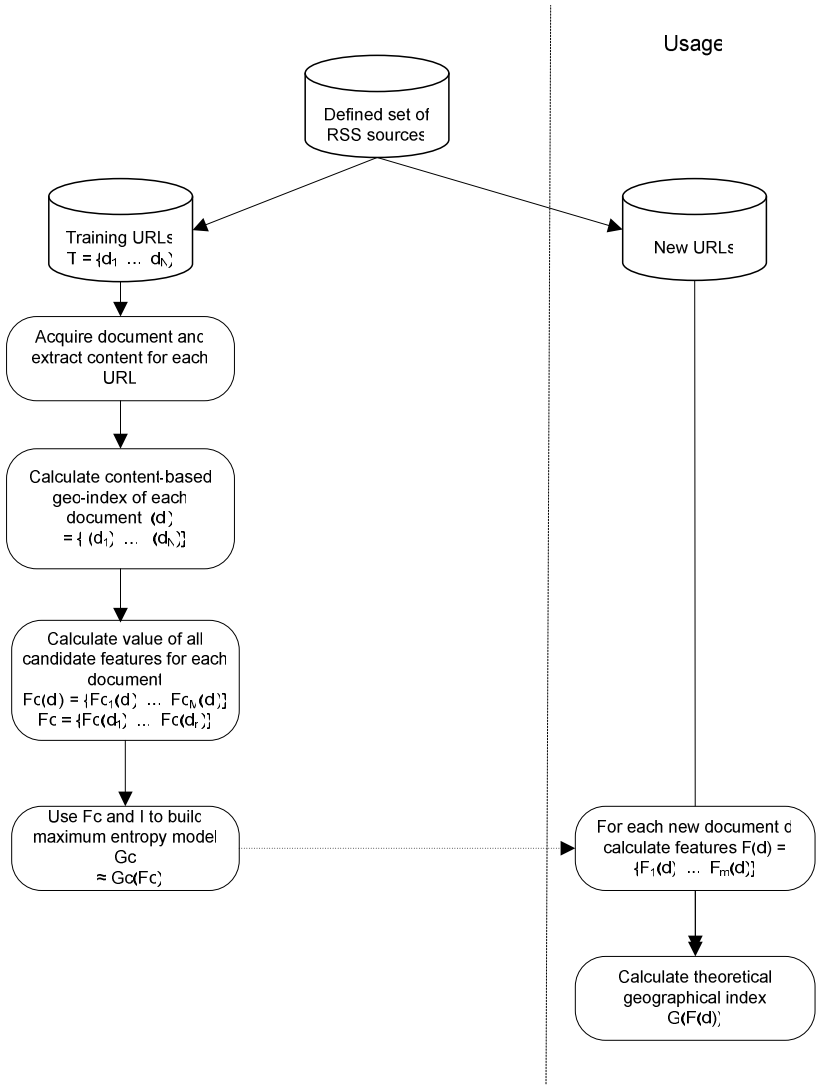



**Figure 37. The source-based document indexing method**

At the same time an analysis of documents retrieved is performed in order to create a reference document index. Documents are analysed using extraction techniques and heuristics described in previous sections on spatial indexing method. The outcome from the document indexing phase are indexes for all documents on the level of counties.

A probability model trained on both documents' source features and their indices is built using the maximum entropy method. The trained model for each new document retrieved assigns the most probable geographical index based on the source features (the URL the document was retrieved from).

It is important to note that indexes created by this method are on the level of counties. This is due to a couple of reasons. The smallest area addressed by various news portals usually refers to a county. Moreover, there is only about 300 counties in Poland whereas our database of cities contains about 50000 of entries. Developing the method for the level of cities would lead to substantial increase in size of a training and evaluation set for the method. County names are also less ambiguous than cities, what also is an issue while developing a probability-based model.

### 4.5.3. Calculation of features

In this method each URL in the system is represented as a structure containing the following values:

- string fullUrl (the complete URL of resource)
- string protocol (lowercase name of protocol used to access resources e.g. http)
- string domain (complete domain name)
- string remainingURL (part of the URL coming after domain)
- array of strings regexTokens (array of tokens within remainingURL, token is understood as any sequence of letters, digits and url-encoded hexadecimal values)
- array of strings sepTokens (array of tokens within remainingUrl, token is understood as any sequence of non-separators, surrounded by separators)
- set of strings parameters (parameter names encoded in HTTP GET way, preceded by "?" or "&" and followed by "=")
- set of strings paramsValues (all pairs of HTTP GET parameters with associated values)

Each URL is also assigned a list of RSS channels (sources) it is present in (called urlSources).

Based on such URL representation features for maximum entropy model training are calculated. The classes of features, altogether with information on a part of URL description that acts as their input, are listed in the following table.

**Table 11. The classes of URL features**

| Code | Description | Input | Variables |
|---|---|---|---|
| P1 | Protocol equal to | protocol | token |
| D1 | Top domain equal to | domain | topDomain |
| D2 | Is in subdomain of | domain | superDomain |
| D3 | Full domain equal to | domain | fullDomain |
| D4 | Domain contains geo name (from gazetteer45) | domain | geoName |
| RU1 | Empty remaining url | remainingUrl | |
| RU2 | Contains specific regex-token | regexTokens | token |
| RU3 | Contains specific separators-token | sepTokens | token |
| RU4 | Contains specific separators-token at specific position | sepTokens | token\|pos |
| RU5 | Contains specific separators-token at specific position (from end) | sepTokens | token\|pos |
| RU6 | Contains specific bigram of separators-tokens | sepTokens | token1\|token2 |
| RU7 | Contains specific regex-token at specific position | regexTokens | token\|pos |
| RU8 | Contains specific regex-token at specific position (from end) | regexTokens | token\|pos |
| RU9 | Contains specific bigram of regex-tokens | regexTokens | token1\|token2 |
| RU10 | Contains specific URL param | params | param |
| RU11 | Contains specific URL param with specific value | paramsValues | param\|value |
| FU1 | Starts with specific prefix (protocol + domain + sequence of separator tokens) | fullUrl | prefix |
| S1 | Present in specific RSS channel (source) | urlSources | sourceId |

Each URL is checked against all features classes and for each possible value of specific variable (e.g. for each token in case of RU2 and RU3) a separate candidate feature is created

[45] For names composed of multiple tokens the gazetteer contained three variants: joined tokens (e.g. szklarskaporeba), hypen-sperated tokens (szklarska-poreba) and dot-separated tokens (szklarska.poreba)

(if does not exist yet). Information on candidate features true for specific URLs is also saved and further used as a part of input for maximum entropy model training.

### 4.5.4. Application details

As an input for training, we use documents with geographical indexes created using the document-based spatial indexing method described in the previous section (excluding the source-based heuristics). Documents with indexes pointing to two or more counties with the same relevance weight are excluded from training i.e. only documents that may be assigned the most probable county based on their document-based index are included into the training set. In case the MEP model produces a probability distribution, where two or more counties are assigned the same level of probability (differing less than 0,01), either the maximum area or maximum population heuristics is applied for evaluation purposes.

In case of features, if there is a feature that a model does not know i.e. it doesn't exist in the training set, it will be ignored by the MEP model at application time.

The proposed source-based spatial indexing method was checked against a set of over 50000 documents. The detailed outcomes of this evaluation are presented in Chapter 5.

## *4.6. Summary*

This section introduced the requirements towards the method of spatial indexing for the needs of creation of company profiles as well as provided description of two methods enabling for spatial indexing of documents retrieved from Internet sources.

Then, an overview of the document-based spatial indexing method was presented. The method developed for the Polish language benefits from information extraction experiences and addresses challenges emerging from the Polish language and the domain of interest. It also applies the Semantic Web ideas i.e. geographical ontology. The application of the geographical ontology enables inter alia for detailed disambiguation of references. Overall, the method proposed is a novel to the topic and may be used by geographical information retrieval while indexing and querying for documents.

The second method presented, namely the source-based document indexing method, builds on top of the previous approach, trying to assign the most probable document index based on features of an URL address the document was retrieved from. This is a novel approach, that to the best of our knowledge was not applied before for the document indexing.

Next chapter presents in detail evaluation outcomes of both proposed methods.

# Chapter 5.
# Evaluation of spatial indexing methods

*Imagination is more important than knowledge.*
*Albert Einstein*

## 5.1. Evaluation approach

This chapter elaborates on evaluation of the document-based spatial indexing method as well as the source-based document indexing method proposed in this dissertation. The aim is to check the efficiency of the proposed methods in terms of precision and recall and therefore prove their usefulness for the public relations domain.

The evaluation that was carried out, consisted of three phases:

1. Preparatory phase: development of resources for the needs of the evaluation.
2. Evaluation of the spatial indexing method (SIM).
3. Evaluation of the source-based indexing method and its comparison with the spatial indexing method.

This chapter is structured according to these three steps and finishes with concluding remarks.

## 5.2. Resources

Both methods to be carefully evaluated needed vast amounts of data. In case of the document-based spatial indexing method this concerned development of an annotated corpus of documents. Whereas in case of the source-based document indexing mechanism a set of documents with URLs they were retrieved from accompanied by geographical indexes needed to be developed.

The procedure of preparation of resources for the needs of evaluation preformed is elaborated below and presented in Figure 38.

The method of spatial indexing is designed to deal with texts written in Polish that are of interest to PR experts. We needed not only articles referring to Poland, but also we were interested in portals that change their content frequently providing news or recent highlights that may be of interest of PR practitioners. We decided to use RSS feeds that are implemented by many news websites and may be associated also with pages that do not provide RSS. In order not to search the Internet for RSS feeds, we decided to use DMOZ[46] – a directory of the

[46] http://www.dmoz.org/

Web, constructed and maintained by a community of editors, to build an initial list of possible sources (however this list could be also delivered by a PR expert). We filtered DMOZ according to category *Top/World/Polska/Regionalne/Europa/Polska/** and then scanned the main pages in search for RSS channels. Overall, we received 2063 addresses of RSS Channels. However, 11 feed channels seem not to work anymore. Therefore, 2052 channels were monitored from 02.06.2008 to 05.08.2008.

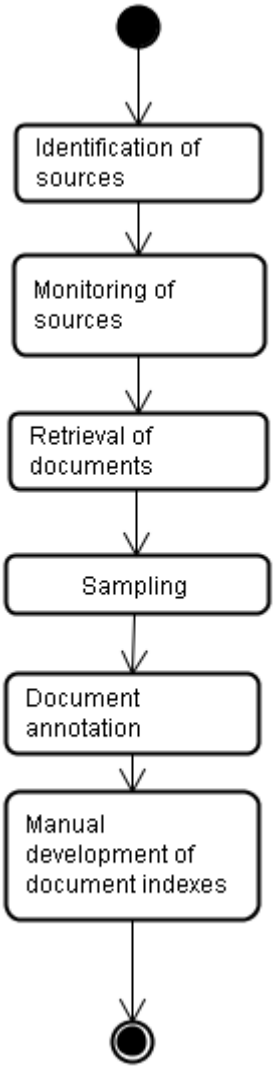



**Figure 38. Evaluation - preparatory phase. UML Activity Diagram**

The statistics for the monitored feeds is as follows:

**Table 12. Statistics for monitoring of feeds**

| **Category** | **Number** |
|---|---|
| Feeds monitored | 2052 |
| Items (number of URLs in all feeds) | 70175 |
| Items per feed (average) | 34,2 |

Then for all feeds' items, URLs and source articles were retrieved and stored in a database. However, not all URLs were valid or still working. Some URL addresses pointed to pages containing viruses, binary or unreadable files. The statistics for URLs is as follows:

**Table 13. Statistics for document retrieval**

| Category | Number |
|---|---|
| All URLs | 70175 |
| Working URLs (items retrieved) | 53193 |

The documents retrieved created our initial corpus of documents. In order to be able to apply the information extraction techniques to build indexes of single documents, the documents were parsed and only the "valid" content (without tags, advertisements, etc.) was stored. Based on this set of documents, two samples were chosen – sample A consisting of 100 documents that were further manually annotated and sample B consisting of 37235 documents for which the geographical indexes were generated automatically using the document-based indexing method. More details on these two samples and their utilization in the evaluation process is presented in the following sections.

### 5.2.1 Annotation guidelines

This section provides an overview of annotation guidelines that were used while annotating corpus of documents for the need of evaluation of the document-based spatial indexing mechanism. In particular, it presents three major issues and problems, which had to be tackled: entity type ambiguity, specifying name mention borders, and finally inner bracketing of the matched text fragments.

Annotation guidelines developed, cover also organizations and persons being the subject of work of (Wieloch, 2010). These guidelines were described in detail in (Abramowicz et al., 2006).

Annotation task focuses on detecting mentions of geographical entities in free-text data. We consider an entity be an object or a set of objects in the real world. Entities can be referenced in a free text by: (a) their name, (b) a common noun phrase (c) a pronoun or (d) an implicit mention in elliptical constructions (e.g., in Polish, subject is often missing in clausal constructions, but it can be inferred from the suffix of a corresponding verb form).

Some ideas while developing guidelines were borrowed from MUC (Chinchor, 1998) and ACE annotation guidelines and taxonomies (ACE, 2009), prepared for other languages and

domains. Originally ACE program specified 7 basic categories: organizations, geo-political entities, locations, persons, facilities, vehicles and weapons. They were used as basis for specifying the annotation task.

Further, we added also annotations for category product, since product names often include valuable clues such as brand and company names, which can be utilized for inferring locations and might implicitly constitute a strong indicator of real estate price level etc. Currently, our annotation guidelines cover four main types of entities:

- Locations (LOC) (natural land forms, water bodies, geographical and political regions, man-made permanent structures, addresses, etc.)
- Organizations (ORG) (companies, government institutions, educational institutions, and other groups of people defined by an organizational structure)
- Persons (PER) (individuals or groups of humans)
- Products (PRD) (brand names, services, goods)

Clearly, LOC is the most structured of the entity types. Its main purpose is to group together entities, which are relevant for the spatial indexing. It groups such entities like: natural land forms (LAN) (e.g. continent names, geographical regions), water bodies (WAT) and administrative regions (ADM). Administrative regions are subdivided into: countries (CRY), provinces (PRO), counties (CNT), communes (CMN), cities (CIT) and other zones (ZON). Cities (CIT) are subdivided into: addresses (ADR) containing zip codes, building numbers, geographical coordinates and URL's or e-mails, districts (DIS) and landmarks (LDM).

Detecting named entities (annotating documents) consists of assigning each name mention in the source document one or more tags corresponding to the type of the mentioned entity, which is accompanied by positional information. Due to eventual type ambiguities, difficulties in specifying name mention borders and subtleties of Polish, we have introduced some annotation guidelines described in more detail below.

**Ambiguities**

Type ambiguity of named-entities is a well-known problem. As it was already described in section 3, there are two types of ambiguities in the geographical domain, namely geo/geo and geo/non-geo. Based on the expert knowledge, we were able to deal with all the ambiguities. The assumption taken for this annotation is that annotator is able to find a correct reference for the mention.

Other ambiguities e.g. concerning tailoring particular name mentions to real-world object, e.g., there are ca. 70 cities in Poland named *Zalesie* and several companies called *POLSOFT*, were also not handled by associating appropriate attributes.

**Name Mentions Border Detection**

Specification of what actually constitutes a name mention in Polish may be somewhat problematic. First of all, we apply the longest-match strategy, i.e., we take as many tokens which are potentially part of the name as possible.

We also consider some nominal constructions, consisting of simple lowercased common noun phrases followed by a proper name as name mentions. Let us consider the phrase *Most Św. Rocha* which is a name of a bridge. It could be alternatively mentioned in the text as *most Św. Rocha*. Without discussing the subtleties of Polish orthography w.r.t. capitalization and the style commonly used in the newspapers etc., we decided to treat both variants as name mentions as far as the leading common noun phrases is potentially a part of the full-name (as in our example).

For solving the problem of name mention borders, we use further rules:

- In case of addresses all keywords, e.g. *ul.*, *Al.*, *al.*, *Plac*, etc. are a part of the name mention (likewise strategy is followed for some other location subtypes). However, specific numbers of buildings are left out of the annotation.
- If deleting a lowercased common noun phrase keyword, e.g., *pomnik* in *pomnik Adama Mickiewicza* (monument of Adam Mickiewicz), results in a name (here: *Adam Mickiewicz*), which does not mach the same entity type (which is the case in our example), then such a keyword is a part of the name mention. Constructions like: *powiat koszaliński* (county of Koszalin), *ocean atlantycki* (Atlantic Ocean) are further examples of this type. As a counter example, consider the keyword *rzeka* (river) in *rzeka Odra*. Here, deleting *rzeka* does not change the type of *Odra* (in the same context). Hence, the keyword *rzeka* is not treated as apart of the name mention.

**Inner Bracketing**

Once name mention boundaries are identified, we eventually add some internal bracketing which reflects the inner structure of the mention to some extent. Consider the following name mentions enriched with inner bracketing.

- [[ul. [Jana III Sobieskiego PER-NAM] LOC-ADM-CIT-ADR] (the street named after Jan III Sobieski, Polish king)

- [[Osiedle [Kopernika PER-NAM] LOC-ADM-CIT-DIS] (the district of buildings named after Copernicus)
- [Kino [Malta LOC-ADM-CIT-DIS & LOC-WAT ] LOC-ADM-CIT-LDM & ORG-REC] (cinema Malta)
- [Giełda Papierów Wartościowych w [Warszawie LOC-ADM-CIT] ORG-COM] (Warsaw Stock Exchange)
- [Uniwersyte Ekonomiczny w [Poznaniu LOC-CIT] ORG-EDU] (university name)

A question arises, how to bracket a given name mention. Intuitively, one would only consider annotations of 'inner' entities which are related to geo-referencing. The following table gives guidelines with examples for entity type combinations (outer – inner), for which inner bracketing is provided.

**Table 14. Entity type combinations**

| | LOC | ORG | PER |
|---|---|---|---|
| LOC | [ul. Biała] 13 (address) | Rondo [ONZ] (United Nations roundabout) | ul. [Jana III Sobieskiego] (street) |
| ORG | AE w [Poznaniu] (university name) | Wydział Prawa [UAM] (Faculty of Law of UAM) | Uniwersytet [Adama Mickiewicza] (university name) |
| PRD | [Warka] Strong | [Microsoft] Exchange | Piwo [Heweliusz] (beer) |

Some complex nominal constructions might pose difficulties while carrying out annotations. Their inner bracketing has to be done carefully. In particular, it is important to differentiate between what we consider a full name and complex noun/prepositional phrases and appositions, which might appear tricky in some context. The following two text fragments clarify the idea:

- [*Szkoła Podstawowa im.* [*Kornela Makuszyńskiego* $_{\text{PER-NAM}}$] *nr. 80 w* [*Poznaniu* $_{\text{LOC-ADM-CIT}}$] $_{\text{ORG-EDU}}$]
- *Siedziba* [*Microsoft* $_{\text{ORG-COM}}$] *w* [*Warszawie* $_{\text{LOC-ADM-CIT}}$] *w* [*Polsce* $_{\text{LOC-ADM-CRY}}$]

The first one happens to be a full-name of the school (with some nested names), whereas the second one constitutes a complex noun phrase consisting of one simple noun phrase followed by two simple preposition phrases, which is unlikely to be a full name. Hence, only *Microsoft*, *Warszawie*, and *Polsce* are tagged.

This corpus of documents was prepared for evaluation of a document-based spatial indexing mechanism. Therefore, we decided not to annotate place names being a part of another named entity. However, for the sake of completeness, integrity and potential utilization of the annotated corpus for other tasks, the same set of documents while annotating was also annotated with all inner entities.

Other issue concerns references such as "na trasie Łódź-Warszawa" (on the way between Łódź and Warsaw). In such case we annotate each city separately.

### 5.2.2 Annotated corpus of documents

Existing language resources for Polish are sparse and inappropriate to tackle the task of evaluation of geoparsing and geotagging mechanism e.g. only morphologically annotated corpora are available (Przepiorkowski, 2007). Therefore, we developed a corpus that besides evaluation of spatial indexing mechanism may be utilized in diverse ways: (a) for automatic learning of patterns for recognition of entities and relations among them (relevant in the geographical domain) and (b) for geographical ontology population.

The corpus annotation was carried out by two persons. The final annotation is a result of two iterations of the process consisting of three phases: (1) definition/tuning of annotation guidelines, (2) annotation, (3) cross-validation. It turned out that ca. 10% of all tags had to be corrected and refined after the first iteration, which reflects the complexity of the annotation task.

As it was already mentioned, based on the initial set of documents retrieved from RSS feeds, we randomly chosen a sample A consisting of 100 documents filtered from different sources, mainly news portals. For these documents we manually developed annotations as well as created document indexes. Statistics for the document corpus are given in Table 14.

Out of 100 HTML documents, only 91 could be successfully parsed. From them 80 documents contained geographical references. In total 1206 annotations in 80 documents were made.

**Table 15. Corpus statistics**

| Volume | Documents | Empty documents | Documents without geographical references | Tags | Tags per document |
|---|---|---|---|---|---|
| 2,69 MB | 100 | 9 | 11 | 1206 | 13,25 |

More fine-grained data accompanied by some examples is given in the table below.

**Table 16. Annotation statistics and examples**

| Category | Total | Examples |
|---|---|---|
| **LOC** | **1206** | |
| ADM CIT | 781 | *Warszawa, Kamionki* |
| ADM CMN | 25 | *gmina Błonie, gminy Lanckorona* |
| ADM CNT | 17 | *powiatu brzeskiego* |
| ADM CRY | 123 | *Polska* |
| ADM PRO | 16 | *woj. wielkopolskie* |
| ADM ZON | 39 | *Puszcza Zielonka* |
| CIT DIS | 73 | *Rataje* |
| CIT ADR | 53 | *ul. Dąbrowskiego 42* |
| CIT LDM | 34 | *Wawel, Zamek Królewski* |
| LAN | 32 | *Europy, Afryki, Kilimandżaro* |
| WAT | 13 | *Bałtyk, Odra* |

For carrying out the annotation task we have chosen Callisto tool (Day et al., 2004) which supports linguistic annotation of textual sources for any Unicode-supported language and allows for defining user-defined domain and task specific tags. Callisto produces a standoff annotation in AIF (ATLAS Interchange Format) format (Laprun et al., 2002). AIF format, implemented as an XML application, offers good properties in respect with extensibility and facilitates widespread exchange and reuse of annotation data.

## 5.3. Ontology validation

In the first phase of our research, before developing both spatial indexing mechanisms, the geographical ontology was validated against a set of competency questions presented in Chapter 4 in an informal way. For the needs of the validation process, all competency questions were translated into WSML. The geographical ontology was validated against these competency questions using the IRIS reasoner[47] integrated with the WSMO Studio[48].

The assumption that was made is that the instance name is a variant of a place name not including Polish diacritics and capital letters.

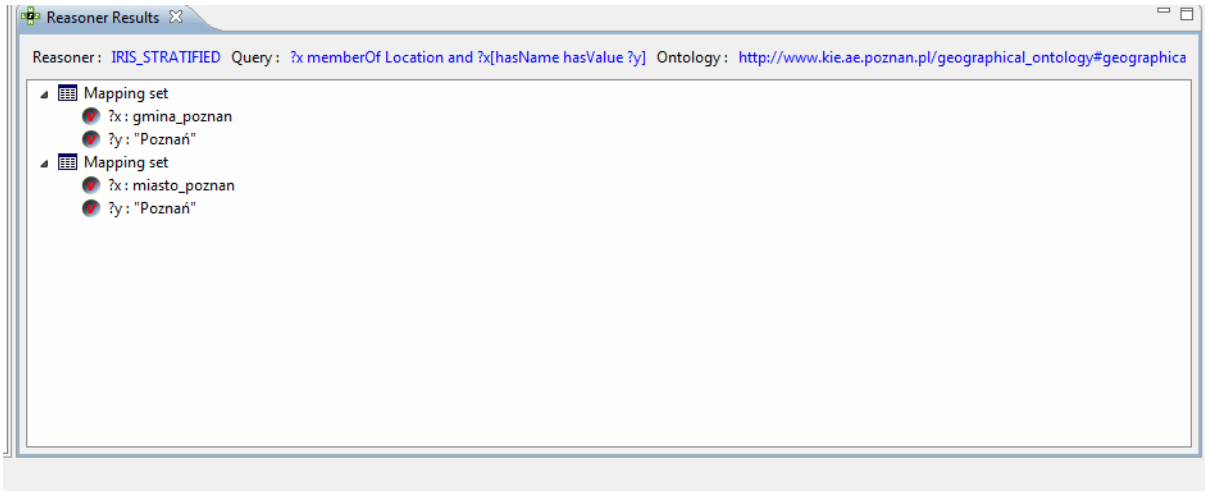
[47] http://www.iris-reasoner.org/

[48] http://www.wsmostudio.org/

Below each competency question is accompanied with its WSML representation and the reasoning outcome. For the needs of presentation clarity, the number of instances in the ontology was decreased, however the questions were tested on the ontology including instances of locations from the whole Poland.

1. What is a name of a given Location (e.g. city or county)?

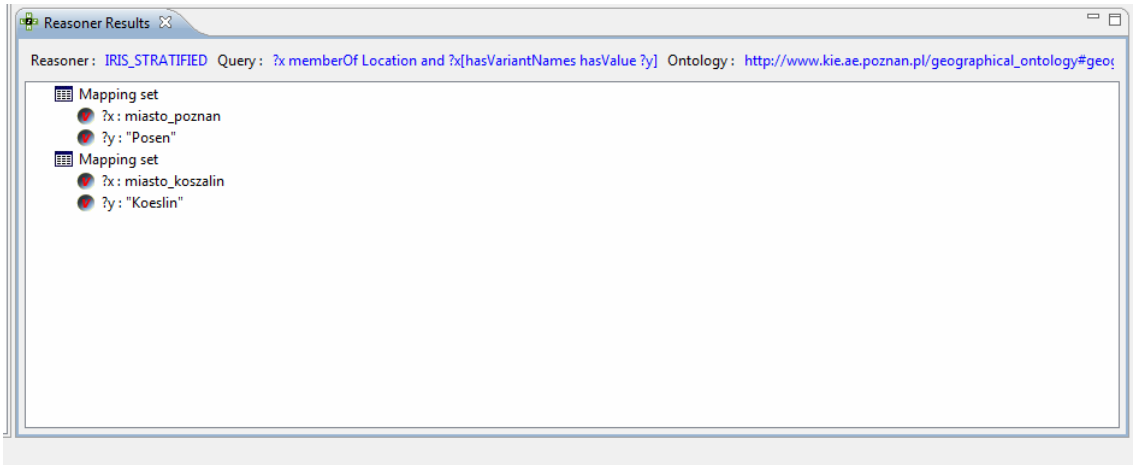
```
?x memberOf Location and ?x[hasName hasValue ?y]
```




In the initial version of the ontology the Location concept was modelled following the SPIRIT guidelines i.e. a name of the location was of type LocationName being a concept in the ontology. This approach however, proved to be incorrect as in Poland there are cities, communes and counties bearing the same name, what led to inconsistencies. After a validation of this approach, the attribute hasName of the Location concept is of type string.

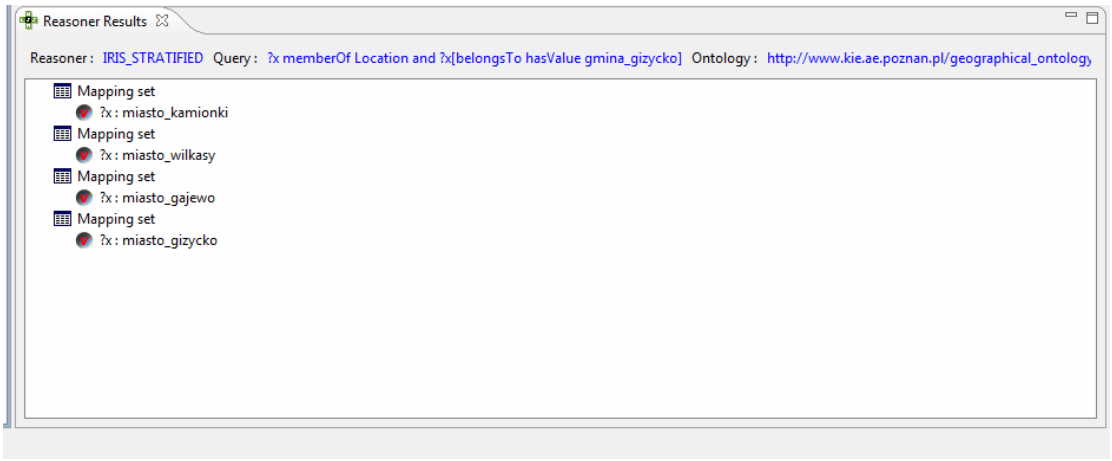
2. What is an alternative name of a Location?

```
?x memberOf Location and ?x[hasVariantNames hasValue ?y]
```



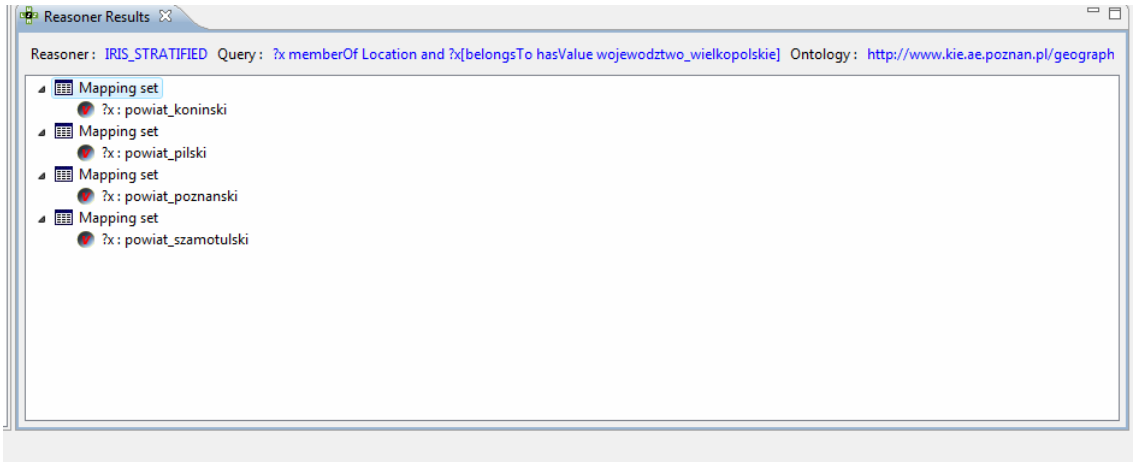
3. What cities are situated in a given commune?

```
?x memberOf Location and ?x[belongsTo hasValue gmina_gizycko]
```



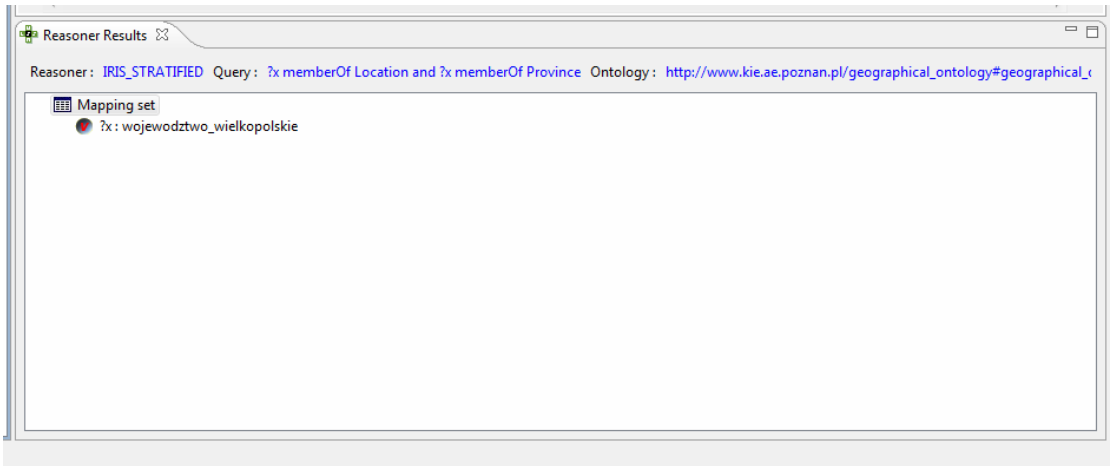

4. What communes belong to a given province?

```
?x memberOf Location and ?x[belongsTo hasValue wojewodztwo_wielkopolskie]
```

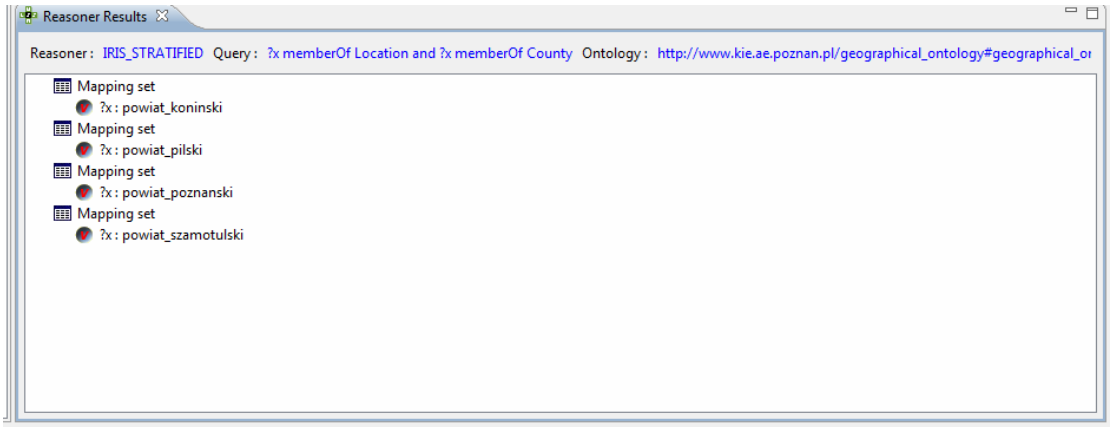



5. What is a type of the place spotted in a document?

```
?x memberOf Location and ?x memberOf Province
```



```
?x memberOf Location and ?x memberOf County
```




6. What is the geographical footprint of the Location?

```
?x memberOf City and  ?x[hasGeometry hasValue ?y]
```


7. What is the time stamp and language of name of a given county?

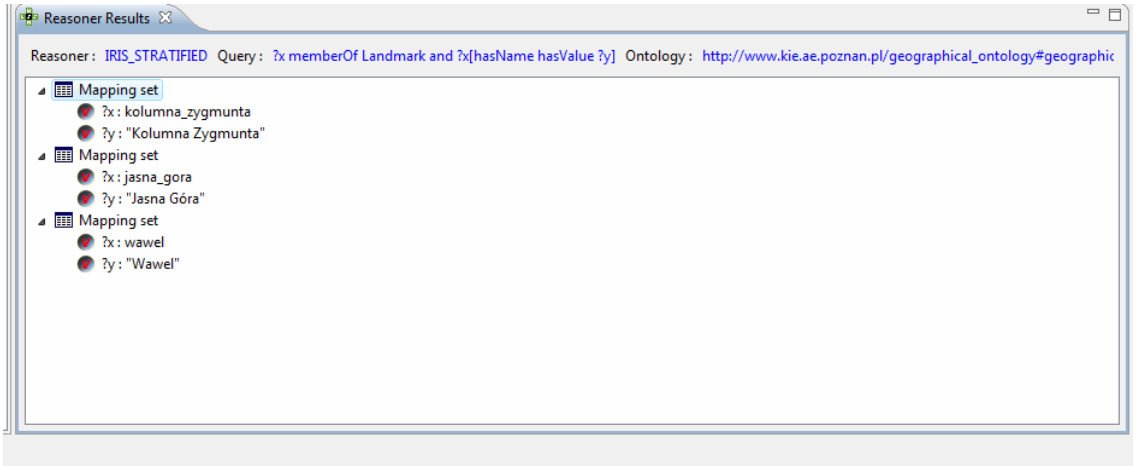
```
?x memberOf County and  ?x[hasLanguageVariant hasValue ?y] and
?x[hasDateOfOrigin hasValue ?z]
```

8. What is the area of a given county?

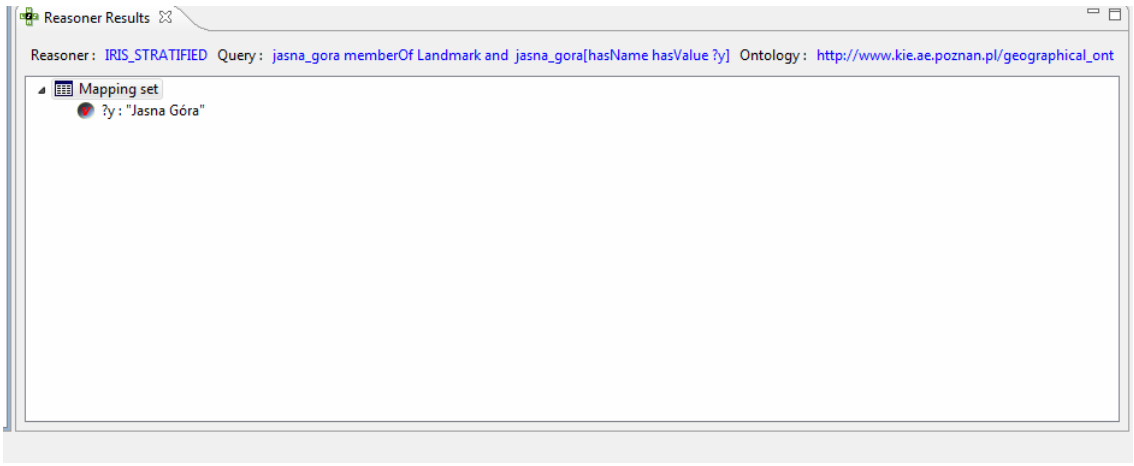
```
?x memberOf County and  ?x[hasAreaInSquareMetres hasValue ?y]
```


9. What is the name of the landmark?

```
?x memberOf Landmark and ?x[hasName hasValue ?y]
```

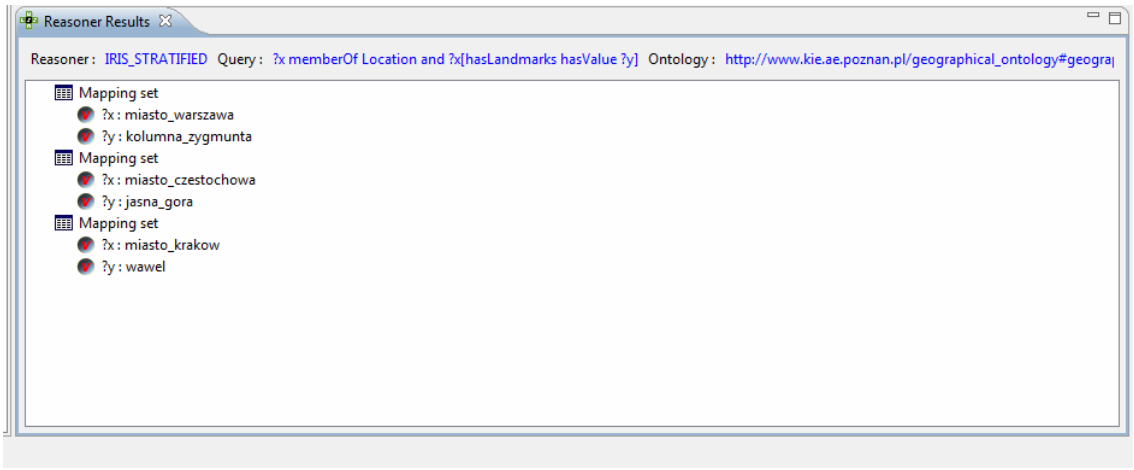



```
jasna_gora memberOf Landmark and  jasna_gora[hasName hasValue ?y]
```



10. What is the name of the city the landmark is situated in?

```
?x memberOf Location and ?x[hasLandmarks hasValue ?y]
```




The ontology evaluation presented above proves that the developed ontology addresses all requirements defined within this research. The evaluation process enabled also an improvement of the geographical ontology especially in case of cardinalities and types of attributes defined for geographical concepts.

Moreover, the ontology meets also the nonfunctional properties defined:

- consistency – all defined concepts are on a similar level of granularity depicting administrative division of Poland. Moreover, the ontology is consistent with resources developed (gazetteer) as well as with type hierarchy defined within SProUT.
- operation – the ontology is operational as it was written in WSML, for which not only ontology modelling tools but also reasoners exist.
- reusability – the ontology is reusable, as it was developed taking into account all previous work in the domain and adjusting it for the needs of the GIR as well as administrative division of Poland.
- clarity – the ontology is readable, well structured (divided into artefacts and location part) and easy interpretable.

As validated the ontology is to be used in the spatial indexing mechanism while preparation of the document index.

## 5.4. Evaluation of the document-based spatial indexing mechanism

This section presents the evaluation outcomes of the document-based spatial indexing method described in Chapter 4. Figure 39 presents shortly the evaluation procedure applied that consists of three steps. Firstly, the spatial indexes using the document-based spatial indexing

method were developed. In the second step these indexes were compared with the reference indexes developed manually in the document annotation process. Further, the analysis of results and common errors was performed. This process was preformed until satisfactory results were achieved. Finally, the method was tested on a set of documents described in the section 5.2.2.

The target accuracy that we defined as our success measure was to outperform the SPIRIT results – one of the most successful initiatives in the area. The SPIRIT initiators achieved the accuracy of around 72% and 25% false positives for all annotations found (Clough, 2005). Then they grounded all geographic references found and achieved the accuracy of about 89%. We may assume that precision of this approach of about 65%.

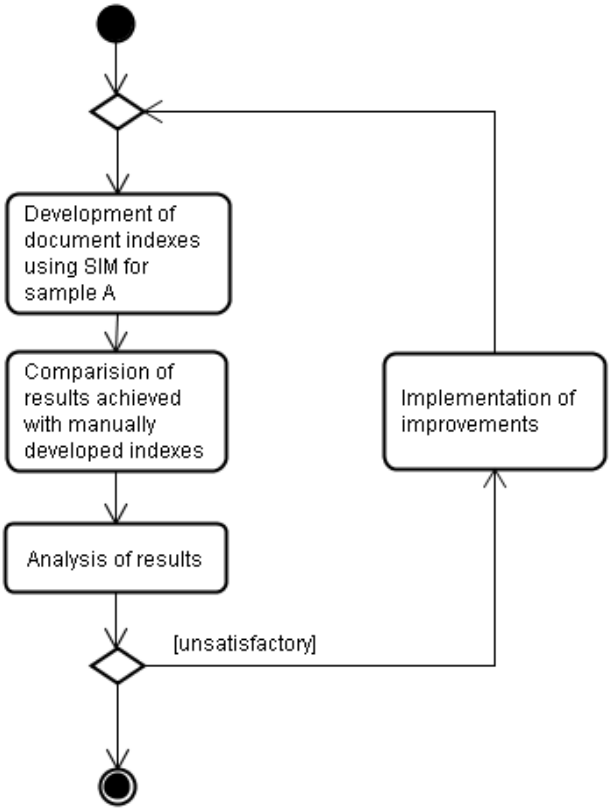



**Figure 39. Evaluation of the document-based spatial indexing method**

Overall, our document-based spatial indexing method achieved the following scores:

Precision: P= 0,9025

Recall: R= 0,8601

The weighted harmonic mean of precision and recall - F-measure: F= 0,8808

Taking into account all challenges for the document-based indexing method presented in the previous chapter, the results obtained for a given collection of documents, are satisfactory. The section below presents the analysis of results achieved.

### 5.4.1 Analysis of results

The evaluation was to identify the quality and accuracy of the indexing mechanism. In order to perform the evaluation we defined four categories of annotations that may be an outcome of indexing:

- Correct (C) being an exact match of the automatically created index with the manual annotation;
- Partially correct (P) being a partial match of the automatically created index with manual annotation. In some information extraction experiments partial and correct annotations are treated as a one set.
- Inserted annotations (I) being annotations of the indexing mechanism that do not have a manual counterpart, they are called also excessive annotations.
- Deleted annotations (D) being annotations that should be automatically created, but were not found within a document.

In the course of research we performed the analysis of indexing results multiple times to improve the measures achieved. The most interesting from our point of view were inserted and deleted annotations.

Based on the performed evaluation we also excluded some rules that for the annotated corpus were providing more false-hits than precise annotations (even when applied with heuristics). Overall out of the set of 36 rules, after evaluation only 15 rules are perceived to be important and are to be used while document indexing. The performance of these rules is presented in table below. For a detailed description of rules please refer to Chapter 4 or to Annex I.

**Table 17. Performance of XTDL rules**

| Rule | Correct annotations | Inserted annotations | Partial annotations |
|---|---|---|---|
| pl_geo_avenue | 2 | 1 | |
| pl_geo_city | 768 | 100 | 47 |
| pl_geo_city2 | | 1 | 22 |
| pl_geo_commune | 16 | | |
| pl_geo_commune2 | 1 | | 1 |
| pl_geo_country | 95 | 14 | |
| pl_geo_county | 5 | | |
| pl_geo_county2 | | | 2 |

| | | | |
|---|---|---|---|
| pl_geo_district | 5 | 2 | |
| pl_geo_lan | | | 1 |
| pl_geo_landmark_3 | 13 | 1 | |
| pl_geo_regions | 26 | 10 | 2 |
| pl_geo_square | 2 | 3 | |
| pl_geo_street | 22 | 1 | 1 |
| pl_geo_streets_2 | 3 | | 1 |

It was mentioned that inserted annotations indicate excessive document indexed. This usually concerns a situation when some words as e.g. Ligi, Lipiec, Małe, Ofiara being not only place names, but also common nouns in Polish are capitalized e.g. at the beginning of a sentence, in a noun phrase. Therefore, they were added to a geographical stop list. Moreover, the geographic indexing was applied without previous annotation with other type of rules, so also place names in names of people and organizations were detected such as Kociołek (surname of Stanisław Kociołek), Kopacz (surname of Ewa Kopacz), Warszawa (in Legia Warszawa – name of a football team), Polska (in Telekomunikacja Polska – Polish telecom). This would not be the case, if extraction of information on organizations and persons would be applied beforehand.

Geographical references in documents that were usually not extracted are inflected geographical names. In gazetteer the inflected forms for the most of the biggest Polish cities and communes, all counties and provinces were included. However, some smaller cities are still not inflected and if they appear without a context in a sentence (there is a reference to a place only using a place name), they are not extracted. Deleted annotations exist also due to the fact that not all landmarks that were mentioned in documents and annotated, were included in gazetteer and in geographical knowledge base that accompanies the ontology developed.

## 5.5. Evaluation of the source-based document indexing method

The following section presents the evaluation of the source-based document indexing method. We compare results achieved using the maximum entropy model with geographical indexes being an outcome of the SProUT annotation as well as compare them with a baseline being a uniform probability distribution.

For the purpose of training and evaluation of the proposed mechanism we used the set of 53193 documents retrieved from RSS feeds. Firstly, we indexed all documents using our document-based indexing method. However, not all documents could be properly parsed. Only 50312 documents were assigned with proper spatial indexes. Documents without a

properly created document index were not excluded from further research because they were retrieved from working URLs. We treat them similarly to documents not pointing to any specific location (for which a geographical index is a vector containing 0 for all counties specified) and in the evaluation phase we assume an equal distribution of probability across all counties.

**Table 18. Statistics for a corpus of documents created using the document-based indexing method**

| **Category** | **Number** |
|---|---|
| Documents | 50 312 |
| Annotations | 717 437 |
| Annotations per document (average) | 14,25 |

Worth to note is that for manually created document indexes for a subset of document corpus the average number of annotations is 13,25. This also underlines validity of the automatically performed annotations. The figure below presents a distribution of references per document.

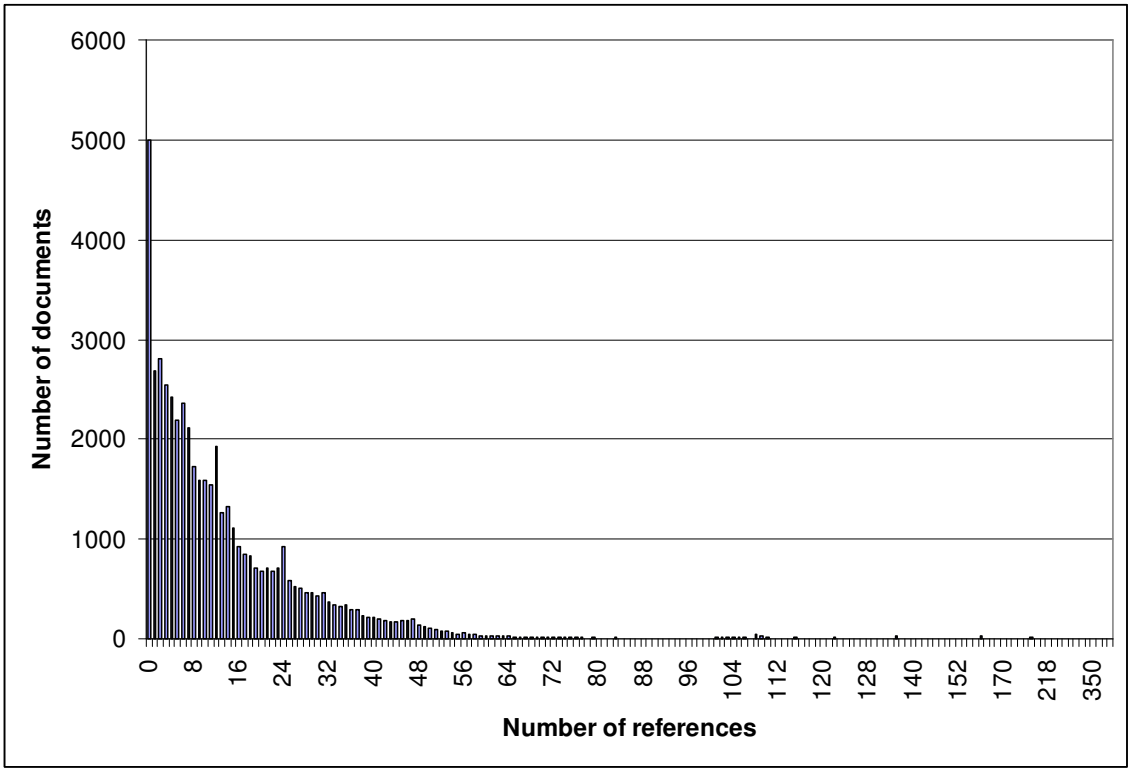



**Figure 40. Distribution of references per document**

For training and evaluation purposes, the set of 50312 documents is divided into two subsets according to 70-30 principle. 37235 form a training set and 15958 are used for evaluation of the mechanism.

The source-based indexing mechanism is to work on the level of counties. In the geographical ontology there are 330 counties in Poland. This is a difference with the official data according to which Poland is divided into 379 counties. This is because the biggest Polish cities have its urban and rural county. From the point of view of public relations these counties may be merged as citizens and companies are interested in news for the whole county area (urban and rural). Therefore, as a result 330 counties instead of existing 379 were included. The mechanism proposed is to associate each URL the most probable county the document under this URL points to. For the reference set of 50312 documents, 37970 were assigned a top spatial index based on the relevance measure defined by the document-based indexing mechanism. It means that within a document one county was assigned a greater relevance for a document that any other existing county. For remaining documents (where two or more counties were assigned with equal importance), either a population or an area-based heuristics was applied i.e. from counties of equal importance for documents we chose a county with a biggest population/area. In 5425 cases outcomes of these heuristics were different, but in 6917 they provided exactly the same result.

Simultaneously, an analysis of URL features was performed. It was decided to not to include in a further research features of URLs that appeared less than 3 and more than 51000 times.

In total 7 experiments were held for which the following success measures were evaluated:

**1. Cosine similarity measure**

The similarity between a reference index and an index being an outcome of the maximum entropy method is measured by means of the cosine angle between vectors that represent the reference index and the maximum entropy index respectively.

**2. Reference place measure**

For each document, the most relevant county is selected based on the document index created. If there are two counties of the same importance, the heuristics based on the area of counties or on number of inhabitants is applied. This measure compares the accuracy of the reference index with an index received using the maximum entropy model.

**3. Geographical distance measure**

This measure assumes that the importance of error of the geographical index of a increases while distance between the reference place and the place being the actual outcome of the mechanism grows. For each document the distance of a capital of the reference county to a

county that is actually the outcome of calculation of the source-based geographical indexing method is calculated. Then, this distance is normalised using the biggest distance between the reference capital and capitals of other counties in Poland (the biggest error possible). The outcome is deducted from 1, and an average for the reference set is produced.

Each of the experiments carried out answered a different research question. Outcomes of all experiments are presented in Table 19. Below a description of each experiment follows.

**Table 19. Outcomes of experiments held**

| Exp | Cosine similarity measure | Reference place measure (accuracy) | Geograph. distance measure | Reference place measure (population heuristics) | Reference place measure (area heuristics) | Geograph. distance measure (population heuristics) | Geograph. distance measure area heuristics) |
|---|---|---|---|---|---|---|---|
| 1 | 0,6431 | 0,7559 | 0,9029 | 0,6732 | 0,6039 | 0,8587 | 0,8291 |
| 2 | 0,5795 | 0,7508 | 0,9004 | 0,6655 | 0,5995 | 0,8544 | 0,8275 |
| 3 | 0,6327 | 0,7396 | 0,8948 | 0,6696 | 0,5989 | 0,8567 | 0,8279 |
| 4 | 0,5922 | 0,7346 | 0,8933 | 0,6615 | 0,5926 | 0,8527 | 0,8241 |
| 5 | 0,6029 | 0,7311 | 0,8913 | 0,6625 | 0,5915 | 0,8536 | 0,8250 |
| 6 | 0,5909 | 0,7500 | 0,8994 | 0,6684 | 0,6017 | 0,8546 | 0,8276 |
| 6 | 0,5909 | 0,7555 | 0,9022 | 0,6677 | 0,6003 | 0,8553 | 0,8283 |
| 6 | 0,5929 | 0,7544 | 0,9051 | 0,6721 | 0,6064 | 0,8596 | 0,8325 |
| *6 AVG* | *0,5916* | *0,7533* | *0,9022* | *0,6694* | *0,6028* | *0,8565* | *0,8295* |
| 7 | 0,5858 | 0,7452 | 0,8963 | 0,6593 | 0,5972 | 0,8508 | 0,8249 |
| 7 | 0,5870 | 0,7524 | 0,9007 | 0,6687 | 0,6037 | 0,8550 | 0,8291 |
| 7 | 0,5920 | 0,7596 | 0,9056 | 0,6741 | 0,6037 | 0,8586 | 0,8300 |
| 7 | 0,5876 | 0,7546 | 0,9032 | 0,6673 | 0,6071 | 0,8555 | 0,8298 |
| 7 | 0,5950 | 0,7678 | 0,9060 | 0,6844 | 0,6139 | 0,8618 | 0,8337 |
| 7 | 0,5916 | 0,7514 | 0,9012 | 0,6720 | 0,6058 | 0,8566 | 0,8297 |
| 7 | 0,5840 | 0,7654 | 0,9090 | 0,6744 | 0,6038 | 0,8586 | 0,8311 |
| *7 AVG* | 0,5890 | 0,7566 | 0,9031 | 0,6715 | 0,6050 | 0,8567 | 0,8298 |

**Experiment 1**

In the first experiment, indexes created using the document-based indexing method were compared with indexes produced by the source-based indexing method. No modifications on

the input data were applied. The research question is as follows: what is the accuracy of the source-based indexing method in comparison with the traditional document-based approach?
Based on the results achieved, it may be summarised that the source-based spatial indexing mechanism provides an acceptable level of quality when creating a document index. It may used as a preprocessing phase for the document-based spatial indexing or in case of processing large amounts of data in a short time, it may even substitute traditional indexing methods.

**Experiment 2**

In the second experiment the research question was: how elimination of the feature concerning the source RSS feed is to affect the indexing results? To address this research question, the feature concerning an RSS feed a document was obtained from was excluded from the training process.
The results of the experiment show however, that this feature has a great impact on quality of a document index. It enables to differentiate different WWW addresses and should be taken into account.

**Experiments 3, 4 and 5**

Experiments 3, 4 and 5 were to check, if modification of the input data underlining differences between importance of different counties in the document index is to affect the initial evaluation outcomes. The assumption was that evaluation measures after such modification should be better than in experiment 1.
To answer such a research question the input data were modified. The source-based indexing method while training takes into account all possible features and only one county (the most relevant for a document), providing as a result distribution of probability. The relevance of a county in the input data may vary from e.g. 0,01 – 1. To underline these differences and importance of a county with a higher relevance measure, it was decided to multiply the input data. If a county weight in a document index exceeded a certain level (multiplication factor), then the input data for this county (and accompanying source features of a document), was multiplied by the outcome of division of the relevance weight for this county by a multiplication factor. The multiplication factor was 1/8 in case of experiment 3, 1/4 in case of experiment 4 and 1/15 in case of experiment 5. The reference index for quantification of the success measures presented in Table 19 was taken from experiment 1.

As it may be easily noticed the evaluation shows, that multiplication worsens the initial results. It may be so because of the overfitting, as the training set of documents is quite small in comparison to a number of source features.

**Experiment 6 and 7**

To check the probable overfitting of the model, experiments 6 and 7 were held. They were to check, if the results achieved are quite stable between the different samples. One of the methods that enable to show how the results of analysis will generalise, is a k-fold cross validation (Mosteller, 1948). The k-fold cross validation is applied e.g. when there is a small training set and a number of parameters in a model is large. In cross-validation input data is divided into disjoint subsets. Then, the training is performed on one subset and validation on the other subset. This process is repeated k-times using different partitions. Finally, achieved validation results are averaged to produce a single estimation.

In experiment 6, a set of documents was divided into 3 subsets. Table 20 presents the average, standard deviation and coefficient of variation for the results achieved. As it may be easily noticed, coefficient of variation is less than 0,001 of average. It means that the results achieved for different samples are very similar.

**Table 20. Experiment 6 - average and standard deviation**

| | **Cosine similarity measure** | **Reference place measure (accuracy)** | **Geograph. distance measure** | **Reference place measure (population heuristics)** | **Reference place measure (area heuristics)** | **Geograph. distance measure (population heuristics)** | **Geograph. distance measure area heuristics)** |
|---|---|---|---|---|---|---|---|
| Avg | 0,5916 | 0,7533 | 0,9022 | 0,6694 | 0,6028 | 0,8565 | 0,8295 |
| Dev | 0,0000009 | 0,0000057 | 0,0000054 | 0,00000373 | 0,0000068 | 0,00000489 | 0,0000047 |
| CV | 0,0002 | 0,0007 | 0,0006 | 0,0006 | 0,0011 | 0,0006 | 0,0006 |

In experiment 7, the data was divided into 7 subsets. Table 21 presents the average, standard deviation and coefficient of variation. In this case also results achieved for each sample were similar, however less similar than in case of 3-fold cross validation.

**Table 21. Experiment 7 - average and standard deviation**

| | Cosine similarity measure | Reference place measure (accuracy) | Geograph. distance measure | Reference place measure (population heuristics) | Reference place measure (area heuristics) | Geograph. distance measure (population heuristics) | Geograph. distance measure area heuristics) |
|---|---|---|---|---|---|---|---|
| Avg | 0,5890 | 0,7566 | 0,9031 | 0,6715 | 0,6050 | 0,8567 | 0,8298 |
| Dev | 0,0000133 | 0,0000558 | 0,0000150 | 0,00005088 | 0,0000214 | 0,00001034 | 0,0000059 |
| CV | 0,0023 | 0,0074 | 0,0017 | 0,0076 | 0,0035 | 0,0012 | 0,0007 |

We may compare the results achieved in all experiments with a baseline. As it was difficult to find similar research outcomes, a uniform distribution of probability across all counties in the document index was taken as a baseline. In this case neither reference place measure nor geometric distance measure could be calculated. All counties were equally probable, so only using appropriate heuristic the most probable county could be chosen. An interesting issue is that applying heuristics for this baseline made all indexes pointing to Warsaw (a county with the biggest population and the biggest area in Poland).

**Table 22. Baseline for experiments held**

| | Cosine similarity measure | Reference place measure (accuracy) | Geograph. distance measure | Reference place measure (population heuristics) | Reference place measure (area heuristics) | Geograph. distance measure (population heuristics) | Geograph. distance measure area heuristics) |
|---|---|---|---|---|---|---|---|
| UD | 0,2327 | | | 0,2055 | 0,1478 | 0,6611 | 0,4763 |

It may be summarised that the performance measures of the source-based indexing mechanism perform much above the baseline, showing also the accuracy of about 70% (that was acceptable in case of document-based indexing in the SPIRIT project elaborated in the previous section). Therefore, this method may be successfully applied when there is a need for a fast document classification.

## 5.6. Validation of the spatial indexing mechanism

The previous sections evaluated the accuracy of both proposed indexing methods. From the application point of view, the source-based indexing method could be used as a preprocessing for the document-based indexing method providing disambiguation heuristics. For validation of a spatial indexing mechanism towards its usefulness for a public relation search engine, the

unified spatial indexing is elaborated. This method is validated against the set of requirements defined in the previous chapter of this dissertation.

Table below presents the comparison of requirements towards and features provided by both indexing mechanisms.

**Table 23. Requirements vs. features of spatial indexing methods developed**

| Requirements | | Spatial indexing mechanism |
|---|---|---|
| **Functional** | Delivery of ontology-based spatial indexing mechanism for news articles retrieved from monitored Web sources, which includes: | The mechanism is designed to work with news articles delivered by RSS feeds (constant stream of news). Each source may be automatically assigned an RSS feed using such tools as http://www.feedity.com. |
| | ○ mechanism for extraction of geographical named entities for the Polish language, | ○ The rule-based mechanism using SProUT for information extraction for documents written in Polish was proposed. |
| | ○ heuristics used for extraction of named entities taking into account specifics of the Polish language, | ○ Disambiguation heuristics were prepared and tested. They benefit e.g. from the word order in the Polish language. |
| | ○ Disambiguation of place names (including geo/geo and geo/non-geo disambiguation), | ○ Disambiguation heuristics were developed and applied. As they are tightly coupled with rules developed, no dedicated accuracy measures for indexing with and without heuristics are provided. |
| | ○ identification of geographical context of web resources. | ○ The source-based indexing mechanism enables to identify geographical scope of a web resource identified by a URL. |
| **Nonfunctional** | • index should be the most precise document surrogate | • The precision and recall of spatial indexing mechanism are about 90% and prove to be more efficient than other known baselines. For the Polish language no comparable mechanism exists. |
| | • development of IE resources for Polish (gazetteer; type hierarchy taking into account specific of administrative division of Poland; extraction rules, corpus of documents) for mechanism evaluation | • IE resources for Polish were developed. These resources may be used in other research. They are to be published on a website of the author. |

| | | |
|---|---|---|
| | • development of geographical ontology for Poland | • The geographical ontology was developed, populated with data and validated. |

The table above shows that all requirements for the spatial indexing mechanism coming from the public relations and GIR domains are satisfied. Next chapter provides exemplary of application of the mechanism for the public relation search engine.

## 5.7. Concluding remarks

The evaluation outcomes of both methods are satisfactory. In case of the document-based spatial indexing, a developed method is much better than a baseline defined. This is especially important when one takes into account that the method was developed for the Polish and not for the English language, and the type of geographical references covered is very similar. The achieved result in comparison to results for other languages also shows that a multilanguage method could be developed and free-word order languages would not much negatively influence its accuracy.

The second, source-based document indexing method proved, that in case a good disambiguation heuristics or a very efficient indexing mechanism is needed, a mechanism based on features of URLs may be applied. This kind of indexing to the best of our knowledge was used previously neither in information retrieval nor in document referencing field. Further feature-related research as well as testing different classification mechanisms could improve the results achieved.

# Chapter 6.
# Application of the spatial indexing methods in the public relation domain

*For a successful technology, reality must take precedence over public relations, for nature cannot be fooled.*

*Richard Feynman*

## 6.1. Introduction

Public relations is "the practice of managing the flow of information between an organization and its publics" (Grunig and Hunt, 1984). Previous chapters described the need for supplying the public relation process both in preparatory and final phases with relevant information on a company and its surroundings retrieved from the Internet sources. They also argued that a new kind of tool support must be provided to the PR analyst.

The proposed spatial indexing methods dealing with description of web resources provide a mechanism that may influence the way today's search engines and public relation portals work.

This chapter presents ideas on how to support a PR analyst with Internet-based tools and concludes with a proposal for a tool benefiting from the spatial indexes of documents.

## 6.2. Public relation support tools and portals

Currently, a public relations practitioner to build an image of a company that is subject to a PR process has to visit a number of different pages usually including news portals, portals aggregating information on different companies, blogs, discussion forums, etc. Then, he aggregates data gathered and produces various reports. This process however, is badly supported by the existing tools. Below some PR support tools are presented and shortly discussed[49].

**Google**

One of the most frequently used tools, is the Google search engine. It is applied while searching the Internet for articles that may influence a company image and that are published on multiple (usually unknown) websites. In Google, a ranked list of documents being a

[49] This list is not supposed to be complete, but provide a general overview of what is available to a PR.

response to a query is presented as a search result (this ranking is based on over 200 different criteria including also a user profile). Then, a user has to manually check potentially interesting documents by clicking on their links. This approach however, is not sufficient. A user is able neither to change the display criteria nor to introduce the semantic understanding into the search process.

Google researchers work (inter alia) on both issues. One of the research results is the way of presentation of a time index enabling for viewing documents only from a given period of time (Figure 42). But a granularity of time cannot be defined.

The support for semantics was recently addressed by Google[50], when it announced support for parsing RDFa statements and microformat properties in HTML web pages and using these statements to enhance the relevance of search results. It may be assumed that using semantic annotations may lead in the future to semi-automatic building of company profiles based on different websites. The first issue regarding changing of display however, is still not addressed. Figure 41 presents the search interface of Google that revolutionarised search engines because of its simplicity.

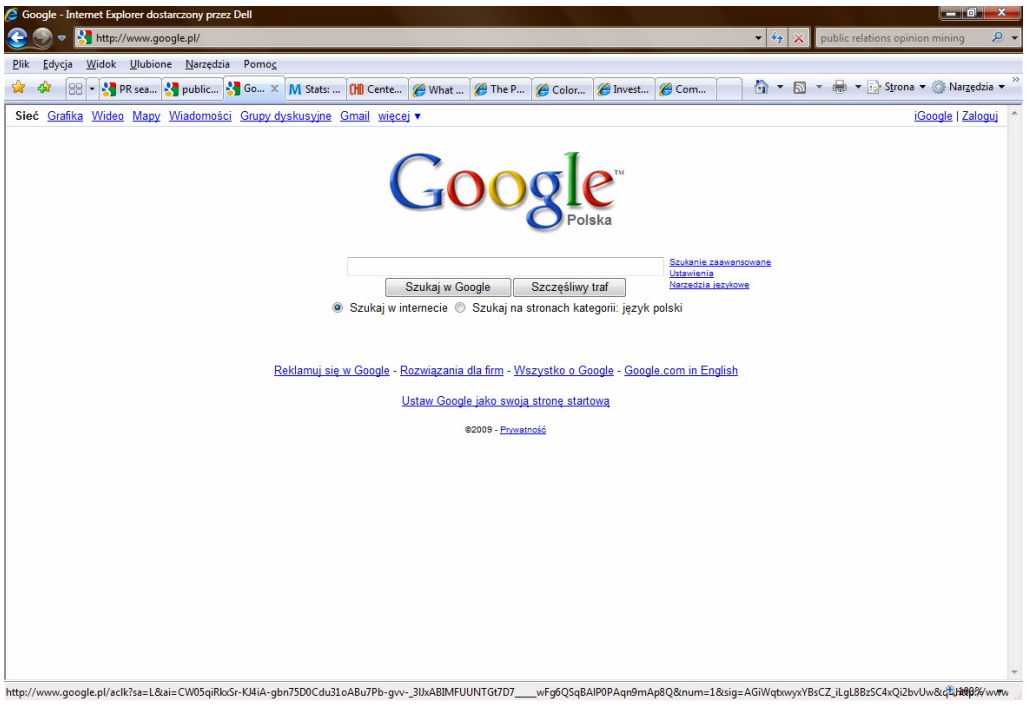



**Figure 41. Google Search Engine. http://www.google.com**

[50] http://googleblog.blogspot.com/2009/05/more-search-options-and-other-updates.html

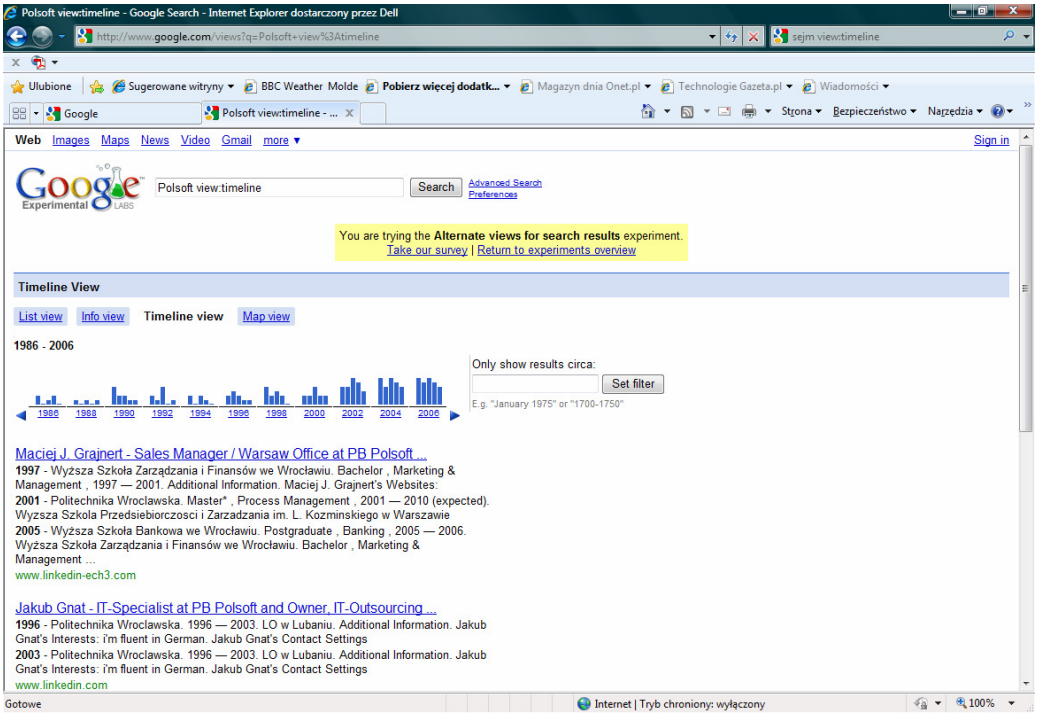



**Figure 42. Google time index. Source: http://www.google.com/views?q=Polsoft+view%3Atimeline**

### Alexandria Digital Library

Alexandria Digital Library (ADL) is a portal with a search engine that uses a combination of time and geographical indexes while searching for documents. The Alexandria Digital Library is an effect of a project held by a consortium of researchers, developers and educators who explore problems related to a distributed digital library with geographically-referenced information. Distributed in this case means that resources may be spread across the Internet. The adjective "geographically-referenced" means that these resources are associated geographical footprints. The searchable information is valuable mainly for supporting basic science including Earth and Social Sciences. The list of included data sources is available at: http://www.alexandria.ucsb.edu/adl/about_adl.html

The ADL supports however only resources that were previously associated with their geographical footprints. Criteria for inclusion in ADL include inter alia:

- the content of included materials should include significant geographical references represented by "graphical footprints" on the map,
- the materials should be of heterogenous data type (multimedia) and enabling heterogenous searching and processing.

This means, that not all types of documents are interesting from the ADL point of view. Moreover, materials to be included in the library are to be initially processed and associated with spatial indexes. The figure below presents the main page of the Alexandria Digital Library.

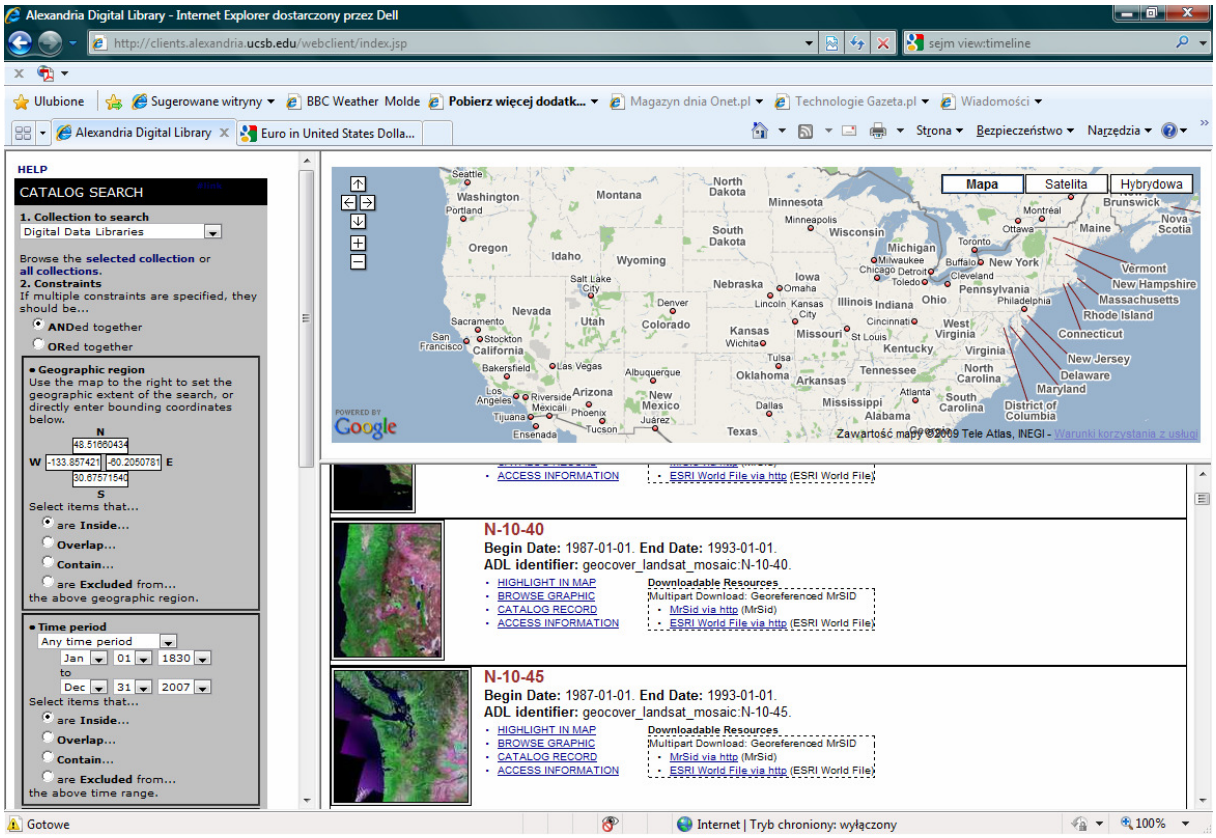



**Figure 43. Alexandria Digital Library.**

**Source: http://clients.alexandria.ucsb.edu/webclient/index.jsp**

### Café News

Other approaches to content searching and presentation include portals such as Café News (http://www.cafenews.pl). Café News enables a user to specify his profile (gender, location, news topic, etc.) and based on this profile, the portal presents the user with personalized information. User is also able to build his portal out of a set of available building blocks. These building blocks present outcomes of monitoring of different, but predefined, information sources. It is important to note that personalization possibilities are quite general. Moreover, the portal does not enable user precisely search or be updated with information only on a company being subject of a PR campaign.

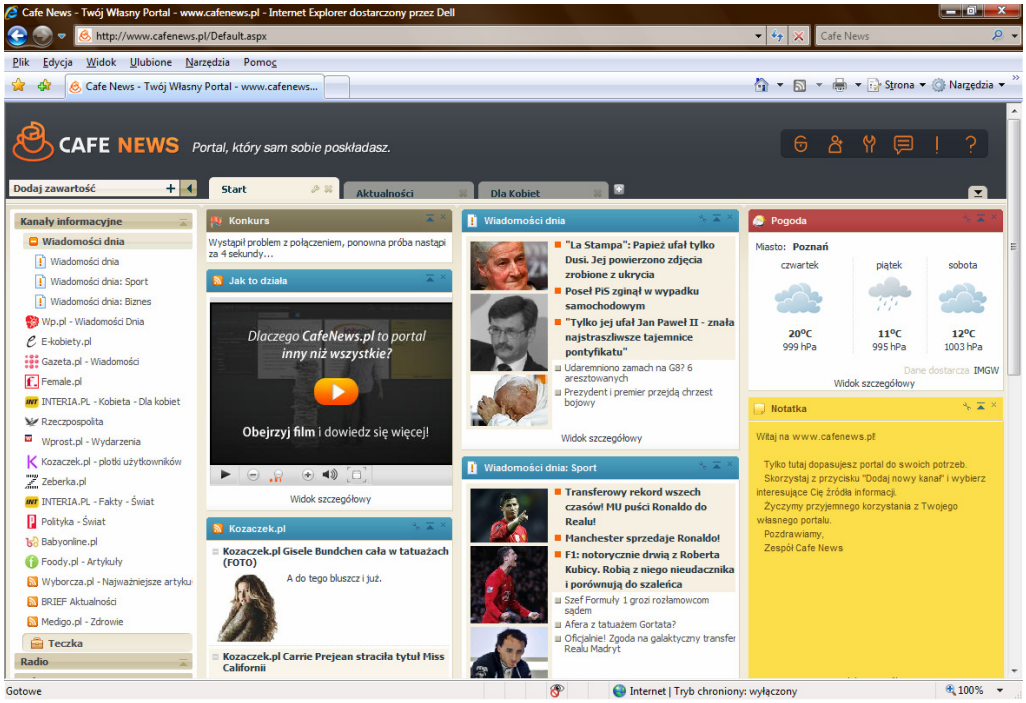



**Figure 44. Café News. Source: http://www.cafenews.pl**

### RadarFarms

An extension of the Café News approach may be found at RadarFarms portal. The originators of the portal aim at monitoring and aggregating information retrieved from various sources in the form of news radars. A news radar is a RSS channel that provides news for a given topic e.g. Cancun, Toronto Real Estate. Full list of radars is available at: http://www.radarfarms.com/?page_id=2

The tool enables monitoring updates of different sources (a user is able also to create a new radar). The portal provides also simple (keyword-based) search capabilities. A user is able to search within a radar (retrieving only interesting news from a radar), within the whole portal or in the Web. Unfortunately, user cannot modify the output ranking, the documents are presented from the most recent to the oldest ones concerning the topic.

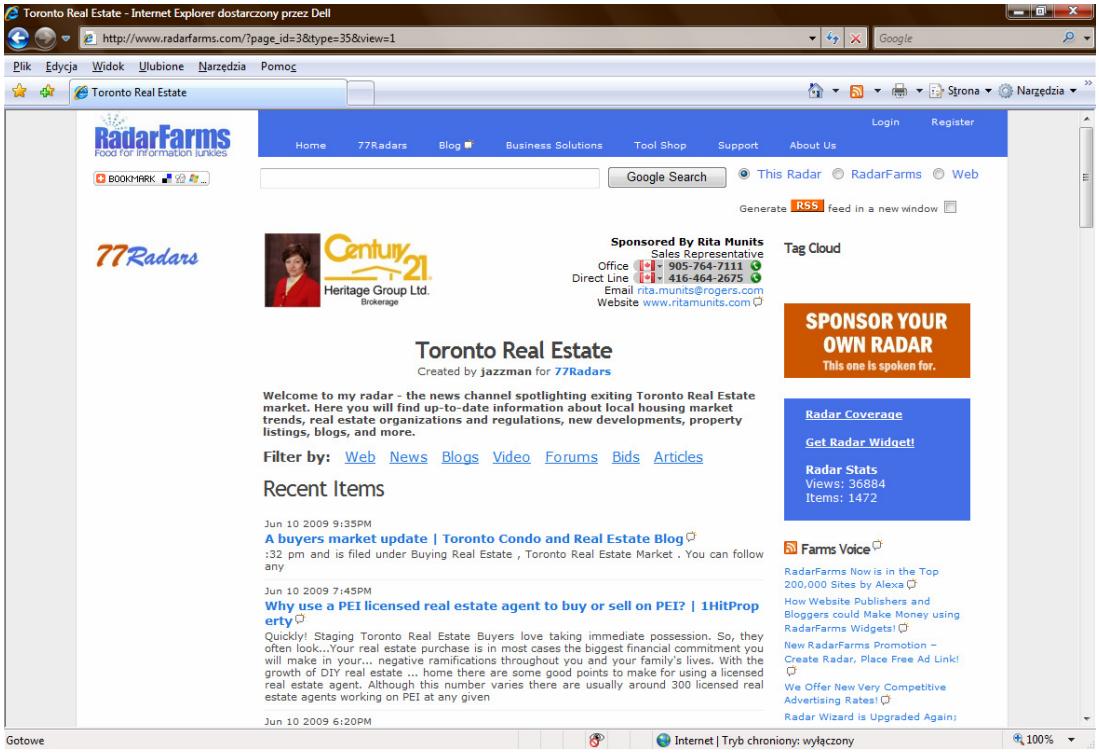



**Figure 45. Radar Farms. Source: http://www. radarfarms.com**

### PR Watch

Another example of a portal supporting PR experts is the PR Watch and its part – the SourceWatch providing news and opinions on PR-related issues in English. PR Watch investigates how the PR industry and other professional institutions manipulate public information, perceptions and opinions. SourceWatch is an "open-content" encyclopaedia of people, groups and issues that shape public agenda. Both of them do not support any personalisation. Regarding the search capabilities, only a keyword-based search is possible. An outcome of this search is a ranked list of documents. User cannot influence this ranking.


**Figure 46. PR Watch. Source: http://www.prwatch.org.**

### Cnet Reviews

There are also portals that are used by PR analysts that perform a kind of opinion mining on different products and services. An example of such a portal is the Cnet Reviews. While searching for a product user besides receiving standard product information, is provided also with user reviews of this product. The portal also aggregates textual comments summarizing the number of positive and negative comments received. The search possibilities are rather limited (keyword-based), however user may influence a ranking of query results (by sorting using the predefined categories e.g. “most-popular”, “lowest price”). The figure below presents the search results for Nokia 6263.

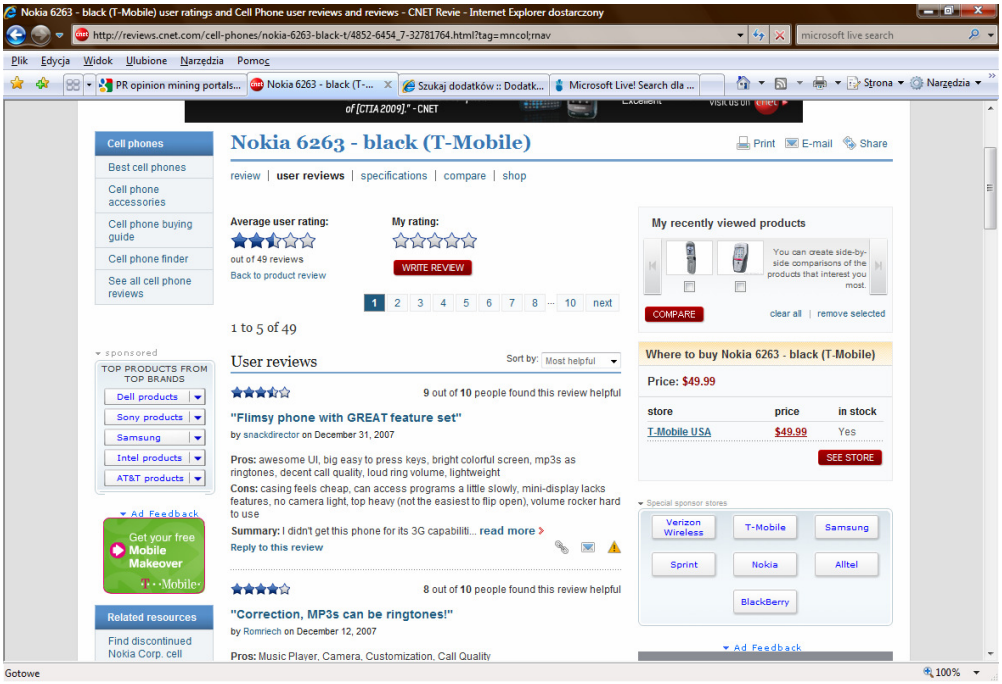



**Figure 47. Cnet Reviews. Source: http://reviews.cnet.com**

**THESEUS TEXO**

An interesting, recent effort in this domain, applying also contextual searching mechanism (based on dimensions of time and geography) is the PR Search developed within the THESEUS Programme funded by the German Federal Ministry of Economy and Technology. The approach taken by TEXO is similar to the one described within this dissertation. A user while searching for information may use a simple keyword-based search. However, there is also a possibility to define additional search dimensions. These search dimensions include inter alia:

- language of a press release,
- time period,
- predefined categories,
- country – based on the document web address.

The user is also able to influence presentation of search results (defining how results should be visualised and sorted). The interface of search engine proposed by TEXO is presented in Figure 48.

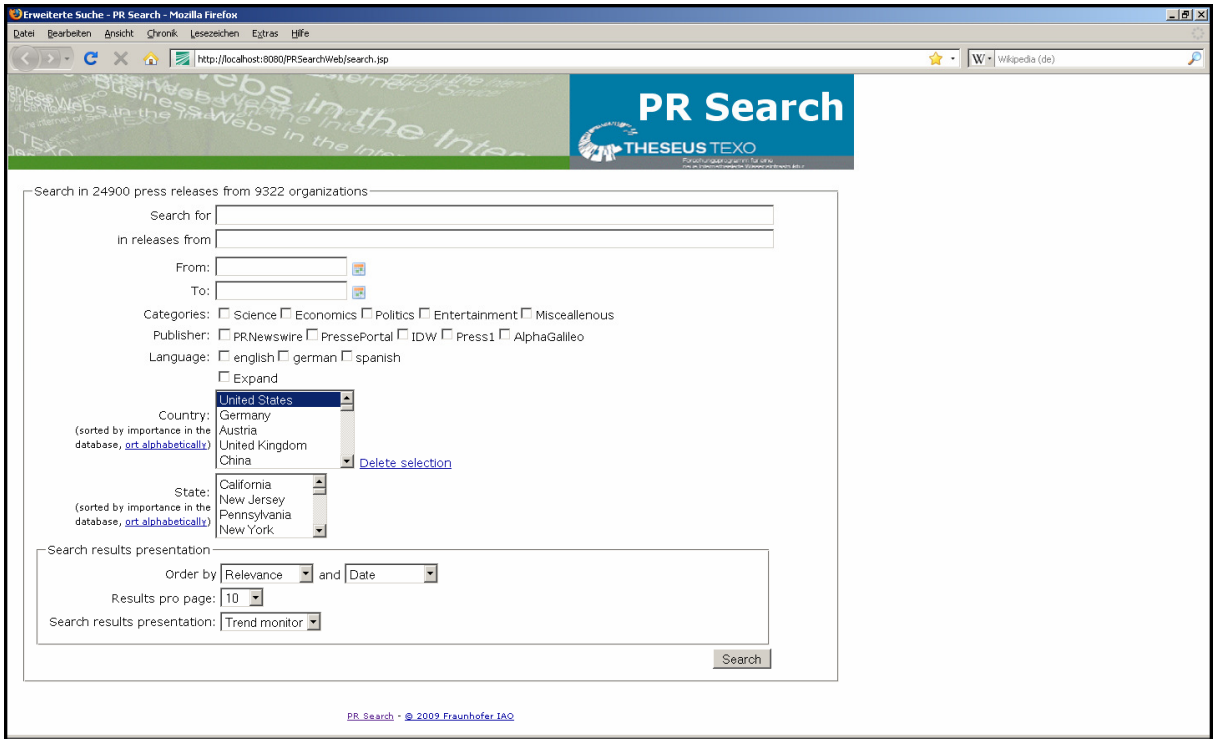



**Figure 48. PR Search. Theseus TEXO**

There is also a number of different portals aggregating press releases that may be further used by public relations analysts. Examples of such portals for the English language are:

- http://www.prnewswire.com
- http://www.alphagalileo.com
- http://www.openpr.com

A summary of approaches with regard to search capabilities is presented in Table 24.

**Table 24. PR tool support**

| Tool / website | Web address | Languages supported | Search features |
|---|---|---|---|
| Alexandria Digital Library | http://clients.alexandria.ucsb.edu/ | English | **IN:** a user may specify information categories within a query – provide address, time period, document type (however only limited resources are indexed and searched)<br>**OUT:** resources matching search criteria are presented and visualised on a map, possibility of browsing the results (for a change of display criteria, a query needs to be reformulated) |

| | | | |
|---|---|---|---|
| Café News | http://www.cafenews.pl | Polish | **IN:** customised and personalised information portal but **no search functionality**<br>**OUT:** aggregated content retrieved from many web sources and RSS feeds; no possibility of visualisation on a map |
| Cnet Reviews | http://reviews.cnet.com | English | **IN:** keyword-based search<br>**OUT:** ranking of items; possibility to narrow search results by choosing one of predefined categories, sorting capabilities (most-popular, review date, lowest price, etc.) |
| Google | http://www.google.com | English, Polish and many other national languages | **IN:** keyword-based search (although a complex mechanism works behind the scenes)<br>**OUT:** document ranking based on features that are unknown to a user and may not be modified (however if a user is logged in, the Google takes into account his profile based on a user's search history) |
| PR Watch | http://www.prwatch.org | English | **IN:** keyword-based search<br>**OUT:** document ranking, no possibility to change a ranking algorithm or sort results differently |
| RadarFarms | http://www.prwatch.org | English | **IN:** keyword-based search<br>**OUT:** possibility of searching within a radar (specific predefined category of news), no possibility to influence the ranking (the time is the only ranking criterion) |
| Theseus TEXO | not available | English, German and Spanish | **IN:** simple and complex (multidimensional) query is possible – advanced search<br>**OUT:** different visualisations of search results are provided, however once displayed may not be modified |

## 6.3. Spatially-enhanced search engine

Section 4 introduced a set of requirements for a search engine that would suit needs of public relations analysts. These requirements included inter alia a possibility to visualize search outcomes on a map, ranking of documents taking into account different search criteria, searching various sources (with an ability to define sources of the utmost importance). Most

of these requirements relate directly to the underlying spatial indexing mechanism that was positively evaluated and validated in the previous chapter.

Taking advantage of this mechanism as well as from methods proposed by (Wieloch, 2010, Bassara, 2009) we may propose a GeoPR search engine that would overcome limitations of the previously described approaches. Figure 49 and Figure 50 present interfaces of a such search engine.

A PR analyst while acquiring and initially processing information on a company, would be offered a tool that would enable him to search the Web taking into account also the criteria of time and geographical area addressed. GeoPR search tool would also enable him to define the most interesting sources of information as well as types of resources being searched.


**Figure 49. Proposed GeoPR search engine interface**

The search results should be presented according to the predefined company profile categories to be useful in the PR campaign. It was discussed, that the usual aim of the PR process is to change the company image. Therefore, the data acquired should be displayed accordingly – taking a company name as a category. The analyst would be also able to visualize and browse search results on a map.

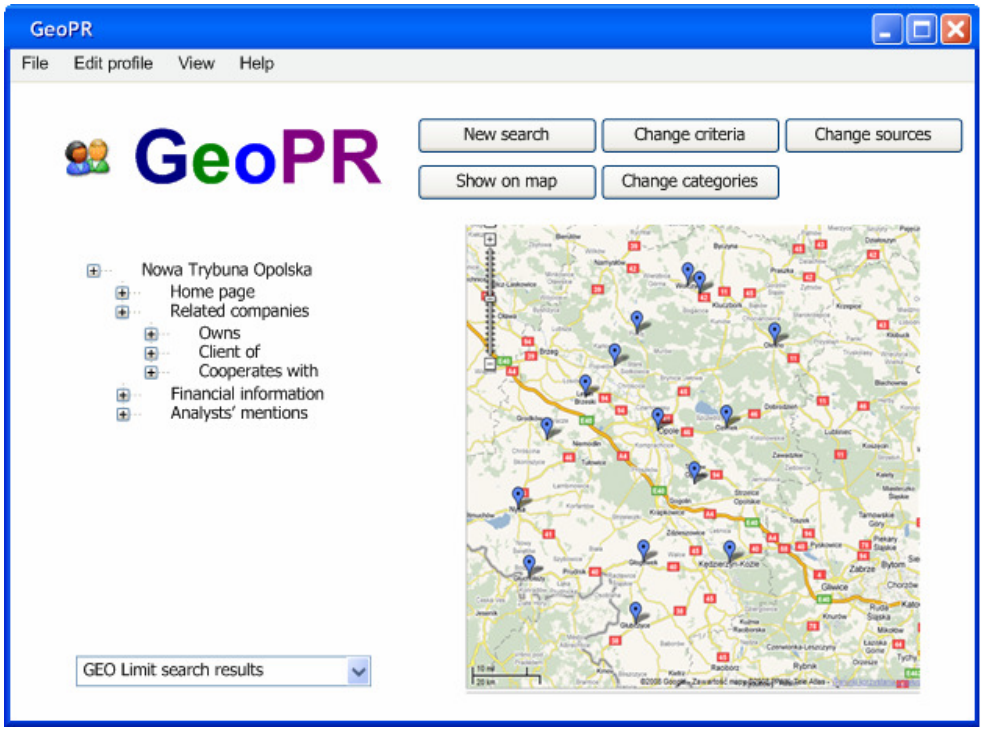



**Figure 50. GeoPR - proposed interface for displaying search results**

Such an approach to query processing and displaying results is possible when resources being searched are properly described. This description includes inter alia creation of a geographical document surrogate, similar to the one produced by mechanisms developed, evaluated and validated within this dissertation.

To summarise - the GeoPR search engine proposed would address all requirements defined for a public relations search engine, as it would enable for:

- viewing results of a query on a map, where points of the map are linked to relevant documents,
- querying for geographic areas which boundaries are not precisely defined but are easily identified by a name (existing in administrative division or only in a common language),
- ranking of documents depending on their spatial and thematic relevance for the query,
- changing ranking of documents based on preferences regarding criteria specified (time, geography, object). This is also enriched by the possibility of ignoring different search criteria and only ranking results according to them.
- handling of different names for the same location especially in case of historical and commonly used names,
- defining high quality sources that should be always presented on top of ranking.

# Conclusions and outlook

The problem addressed in this dissertation concerns development of spatial indexing methods for the needs of creation of company profiles. Currently, to develop a profile of a company, a public relation analyst has to merge opinions of different people and organizations published in multiple forms, most recently including also the Internet. This usually leads to an information overload. To deal with this problem and to benefit from all data acquired, new tools offering search possibilities need to be introduced. These tools however, to present the most relevant data, need to be applied to properly described documents. Such as the ones provided by the spatial indexing mechanisms.

This dissertation investigates in detail requirements of the PR analysts and provides indexing mechanisms that not only address these requirements but also are novel in comparison to other approaches in the domain.

## *Main contribution*

The main aim of is this dissertation was to develop a method of spatial indexing of free text documents filtered from the Internet and written in the Polish language that would support PR practitioners in their everyday work.

Therefore, the major contribution of this dissertation is two fold. On one hand, the dissertation provides a set of requirements towards a PR search engine. The most important conclusion from the analysis is that such a tool should take into account contexts of information and provide search dimensions applying these contexts. Although, the analysis could be broadened using a questionnaire or interviewing a statistically valid set of PR analysts, it still provides important remarks that enable for shaping directions of future research in the domain.

On the other hand, the dissertation provides two novel spatial indexing methods. The document-based spatial indexing method while analysing the document using information extraction approaches, produces a precise and accurate document surrogate. The source-based indexing method provides an acceptable level of quality of a geographical index developed focusing only on features of a document source.

The research hypothesis defined is as follows: *the introduction of a semantic-based accurate and precise geographical indexing mechanism will provide functionalities needed for creation of a new kind of a search engine that may be used within the entry and output phases of the PR process.*

Moreover, the following research goals are defined:

- structuring of public relation analysts' requirements towards a search engine to be used in the entry and output phases of the PR process,
- development of a corpus of news articles written in Polish, which can be used e.g. during an evaluation of the spatial indexing mechanism,
- development of geographical ontology and gazetteer for specification of artefacts emerging from geographical resources for Poland, that may be processed automatically,
- definition of the spatial indexing methods allowing for producing accurate surrogates of documents.

To address the research hypothesis and defined goals, according to the applied methodology several research methods were applied. Based on a detailed literature review from the area of geographical information retrieval, an initial version of the spatial indexing method was proposed. Then, based on the requirements for a PR search engine gathered in the process of analysis of materials from the PR domain as well as during workshops organized with representatives of the Public Relations Department at Poznan University of Economics, the method was modified and further detailed. In the third step, an initial prototype was developed and tested. Results achieved enabled for further alignment of the spatial document indexing method. Finally, the empirical evaluation of the method with regard to its precision and recall on the corpus of manually annotated documents was carried out.

After the first method was validated, a second method enabling indexing of sources was proposed. This method was to be initially used as one of heuristics in the document-based indexing method. Based on the assumption that RSS feeds as well as websites of local newspapers are location-oriented, it was to be applied for disambiguation of geographical name mentions. However, after development of a prototype and first tests, it appeared that this method being a more efficient one in terms of computing resources, may substitute the first one. Final tests proved that the second, source-based document indexing method in about 70% of cases provides the same results as document-based document indexing method.

## Detailed remarks

According to the applied methodology, the dissertation was divided in three parts. In the first part, a detailed overview of the problem domains including public relations, geographical information retrieval and information extraction is provided. The analysis of the domain of public relations shows a need for introduction of sophisticated tools to support the PR

campaign. PR analysts in their research catalogue different types of sources that may be used in introductory as well as in the evaluation phases of the PR process. Between these sources are Internet sources, that constantly gain on importance. These sources also include RSS feeds that may be associated even to web sites that do not include them. PR practitioners claim that employing the most recent ICT developments into tools supporting PR process would lead to a substantial improvement of efficiency of campaigns being carried out.

In addition, current approaches to extracting geography from web resources were introduced and discussed.

In the second part, novel mechanisms for spatial document indexing were proposed. The first method takes advantage of information extraction techniques and information retrieval experience. The second one, proposed a novel approach to document indexing taking into account source features.

In the third part of the dissertation, both methods were evaluated, validated and examples of their application were presented. The implementation of both methods was carefully evaluated using resources prepared within a scope of this research. The document-based indexing mechanism however being 90% as accurate as manual annotation, is quite expensive in terms of computing resources. However, document indices created this way may be substituted by indices produced by the source-based indexing method. The source-based indexing method was initially to provide information on a source that was to be used for disambiguation of named entities within the document-based spatial indexing method. It the course of research it was proven that for our sources, the proposed method in 70% of cases is as accurate as the document-based spatial indexing method and may substitute it.

Worth to mention is that this dissertation also contributes to establishing foundation for analysis of geographical name mentions as well as anaphora resolution for Polish. The corpus developed within this research is to be made public on the author's website.

## *Future work*

Information retrieval is a broad research domain and application of its tools for addressing the PR requirements brings a lot of challenges. This dissertation addresses these issues with regard to the spatial dimension of information. Linking extracted geographical information with other contexts of information such as time, people and organizations, would significantly improve accuracy of methods and usefulness of tools for PR experts. This concerns development of contextual search engines that automatically build profiles of people and organizations based on the free text documents.

Other challenges emerge from the domain of Future Internet initiative supported recently by the European Commission. Future Internet focuses on collaboration between people, things and services in the future pan-european network. In this network an identity of thing, service or a person is to play a key role especially with regard to personalization, authentication and trust. Spatial indexing techniques may enable building identities of people and organizations based on free text of documents available in the Web. Outcomes of these techniques when combined with a user behaviour may enable gathering of data that may be further used for user's convenience e.g. while personalizing content, presenting information of utmost importance, remembering about an important anniversary.

Another domain where spatial indexing might be used is e-government. Today administration wants to react quickly to all kinds of problems. Some of these problems are related to the cyberreality, some of them are only reported in the Internet. For these ones including e.g. increasing crime or traffic problems in a certain area, if the information is collected and delivered in a certain way a Government may react quickly and efficiently. An example of such system is XMedia developed at Sheffield University in a 6th Framework Programme project[51].

## *Final remarks*

Research carried out in the dissertation leads to a conclusion that the public relation practitioners may benefit from application of tools described in the dissertation as well as from other technologies mentioned. Application of these technologies enables an effective usage of such resource as the Internet and in consequence leads to increase in productivity and effectiveness of PR campaigns held.

[51] http://www.x-media-project.org/

# Appendix 1. XTDL SProUT rules (ver. 01.2009)

```
;; GEO_REFERENCES Grammar for Polish
;; version September 2008

;;;
;;; SUPPORTING GRAMMARS - for extraction of names, phrases to geographical names
;;;

;;
;; extraction of given names
;;

pl_person_gaz_givenname :/ (
gazetteer & [G_CONCEPT #name, GTYPE gaz_given_name, G_GENDER #gen, G_CASE #case,
SURFACE #surf, CSTART #cs,
CEND #ce]
) -> ne-person & [GIVEN_NAME #name, SEX #gen, SURFACE #surf, NCSTART #cs, NCEND
#ce].

;;
;; extraction of initial
;; "M." and middle name

pl_person_initial :/ (
token & [TYPE comma] ?
gazetteer & [GTYPE gaz_initial, SURFACE #initial, CSTART #cs]
token & [SURFACE "."] ?
) -> sign & [SURFACE #initial, CSTART #cs], where #middle = Append(#initial, ".").

;;
;; extraction of surnames
;;

pl_person_gaz_surname :/ (
gazetteer & [G_CONCEPT #surname, GTYPE gaz_surname, SURFACE #surf, CSTART #cs, CEND
#ce]
) -> ne-person & [SURNAME #surname, SURFACE #surf, NCSTART #cs, NCEND #ce].

;;
;; a rule for recognizing a sequence of nouns in genitive
;;

pl_noun_genitive_sequence :/
```

```
(morph & [POS noun, SURFACE #surf_1, INFL infl_noun & [CASE_NOUN gen]])
(morph & [POS noun, SURFACE #surf_2, INFL infl_noun & [CASE_NOUN gen]])  ?
(morph & [POS noun, SURFACE #surf_3, INFL infl_noun & [CASE_NOUN gen]])  ?
(morph & [POS noun, SURFACE #surf_4, INFL infl_noun & [CASE_NOUN gen]])  ?
(morph & [POS noun, SURFACE #surf_5, INFL infl_noun & [CASE_NOUN gen]])  ?
-> #surf, where #surf=ConcWithBlanks(#surf_1,#surf_2,#surf_3,#surf_4,#surf_5),
Capitalized(#surf_1),          Capitalized(#surf_2),          Capitalized(#surf_3),
Capitalized(#surf_4), Capitalized
(#surf_5).

;;
;; a rule for capitalized adjectives
;;

pl_capitalized_adjective :/ (morph & [POS adjective, STEM #stem, SURFACE #surface1,
INFL infl_adjective & [CASE_ADJECTIVE #c,
NUMBER_ADJECTIVE #n, GENDER_ADJECTIVE #g]])
->morph & [SURFACE #surface1, STEM #name, INFL infl_adjective & [CASE_ADJECTIVE #c,
NUMBER_ADJECTIVE #n, GENDER_ADJECTIVE #g]],
where Capitalized(#surface1), #name=CapitalizeWord(#stem).

;;
;; a rule for capitalized adjectives with stem correction according to gender
;;

pl_capitalized_adjective_with_special_stem :/
@seek(pl_capitalized_adjective)   &   [SURFACE   #surface1,   STEM   #stem,   INFL
infl_adjective & [CASE_ADJECTIVE #c, NUMBER_ADJECTIVE #n, GENDER_ADJECTIVE #g]]
->morph & [SURFACE #surface1, STEM #name, INFL infl_adjective & [CASE_ADJECTIVE #c,
NUMBER_ADJECTIVE #n, GENDER_ADJECTIVE #g]], where #name=CorrectSuffixPL(#stem,#g).

;;
;;phrases  that  may  appear  in  organisation  names  e.g.  Uniwersytet  im.  Adama
Mickiewicza
;;

pl_org_name_innerwords :/ (
morph & [STEM "w", SURFACE #w1] |
morph & [STEM "i", SURFACE #w1] |
morph & [STEM "na", SURFACE #w1] |
morph & [STEM "nad", SURFACE #w1] |
(token & [SURFACE #w1 & "im"] token & [TYPE dot, SURFACE #w2]) |
morph & [STEM "imię", SURFACE #w1]
) -> #n, where #n=Conc(#w1,#w2).
```

```
;;
;; organisation names and abbreviations
;; e.g. TCH, mBank
;;

pl_entity_single_name :/ (
(
token & [TYPE first_capital_word, SURFACE #n1] |
token & [TYPE all_capital_word, SURFACE #n1] |
token & [TYPE mixed_word_first_lower, SURFACE #n1] |
token & [TYPE mixed_word_first_capital, SURFACE #n1] |
token & [TYPE number_word_first_capital, SURFACE #n1] |
token & [TYPE number_word_first_lower, SURFACE #n1] |
token & [TYPE word_number_first_capital, SURFACE #n1] |
token & [TYPE word_number_first_lower, SURFACE #n1] |
token & [TYPE word_with_hyphen_first_capital, SURFACE #n1] |
token & [TYPE other_symbol, SURFACE #n1]
)
gazetteer & [GTYPE gaz_domain, SURFACE #n2] ?
) -> sign & [SURFACE #name], where #name = Conc(#n1,#n2).

;;
;; extraction of nominal phrases e.g. Nad Wierzbakiem
;;

pl_phrase_PN :/
(
morph & [POS prep, SURFACE #surface1, INFL infl_prep & [CASE_PREP #case]]
morph & [POS noun, SURFACE #surface2, INFL infl_noun & [CASE_NOUN #case]]
)
-> #geo_fraza_PN, where #geo_fraza_PN=ConcWithBlanks(#surface1, #surface2),
Capitalized(#surface1), Capitalized(#surface2).

;;
;; rules for extraction of "advanced" nominal phrases e.g. Nad Zielonym Wierzbakiem
;;

pl_phrase_PAN :/
(
morph & [POS prep, SURFACE #surface1, INFL infl_prep & [CASE_PREP #case]]
morph & [POS adjective, SURFACE #surface2, INFL infl_adjective & [CASE_ADJECTIVE
#case, GENDER_ADJECTIVE #gen, NUMBER_ADJECTIVE #num]]
morph & [POS noun, SURFACE #surface3, INFL infl_noun & [CASE_NOUN #case,
GENDER_NOUN #gen, NUMBER_NOUN #num]]
)
```

```
->   #geo_fraza_PAN,   where   #geo_fraza_PAN=ConcWithBlanks(#surface1,   #surface2,
#surface3),
Capitalized(#surface1), Capitalized(#surface2), Capitalized(#surface3).


;;
;; extraction of dates e.g. 26 czerwca 1956
;;

pl_date :/
(
 token & [TYPE any_natural_number, SURFACE #dzien]
  (        morph & [STEM "styczeń", SURFACE #miesiac]
           | morph & [STEM "luty", SURFACE #miesiac]
           | morph & [STEM "marzec", SURFACE #miesiac]
           | morph & [STEM "kwiecień", SURFACE #miesiac]
           | morph & [STEM "maj", SURFACE #miesiac]
           | morph & [STEM "czerwiec", SURFACE #miesiac]
           | morph & [STEM "lipiec", SURFACE #miesiac]
           | morph & [STEM "sierpień", SURFACE #miesiac]
           | morph & [STEM "wrzesień", SURFACE #miesiac]
           | morph & [STEM "październik", SURFACE #miesiac]
           | morph & [STEM "listopad", SURFACE #miesiac]
           | morph & [STEM "grudzień", SURFACE #miesiac]
   )
)
->#geo_date, where #geo_date=ConcWithBlanks(#dzien, #miesiac, #rok).


pl_date_long :/
(
 token & [TYPE any_natural_number, SURFACE #dzien]
  (        morph & [STEM "styczeń", SURFACE #miesiac]
           | morph & [STEM "luty", SURFACE #miesiac]
           | morph & [STEM "marzec", SURFACE #miesiac]
           | morph & [STEM "kwiecień", SURFACE #miesiac]
           | morph & [STEM "maj", SURFACE #miesiac]
           | morph & [STEM "czerwiec", SURFACE #miesiac]
           | morph & [STEM "lipiec", SURFACE #miesiac]
           | morph & [STEM "sierpień", SURFACE #miesiac]
           | morph & [STEM "wrzesień", SURFACE #miesiac]
           | morph & [STEM "październik", SURFACE #miesiac]
           | morph & [STEM "listopad", SURFACE #miesiac]
           | morph & [STEM "grudzień", SURFACE #miesiac]
   )
 (
```

```
   (token & [TYPE any_natural_number, SURFACE #rok] (token & [SURFACE "r"] token &
[TYPE dot])?)
  | (token & [TYPE any_natural_number, SURFACE #rok] (token & [SURFACE "Roku"]
token & [TYPE dot])))?
)
->#geo_date,     where     #geo_date=ConcWithBlanks(#dzien,     #miesiac,     #rok),
StringLength(#rok, "gt", "3").

;;
;; extraction of noun phrases, extraction of noun-adjective phrases
;;

pl_phrase_NN :/
(
@seek(pl_noun_genitive_sequence) & #geo_fraza_NN
)
-> #geo_phrase_NN.

pl_phrase_A :/
(
 morph & [POS adjective, STEM #stem, SURFACE #surface1, INFL infl_adjective &
[GENDER_ADJECTIVE #g]]
)
->        #geo_phrase_A,        where        #geo_phrase_A=CorrectSuffixPL(#stem,#g),
Capitalized(#surface1).

pl_phrase_NA :/
(
morph & [POS noun, SURFACE #surface1, INFL infl_noun & [CASE_NOUN gen]]
morph & [POS adjective, SURFACE #surface2, INFL infl_adjective & [CASE_ADJECTIVE
gen]]
)
-> #geo_phrase_NA, where #geo_fraza_NA=ConcWithBlanks(#surface1, #surface2).

pl_phrase_AN :/
(
morph & [POS adjective, SURFACE #surface1, INFL infl_adjective & [CASE_ADJECTIVE
gen]]
morph & [POS noun, SURFACE #surface2, INFL infl_noun & [CASE_NOUN gen]]
)
-> #geo_phrase_AN, where #geo_fraza_AN=ConcWithBlanks(#surface1, #surface2).

;;
;; extraction of person names e.g. Jan Maria Rokita
;;
```

```
pl_persons :/
(
(gazetteer & [SURFACE #posit, GTYPE gaz_position]) ?
    ((((@seek(pl_person_gaz_givenname) & [SURFACE #given_name])
     (@seek(pl_person_gaz_givenname) & [SURFACE #given_name_2]) ?)
       | ((@seek(pl_person_initial) & [SURFACE  #initial])(@seek(pl_person_initial)
& [SURFACE #initial2])? ))
       (token & [SURFACE roman_number & #rzymskie])?
      (@seek(pl_person_gaz_surname) & [SURFACE #last_name] )
)
->#pl_geo_osoby,                                                                where
#pl_geo_osoby=ConcWithBlanks(#posit,#given_name,#given_name_2,#initial,#initial2,#r
zymskie,#last_name).

;;
;; extraction of reverse person names e.g. Rokita Jan Maria
;;

pl_persons_reverse :/
(
       (@seek(pl_person_gaz_surname) & [SURFACE #last_name] )
        ((((@seek(pl_person_gaz_givenname) & [SURFACE #given_name])
         (@seek(pl_person_gaz_givenname) & [SURFACE #given_name_2]) ?)
         |             ((@seek(pl_person_initial)             &             [SURFACE
#initial])(@seek(pl_person_initial) & [SURFACE #initial2])? ))
       (gazetteer & [SURFACE #posit, GTYPE gaz_position]) ?
)
->#pl_geo_osoby,                                                                where
#pl_geo_osoby=ConcWithBlanks(#posit,#given_name,#given_name_2,#initial,#initial2,#l
ast_name).

;;
;; extraction of ancient person names e.g. Zygmunt III Waza
;;

pl_roman_names_long :/
(
   (gazetteer & [SURFACE #posit, GTYPE gaz_position]) ?
    ((@seek(pl_person_gaz_givenname) & [SURFACE #given_name])
    (@seek(pl_person_gaz_givenname) & [SURFACE #given_name_2]) ?)
    (token & [SURFACE roman_number & #rzymskie])?
    (@seek(pl_person_gaz_surname) & [SURFACE #last_name] )
)
->#pl_geo_osoby,                                                                where
#pl_geo_osoby=ConcWithBlanks(#posit,#given_name,#given_name_2,#rzymskie,#last_name)
.
```

```
;;
;; extraction of ancient person names e.g. Bolesław II
;;

pl_roman_names :/
(
   (gazetteer & [SURFACE #posit, GTYPE gaz_position]) ?
    ((@seek(pl_person_gaz_givenname) & [SURFACE #given_name])
    (@seek(pl_person_gaz_givenname) & [SURFACE #given_name_2]) ?)
    (token & [SURFACE roman_number & #rzymskie])?
)
->#pl_geo_osoby,                                                                    where
#pl_geo_osoby=ConcWithBlanks(#posit,#given_name,#given_name_2,#rzymskie,#last_name)
.

;;
;;
;; CONSOLIDATING rule for supporting rules
;;
;;

pl_geo_all_names :/
(
 (@seek(pl_persons) & #nazwa_do_adresu)
 | (@seek(pl_persons_reverse) & #nazwa_do_adresu)
 | (@seek(pl_roman_names_long) & #nazwa_do_adresu)
 | (@seek(pl_roman_names) & #nazwa_do_adresu)
 | (@seek(pl_phrase_NN) & #nazwa_do_adresu)
 | (@seek(pl_phrase_A) & #nazwa_do_adresu)
 | (@seek(pl_phrase_NA) & #nazwa_do_adresu)
 | (@seek(pl_phrase_AN) & #nazwa_do_adresu)
 | (@seek(pl_date) & #nazwa_do_adresu)
 | (@seek(pl_date_long) & #nazwa_do_adresu)
 | (@seek(pl_phrase_PN) & #nazwa_do_adresu)
 | (@seek(pl_phrase_PAN) & #nazwa_do_adresu)
 | token & [TYPE first_capital_word, SURFACE #nazwa_do_adresu]
)
->         sign         &         [SURFACE         #nazwa_do_adresu1],         where
#nazwa_do_adresu1=CapitalizeWord(#nazwa_do_adresu).

;;
;; extraction of countries
;;

pl_geo_country :>
```

```
 (gazetteer & [SURFACE #panstwo, GTYPE gaz_country, CSTART #cs, CEND #ce, G_CONCEPT
#koncept])
-> ne-location & [LOCTYPE adm, LOCSUBTYPE cry, DESCRIPTOR #koncept, LOCNAME
#panstwo, NCSTART #cs, NCEND #ce].

;;
;; extraction of provinces
;;

pl_geo_province :>
 (gazetteer & [SURFACE #wojewodztwo, GTYPE gaz_wojewodztwo, CSTART #cs, CEND #ce,
G_CONCEPT #koncept])|
 ((token & [SURFACE "woj", CSTART #cs] token & [TYPE dot] ) (gazetteer & [SURFACE
#wojewodztwo, GTYPE gaz_wojewodztwo, CEND #ce, G_CONCEPT #koncept]))|
 ((morph & [STEM "województwo", CSTART #cs]) (gazetteer & [SURFACE #wojewodztwo,
GTYPE gaz_wojewodztwo, CEND #ce, G_CONCEPT #koncept]))
-> ne-location & [LOCTYPE adm, LOCSUBTYPE pro, LOCNAME #wojewodztwo, DESCRIPTOR
#koncept, NCSTART #cs, NCEND #ce].

;;
;; extraction of regions
;;

pl_geo_regions :>
 (gazetteer & [SURFACE #region, GTYPE gaz_region, CSTART #cs, CEND #ce, G_CONCEPT
#koncept])
-> ne-location & [LOCTYPE lan, LOCNAME #region, DESCRIPTOR #koncept, NCSTART #cs,
NCEND #ce].

;;
;; extraction of counties
;;

pl_geo_county :>
 ((token & [SURFACE "pow", CSTART #cs] token & [TYPE dot] ) (gazetteer & [SURFACE
#powiat, GTYPE gaz_powiat, CEND #ce, G_CONCEPT #koncept]))|
 ((morph & [STEM "powiat", CSTART #cs]) (gazetteer & [SURFACE #powiat, GTYPE
gaz_powiat, CEND #ce, G_CONCEPT #koncept]))
-> ne-location & [LOCTYPE adm, LOCSUBTYPE cnt, LOCNAME #powiat, DESCRIPTOR
#koncept, NCSTART #cs, NCEND #ce].

pl_geo_county2 :>
 ((token & [SURFACE "w"])?) (morph & [STEM "powiat", CSTART #cs1])(((morph & [STEM
"grodzki"])|(morph & [STEM "ziemski"]))?)
 (((token & [TYPE colon])|(token & [TYPE opening_bracket])|(token & [TYPE
hyphen]))?)
```

```
 (gazetteer & [SURFACE #powiat1, GTYPE gaz_powiat, CEND #ce1, G_CONCEPT
#koncept1])(token & [TYPE comma])?
 ((gazetteer & [SURFACE #powiat2, GTYPE gaz_powiat, CSTART #cs2, CEND #ce2,
G_CONCEPT #koncept2])(token & [TYPE comma]))?
 (gazetteer & [SURFACE #powiat3, GTYPE gaz_powiat, CSTART #cs3, CEND #ce3,
G_CONCEPT #koncept3])?
 (token & [SURFACE "i"]|(token & [TYPE comma]))?
 (gazetteer & [SURFACE #powiat4, GTYPE gaz_powiat, CSTART #cs4, CEND #ce4,
G_CONCEPT #koncept4])((token & [TYPE closing_bracket])?)
-> ne-location & [LOCTYPE adm, LOCSUBTYPE cnt, LOCNAME #powiat1, DESCRIPTOR
#koncept1, NCSTART #cs1, NCEND #ce1].

pl_geo_county21 :>
 ((token & [SURFACE "w"])?) (morph & [STEM "powiat", CSTART #cs1])(((morph & [STEM
"grodzki"])|(morph & [STEM "ziemski"]))?)
 (((token & [TYPE colon])|(token & [TYPE opening_bracket])|(token & [TYPE
hyphen]))?)
 (gazetteer & [SURFACE #powiat1, GTYPE gaz_powiat, CEND #ce1, G_CONCEPT
#koncept1])(token & [TYPE comma])?
 ((gazetteer & [SURFACE #powiat2, GTYPE gaz_powiat, CSTART #cs2, CEND #ce2,
G_CONCEPT #koncept2])(token & [TYPE comma]))?
 (gazetteer & [SURFACE #powiat3, GTYPE gaz_powiat, CSTART #cs3, CEND #ce3,
G_CONCEPT #koncept3])?
 (token & [SURFACE "i"]|(token & [TYPE comma]))?
 (gazetteer & [SURFACE #powiat4, GTYPE gaz_powiat, CSTART #cs4, CEND #ce4,
G_CONCEPT #koncept4])((token & [TYPE closing_bracket])?)
-> ne-location & [LOCTYPE adm, LOCSUBTYPE cnt, LOCNAME #powiat2, DESCRIPTOR
#koncept2, NCSTART #cs2, NCEND #ce2].

pl_geo_county22 :>
 ((token & [SURFACE "w"])?) (morph & [STEM "powiat", CSTART #cs1])(((morph & [STEM
"grodzki"])|(morph & [STEM "ziemski"]))?)
 (((token & [TYPE colon])|(token & [TYPE opening_bracket])|(token & [TYPE
hyphen]))?)
 (gazetteer & [SURFACE #powiat1, GTYPE gaz_powiat, CEND #ce1, G_CONCEPT
#koncept1])(token & [TYPE comma])?
 ((gazetteer & [SURFACE #powiat2, GTYPE gaz_powiat, CSTART #cs2, CEND #ce2,
G_CONCEPT #koncept2])(token & [TYPE comma]))?
 (gazetteer & [SURFACE #powiat3, GTYPE gaz_powiat, CSTART #cs3, CEND #ce3,
G_CONCEPT #koncept3])?
 (token & [SURFACE "i"]|(token & [TYPE comma]))?
 (gazetteer & [SURFACE #powiat4, GTYPE gaz_powiat, CSTART #cs4, CEND #ce4,
G_CONCEPT #koncept4])((token & [TYPE closing_bracket])?)
-> ne-location & [LOCTYPE adm, LOCSUBTYPE cnt, LOCNAME #powiat3, DESCRIPTOR
#koncept3, NCSTART #cs3, NCEND #ce3].
```

```
pl_geo_county23 :>
 ((token & [SURFACE "w"])?) (morph & [STEM "powiat", CSTART #cs1])(((morph & [STEM
"grodzki"])|(morph & [STEM "ziemski"]))?)
 (((token & [TYPE colon])|(token & [TYPE opening_bracket])|(token & [TYPE
hyphen]))?)
 (gazetteer & [SURFACE #powiat1, GTYPE gaz_powiat, CEND #ce1, G_CONCEPT
#koncept1])(token & [TYPE comma])?
 ((gazetteer & [SURFACE #powiat2, GTYPE gaz_powiat, CSTART #cs2, CEND #ce2,
G_CONCEPT #koncept2])(token & [TYPE comma]))?
 (gazetteer & [SURFACE #powiat3, GTYPE gaz_powiat, CSTART #cs3, CEND #ce3,
G_CONCEPT #koncept3])?
 (token & [SURFACE "i"]|(token & [TYPE comma]))?
 (gazetteer & [SURFACE #powiat4, GTYPE gaz_powiat, CSTART #cs4, CEND #ce4,
G_CONCEPT #koncept4])((token & [TYPE closing_bracket])?)
-> ne-location & [LOCTYPE adm, LOCSUBTYPE cnt, LOCNAME #powiat4, DESCRIPTOR
#koncept4, NCSTART #cs4, NCEND #ce4].

;;
;; extraction of communes
;;

pl_geo_commune :>
 ((token & [SURFACE "gm", CSTART #cs] token & [TYPE dot] ) (gazetteer & [SURFACE
#gmina, GTYPE gaz_gmina, CEND #ce, G_CONCEPT #koncept]))|
 ((morph & [STEM "gmina", CSTART #cs]) (gazetteer & [SURFACE #gmina, GTYPE
gaz_gmina, CEND #ce, G_CONCEPT #koncept]))
-> ne-location & [LOCTYPE adm, LOCSUBTYPE cmn, LOCNAME #gmina, DESCRIPTOR #koncept,
NCSTART #cs, NCEND #ce].

pl_geo_commune2 :>
 ((morph & [STEM "gmina", CSTART #cs1])|(token & [SURFACE "gm", CSTART #cs] token &
[TYPE dot] ))
 (((morph & [STEM "miejski"])|(morph & [STEM "wiejski"]))?)(((token & [TYPE
colon])|(token & [TYPE hyphen])?))
 (gazetteer & [SURFACE #gmina1, GTYPE gaz_gmina, CEND #ce1, G_CONCEPT
#koncept1])(token & [TYPE comma])?
 ((gazetteer & [SURFACE #gmina2, GTYPE gaz_gmina, CSTART #cs2, CEND #ce2, G_CONCEPT
#koncept2])(token & [TYPE comma]))?
 (gazetteer & [SURFACE #gmina3, GTYPE gaz_gmina, CSTART #cs3, CEND #ce3, G_CONCEPT
#koncept3])?
 (token & [SURFACE "i"]|(token & [TYPE comma]))?
 (gazetteer & [SURFACE #gmina4, GTYPE gaz_gmina, CSTART #cs4, CEND #ce4, G_CONCEPT
#koncept4])
-> ne-location & [LOCTYPE adm, LOCSUBTYPE cmn, LOCNAME #gmina1, DESCRIPTOR
#koncept1, NCSTART #cs1, NCEND #ce1].
```

```
pl_geo_commune21 :>
 ((morph & [STEM "gmina", CSTART #cs1])|(token & [SURFACE "gm", CSTART #cs] token &
[TYPE dot] ))
 (((morph & [STEM "miejski"])|(morph & [STEM "wiejski"]))?)(((token & [TYPE
colon])|(token & [TYPE hyphen])?))
 (gazetteer & [SURFACE #gmina1, GTYPE gaz_gmina, CEND #ce1, G_CONCEPT
#koncept1])(token & [TYPE comma])?
 ((gazetteer & [SURFACE #gmina2, GTYPE gaz_gmina, CSTART #cs2, CEND #ce2, G_CONCEPT
#koncept2])(token & [TYPE comma]))?
 (gazetteer & [SURFACE #gmina3, GTYPE gaz_gmina, CSTART #cs3, CEND #ce3, G_CONCEPT
#koncept3])?
 (token & [SURFACE "i"]|(token & [TYPE comma]))?
 (gazetteer & [SURFACE #gmina4, GTYPE gaz_gmina, CSTART #cs4, CEND #ce4, G_CONCEPT
#koncept4])
-> ne-location & [LOCTYPE adm, LOCSUBTYPE cmn, LOCNAME #gmina2, DESCRIPTOR
#koncept2, NCSTART #cs2, NCEND #ce2].


pl_geo_commune22 :>
 ((morph & [STEM "gmina", CSTART #cs1])|(token & [SURFACE "gm", CSTART #cs] token &
[TYPE dot] ))
 (((morph & [STEM "miejski"])|(morph & [STEM "wiejski"]))?)(((token & [TYPE
colon])|(token & [TYPE hyphen])?))
 (gazetteer & [SURFACE #gmina1, GTYPE gaz_gmina, CEND #ce1, G_CONCEPT
#koncept1])(token & [TYPE comma])?
 ((gazetteer & [SURFACE #gmina2, GTYPE gaz_gmina, CSTART #cs2, CEND #ce2, G_CONCEPT
#koncept2])(token & [TYPE comma]))?
 (gazetteer & [SURFACE #gmina3, GTYPE gaz_gmina, CSTART #cs3, CEND #ce3, G_CONCEPT
#koncept3])?
 (token & [SURFACE "i"]|(token & [TYPE comma]))?
 (gazetteer & [SURFACE #gmina4, GTYPE gaz_gmina, CSTART #cs4, CEND #ce4, G_CONCEPT
#koncept4])
-> ne-location & [LOCTYPE adm, LOCSUBTYPE cmn, LOCNAME #gmina3, DESCRIPTOR
#koncept3, NCSTART #cs3, NCEND #ce3].

pl_geo_commune23 :>
 ((morph & [STEM "gmina", CSTART #cs1])|(token & [SURFACE "gm", CSTART #cs] token &
[TYPE dot] ))
 (((morph & [STEM "miejski"])|(morph & [STEM "wiejski"]))?)(((token & [TYPE
colon])|(token & [TYPE hyphen])?))
 (gazetteer & [SURFACE #gmina1, GTYPE gaz_gmina, CEND #ce1, G_CONCEPT
#koncept1])(token & [TYPE comma])?
 ((gazetteer & [SURFACE #gmina2, GTYPE gaz_gmina, CSTART #cs2, CEND #ce2, G_CONCEPT
#koncept2])(token & [TYPE comma]))?
```

```
 (gazetteer & [SURFACE #gmina3, GTYPE gaz_gmina, CSTART #cs3, CEND #ce3, G_CONCEPT
#koncept3])?
 (token & [SURFACE "i"]|(token & [TYPE comma]))?
 (gazetteer & [SURFACE #gmina4, GTYPE gaz_gmina, CSTART #cs4, CEND #ce4, G_CONCEPT
#koncept4])
-> ne-location & [LOCTYPE adm, LOCSUBTYPE cmn, LOCNAME #gmina4, DESCRIPTOR
#koncept4, NCSTART #cs4, NCEND #ce4].

;;
;; extraction of cities
;;

pl_geo_city :>
 (gazetteer & [SURFACE #miasto, GTYPE gaz_city, CSTART #cs, CEND #ce, G_CONCEPT
#koncept])
-> ne-location & [LOCTYPE adm, LOCSUBTYPE cit, LOCNAME #miasto, DESCRIPTOR
#koncept, NCSTART #cs, NCEND #ce].

pl_geo_city2 :>
  ((morph & [STEM "miasto"])|(morph & [STEM "miasteczko"])|(morph & [STEM "wieś"]))
  (gazetteer & [SURFACE #miasto1, GTYPE gaz_city, CSTART #cs1, CEND #ce1, G_CONCEPT
#koncept1])
-> ne-location & [LOCTYPE adm, LOCSUBTYPE cit, LOCNAME #miasto, DESCRIPTOR
#koncept1, NCSTART #cs, NCEND #ce].

;;
;; extraction of streets e.g. ul. Paderewskiego, ulica Paderewskiego
;;

pl_geo_street :>
(
  morph & [STEM "ulica", CSTART #cs]  |
  ( ( token & [SURFACE "Ul", CSTART #cs] | token & [SURFACE "ul", CSTART #cs] |
token & [SURFACE "UL", CSTART #cs] ) (token & [TYPE dot] | token & [TYPE comma]))
 @seek(pl_geo_all_names) &  [SURFACE #nazwa_do_adresu, CEND #ce]
)
-> ne-location-postal & [LOCTYPE adm, LOCSUBTYPE str, STREET #nazwa_do_adresu,
LOCNAME #cala_nazwa, NCSTART #cs, NCEND #ce],
where #cala_nazwa=ConcWithBlanks("ul.", #nazwa_do_adresu).

;;
;; advanced rules for streets (e.g. przy zbiegu ulic Serbskiej, Słowiańskiej i
Lechickiej)
;;

pl_geo_streets_1 :> (token & [SURFACE "zbiegu"])
```

```
                  ((morph & [STEM "ulica", CSTART #cs])|
                   (((token & [SURFACE "Ul", CSTART #cs]) | (token & [SURFACE "ul",
CSTART #cs]) | (token & [SURFACE "UL", CSTART #cs])) token &
[TYPE dot]))
                       (token & [TYPE colon]) ?
                       (@seek(pl_geo_all_names)  &    [SURFACE  #nazwa_ulicy,  CEND
#ce])
                       ((token & [TYPE comma])(@seek(pl_geo_all_names)))?
                       (token & [SURFACE "i"])
                       (@seek(pl_geo_all_names))
-> ne-location-postal & [LOCTYPE adm, LOCSUBTYPE str, LOCNAME #nazwa_ulicy, NCSTART
#cs, NCEND #ce].

pl_geo_streets_12 :> (token & [SURFACE "zbiegu"])
                  ((morph & [STEM "ulica", CSTART #cs])|
                   (((token & [SURFACE "Ul", CSTART #cs]) | (token & [SURFACE "ul",
CSTART #cs]) | (token & [SURFACE "UL", CSTART #cs])) token &
[TYPE dot]))
                       (token & [TYPE colon]) ?
                       (@seek(pl_geo_all_names))
                       ((token & [TYPE comma])(@seek(pl_geo_all_names) &   [SURFACE
#nazwa_ulicy, CEND #ce]))
                       (token & [SURFACE "i"])
                       (@seek(pl_geo_all_names))
->  ne-location-postal  &  [LOCTYPE  adm,  LOCSUBTYPE  str,    LOCNAME  #nazwa_ulicy,
NCSTART #cs, NCEND #ce].

pl_geo_streets_13 :> (token & [SURFACE "zbiegu"])
                  ((morph & [STEM "ulica", CSTART #cs])|
                   (((token & [SURFACE "Ul", CSTART #cs]) | (token & [SURFACE "ul",
CSTART #cs]) | (token & [SURFACE "UL", CSTART #cs])) token &
[TYPE dot]))
                       (token & [TYPE colon]) ?
                     (@seek(pl_geo_all_names)) ?
                       ((token & [TYPE comma])(@seek(pl_geo_all_names)))?
                       (token & [SURFACE "i"])
                       (@seek(pl_geo_all_names)  &    [SURFACE  #nazwa_ulicy,  CEND
#ce])
-> ne-location-postal & [LOCTYPE adm, LOCSUBTYPE str, LOCNAME #nazwa_ulicy, NCSTART
#cs, NCEND #ce].

pl_geo_streets_2 :>
((token & [SURFACE "przy"])|(morph & [STEM "skrzyżowanie"]))
((morph & [STEM "ulica", CSTART #cs1]) | (((token & [SURFACE "Ul", CSTART #cs1]) |
(token & [SURFACE "ul", CSTART #cs1])
| (token & [SURFACE "UL", CSTART #cs1])) token & [TYPE dot]))
```

```
((token & [TYPE colon])?)
(@seek(pl_geo_all_names) & [SURFACE #nazwa_ulicy1, CEND #ce1])((token & [TYPE
comma])?)
((@seek(pl_geo_all_names) & [SURFACE #nazwa_ulicy2, CSTART #cs2, CEND #ce2])
(token & [TYPE comma]))?
(@seek(pl_geo_all_names) & [SURFACE #nazwa_ulicy3, CSTART #cs3, CEND #ce3])?
((token & [TYPE comma]) |(token & [SURFACE "i"])|(token & [SURFACE "oraz"]))?
(@seek(pl_geo_all_names) & [SURFACE #nazwa_ulicy4, CSTART #cs4, CEND #ce4])
-> ne-location-postal & [LOCTYPE adm, LOCSUBTYPE str, LOCNAME #nazwa_ulicy1,
NCSTART #cs1, NCEND #ce1].

pl_geo_streets_21 :>
((token & [SURFACE "przy"])|(morph & [STEM "skrzyżowanie"]))
((morph & [STEM "ulica", CSTART #cs1]) | (((token & [SURFACE "Ul", CSTART #cs1]) |
(token & [SURFACE "ul", CSTART #cs1])
| (token & [SURFACE "UL", CSTART #cs1])) token & [TYPE dot]))
((token & [TYPE colon])?)
(@seek(pl_geo_all_names) & [SURFACE #nazwa_ulicy1, CEND #ce1])((token & [TYPE
comma])?)
((@seek(pl_geo_all_names) & [SURFACE #nazwa_ulicy2, CSTART #cs2, CEND #ce2])
(token & [TYPE comma]))?
(@seek(pl_geo_all_names) & [SURFACE #nazwa_ulicy3, CSTART #cs3, CEND #ce3])?
((token & [TYPE comma]) |(token & [SURFACE "i"])|(token & [SURFACE "oraz"]))?
(@seek(pl_geo_all_names) & [SURFACE #nazwa_ulicy4, CSTART #cs4, CEND #ce4])
-> ne-location-postal & [LOCTYPE adm, LOCSUBTYPE str, LOCNAME #nazwa_ulicy2,
NCSTART #cs2, NCEND #ce2].

pl_geo_streets_22 :>
((token & [SURFACE "przy"])|(morph & [STEM "skrzyżowanie"]))
((morph & [STEM "ulica", CSTART #cs1]) | (((token & [SURFACE "Ul", CSTART #cs1]) |
(token & [SURFACE "ul", CSTART #cs1])
| (token & [SURFACE "UL", CSTART #cs1])) token & [TYPE dot]))
((token & [TYPE colon])?)
(@seek(pl_geo_all_names) & [SURFACE #nazwa_ulicy1, CEND #ce1])((token & [TYPE
comma])?)
((@seek(pl_geo_all_names) & [SURFACE #nazwa_ulicy2, CSTART #cs2, CEND #ce2])
(token & [TYPE comma]))?
(@seek(pl_geo_all_names) & [SURFACE #nazwa_ulicy3, CSTART #cs3, CEND #ce3])?
((token & [TYPE comma]) |(token & [SURFACE "i"])|(token & [SURFACE "oraz"]))?
(@seek(pl_geo_all_names) & [SURFACE #nazwa_ulicy4, CSTART #cs4, CEND #ce4])
-> ne-location-postal & [LOCTYPE adm, LOCSUBTYPE str, LOCNAME #nazwa_ulicy3,
NCSTART #cs3, NCEND #ce3].

pl_geo_streets_23 :>
((token & [SURFACE "przy"])|(morph & [STEM "skrzyżowanie"]))
```

```
((morph & [STEM "ulica", CSTART #cs1]) | (((token & [SURFACE "Ul", CSTART #cs1]) |
(token & [SURFACE "ul", CSTART #cs1])
| (token & [SURFACE "UL", CSTART #cs1])) token & [TYPE dot]))
((token & [TYPE colon])?)
(@seek(pl_geo_all_names) & [SURFACE #nazwa_ulicy1, CEND #ce1])((token & [TYPE
comma])?)
((@seek(pl_geo_all_names) & [SURFACE #nazwa_ulicy2, CSTART #cs2, CEND #ce2])
(token & [TYPE comma]))?
(@seek(pl_geo_all_names) & [SURFACE #nazwa_ulicy3, CSTART #cs3, CEND #ce3])?
((token & [TYPE comma]) |(token & [SURFACE "i"])|(token & [SURFACE "oraz"]))?
(@seek(pl_geo_all_names) & [SURFACE #nazwa_ulicy4, CSTART #cs4, CEND #ce4])
-> ne-location-postal & [LOCTYPE adm, LOCSUBTYPE str, LOCNAME #nazwa_ulicy4,
NCSTART #cs4, NCEND #ce4].

;;
;; extraction of avenues (Aleja Niepodległości, Al. Niepodległości)
;;

pl_geo_avenue :>
(
  ((morph & [STEM "aleja", CSTART #cs]) | (morph & [STEM "Aleja", CSTART #cs]))
  |
  (
    ((token & [SURFACE "Al", CSTART #cs]) | (token & [SURFACE "al", CSTART #cs]) |
(token & [SURFACE "AL", CSTART #cs]))
    token & [TYPE dot]
  )
)
(
 @seek(pl_geo_all_names) & [SURFACE #nazwa_do_adresu, CEND #ce]
)
-> ne-location-postal & [LOCTYPE adm, LOCSUBTYPE str, STREET #nazwa_do_adresu,
LOCNAME #cala_nazwa, NCSTART #cs, NCEND #ce],
where #cala_nazwa=ConcWithBlanks("Aleja", #nazwa_do_adresu).

;;
;; extraction of districts (osiedle Tysiąclecia)
;;

pl_geo_district :>
(
  ((morph & [STEM "osiedle", CSTART #cs]) |(morph & [STEM "Osiedle", CSTART #cs]))
  |
  (
    ((token & [SURFACE "Os", CSTART #cs]) | (token & [SURFACE "os", CSTART #cs]) |
(token & [SURFACE "OS", CSTART #cs]))
```

```
    token & [TYPE dot]
  )
)
(
 @seek(pl_geo_all_names) &  [SURFACE #nazwa_do_adresu, CEND #ce]
)
-> ne-location-postal & [LOCTYPE adm, LOCSUBTYPE dis, STREET #nazwa_do_adresu,
LOCNAME #cala_nazwa, NCSTART #cs, NCEND #ce],
where #cala_nazwa=ConcWithBlanks("Osiedle", #nazwa_do_adresu).

;;
;; extraction of names of squares (Pl. Wiosny Ludów, Pl. Tysiąclecia)
;;

pl_geo_square :>
(
  ((morph & [STEM "plac", CSTART #cs]) | (morph & [STEM "Plac", CSTART #cs]))
  |
  (
    ((token & [SURFACE "Pl", CSTART #cs]) | (token & [SURFACE "pl", CSTART #cs]) |
(token & [SURFACE "PL", CSTART #cs]))
    token & [TYPE dot]
  )
)
(
 @seek(pl_geo_all_names) &  [SURFACE #nazwa_do_adresu, CEND #ce]
)
-> ne-location-postal & [LOCTYPE adm, LOCSUBTYPE str, STREET #nazwa_do_adresu,
LOCNAME #cala_nazwa, NCSTART #cs, NCEND #ce],
where #cala_nazwa=ConcWithBlanks("Plac", #nazwa_do_adresu).

;;
;; extraction of landmarks
;;

pl_geo_landmark :> (
@seek(pl_entity_single_name) & [SURFACE #n0, CSTART #cs]
(@seek(pl_org_name_innerwords) &#i1 ? @seek(pl_entity_single_name) & [SURFACE #n1])
?
@seek(pl_org_name_innerwords) &#i2 ?
morph & [STEM "hotel", SURFACE #inner] |
morph & [STEM "pomnik", SURFACE #inner] |
morph & [STEM "stadion", SURFACE #inner] |
morph & [STEM "teatr", SURFACE #inner] |
morph & [STEM "kino", SURFACE #inner] |
morph & [STEM "park", SURFACE #inner] |
```

```
morph & [STEM "twierdza", SURFACE #inner] |
morph & [STEM "pałac", SURFACE #inner]
@seek(pl_org_name_innerwords) &#i3 ? @seek(pl_entity_single_name) & [SURFACE #n3,
CEND #ce]
) -> ne-location & [
        LOCTYPE adm,
        LOCSUBTYPE ldm,
        LOCNAME #n,
        NCSTART #cs,
        NCEND #ce],
        where #n = ConcWithBlanks
(#n0,#i1,#n1,#i2,#i3,#n3).

pl_geo_landmark_2 :> (
morph & [STEM "hotel", SURFACE #inner, CSTART #cs] |
morph & [STEM "pomnik", SURFACE #inner, CSTART #cs] |
morph & [STEM "stadion", SURFACE #inner, CSTART #cs] |
morph & [STEM "teatr", SURFACE #inner, CSTART #cs] |
morph & [STEM "kino", SURFACE #inner, CSTART #cs] |
morph & [STEM "park", SURFACE #inner, CSTART #cs] |
morph & [STEM "twierdza", SURFACE #inner, CSTART #cs] |
morph & [STEM "pałac", SURFACE #inner, CSTART #cs]
@seek(pl_org_name_innerwords) &#i1 ? @seek(pl_entity_single_name) & [SURFACE #n1,
CEND #ce]
) -> ne-location & [
        LOCTYPE adm,
        LOCSUBTYPE ldm,
        LOCNAME #n,
        NCSTART #cs,
        NCEND #ce],
        where #n = ConcWithBlanks(#inner,#i1,#n1).


pl_geo_landmark_3 :>
(gazetteer & [SURFACE #landmark, GTYPE gaz_landmark, CSTART #cs, CEND #ce,
G_CONCEPT #koncept, G_CITY #miasto ])
-> ne-location & [LOCTYPE adm, LOCSUBTYPE cit, LOCNAME #miasto, DESCRIPTOR
#koncept, NCSTART #cs, NCEND #ce].

;;
;; extraction of waterbodies
;;

pl_geo_wat :> (
@seek(pl_entity_single_name) & [SURFACE #n0, CSTART #cs]
```

```
(@seek(pl_org_name_innerwords) &#i1 ? @seek(pl_entity_single_name) & [SURFACE #n1])
?
@seek(pl_org_name_innerwords) &#i2 ?
morph & [STEM "ocean", SURFACE #inne] |
morph & [STEM "morze", SURFACE #inner] |
morph & [STEM "jezioro", SURFACE #inner]|
morph & [STEM "rzeka", SURFACE #inner] |
morph & [STEM "kanał", SURFACE #inner] |
morph & [STEM "strumień", SURFACE #inner]
(@seek(pl_org_name_innerwords) &#i3 ? @seek(pl_entity_single_name) & [SURFACE #n3])
?
@seek(pl_org_name_innerwords) &#i4 ? @seek(pl_entity_single_name) & [SURFACE #n4,
CEND #ce]
) -> ne-location & [
        LOCTYPE wat,
        LOCNAME #n,
        NCSTART #cs,
        NCEND #ce],
        where #n = ConcWithBlanks
(#n0,#i1,#n1,#i2,#i3,#n3,#i4, #n4).

pl_geo_wat_first :> (
morph & [STEM "ocean", SURFACE #inner, CSTART #cs] |
morph & [STEM "morze", SURFACE #inner, CSTART #cs] |
morph & [STEM "jezioro", SURFACE #inner, CSTART #cs]|
morph & [STEM "rzeka", SURFACE #inner, CSTART #cs] |
morph & [STEM "kanał", SURFACE #inner, CSTART #cs] |
morph & [STEM "strumień", SURFACE #inner, CSTART #cs]
(@seek(pl_org_name_innerwords) &#i1 ? @seek(pl_entity_single_name) & [SURFACE #n1])
?
@seek(pl_org_name_innerwords) &#i2 ? @seek(pl_entity_single_name) & [SURFACE #n2,
CEND #ce]
) -> ne-location & [
        LOCTYPE wat,
        LOCNAME #n,
        NCSTART #cs,
        NCEND #ce],
        where #n = ConcWithBlanks(#inner,#i1,#n1,#i2,#n2).

pl_geo_see :>
  (gazetteer & [SURFACE #morze, GTYPE gaz_sea, CSTART #cs, CEND #ce, G_CONCEPT
#koncept])
-> ne-location & [LOCTYPE wat, LOCNAME #morze, DESCRIPTOR #koncept, NCSTART #cs,
NCEND #ce].

pl_geo_lake :>
```

```
 (gazetteer & [SURFACE #jezioro, GTYPE gaz_lake, CSTART #cs, CEND #ce, G_CONCEPT
#koncept])
-> ne-location & [LOCTYPE wat, LOCNAME #jezioro, DESCRIPTOR #koncept, NCSTART #cs,
NCEND #ce].

pl_geo_river :>
 (gazetteer & [SURFACE #rzeka, GTYPE gaz_river, CSTART #cs, CEND #ce, G_CONCEPT
#koncept])
-> ne-location & [LOCTYPE wat, LOCNAME #rzeka, DESCRIPTOR #koncept, NCSTART #cs,
NCEND #ce].

pl_geo_lan :> (
morph & [STEM "wyspa", SURFACE #inner, CSTART #cs] |
morph & [STEM "kontynent", SURFACE #inner, CSTART #cs] |
morph & [STEM "góra", SURFACE #inner, CSTART #cs]
@seek(pl_org_name_innerwords) &#i1 ? @seek(pl_entity_single_name) & [SURFACE #n1,
CEND #ce]
) -> ne-location & [
        LOCTYPE lan,
        LOCNAME #n,
        NCSTART #cs,
        NCEND #ce],
        where #n = ConcWithBlanks(#inner,#i1,#n1).
```

# Appendix 2. Geographical ontology

```
wsmlVariant _"http://www.wsmo.org/wsml/wsml-syntax/wsml-flight"
namespace { _"http://www.kie.ae.poznan.pl/geographical_ontology#",
     wsmostudio _"http://www.wsmostudio.org#" }

ontology geographical_ontology
     nonFunctionalProperties
          wsmostudio#version hasValue "0.7.3"
     endNonFunctionalProperties

concept Artefacts

concept LocationGeometry subConceptOf Artefacts
hasCentroid ofType Coordinates

concept SpatialRelationship subConceptOf Artefacts
relates ofType Location

concept AdjacentToRelation subConceptOf SpatialRelationship

concept OverlappingRelation subConceptOf SpatialRelationship

concept Coordinates subConceptOf Artefacts
hasLatitude ofType _decimal
hasLongitude ofType _decimal

concept Location
hasName ofType (1 1) _string
hasVariantNames ofType _string
hasDescription ofType (0 1) _string
hasAreaInSquareMetres ofType (0 1) _decimal
hasGeometry ofType (0 1) LocationGeometry
hasModificationDate ofType (0 1) _date
hasLanguageVariant ofType (0 1) _string
hasDateOfOrigin ofType (0 1) _string

concept Landmark subConceptOf Location

concept AdministrativeRegion subConceptOf Location
hasInhabitants ofType _integer
```

```
concept Country subConceptOf AdministrativeRegion
belongsTo ofType (1 1) LandForm

concept Province subConceptOf AdministrativeRegion
belongsTo ofType (1 1) Country

concept County subConceptOf AdministrativeRegion
belongsTo ofType (1 1) Province

concept Commune subConceptOf AdministrativeRegion
belongsTo ofType (1 1) County

concept Zone subConceptOf AdministrativeRegion
belongsTo ofType (1 1) Location

concept District subConceptOf AdministrativeRegion
belongsTo ofType (1 1) City

concept LandForm subConceptOf Location

concept WaterBody subConceptOf Location

concept City subConceptOf AdministrativeRegion
hasLandmarks ofType (0 1) Landmark
hasAddress ofType (0 1) _string
hasDistrict ofType (0 1) District
belongsTo ofType (1 1) Commune
```

# References


ABRAMOWICZ, W. (Ed.) (2003). *Knowledge-Based Information Retrieval and Filtering from the Web*, Kluwer Academic Publishers.

ABRAMOWICZ, W. (2008). *Filtrowanie Informacji,* Poznań, Wydawnictwo Akademii Ekonomicznej w Poznaniu.

ABRAMOWICZ, W., FILIPOWSKA, A., KACZMAREK, T. & KOWALKIEWICZ, M. (2007a). IT Tools supporting Public Relations campaigns. IN KHOSROW-POUR, M. (Ed.) *Managing Worldwide Operations and Communications with Information Technology.* Vancouver, Canada, Hershey: IGI Publishing.

ABRAMOWICZ, W., FILIPOWSKA, A., PISKORSKI, J., WĘCEL, K. & WIELOCH, K. (2006). Linguistic Suite for Polish Cadastral System. *5th International Conference on Language Resources and Evaluation.* Genoa, Italy.

ABRAMOWICZ, W., FLEJTER, D. & KACZMAREK, T. (2007b). Architectures for Deep Web Data Extraction and Integration. IN BORZEMSKI, L., GRZECH, A., ŚWIĄTEK, J. & WILIMOWSKA, Z. (Eds.) *Information Systems Architecture and Technology.*

ABRAMOWICZ, W., KALCZYŃSKI, P. J. & WĘCEL, K. (2002). *Filtering the Web to Feed Data Warehouses,* London, Springer-Verlag.

ABRAMOWICZ, W. & WIŚNIEWSKI, M. (2008). Proximity Window Context in N-Gram Models for Term Extraction in Ontology Learning from Text. *DEXA '08: 19th International Conference on Database and Expert Systems Applications.* Washington, DC, USA, IEEE Computer Society.

ACE, H. (2009). Automatic Content Extraction (ACE) Evaluation. IN TECHNOLOGY, N. I. O. S. (Ed.).

ADAMUS-MATUSZYŃSKA, A. (1999). Public Relations - Komunikowaniem w społecznej przestrzeni. IN ADAMUS-MATUSZYŃSKA, A. (Ed.) *Wybrane problemy Public Relations.* Katowice, Kolegium Zarządzania Akademii Ekonomicznej w Katowicach.

AGENCY, A. R. P. (1996). Proceedings of the TIPSTER Text Program (phase II). California, USA, Morgan Kaufmann.

AKERLOF, G. A. (1970). The Market for 'Lemons': Quality Uncertainty and the Market Mechanism. *Quarterly Journal of Economics,* 84**,** 488-500.

ALTKORN, J. (2004). *Wizerunek firmy,* Dąbrowa Górnicza, Wyższa Szkoła Biznesu w Dąbrowie Górniczej.

AMITAY, E., HAR'EL, N., SIVAN, R. & SOFFER, A. (2004). Web-a-Where: Geotagging Web Content. *SIGIR.* Sheffield, South Yorkshire, UK.

ANDRADE, L. & SILVA, M. J. (2006). Relevance Ranking for Geographic IR. IN ACM (Ed.) *SIGIR 1006.* USA.

APPELT, D. E. & ISRAEL, D. J. (1999). Introduction to Information Extraction Technology. A tutorial prepared for IJCAI-99. *IJCAI-99.* Stockholm, Sweden.

ARIKAWA, M., SAGARA, T. & OKAMURA, K. (2000). Spatial Media Fusion Project. *Digital Libraries: Research and Practice.* Kyoto, Japan.

ARROW, K. J. (1963). Uncertainty and the Welfare Economics of Medical Care. *The American Economic Review,* LIII**,** 941-973.

ASADI, S., XU, J., SHI, Y., DIEDERICH, J. & ZHOU, X. (2006). Calculation of Target Locations for Web Resources. *7th International Conference on Web Information Systems Engineering (WISE 2006).* Wuhan, China.

AUDI, R. (2001). *Cambridge Dictionary of Philosophy*, Press Syndicate of the University of Cambridge.

AUSTIN, E. W. & PINKLETON, B. E. (2006). *Strategic Public Relations Management: Planning and Managing Effective Communication Programs*, Routledge.

BAEZA-YATES, R. & RIBEIRO-NETO, B. D. A. (1999). *Modern information retrieval,* New York, Harlow, England, ACM Press, Addison-Wesley.

BAGGA, A. (1998). Evaluation of Coreferences and Coreference Resolution Systems. *First Language Resource and Evaluation Conference.*

BALCEROWICZ, L. (1997). *Socjalizm kapitalizm transformacja. Szkice z przełomu epok.,* Warszawa, Wydawnictwo Naukowe PWN.

BASSARA, A. (2009). Temporalne indeksowanie dokumentów dla analizy rynków finansowych, PhD Thesis. *Department of Information Systems.* Poznan, Poznan University of Economics.

BEARD, K. & SHARMA, V. (1997). Multidimensional ranking in digital spatial libraries. *Journal of Digital Libraries. Special issue on Meta-data.*, 153-160.

BELKIN, N. J. & CROFT, W. B. (1992). Information filtering and information retrieval: Two sides of the same coin? *Communications of the ACM,* 35**,** 29-39.

BERGER, A. L., PIETRA, S. A. D. & PIETRA, V. J. D. (1996). A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics* 22**,** 39-71.

BERGHEL, H. (1997). Cyberspace 2000: Dealing with Information Overload. *Communications of the ACM,* 40**,** 19-24.

BERNERS-LEE, T., HENDLER, J. & LASSILA, O. (2001). The Semantic Web. *Scientific American,* 2001**,** 34-43.

BIKEL, D. M., SCHWARTZ, R. & WEISCHEDEL, R. M. (1999). An Algorithm that Learns What's in a Name. *Machine Learning. Special issue on natural language learning.,* 34**,** 211-231.

BILHAUT, F., CHARNOIS, T., ENJALBERT, P. & MATHET, Y. (2003). Geographic reference analysis for geographic document querying. IN KORNAI, A. & SUNDHEIM, B. (Eds.) *Workshop on the Analysis of Geographic References at the NAACL-HLT 2003 Conference.* Edmonton, Canada.

BOORSTIN, D. (1963). *The Image of What Happened to the American Dream*, Pengguin Books.

BRIN, S. & PAGE, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems* 30**,** 107-117.

BROCKMANS, S., VOLZ, R., EBERHART, A. & LOEFFLER, P. (2004). Visual modelling of OWL-DL ontologies using UML. *3rd International Semantic Web Conference.* Hiroshima, Japan, Springer Verlag.

BRUNSTEIN, A. (2002). Annotation Guidelines for Answer Types, http://www.ldc.upenn.edu/Catalog/docs/LDC2005T33/BBN-Types-Subtypes.html.

BUCHER, B., CLOUGH, P., JOHO, H., PURVES, R. & SYED, A. (2005). Geographic IR Systems: requirements and evaluation. *22nd International Cartographic Conference.*

BUDZYNSKI, W. (2002). *Wizerunek firmy. Kreowanie, zarządzanie, efekty.,* Warszawa, Wydawnictwo Poltext.

BUSCALDI, D. & ROSSO, P. (2008). A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Science,* 22**,** 301-313.

BUYUKKOKTEN, O., CHO, J., GARCIA-MOLINA, H., GRAVANO, L. & SHIVAKUMAR, N. (1999). Exploiting Geographical Location Information of Web Pages. *ACM SIGMOD Workshop on the Web and Databases (WebDB'99).*

CAI, G. (2007). Contextualisation of Geospatial Database Semantics for Human-GIS Interaction. *GeoInformatica,* 11**,** 217-237.

CENKER, E. M. (2000). *Public relations,* Poznań, Wydawnictwo Wyższej Szkoły Bankowej.

CHAKRABARTI, S. (1999). Mining the Link Structure of the World Wide Web. *IEEE Computer*.

CHAVES, M. S., RODRIGUES, C. & SILVA, M. J. (2007). Data Model for Geographic Ontologies Generation. IN RAMALHO, J. C., LOPES, J. C. & CARRICO, L. (Eds.) *XATA2007 : XML : aplicacoes e tecnologias associadas.* Lisboa.

CHINCHOR, N. A. (1998). Overview of MUC-7. *Message Understanding Conference.*

CLOUGH, P. (2005). Extracting metadata for spatially aware information retrieval on the internet. *Workshop on Geographic Information Retrieval.* Bremen, Germany, ACM Press.

COLLEY, R. H. (1961). *Defining Advertising Goals for Measured Advertising Results,* New York, Association of National Advertisers, Inc.

COLLINS, M., HAJIC, J., RAMSHAW, L. & TILLMANN, C. (1999). A statistical parser for Czech. IN MARYLAND, U. O. (Ed.) *ACL*

CRISFORD, J. (1974). *Public relation advances,* London.

CUTLIP, S. M., CENTER, A. H. & BROOM, G. M. (2005). *Effective Public Relations*, Prentice Hall.

DAVIS, C., VIXIE, P., GOODWIN, T. & DICKINSON, I. (1996). Means for Expressing Location Information in the Domain Name System. *RFC 1876, The Internet Society.*

DAY, D., MCHENRY, C., KOZIEROK, R. & RIEK, L. (2004). Callisto: A Configurable Annotation Workbench. *LREC 2004: Fourth International Conference on Language Resources and Evaluation.* Lissbon, Portugal.

DELBONI, T. M., BORGES, K. A. V. & LAENDER, A. H. F. (2005). Geographic web search based on positioning expressions. *Workshop On Geographic Information Retrieval.* ACM Press.

DENBER, M. (1998). Automatic resolution of anaphora in English. Technical report. IN EASTMAN KODAK CO, I. S. D. (Ed.).

DENSHAM, I. & REID, J. (2003). A Geo-coding service encompassing a geo-parsing tool and integrated digital gazetteer service. *Workshop on the Analysis of Geographic References at the NAACL-HLT 2003 conference.* Edmonton, Canada.

DEY, A. K. & ABOWD, G. D. (2000). Towards a better understanding of context and context-awareness. *Conference on Human Factors in Computing Systems.* Hague, The Netherlands.

DIMITROV, M. (2002). A Light-weight Approach to Coreference Resolution for Named Entities in Text. *Faculty of Mathematics and Informatics, Department of Information Technologies.* Sofia, University of Sofia "St. Kliment Ohridski".

DING, J., GRAVANO, L. & SHIVAKUMAR, N. (2000). Computing Geographical Scopes of Web Resources. *VLDB.* Cairo, Egypt.

DOKTOROWICZ, K. (1999). Współpraca z mediami jako element działań public relations. IN ADAMUS-MATUSZYŃSKA, A. (Ed.) *Wybrane Problemy Public Relations.* Katowice, Kolegium Zarządzania Akademii Ekonomicznej w Katowicach.

DROZDZYNSKI, W., KRIEGER, H.-U., PISKORSKI, J., SCHÄFER, U. & XU, F. (2004). Shallow Processing with Unification and Typed Feature Structures - Foundations and Applications. *Kuenstliche Intelligenz,* 1**,** 17-23.

EDMUNDS, A. & MORRIS, A. (2000). The problem of information overload in business organisations: a review of the literature. *International Journal of Information Management,* 20**,** 17-28.

EGENHOFER, M. J. (2002). Toward the semantic geospatial web. *10th Symposium on Advances in Geographic Information Systems (GIS-02).*

EUROGEOGRAPHICS (2004). Seamless Administrative Boundaries of Europe (SABE).

FARRELL, C., SCHULZE, M., PLEITNER, S. & BALDONI, D. (1994). DNS Encoding of Geographical Location. *RFC 1712, The Internet Society.*

FILIPOWSKA, A. & WĘCEL, K. (2004). Wykorzystanie wizualizacji i algorytmów data mining do analizy indeksów geograficznych dokumentów. *SAS Forum 2004.*

FLEJTER, D. & KACZMAREK, T. (2007). Wybrane aspekty integracji informacji z głębokiego Internetu (Choosen apsects of Deep Web information integration). *Kapitał ludzki i wiedza w gospodarce.*

FLORIDI, L. (2005). Semantic Conceptions of Information. IN ZALTA, E. N. (Ed.) *The Stanford Encyclopedia of Philosophy.*

FRANTZI, K., ANANIADOU, S. & MIMA, H. (2000). Automatic recognition of multi-word terms:. the C-value/NC-value method. *International Journal on Digital Libraries***,** 115-130.

FRIBURGER, N. & MAUREL, D. (2002). Textual Similarity Based on Proper Names. *MFIR 2002 at the 25th ACM SIGIR Conference.* Tampere, Finland.

FU, G., ABDELMOTY, A. I. & JONES, C. B. (2003). Design of a Geographical Ontology, D5 3101. Public Deliverable SPIRIT Project. .

FU, G., JONES, C. B. & ABDELMOTY, A. I. (2005). Building a Geographical Ontology for Intelligent Spatial Search on the Web. IN VERLAG, S. (Ed.) *IASTED International Conference on Databases and Applications*

GALE, W. A., CHURCH, K. W. & YAROWSKY, D. (1992). One sense per discourse. *Human Language Technology Conference archive, Proceedings of the workshop on Speech and Natural Language.* Association for Computational Linguistics

GANTZ, J. F., CHUTE, C., MANFREDIZ, A., MINTON, S., REINSEL, D., SCHLICHTING, W. & TONCHEVA, A. (2008). The Diverse and Exploding Digital Universe. An Updated Forecast of Worldwide Information Growth Through 2011. An IDC White Paper - sponsored by EMC. IN IDC (Ed.), IDC.

GARDENT, C. & WEBER, B. (2001). Towards the use of automated reasoning in doscourse disambiguation. *Logic, Languages and Information,* 10 (4)**,** 487-509.

GETTY (2004). Getty. Thesaurus of Geographic Name.

GEY, F. (2000). Research to Improve Cross-Language Retrieval - Position Paper. IN PETERS, C. (Ed.) *Cross-Language Information Retrieval and Evaluation, Workshop of Cross-Language Evaluation Forum, CLEF.* Lisbon, Portugal, Springer.

GEY, F., LARSON, R., SANDERSON, M., BISCHOFF, K., MANDL, T., WOMSER-HACKER, C., SANTOS, D. & ROCHA, P. (2006). GeoCLEF 2006: the CLEF 2006 cross-language geographic information retrieval track overview. *CLEF 2006.*

GRAVANO, L., HATZIVASSILOGLOU, V. & LICHTENSTEIN, R. (2003). Categorizing Web Queries According to Geographical Locality. *12th ACM Conference on Information and Knowledge Management (CIKM 2003).*

GREEN, J. J. (2005). Google PageRank And Related Technologies. An extensive discussion of Google's PageRank technology.

GRISE, M. & GALLUPE, R. B. (2000). Information overload: Addressing the productivity paradox in face-to-face electronic meetings. *Journal of Management Information Systems,* 16**,** 157-185.

GRISHMAN, R. & SUNDHEIM, B. (1996). Message Understanding Conference - 6: A Brief History. *COLING.*

GROUP, U. N. G. I. W. Report of the 3rd UNGIWG Plenary Meeting.

GRUBER, T. R. (1993). A translation approach to portable ontologies. *Knowledge Acquisition,* 5(2)**,** 199-220.

GRUNIG, J. E. (2001). Two-way symmetrical public relations. Past, present and future. IN HEATH, R. L. (Ed.) *Handbook of public relations.* Sage.

GRUNIG, J. E., DOZIER, D., EHLING, W., GRUNIG, L., REPPER, F. & WHITE, J. (1992). *Excellence In Public Relations and Communications Management,* Mahwah, NJ, IABC Research Foundation.

GRUNIG, J. E. & HUANG, Y. H. (2000). From organizational effectiveness to relationship indicators: antecedents of relationships, public relations strategies, and relationship outcomes. IN LEDINGHAM, J. A. & BRUNING, S. D. (Eds.) *Public Relations as Relationship Management: A Relational Approach to the Study and Practice of Public Relations.* Mahwah, NJ, Lawrence Erlbaum Associates.

GRUNIG, J. E. & HUNT, T. (1984). *Managing Public Relations,* Orlando, FL, Harcourt Brace Jovanovich.

GUARINO, N. & GIARETTA, P. (1995). Ontologies and Knowledge Bases. Towards a Terminological Clarification. IN MARS, N. (Ed.) *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing.* Amsterdam, IOS Press.

HARPING, P. (1997). The Limits of the World: Theoretical and Practical Issues in the construction of the Getty Thesaurus of Geographical Names. *ICHIM 97: The Fourth International Conference on Hypermedia and Interactivity in Museums.* Paris, France.

HARPRING, P. (1997). Proper words in proper places: The Thesaurus of Geographic Names. *MDA Information,* 2 (3)**,** 5-12.

HAUPTMANN, A. G. & OLLIGSCHLAEGER, A. M. (1999). Using location information from speech recognition of television news broadcasts. *Access-Audio-1999***,** 102-106.

HE, B., ZHANG, Z. & CHANG, K. C. (2004). Knocking the Door to the Deep Web: Integrating Web Query Interfaces. *Proceedings of the 2004 ACM SIGMOD Conference (SIGMOD 2004).*

HEATH, R. L. (2004). *Encyclopedia of Public Relations*, Sage Publications, Inc

HEPP, M. (2007). Ontologies: State of the Art, Business Potential, and Grand Challenges. IN MARTIN HEPP, P. D. L., ALDO DE MOOR, YORK SURE (Ed.) *Ontology Management: Semantic Web, Semantic Web Services, and Business Applications.* Springer.

HEVNER, A. R., MARCH, S. T., PARK, J. & RAM, S. (2004). Design Science in Information Systems Research. *MIS Quarterly,* 28**,** 75-105.

HILL, L. (1990). Access to Geographic Concepts in Online Bibliographic Files: Effectiveness of Current Practices and the Potential of a Graphic Inteface. University of Pittsburgh.

HILL, L. L. (2000). Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. *4th European Conference on Research and Advanced Technology for Digital Libraries.* Lecture Notes in Computer Science.

HIMMELSTEIN, M. (2005). Local Search: The Internet Is the Yellow Pages. *Computer,* 38**,** 26-34.

HOLTZ, S. (2002). *Public Relations on the Net: Winning Strategies to Inform, & Influence the Media, the Investment Community, the Government, the Public, & More*, American Management Association.

IGNAT, C., POULIQUEN, B., RIBEIRO, A. & STEINBERGER, R. (2003). Extending an Information Extraction Tool Set to Central and Eastern European Languages. *International Workshop "Information Extraction for Slavonic and other Central and Eastern European Languages" (IESL'2003), held at RANLP'2003.* Borovets, Bulgaria.

JANUSZKO, W. (2001). *Systemy Informacji gospodarczej*, Wydawnictwo SBP.

JAROSZ-PALACH, A. (2005). Zarządzanie marketingowe przedsiębiorstwem w erze informacyjnej. IN NOWICKI, A. (Ed.) *System informacyjny marketingu przedsiębiorstw.* Warszawa, Polskie Wydawnictwo Ekonomiczne S.A.

JAYNES, E. T. (1957). Information Theory and Statistical Mechanics. *Physical Review,* 106**,** 620-630.

JIN, R. & DUMAIS, S. T. (2001). Probabilistic combination of content and links. *24th Conference on Research and Development in Information Retrieval.*

JONES, C. B., ALANI, H. & TUDHOPE, D. (2001). Geographical Information Retrieval with ontologies of place. *International Conference on Spatial Information Theory: Foundations of Geographic Information Science* Springer Verlag, London.

JONES, R., ZHANG, W. V., REY, B., JHALA, P. & STIPP, E. (2008). Geographic intention and modification in web search. *International Journal of Geographical Information Science,* 22**,** 229-246.

KABRA, G., LI, C. & CHANG, K. (2005). Query Routing: Finding Ways in the Maze of the Deep Web. *Proceedings of the ICDE International Workshop on Challenges in Web Information Retrieval and Integration (ICDE-WIRI 2005).*

KACZMAREK-ŚLIWIŃSKA, M. (2005a). Efektywność Internet PR - próba ujęcia metodologicznego w świetle badań. *Piar.pl,* 2005**,** 96-102.

KACZMAREK-ŚLIWIŃSKA, M. (2005b). Internet jako narzędzie public relations przedsiębiorstw okresu transformacji w Polsce. *Wydział Ekonomii.* Poznań, Akademia Ekonomiczna w Poznaniu.

KACZMAREK-ŚLIWIŃSKA, M. (2006). Monitoring Internetu na potrzeby działań Public Relations. Koszalin.

KACZMAREK, T. (2006). Integracja danych z głębokiego Internetu dla potrzeb analizy otoczenia przedsiębiorstwa, PhD Thesis. *Wydział Ekonomii.* Poznań, Akademia Ekonomiczna w Poznaniu.

KAISER, K. & MIKSCH, S. (2005). Information Extraction. A Survey. IN VIENNA UNIVERSITY OF TECHNOLOGY, I. O. S. T. A. I. S. (Ed.) Vienna, Austria.

KELLY, K. (1997). New rules for the internet period. Explanation of Twelve Principles of the Network Economy.

KIMLER, M. (2004). Geo-Coding: Recognition of geographical references in unstructured text and their visualisation. *Department of Computer Science and Technology.* Ispra, University of Applied Sciences Hof.

KISIELNICKI, J. & SROKA, H. (2001). *Systemy informacyjne biznesu,* Warszawa, Agencja Wydawnicza "Placet".

KLEINBERG, J. M. (1999). Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM (JACM),* 46**,** 604 - 632.

KNOWLEDGESTORM, I. (2006). Emerging Media Series: The Influence of Podcasts on B2B Technology Purchase Decisions.

KOTLER, P. (1999). *Marketing. Analiza, planowanie, wdrażanie, kontrola.,* Warszawa.

KOWALKIEWICZ, M., KACZMAREK, T. & ABRAMOWICZ, W. (2006a). myPortal: Robust Extraction and Aggregation of Web Content. *Proceedings of the 32nd International Conference on Very Large Data Bases.* Seoul, Korea, ACM Press.

KOWALKIEWICZ, M., ORLOWSKA, M., KACZMAREK, T. & ABRAMOWICZ, W. (2006b). Towards more personalized Web: Extraction and integration of dynamic content from the Web. *Proceedings of the 8th Asia Pacific Web Conference APWeb 2006.* Harbin, China, Springer Verlag.

KRIEGER, H.-U., DROZDZYNSKI, W., PISKORSKI, J., SCHÄFER, U. & XU, F. (2004). A Bag of Useful Techniques for Unification-Based Finite-State Transducers. *7th KONVENS.* Vienna, Austria.

KRUPKA, G. R. & HAUSMAN, K. (1998). Isoquest, Inc: Description of the NetOwl(TM) extractor system as used for MUC-7. *Seventh Message Understanding Conference (MUC-7).* Fairfax, Virginia, USA.

LALMAS, M. & RUTHVEN, I. (1998). Representing and retrieving structured documents with Dempster-Shafer's theory of evidence: From theory to practice. *Journal of Documentation*.

LAPRUN, C., FISCUS, J., GAROFOLO, J. & PAJOT, S. (2002). A Practical Introduction to Atlas. *LREC 2002: Third International Conference on Language Resources and Evaluation.* La Palma, Canary Islands, Spain.

LARMAN, C. (2004). *Applying UML and Patterns*, Prentice Hall PTR.

LARSON, R. R. (1996). Geographic Information Retrieval and Spatial Browsing. *GIS and Libraries: Patrons, Maps and Spatial Information.* University of Illinois.

LAVELLI, A., CALIFF, M., CIRAVEGNA, F., FREITAG, D., GIULIANO, C., KUSHMERIK, N. & ROMANO, L. (2004). IE evaluation: Criticisms and recommendations. *AAAI Workshop on Adaptive Text Extraction and Mining.*

LEE, F., BRESSAN, S. & QOOI, B. C. (2000). Global atlas: calibrating and indexing documents from the internet in the cartographical paradigm. *1st International Conference on Web Information Systems Engineering (WISE 2000).* Hong Kong, China, IEEE Computer Society.

LEHNERT, W., CARDIE, C., FISCHER, D., MCCARTHY, J., RILOFF, E. & SODERLAND, S. (1992). Description of the CIRCUS system as used for MUC–4. *3rd Message Understanding Conference (MUC–4).*

LEIDNER, J. L. (2004). Towards a reference corpus for automatic toponym resolution evaluation. *Workshop on Geographic Information Retrieval, SIGIR.*

LEIDNER, J. L. (2007a). Toponym Resolution in Text Annotation, Evaluation and Applications of Spatial Grounding of Place Names. *Institute for Communicating and Collaborative Systems, School of Informatics.* Edinburgh, University of Edinburgh.

LEIDNER, J. L. (2007b). Toponym Resolution: A First Large-Scale Comparative Evaluation.

LEIDNER, J. L., SINCLAIR, G. & WEBBER, B. (2003). Grounding spatial named entities for information extraction and question answering. *Workshop on the Analysis of Geographic References held at the Joint Conference for Human Language Technology and the Annual Meeting of the Noth American Chapter of the Association for Computational Linguistics 2003 (HLT/NAACL'03).* Edmonton, Canada.

LEVITT, S. & DUBNER, S. J. (2005). *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything*, William Morrow.

LI, H., SRIHARI, R. K., NIU, C. & LI, W. (2002). Location Normalization for Information Extraction. *19th Conference on Computational Linguistics.* Taipei, Taiwan.

LI, H., SRIHARI, R. K., NIU, C. & LI, W. (2003a). infoXtract location normalization: a hybrid approach to geographical referenes in information extraction. *Workshop on the Analysis of Geographic References.* Edmonton, Canada.

LI, Y., BANDAR, Z. A. & MCLEAN, D. (2003b). An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering,* 15 (4).

LI, Z., WANG, C., XIE, X., WANG, X. & MA, W.-Y. (2006). Indexing implicit locations for geographical information retrieval. *The 3rd International Workshop on Geographic Information Retrieval, GIR 2006.*

LINDENMANN, W. K. (2003). Guidelines and Standards for Measuring the Effectiveness of PR Programs and Activities.

LINE, M. B. (1969). Information requirements in the social sciences: some considerations. *Journal of Librairanship,* 1**,** 1-19.

LOEB, S. & TERRY, D. (1992). Information filtering. *Communications of the ACM,* 35**,** 26-28.

ŁAWNICZAK, R. (2003a). Public Relations and Communication Management - Poland. IN RULER, B. V. & VERCIC, D. (Eds.) *Public Relations and Communication Management in Europe. A Nation-by-Nation Introduction to Public Relations Theory and Practice.* Berlin, Mouton de Gruyter.

ŁAWNICZAK, R. (2003b). The Transitional Approach to Public Relations. IN RULER, B. V. & VERCIC, D. (Eds.) *Public Relations and Communication Management in Europe. A Nation-by Nation Introduction to Public Relations Theory and Practice.* Berlin, Mouton de Gruyter.

ŁAWNICZAK, R. & KACZMAREK-ŚLIWIŃSKA, M. (2004). Internet PR in a Transition Economy – The Case of Poland. *BledCom 2004 The 11th International Public Relations Research Symposium.* Bled, Slovenia, BledCom.

MA, Q. & TANAKA, K. (2004). Retrieving regional information from web by contents localness and user location. *Asia Information Retrieval Symposium (AIRS 2004).* Beijing, China, Lecture Notes in Computer Science.

MACIASZEK, L. (2007). *Requirements Analysis and System Design,* Harlow, England, Addison Wesley.

MACNAMARA, J. R. (1999). Research in Public Relations. A review of the use of evaluation and formative research. *Asia Pacific Public Relations Journal,* 1**,** 107-134.

MAEDCHE, A. & STAAB, S. (2001). Ontology Learning for the Semantic Web. *IEEE Intelligent Systems,* 16 (2)**,** 72-79.

MARK, D. M., SMITH, B. & TVERSKY, B. (1999). Ontology and Geographic Objects: An Empirical Study of Cognitive Categorization. *Lecture Notes in Computer Science,* 1661.

MARKOWETZ, A., BRINKHO, T. & SEEGER, B. (2004). Geographic Information Retrieval. *3rd International Workshop on Web Dynamics.*

MARKOWETZ, A., CHEN, Y., SUEL, T., LONG, X. & SEEGER, B. (2005). Design and implementation of a geographic search engine. *Technical Report TR-CIS-2005-03.* Brooklyn, New York, USA, Polytechnic University.

MARSH, E. & PERZANOWSKI, D. (1998). MUC-7 Evaluation of IE Technology: Overview of Results. *Seventh Message Understanding Conference (MUC-7).*

MARTINS, B., CHAVES, M. & SILVA, M. J. (2005a). Assigning Geographical Scopes To Web Pages? *Advances in Information Retrieval.* Springer Berlin / Heidelberg.

MARTINS, B., SILVA, M. J. & ANDRADE, L. (2005b). Indexing and Ranking in GeoIR Systems. *Geographic Information Retrieval.* Bremen, Germany, ACM.

MARTINS, B., SILVA, M. J. & CHAVES, M. S. (2005c). Challenges and Resources for Evaluating Geographical IR. *GIR.* ACM.

MARTINS, B., SILVA, M. J., FREITAS, S. & AFONSO, A. P. (2006). Handling Locations in Search Engine Queries. IN ACM (Ed.) *GIR.* Seattle, USA, ACM.

MAYNARD, D., BONTCHEVA, K. & CUNNINGHAM, H. (2003). Towards a semantic extraction of named entities. IN ARCHIVE, A. E. (Ed.) UK.

MCCURLEY, K. S. (2001). Geospatial Mapping and Navigation of the Web. *WWW10.* Hong Kong.

MCKEONE, D. H. (1995). *Measuring Your Media Profile,* Aldershot, Gower Publishing Limited.

MICHIGAN, S. O. (2007). Official State of Michigan Portal. Michigan.

MIKHEEV, A., MOENS, M. & GOVER, C. (1999). Named Entity Recognition without Gazetteers. *EACL.* Bergen, Norway.

MITKOV, R. (1998). Robust Anaphora Resolution with Limited Knowledge. *COLING.*

MIZZARO, S. (1998). How many relevances in information retrieval? *Interacting With Computers,* 10**,** 305-322.

MOSTELLER, F. (1948). A k-sample slippag test for an extreme population. *Annals of Mathematical Statistics,* 19**,** 58-65.

MUSLEA, I. (1999). Extraction Patterns for Information Extraction Tasks: A Survey. *AAAI-99 Workshop on Machine Learning for Information Extraction.*

NADO, R. A. & HUFFMAN, S. B. (1997). Extracting entity profiles from semistructured information spaces. *SIGMOD Record,* 26**,** 32-38.

NITSCH, H. (1975). *Dynamische Public Relations: Unternehmerische Öffentlichkeitsarbeit, Strategie f.d. Zukunft*, Taylorix-Fachverlag.

OLEŃSKI, J. (2001). *Ekonomika Informacji. Podstawy,* Warszawa, Polskie Wydawnictwo Ekonomiczne.

OLIVER, S. (2005). *Strategia Public Relations,* Warszawa, Polskie Wydawnictwo Ekonomiczne.

PARSONS, P. J. (2004). *Ethics In Public Relations: A Guide To Best Practice,* Norfolk.

PASSONNEAU, R., HABASH, N. & RAMBOW, O. (2006). Inter-annotator Agreement on a Multilingual Semantic Annotation Task. *Fifth International Conference on Language Resources and Evaluation (LREC).*

PERIAKARUPPAN, R. & NEMETH, E. (1999). GTrace - A Graphical Traceroute Tool. *13th Systems Administration Conference - LISA '99.* Seattle, Washington, USA.

PESHKIN, L. & PFEFFER, A. (2003). Bayesian information extraction network. *18th International Joint Conference on Artificial Intelligence (IJCAI).*

PIASECKI, M. (2007). Cele i zadania lingwistyki informatycznej. IN STALMASZCZYK, P. (Ed.) *Metodologie jezykoznawstwa. Wspolczesne tendencje i kontrowersje.*

PILGRIM, M. (2002). What Is RSS?, http://www.xml.com/pub/a/2002/12/18/dive-into-xml.html.

PISKORSKI, J. (2002). Shallow text processor applied to information retrieval of business documents, PhD Thesis. Warszawa, Instytut Podstaw Informatyki Polskiej Akademii Nauk.

PISKORSKI, J., WIELOCH, M., PIKUŁA, M. & SYDOW, M. (2008). Towards Person Name Matching for Inflective Languages. *NLPIX.* ACM.

PITON, O. & MAUREL, D. (2001). Les Noms Propres Geographiques et le Dictionnaire Prolintex, les lieux situes hors de France. *4th INTEX workshop.* Bordeaux, France.

POULIQUEN, B., KIMLER, M., STEINBERGER, R., IGNAT, C., OELLINGER, T., BLACKLER, K., FLUART, F., ZAGHOUANI, W., WIDIGER, A., FORSLUND, A.-C. & BEST, C. (2006). Geocoding Multilingual Texts: Recognition, Disambiguation and Visualisation. *5th International Conference on Language Resources and Evaluation (LREC'2006).* Genoa, Italy.

POULIQUEN, B., STEINBERGER, R., IGNAT, C. & GROEVE, T. D. (2004). Geographical Information Recognition and Visualisation in Texts Written in Various Languages. *ACM Symposium on Applied Computing.* ACM.

PRZEPIORKOWSKI, A. (2007). Slavonic Information Extraction and Partial Parsing. *Balto-Slavonic Natural Language Processing 2007.* Prague, Czech Republic, Association for Computational Linguistics.

PURVES, R. S., CLOUGH, P., JONES, C. B., ARAMPATZIS, A., BUCHER, B., FINCH, D., FU, G., JOHO, H., SYED, A. K., VAID, S. & YANG, B. (2007). The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. *International Journal of Geographical Information Science,* 21**,** 717-745.

RATNAPARKHI, A. (1997). A Simple Introduction to Maximum Entropy Models for Natural Language Processing IN SCIENCE, N. S. A. T. C. F. R. I. C. (Ed.).

RAUCH, E., BUKATIN, M. & BAKER, K. (2003). A confidence-based framework for disambiguating geographic terms. *Workshop on the Analysis of Geographic References.* Edmonton, Canada.

REICH, B. & SOLOMON, D. (2008). *Media Rules: Mastering Today's Technology,* Hoboken, NJ, John Wiley & Sons.

RICHARDSON, W. (2005). The ABCs of RSS, http://www.techlearning.com/story/showArticle.php?articleID=163100414.

RIJSBERGEN, C. J. V. (1979). *Information Retrieval,* London, Butterworths.

RIJSBERGEN, C. J. V. V., CRESTANI, F. & LALMAS, M. (1998). *Information Retrieval: Uncertainty and Logics: Advanced Models for the Representation and Retrieval of Information*, Springer.

SANDERSON, M. & KOHLER, J. (2004). Analyzing Geographic Queries. IN ACM (Ed.) *ACM SIGIR Workshop on Geographic Information Retrieval.* Sheffield, UK.

SANG, E. F. T. K. & MEULDER, F. D. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *CoNLL-2003.* Edmonton, Canada.

SCHARL, A. (2007). Towards the Geospatial Web: Media Platforms for Managing Geotagged Knowledge Repositories. IN SCHARL, A. & TOCHTERMANN, K. (Eds.) *The Geospatial Web – How Geo-Browsers, Social Software and the Web 2.0 are shaping the Network Society.* London, Springer.

SCHILIT, W. N. (1995). System architecture for context-aware mobile computing. Columbia University.

SCHMIDT, A., GELLERSEN, H.-W. & BEIGL, M. (1999). Matching Information and Ambient Media. *CoBuild***,** 140-149.

SCHOCKAERT, S. & COCK, M. D. (2007). Neighborhood Restrictions in Geographic IR. *SIGIR 2007.* ACM.

SEITEL, F. P. (2003). *Public relations w praktyce (The Practice of Public Relations),* Warszawa, Felberg SJA.

SEITEL, F. P. (2006). *The Practice of Public Relations*, Pearson Publishing.

SEKINE, S. (2003). Definition of Sekine's Extended Named Entity. Version 6.1.0 (English).

SEKINE, S., SUDO, K. & NOBATA, C. (2002). Extended Named Entity Hierarchy. *LREC.*

SHANNON, C. E. (1948). A mathematical theory of communication. *Bell System Technology Journal,* 27**,** 379-423, 623-656.

SHANNON, C. E. (1993). *Collected Papers,* Piscataway, NJ, USA, IEEE Press.

SHAPIRO, C. & VARIAN, H. R. (1999). *Information Rules. A Strategic Guide to the Network Economy,* Boston, Massachusetts, Harvard Business School Press.

SILVER, G. A. & SILVER, M. L. (1989). *Systems Analysis and Design,* Reading, MA, Addison Wesley.

SILVER, M. S., MARKUS, M. L. & BEATH, C. M. (1995). The Information Technology Interaction Model: a Foundation for the MBA Core Course. *MIS Quarterly,* 19**,** 361-390.

SINNOTT, R. W. (1984). Virtues of the Haversine. *Sky and Telescope,* 68 (2).

SMITH, B. (1995). On drawing lines on a map. IN FRANK, A. & KUHN, W. (Eds.) *Spatial Information Theory. A Theoretical Basis for GIS.* Berlin, Springer.

SMITH, B. & MARK, D. M. (2001). Geographic categories: an Ontological Inverstigation. *International Journal of Geographical Information Science,* 15**,** 591-612.

SMITH, B. & WELTY, C. (2001). Ontology: Towards a New Synthesis. *International conference on Formal Ontology in Information Systems.* Ogunquit, Maine, USA

SMITH, D. A. & CRANE, G. (2001). Disambiguating geographic names in a historical digital library. *5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL '01).* Darmstadt, Springer.

SMITH, R. D. (2004). *Strategic Planning for Public Relations* Routledge.

SMITH, T., ANDRESEN, D., CARVER, L., DOLIN, R., FISCHER, C., FREW, J., GOODCHILD, M., IBARRA, O., KEMP, R., KOTHURI, R., LARSGAARD, M., MANJUNATH, B., NEBERT, D., SIMPSON, J., WELLS, A., YANG, T. & ZHENG, Q. (1996). A digital library for geographically referenced materials. *IEEE Computer,* 29**,** 54–60.

SOUZA, L. A., JR, C. A. D., BORGES, K. A. V., DELBONI, T. M. & LAENDER, A. H. F. (2005). The Role of Gazetteers in Geographic Knowledge Discovery on the Web. *Third Latin American Web Congress table of contents.*

SOWA, J. F. (1984). *Conceptual Structures. Information Processing in Mind and Machine.,* Reading, MA, USA, Addison Wesley.

SRIHARI, R. K., LI, W., NIU, C. & CORNELL, T. (2003). InfoXtract: A Customizable Intermediate Level Information Extraction Engine. *HLT-NAACL 2003 Workshop on Software Engineering and Architecture of Language Technology Systems (SEALTS).*

STIGLITZ, J. E. (2000). The Contributions of the Economics of Information to Twentieth Century Economics. *The Quarterly Journal of Economics***,** 1441-1478.

STONEBRAKER, M. & HELLERSTEIN, J. M. (2001). Content Integration for EBusiness. *ACM SIGMOD Record,* 30**,** 552-560.

STRZALKOWSKI, T. (1999). *Natural Language Information Retrieval*, Kluwer, Academic Publishing.

SZNAJDER, A. (1993). *Sztuka Promocji,* Warszawa, Business Press Ltd.

SZYFTER, J. P. (2005). *Public Relations w Internecie*, Helion.

TOBLER, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography,* 46**,** 234-240.

TOMAI, E. & KAVOURAS, M. (2004). From "Onto-GeoNoesis" to "Onto-Genesis". The Design of Geographic Ontologies. *GeoInformatica,* 8:3**,** 285-302.

TREADWELL, D. & TREADWELL, J. B. (2005). *Public Relations Writing: Principles in Practice*, Sage Publications Inc.

TURMO, J., AGENO, A. & CATALA, N. (2006). Adaptive Information Extraction. *ACM Computing Surveys (CSUR)* 38.

VAID, S. & JONES, C. B. (2004). Report on Spatial Indexing Methods. SPIRIT D 12 2201.

VAID, S., JONES, C. B., JOHO, H. & SANDERSON, M. (2005). Spatio-textual indexing for geographic search on the web. *SSTD-05, the 9th Symposium on Spatial and Temporal Databases.*

VESTAVIK, O. (2004). Geographic Information Retrieval: An Overview. IN TECHNOLOGY, N. U. O. S. A. (Ed.) Trondheim, Norway

WANG, C., XIE, X., WANG, L., LU, Y. & MA, W.-Y. (2005a). Detecting Geographic Locations from Web Resources. *GIR.* Bremen, Germany.

WANG, C., XIE, X., WANG, L., LU, Y. & MA, W. Y. (2005b). Web resource geographic location classification and detection. *14th International World Wide Web Conference (WWW 2005).* Chiba, Japan.

WEBER, L. (2007). *Marketing to the Social Web: How Digital Customers Buil Your Business,* Hoboken, NJ, John Wiley & Sons.

WELTY, C., LEHMANN, F., GRUNINGER, G. & USCHOLD, M. (1999). Ontology: Expert Systems All Over Again? *National Conference on Artificial Intelligence.* Austin, Texas.

WĘCEL, K. (2002). Profilowanie Hurtowni Danych dla Potrzeb Filtrowania Informacji Ekonomicznej. *Wydział Ekonomii.* Poznań, Akademia Ekonomiczna.

WIELOCH, K. (2010). Budowanie profili przedsiębiorstw z wykorzystaniem utożsamiania odwołań sementycznych do przedsiębiorstw w polskich tekstach ekonomicznych, PhD Thesis. Poznań, Poznań University of Economics.

WIŚNIEWSKI, M. (2009). Uczenie ontologii z tekstu. *Department of Information Systems.* Poznan, Poznan University of Economics.

WOJCIK, K. (2001). *Public Relations od A do Z,* Warszawa, Agencja Wydawnicza "Placet".

WOJCIK, K. (2005). *Public Relations - wiarygodny dialog z otoczeniem,* Warszawa, Placet.

WOLINSKI, M. (2006). Morfeusz - a practical tool for the morphological analysis of Polish. IN KLOPOTEK, M., WIERZCHON, S. T. & TROJANOWSKI, K. (Eds.) *Intelligent Information and Web Mining.* Ustron, Poland, Springer-Verlag.

WOODRUFF, A. & PLAUNT, C. (1994). GIPSY: Georeferenced Information Processing System. *Journal of the American Society for Information Science,* 45**,** 645-655.

WRIGHT, D. K. & HINSON, M. D. (2008). How Blogs and Social Media are Changing Public Relations and the Way it is Practiced. *Public Relations Journal,* 2.

YANGARBER, R. & GRISHMAN, R. (1997). Customization of Information Extraction Systems. *International Workshop on Lexically Driven Information Extraction.* Frascati, Italy.

YOKOJI, S., TAKAHASHI, K. & MIURA, N. (2001). Kokono Search: A Location Based Search Engine. *Tenth International World Wide Web Conference.* Hong Kong.

YOUNG GEUN, H., SANG HO, L., JAE HWI, K. & YANGGON, K. (2008). A new aggregation policy for RSS services. *Proceedings of the 2008 international workshop on Context enabled source and service selection, integration and adaptation: organized with the 17th International World Wide Web Conference (WWW 2008).* Beijing, China, ACM.

ZHANG, Q., XIE, X., WANG, L., YUE, L. & MADEFI, W.-Y. (2008). Detecting Geographical Serving Area of Web Resources. IN CITESEERX (Ed.) *Scientific Literature Digital Library and Search Engine (United States).*

ZHOU, Y., XIE, X., WANG, C., GONG, Y. & MA, W.-Y. (2005). Hybrid Index Structures for Location-based Web Search. *CIKM.* Bremen, Germany.