Ontology Structure, Reasoning Approach and Querying Mechanism in a Semantic-Enabled Efficient and Scalable Retrieval of Experts

Witold Abramowicz, Elżbieta Bukowska, Monika Kaczmarek, Monika Starzecka
Department of Information Systems, Faculty of Informatics and Electronic Commerce, Poznan University of Economics, Poznań, Poland
{w.abramowicz; e.bukowska; m.kaczmarek; m.starzecka}@kie.ue.poznan.pl

Abstract—**Efficient utilization of knowledge became a key to the success of an organization. The need to identify experts within or outside an organization has been for a long time inspiration for various initiatives undertaken by academia and industry. The eXtraSpec system developed in Poland is an example of such initiatives. In order to realize its tasks, the eXtraSpec system needs not only to be able to acquire and extract information from various sources, but also requires an appropriate representation of information, supporting reasoning over person's characteristics. The considered mechanism should allow for a precise identification of required data, but simultaneously, be efficient and scalable. The main goal of this paper, is to present the ontology structure, reasoning approach as well as querying mechanism applied in the eXtraSpec project, and discuss the underlying motivation, which led to the development of a semantic-based mechanism to retrieve experts in its current state.**

Keywords - Expert finding system; knowledge representation; expert characteristic, reasoning, querying

I. Introduction

An efficient acquisition and utilisation of knowledge is considered to be a key element contributing to a success of an organization operating in the competitive settings of the knowledge-based economy [49]. The organizations need not only to know the skills and expertise of their employees, but also need to be able to conduct an appropriate recruitment process. More and more often organizations, in order to locate expertise they require, take advantage of various Internet portals, including social portals, as well as other artefacts available on the Internet [50]. As the data and information on various experts available on WWW is very dispersed and of distributed nature, a need appears to support the processes of human resources management using the IT-based solutions e.g., information extraction and retrieval systems, especially expert retrieval systems.

Within an information retrieval (IR) process, a single query is executed on a set of documents in order to identify the relevant ones [2]. In general, a typical retrieval system encompasses three main components:

- a module responsible for collecting data (documents) and creating their easily processable representation in the form of an index;
- an interface allowing formulating queries reflecting the current information needs of a user and usually consisting of a set of keywords and finally,
- a mechanism matching a query to created indexes in order to identify the relevant documents.

All three elements affect the quality of the retrieval process, i.e., values of precision and recall metrics.

The traditional expert retrieval systems, being a subset of information retrieval systems focusing on identification of required experts, face the same problems as traditional IR systems. The mentioned problems are caused by usage of different keywords and different levels of abstraction by users when formulating queries on the same topic, or by using different words and phrases in the description of a phenomenon based on which indexes are created. In order to address these issues, very often semantics is applied, so as in response to a user query, a retrieval system returns documents, which do not contain words included in the query, but are still relevant to the user's information needs.

There are many research and commercial initiatives aiming at the development of retrieval systems in general and expert retrieval systems in particular, supported by semantics. They are to provide interested parties with detailed information on people's experience and skills. One of such initiatives is the Polish project eXtraSpec [23]. Its main goal is to combine company's internal electronic documents and information sources available on the Internet in order to provide an effective method of searching experts with competencies in the given field.

The main process in the eXtraSpec system flows as follows: the system acquires data from dedicated sources (on the Web or from the inside of the company) and saves it as an extracted profile (PE), whose structure is based on the European Curriculum Vitae Standard [38]. In the next step, data in PE is normalized. As a result of the normalization process, the normalized profile is generated (PN). Finally, PN are analysed and aggregated to the form of aggregated profile (PA) (i.e., one person is described by one and only one PA) serving as a source of information on experts. Based on the information provided by the aggregated profiles, the eXtraSpec system is to support three main scenarios:

- finding experts with desired characteristic,
- defining teams of experts, and
- verifying data on a person in question.

In order to support the above mentioned flow as well as three scenarios identified, the eXtraSpec system needs not only to be able to acquire and extract information from various sources, but also requires an appropriate representation of information that would support reasoning over person's characteristics. In addition, the reasoning and querying mechanism should on the one hand, precisely identify required data, and, on the other, be efficient and scalable.

The main goal of this paper, being an extended version of [1], is to present more in depth the ontology structure, reasoning approach as well as querying mechanism applied in the eXtraSpec project, and discuss the underlying motivation, which led to the development of a semantic-based mechanism to retrieve experts in its current state.

In order to fulfill the mentioned goals, the paper is structured as follows. First, the related work in the area of expert's retrieval and using semantics to describe experts is discussed. Then, the description of the identified querying strategies constituting requirements for the defined solution follows. Next, the ontology developed for the needs of the eXtraSpec project to support retrieval of experts is presented. Then, the short description of the considered scenarios regarding the application of the reasoning infrastructure, as well as the description of the selected one, follows. Finally, the system architecture as well as implementation details of the reasoning mechanisms are given. The paper concludes with final remarks.

II. Related work

The need to find expertise within an organization has been for a long time inspiration for initiatives aiming at the development of a class of search engines, being a subset of information retrieval systems, called expert finders or expert retrieval systems [3].

There are several aspects connected with the expert finding task, for instance, following McDonald and Ackerman [4], those may be:

- expertise identification aiming at answering a question - who is an expert on a given topic?, and
- expertise selection aiming at answering a question - what does X know?

Within our research, we focus on the first aspect i.e., identifying a relevant person given a concrete need.

First systems focusing on the expertise identification task relied on a database like structure containing a description of experts' skills (e.g., [5]). However, such systems faced many problems, e.g.:

- how to ensure precise results given a generic description of expertise and simultaneously fine-grained and specific queries [6], or
- how to guarantee the accuracy and validity of stored information given the static nature of a database and volatile nature of person's characteristics.

To address these and similar problems other systems were proposed focusing on automated discovery of up-to-date information from specific sources such as e.g., e-mail communication [7]. In addition, instead of focusing only on specific document types, systems that index and mine published intranet documents [8] or analyse social networks [45], were proposed. An example may be the Spree project [9] aiming at providing automatic expert finding facility, able to answer a given question. The system automatically builds qualification profiles from documents and uses communities and the social software in order to provide efficient searching capabilities.

In addition, currently the Web itself offers many other possibilities to find information on experts, as there are a number of contact management portals or social portals, where users can search for experts, potential employees or publish their curricula in order to be found by future employers (e.g., [25][26][27]).

When it comes to the algorithms applied to assess whether a given person is suitable to carry our a given task, at first, standard information retrieval techniques to locate an expert on a given topic were applied [10][11]. Usually, expertise of a person was represented in a form of a term vector and a query result was represented as a list of relevant persons.

If matching a query to a document relies on a simple mechanism checking whether a document contains the given keywords, then the well-known IR problems occur:

- low precision of returned results (there is a word, but not in this context);
- low value of recall (relevant documents described using a different set of keywords, are not identified);
- a large number of documents returned by the system (especially in a response to a general query) the processing of which is impossible (e.g., due to the time constraints).

Therefore, a few years ago, the Enterprise Track at the Text Retrieval Conference (TREC) was started in order to study the expert-finding topic. It resulted in further advancements of the expert finding techniques and application of numerous methods, such as probabilistic techniques or language analysis techniques to improve the quality of finding systems (e.g., [12][13][14][15]).

As the Semantic Web technology [42] is getting more and more popular [43], it is not surprising that it has been used to enrich descriptions within expert finding systems. The introduction of semantics into search systems may take two forms:

- the use of semantics in order to analyze indexed documents or queries (query expansion [44]),
- operating on semantically described resources with use of reasoners (e.g., operating on contents of RDF (Resource Description Framework [46]) files and ontologies represented in e.g., the OWL (Web Ontology language [47])).

Within the expert finding systems both approaches have been applied, as well as a number of various ontologies used to represent competencies and skills were developed.

For instance, the goal of a Single European Employment Market-Place (SEEMP) [16] was to provide interoperable architecture for e-Employment services. The mentioned project used an ontology in order to provide a semantic description of job offers and people's CV. The main ontology developed within this project is called Reference Ontology and it consists of thirteen sub-ontologies: Competence, Compensation, Driving License, Economic Activity, Education, Geography, Job Offer, Job Seeker, Labour Regulatory, Language, Occupation, Skill and Time. The Reference Ontology has been built based on the commonly used standards, e.g., ISO 4217 [28], ISCO-88 COM [29], ONET [30] or DAML ontology [31].

In turn, in [17] authors describe requirements and a process of ontology creation for the needs of human resources management. They developed an ontology that is used in two projects: a meta-search engine for searching jobs on job portals [18] and by a university competence management system [19]. The ontology was created in the OWL formalism. It consists of sub-ontologies for competencies, occupations and learning objects.

Another example is the ExpertFinder system [20] being a framework for reuse of already existing vocabularies in order to apply them in semantically supported systems. It provides terms and best practices for describing web pages, persons, institutions, events, areas of expertise, relations between persons, educational aspects etc. ExpertFinder uses such vocabularies as: FOAF (Friend of a Friend) [32], SIOC [33], vCard [34] or Dublin Core [35].

In addition, numerous ontologies, taxonomies and classifications have been created in the human resources management area, e.g., taxonomies for job descriptions such as e.g., the Standard Occupational Classification (SOC) [36] of the Unites States Federal statistical agencies or taxonomy of skills developed within the KOWIEN project [21].

The problem tackled within this paper is related to the semantic-based expert finding. The eXtraSpec system acquires information from outside and assumes that one can build a profile of a person based on the gathered information. It is important for the users of an expert finding system that the system operates on a large set of experts. More experts imply bigger topic coverage and increased probability of a question being answered. However, it simultaneously causes problems connected to the heterogeneity of information as well as low values of both precision and recall of the system. The application of semantics may help to normalize the gathered data and ensure an appropriate level of precision and recall, however, it generates problems with scalability and efficiency of the designed mechanisms that need to be addressed.

When it comes to the ontology, the eXtraSpec system differs from other projects under a few aspects:

- it is not limited only to hierarchical relations;
- it has been developed for the Polish language and relates to Polish standards;
- it has been built in accordance to the Simple Knowledge Organization System (SKOS) [37] standard.

Applying semantics undoubtedly offers a way to handle the precision, recall, and helps to normalize data, however, the application of semantics impacts the performance as well as scalability of the system. Therefore, a design decision needed to be taken regarding the way the semantics should be applied in order to ensure the required quality of the system. In the next section, we present the considered querying strategies, developed ontology, reasoning scenarios and the underlying motivation.

III. Querying strategies

In order to identify the requirements towards the persons' characteristics, scope of information needed to be covered by ontologies, as well as the querying and reasoning mechanism developed within the eXtraSpec system, first, exemplary searching strategies a user looking for experts may be interested in were considered. The strategies have been specified based on the conducted studies of the literature and interviews with employers conducting recruitment processes. The six most common searching goals are as follows:

1. To find an expert with some experience at a position/role of interest.
2. To find an expert having some specific language skills on a desired level.
3. To find an expert having some desired competencies.
4. To find students who graduated recently/will graduate soon in a given domain of interest.
5. To find a person having expertise in a specific domain.
6. To find a person with specific education background, competencies, fulfilled roles, etc. Although the enumerated goals (1-5) sometimes are used separately, usually though, they constitute building blocks of more complex scenarios within which they are freely combined using various logical operators.

As already mentioned, the above querying goals imposed some requirements on the information on experts that should be available, and in consequence, also ontologies that needed to be developed for the project's needs, as well as the reasoning and querying mechanism. Tables 1-3 summarize the requirements on the scope of information required to describe an expert, on querying and reasoning mechanism, as well as on the ontology itself.

TABLE 1 Querying strategies and resulting requirements on the scope of information

Scenario No.	Requirements on the scope of information
1. To find an expert with some experience on a position of interest.	An expert description MUST include information on positions and jobs undertaken so far as well as their duration.
2 To find an expert having some specific language skills on a desired level.	An expert description MUST provide information on: known languages, obtained certificates and a level of language skills.
3 To find an expert having some competencies.	An expert description MUST include information on soft and tangible competencies of a person.

Scenario No.	Requirements on the scope of information
4 To find students who graduated recently or will graduate soon in a given domain.	An expert description MUST include information on educational background of a person, especially: educational organization, date of graduation and educational result.
5 To find a person having expertise in a specific domain.	An expert description MUST include information on organizations a person worked for. Please note that the information on the domains the organizations operate in should be provided by ontology (see Table 3)
6 To find a person with specific education, competencies, jobs, etc.	Features of interests for this scenario include all previously mentioned.

TABLE 2 QUERYING STRATEGIES AND RESULTING REQUIREMENTS ON REASONING AND QUERYING MECHANISM

Scenario No.	Requirements on reasoning and querying mechanism
1. To find an expert with some experience on a position of interest.	The querying and reasoning mechanism MUST be able to integrate experience history (e.g., add the length of duration from different places, but gained on the same or similar position) and then reason on a position's hierarchy (i.e., taking into account narrower or broader concepts).
2 To find an expert having some specific language skills on a desired level.	If the information is not explicitly given, the querying and reasoning mechanism SHOULD be able to associate different certificates with languages and proficiency levels.
3 To find an expert having some competencies.	The querying and reasoning mechanism SHOULD be able to operate not only on implicitly given competencies, but also reason on jobs and then on connected competencies. Thus, the querying and reasoning mechanism SHOULD tackle also other relations than is-a.
4 To find students who graduated recently/will graduate in a given domain.	The querying and reasoning mechanism MUST be able to reason on the hierarchy of educational organizations, on dates and results.
5 To find a person having expertise in a specific domain.	The querying and reasoning mechanism SHOULD be able to associate organizations with domains they operate in.
6 To find a person with specific education, competencies, jobs, etc.	The querying and reasoning mechanism MUST be able to combine results from various querying strategies using different logical operators.

TABLE 3 QUERYING STRATEGIES AND RESULTING REQUIREMENTS ON ONTOLOGY

Scenario No.	Requirements on ontology
1. To find an expert with some experience on a position of interest.	The ontology MUST represent a is-a hierarchy of different positions and jobs allowing for their categorization and reasoning on their hierarchical relations.
2 To find an expert having some specific language skills on a desired level.	The ontology MUSTt represent languages certificates (is-a hierarchy) together with information on the language and the proficiency level, mapped to one scale.
3 To find an expert having some competencies	The ontology MUST represent skills and competencies and their hierarchical dependencies as well as some additional relations as appropriate.
4 To find students who graduated recently/will graduate in a given domain	The ontology MUST provide a hierarchy of educational organizations allowing for their categorization and reasoning on their hierarchical dependencies.
5 To find a person having expertise in a specific domain	The ontology SHOULD provide information on organizations allowing for their categorization (is-a relation) as well as provide information on the domains they operate in.
6 To find a person with specific education, competencies, jobs, etc.	Requirements on ontologies are the same as in scenarios 1-5.

The next section presents the developed ontology meeting the above enumerated requirements.

IV. ONTOLOGIES IN THE EXTRASPEC PROJECT

A. *Requirements*

The ontology developed for the system, as already mentioned in the previous section, needed to support the defined requirements resulting from the identified strategies. However, also some additional requirements, resulting from the already presented system flow, have been identified.

The eXtraSpec system acquires automatically data from dedicated sources, both company external and internal ones. The extracted content is saved as an extracted profile (PE), which is an XML file compliant with the defined structure of an expert profile based on the European Curriculum Vitae Standard [38]. Therefore, it consists of a number of attributes, such as e.g., education level, position, skill, that are assigned to different profile's categories such as e.g., personal data, educational history, professional experience. Vocabulary in the extracted content is then processed and normalized using the developed ontology. The result of the normalization process is a normalized profile (PN). An important assumption is: one standardized profile describes one person, but one person may be described by a number of standardized profiles (e.g., information on a given person at different points of time or information acquired from different sources). Thus, normalized profiles are analysed and then aggregated, in order to create an aggregated profile (PA) of a person. Finally, the reasoning mechanism is fed with the created aggregated profiles and answers user queries on experts. Thus, the additional requirements the ontology should address are as follows:

1. The ontology MUST enable semantic annotation of all elements of aggregated profile.
2. The ontology MUST support the normalization process of extracted profiles.

The creation of ontology for the needs of the eXtraSpec project was preceded by thorough analysis of the requirements resulting from the scenarios supported by the system as well as those mentioned above. In addition, the consequences of applying various formalisms and data models for the ontology modelling, and its further application, were investigated. In consequence, three assumptions were formulated:

- only few relations will be needed and thus, represented,
- developed ontologies should be easy to translate into other formalisms,
- the expressiveness of used ontology language is important, however, the efficiency of the reasoning mechanism is also crucial.

B. Formalism

As the result of the conducted analysis of different formalisms and data models, the decision was taken to apply the OWL language as the underlying formalism and the Simple Knowledge Organization System (SKOS) [37] model as a data model. The criteria that influenced our choice were as follows:

- relatively easy translation into other formalisms;
- simplicity of representation;
- expressiveness of used ontology language;
- efficiency of the reasoning mechanism.

Many knowledge representations, such as thesauri, taxonomies and classifications, share some structure elements and are used in similar applications. SKOS gathers most of those similarities and explicitly enables data and technology exchange between different applications. The SKOS data model enables low cost migration that allows making a connection between existing SKOS and the Semantic Web. Ontologies developed in accordance to the SKOS model can be expressed in any known ontology language.

Because of the strong software support and a wide usage of OWL, we decided to use that formalism within our work.

C. Model

The basic element of the eXtraSpec system is an already mentioned profile of an expert. Each expert is described with series of information, for example: name and family name, history of education, career history, hobby, skills, and obtained certificates. For the needs of the project, a data structure to hold all that information was designed. To make the reasoning possible, a domain knowledge for each of those attributes is needed. The domain knowledge is represented by the ontology. Ten attributes from the profile of an expert were selected to be a ‘dictionary reference', i.e., the attributes, whose values are references to instances from the ontology. Those attributes are:

- educational organization – name of organization awarding a particular level of education or educational title;
- certifying organization – name of an organization that issued the particular certificate;
- client, employer and role – those three attributes are used to describe the history of employment. A single step in the employment history is described as a business relation. Each relation consists of three basic elements: client (i.e., an employer) and a role (i.e., profession) that an expert fulfilled in this relation;
- scope of education – the domain of education (for example: IT, construction, transportation);
- topic of education – for a higher education description, it will be a name of the specialization, for trainings or courses etc. – their topic;
- result of education – the obtained title;
- skill – an ability to perform an activity or job well, especially because someone has practiced it;
- name of a certificate;
- degree of a skill.

Performed analysis of the requirements imposed on the ontology for the needs of reasoning, concluded with the definition of a set of relations that should be defined. They are as follows:

- hasSuperiorLevel - representing hierarchical relations between concepts,
- isEquivalent – representing the substitution between concetps,
- isLocatedIn – representing various geographical dependencies,
- isLocatedInCity – representing geographical dependencies,
- isLocatedInVoivodeship – representing geographical dependencies,
- provesSkillDegree – connecting skills and certificates,
- worksInLineOfBusiness - representing dependencies between organizations and lines of business,
- isPartOf – representing a composition of elements, for example: ability of using MSWord is a part of ability of using MSOffice (however, knowing MSWord does not imply that a person knows the entire MSOffice suit).

Additionally, various built-in SKOS relations have been used, namely:

- broader,
- hasTopConcept,
- inScheme,
- narrower,
- topConceptOf.

The SKOS model, while providing simplicity and easy translation into many different formalisms, imposes some restrictions. The most important one is the lack of support for some features and facilities provided by the OWL language. An overall idea of an ontology stack apart of concepts and data properties, assumed definition of some object properties. The designed ontology needed to be coherent with the SKOS model specification, processable by the used SKOS API and still represent all above mentioned areas and relations. To meet all those assumptions, the designed data structure is one SKOS ontology with eight concept schemas for each area of interest: Organizations (for organizational organizations, certifying organizations, Employer and Client), SkillName, SkillDegree, Certificate, Role, EducationScope, EducationTopic, EducationResult as well as complementary schemas for Cities and Voivodeship, Languages and Line of Business.

In the process of profile normalization the values from the extracted profile are linked to the concepts from the ontology. It is possible that the normalization mechanism will not be able to find the extracted value within the ontology. In this case, we assume that the extracted value should not be discarded; instead, it should be added to the appropriate Concept Schema. Therefore, every Concept Schema has a top concept TMP. Possible candidates for new concepts are added as a subConceptOf TMP, and later can be resolved by an expert. In this way, we make it possible to extend the ontology with new concepts found in the Internet or other sources describing expert's profiles.

D. *Sources of information*

While building the ontology for the needs of the eXtraSpec system, a wide range of taxonomies and classifications has been analyzed in order to indentify the best practices and solutions. As the eXtraSpec system is a solution designed for the Polish language, so is also the developed ontology. In order to develop particular Concept Schemas information from series of sources was incorporated. Table 4 shows the exemplary sources used to create the ontology structure as well as instances of numerous concepts.

TABLE 4 SOURCES OF INFORMATION

Concept schema	Sources of information
Organizations	The branch with Educational Organizations currently includes all Polish academic organizations, according to the official list published by the Polish Ministry of Science and Higher Education [39]. Additionally, a branch with employer organizations has been prepared based on the publicly available Internet sources.
Role	As this concept schema includes the classification of legally named professions in Poland, the source in this case was the official Polish Classification of Occupations [40] published by the Polish Ministry of Labor and Social Policy.
EducationScope	The data to create this Concept Schema was obtained from a number of Polish online job portals. The list of topical areas of education was slightly different in every portal. The final list of concepts in EducationScope Concept Schema is a combination of all of them.
EducationTopic	Currently this concept schema includes a list of specializations that a student may graduate in at Polish Higher Education Organizations based on the official register published by the Polish Ministry of Science and Higher Education.
EducationResult	This concept schema includes scientific titles, occupational tittles and academic degrees that may be obtained in Poland based on the appropriate ordinances of the Polish Ministry of Science and Higher Education.
CertificateName	On-going analysis focuses on language certificates possible to be obtained by Polish citizens.
SkillName	This concept schema was based on series of skill classifications provided by Polish job portals, as well as international scientific publications from the area of human resources management and IT solutions for human resources management area.
SkillDergee	On-going analysis focuses on solutions used in the mentioned job portals.
City	In this case the list of Polish cities was used.
Language	In this case a list of languages a good command of which can be proved by a certificate was utilised.
LineOfBusiness	In this case a list of lines of business that are used by job portals was prepared.
Voivodeship	In this case a list of Polish Voivodeships was used.

V. QUERYING AND REASONING MECHANISM

One of the most important functionalities of the eXtraSpec system is the identification of persons having the desired expertise. The application of the Semantic Web technologies in order to ensure the appropriate quality of returned results implies application of a reasoning mechanism to answer user queries.

In order to support the querying and reasoning scenarios, the eXtraSpec system needs not only an appropriate representation of information supporting reasoning over person's characteristics (as described within the previous section), but also the querying and reasoning mechanism itself supporting on the one hand, precise identification of required data, and on the other hand, being efficient and scalable.

A. *Approaches to semantic-enabled reasoning*

Given the above criteria (precision and recall on the one hand, and efficiency and scalability on the other), three possible approaches were considered.

The *first* approach involves using the fully-fledged semantics by expressing all expert profiles as instances of an ontology, formulating queries using the defined ontology, and then, executing a query using the reasoning mechanism. This approach involves the need to load all ontologies into the reasoning engine and representing all individual profiles as ontology instances. The performed experiments showed that querying the reasoning infrastructure, even while using only a small set of gathered profiles, is a resource (large memory consumption) and time consuming task (up to a few minutes). Therefore, although having a high precision and recall, it has poor performance and scalability.

The *second* approach relies on the query expansion using an ontology, i.e., adding keywords to the query by using an ontology to narrow or broaden the meaning of the original query. It allows getting answers faster than the previous approach, however, it could not take into account additional relations expressed in the ontology, and therefore, did not always allow for an increased precision. In addition, each user query needs to be normalized and then expanded using the ontology, therefore, the application of a reasoner was necessary. The experiments showed that it affected the values of the system performance and scalability.

The *third* approach called pre-reasoning involves two independent processes:

- creation of enriched profiles (indexes), to which additional information reasoned from the ontology is added and saved within the repository as syntactic data;
- formulating a query with the help of the appropriate GUI using the defined ontology serving as a controlled vocabulary. Then, the query is executed directly on a set of profiles using the traditional mechanisms of IR. There is no need to use the reasoning engine while executing a query.

This approach allows circumventing the drawbacks associated with the first approach, shifting the burden of an operation on the stage of indexing using ontologies.

Our experiments proved that applying the fully-fledged semantics is a precise, but neither efficient nor scalable solution. The query expansion provides an increased precision of the results (in comparison to the traditional IR mechanisms) and has better scalability and efficiency than the fully-fledged semantics, however, does not allow to take full advantage of the developed ontologies and existing relations between concepts. Only application of the third considered approach allows taking advantage of the mature IR mechanisms while increasing the accuracy and completeness of the returned results by: introducing a preliminary stage called pre-reasoning in order to create enriched indexes and the minimum use of the reasoning engine during the search.

B. *Querying and reasoning component – architecture*

The eXtraSpec system consists of a number of modules specialized for different tasks. Its architecture is described in [23], in this paper we focus on the REA component (REAsoning) presented in Figure 1.

REA consists of an indexing mechanism (indexer), a searching mechanism (searcher), a composition mechanism (composer) and a reasoning engine with a set of ontologies loaded.

The selected approach requires the support of two independent processes:

- First, creating indexes of profiles - optimized for search, i.e., structured so as to enable a very fast search based on criteria pre-set by a user. The aggregated profile is analysed, divided into relevant sections, and then enriched with additional information using the ontology (pre-reasoning). Any modification of the ontology forces the need to change indexes.
- The second process that needs to be supported is defining the query matching mechanism on the enriched indexes - this process is initiated by the task of a user formulating queries using a graphical interface that is also discussed later within this section. An employer, constructing a query points to interesting criteria and values they should meet. In the background, the desired values of various features from the lists and combo boxes, point to specific elements from the ontology [48].

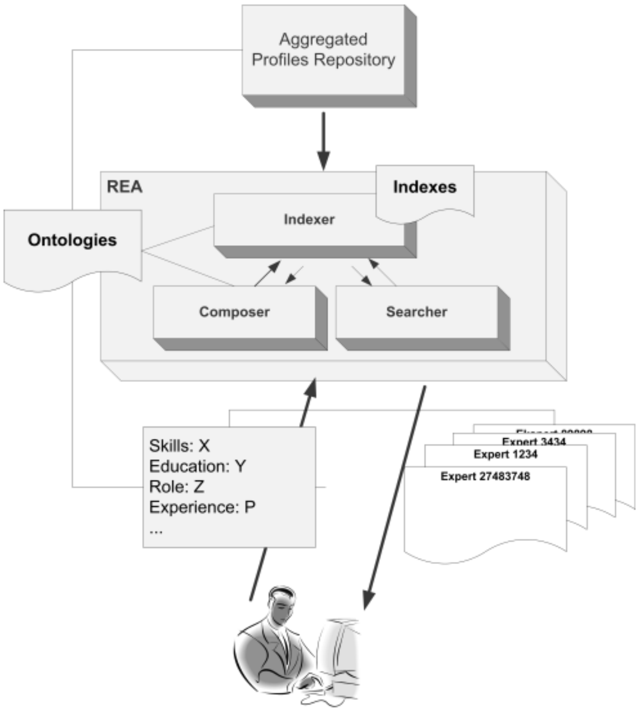


Figure 1. REA Overview

C. *Profile structure*

To realize the information retrieval side of the mechanism, the open-source java library Lucene [41], supported by the Apache Software Foundation, was selected. Instead of searching text documents directly, Lucene searches the previously prepared index. This speeds up the searching process and makes it more efficient. An index consists of at least one document. A document is a basic unit that is indexed and searchable, and represents text files, HTML code or database tables. A single document consists of fields. Each field has a unique name (used as a key) and a value. The result of the search process is a list of all relevant documents.

Fields in the Lucene documents cannot be grouped together nor stored as hierarchical structures. However, within an aggregated profile (PA), which is a base profile for searching, some hierarchies and groups might be found. Since an explicit mapping from PA to the Lucene document is not possible, during the indexing process profiles are divided into a number of separate documents as also shown in Figure 2.

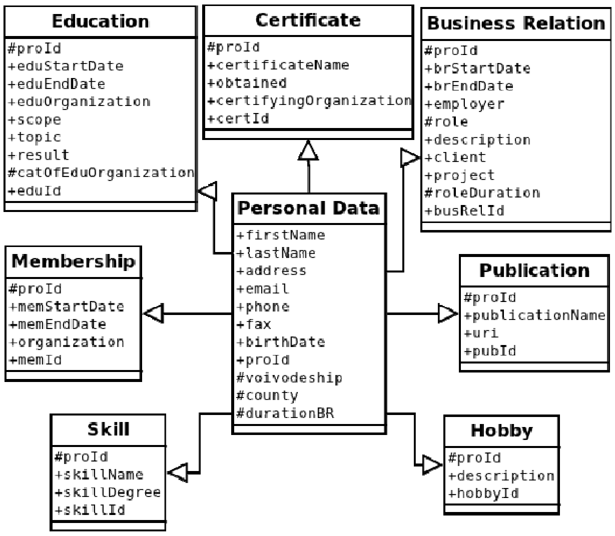


Figure 2. Data model overview

Each person is represented by exactly one Personal Data document and a number of corresponding documents that represent different groups of information. Each document contains an additional field with the profile ID that enables binding documents with the expert's main profile. Thus, for each listed category from PA, the separate Lucene document is created, e.g.., for one obtained certificate, one document is created. The mentioned documents are as follows:

- personal data (e.g., first name, last name, phone number, address),
- history of education,
- certificates,
- skills,
- publications,
- mentions,
- history of employment,
- organisations,
- hobby.

Concurrently with the indexing process, pre-reasoning takes place, in order to enhance the profile with the implied facts. The documents contain fields generated directly from PA (marked with +) as well as additional fields (marked with #). Moreover, fields such as e.g., role, skillName, catOfEduOrganization contain not only the concept from PA but also a hierarchy of its super-concepts from the ontology. Super-concepts are indexed as additional values for the given document field: these values are saved as next array elements and it is assumed that the higher array index number, the smaller weight the concept has. The assigned weight affects the ranking procedure.

As already mentioned, if the returned super-concepts do not correspond with the PA elements conceptually, additional fields are added to the document being indexed. For example, PA element 'address' might be divided into data that is more detailed, i.e., zip code, city, street, etc. Based on the zip code it is possible to specify the county and the province, and search for experts using the spatial criteria. Since PA does not contain such elements, we add fields to the personal data document during the indexing process.

D. Query structure

Lucene provides a very flexible but simple query structure. Therefore, in the eXtraSpec system it had to be extended in order to correspond to the defined requirements that result from the querying scenarios. They are as follows:

1. The querying and reasoning mechanism MUST allow building queries in a structured way (i.e., feature: desired value).
2. The querying and reasoning mechanism MUST support definition of desired values of attributes in a way suitable to the type of data stored within the given feature (i.e., text fields using wild-cards, date fields - after of before certain dates; numbers - less than..).
3. The querying and reasoning mechanism MUST allow to join a subset of selected criteria within the same category into one complex requirement (e.g., category: education; {education level: university AND finished date: after 2010 year}) using different logical operators.
4. The querying and reasoning mechanism MUST allow formulating a set of complex requirements within one category with different logical operators.
5. The querying and reasoning mechanism MUST allow joining complex requirements formulated in various profile categories into one criteria with different logical operators.

The logical operators between different sets of criteria and criteria themselves include such operators as: must, should, must not.

In order to answer more sophisticated queries encompassing several criteria from various documents, users' queries are executed on the index using a set of QueryObjects for different categories. Those QueryObjects are in turn sets of QueryObjects within the given category, each consisting of a set of QueryObject's structures consisting of a query string and a query operator. A query string is a Lucene compliant phrase that includes the field name and the relative value. A query operator is a logical operator: MUST, SHOULD, MUST_NOT, that defines whether the specified criteria should be included or excluded from the result set.

The performed tests have shown that the defined query object fulfils the formulated requirements.

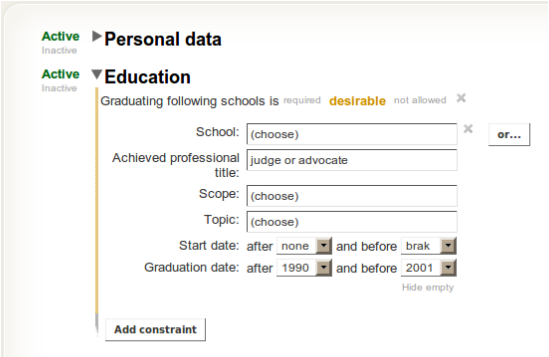
The application of semantics in the form of a pre-reasoning phase allowed achieving precise results, simultaneously allowing taking advantage of the matured IR mechanisms guaranteeing scalability and good performance of the system. Such a structure of the query together with the set of defined methods allow to address the scenarios defined above, however, makes formulating queries more complicated for users. Thus, a challenge of designing a user-friendly interface has appeared. The developed interface is shortly described in the next subsection.

E. GUI

The front-end to the eXtraSpec system should enable users to build complex queries describing characteristics of desired experts. During the analysis phase the main requirements for the system interface have been defined, namely [48]:

1. The interface MUST enable a user to specify constraints on expert's attributes and select whether the value of an attribute is required, desired (but not required) or not allowed.
2. The interface SHOULD enable grouping of constraints e.g., it should be possible to specify a graduated school and graduation date as one criterion.
3. The interface SHOULD provide a possibility to build queries which include complementary and alternative constraints.
4. The interface SHOULD enable providing some of criteria values typed-in as free text (with wildcards) and some of them to be selected from the eXtraSpec system knowledge base.
5. The interface SHOULD be loosely coupled with the system.
6. The interface SHOULD be understandable and easy to use.

The conceptual model of the interface is determined by the scheme of querying the experts finding system and the structure of the aggregated profile. The search criteria are divided into the following categories: personal data, education, professional experience, foreign languages, courses, certificates, additional skills, organization membership and interests.



Active
Inactive
▶Personal data

Active
Inactive
▼Education

Graduating following schools is required **desirable** not allowed ×

School: (choose) × or...

Achieved professional title: judge or advocate

Scope: (choose)

Topic: (choose)

Start date: after none and before brak

Graduation date: after 1990 and before 2001

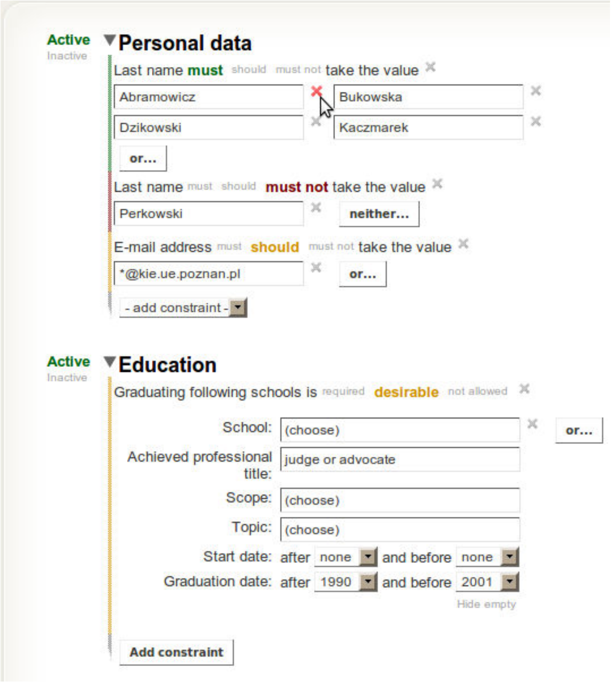
Hide empty

Add constraint

Figure 3. The eXtraSpec GUI (1)

Categories consist of groups of fields. Desired values of these fields are specified in the interface by criteria values, and field groups by criteria groups. Each criterion has a label and a value typed by the user, selected from list or from values tree loaded from the ontology.

As a result eXtraSpec system front-end is a dynamic web user interface with cross-browser compatibility.



Active
Inactive
▼Personal data

Last name **must** should must not take the value ×

Abramowicz × Bukowska ×

Dzikowski × Kaczmarek ×

or...

Last name must should **must not** take the value ×

Perkowski × neither...

E-mail address must **should** must not take the value ×

*@kie.ue.poznan.pl × or...

- add constraint -

Active
Inactive
▼Education

Graduating following schools is required **desirable** not allowed ×

School: (choose) × or...

Achieved professional title: judge or advocate

Scope: (choose)

Topic: (choose)

Start date: after none and before none

Graduation date: after 1990 and before 2001

Hide empty

Add constraint

Figure 4. The eXtraSpec system GUI (2)

The developed interface has been successfully evaluated. See [48] for more details.

VI. CONCLUSIONS

The main goal of the eXtraSpec project is to develop a system supporting analysis of company documents and selected Internet sources for the needs of searching for experts from a given field or with specific competencies. The provided system focuses on processing texts written in the Polish language. The obtained information is stored in the system in the form of experts' profiles and may be consolidated when needed. The system aims to offer a user friendly interface to perform queries that allow to find persons with specific characteristics. Realisation of this goal requires interconnection between the developed interface and underlying ontologies. Within this paper, we have discussed the concept and considered scenarios regarding the implementation of the querying and reasoning mechanism for the needs of the eXtraSpec system. We argue that by introducing the pre-reasoning phase, the application of semantics may be used to achieve precise results when searching for experts and at the same time, ensure the proper performance and scalability.

The set of developed ontologies discussed within this paper was designed specially for the Polish language, however, the main structure and model as well as defined relations may be reused also for other languages. The ontology in question is still under development, however, in the current state of affairs the reasoning about competencies in order to complete the expert profile with additional data on education, work experience is successfully performed by the REA component described within this paper. Our current work focuses on the implementation of the second scenario supported by the eXtraSpec system i.e., composition of teams of experts using the developed ontology.

Acknowledgment

The work published in this article was supported by the project titled: "Advanced data extraction methods for the needs of expert search" financed under the Innovative Economy Framework and partially supported by European Union in the European Regional Development Fund programme (contract no. UDA-POIG.01.03.01-30-150/08-01).

References

[1] Abramowicz, W., Bukowska, E., Kaczmarek, M., Starzecka, M. "Semantic-enabled Efficient and Scalable Retrieval of Experts", Proceedings of Third International Conference on Information, Process, and Knowledge Management (eKNOW), 2011

[2] van Rijsbergen, C. J.; "Information Retrieval and Information Reasoning". Computer Science Today 1995,pages 549-559

[3] Yimam, D.; "Expert finding systems for organizations: Domain analysis and the demoir approach" in: ECSCW 999 Workshop: Beyond KNowledge Management: Managing Expertise, pages 276–283, New York, NY, USA, 1996. ACM Press

[4] McDonald, D. W. and Ackerman, M. S.; "Expertise recommender: a flexible recommendation system and architecture" in: CSCW '00: Proceedings of the 2000 ACM conference on Computer supported cooperative work, pages 231–240. ACM Press, 2000.

[5] Yimam-Seid, D. and Kobsa, A. "Expert finding systems for organizations: Problem and domain analysis and the demoir approach". Journal of Organizational Computing and Electronic Commerce, 13(1):1–24, 2003

[6] Kautz, H., Selman, B., and Milewski, A.; "Agent amplified communication" in: Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96), pages 3–9, 1996

[7] Campbell, C. S., Maglio, P. P., Cozzi, A., and Dom, B.; "Expertise identification using email communications" in: CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management, pages 528–531. ACM Press, 2003

[8] Hawking, D.; "Challenges in enterprise search" in: Proceedings Fifteenth Australasian Database Conference, 2004

[9] Metze, F., Bauckhage, Ch., and Alpcan, T., "The "Spree" Expert Finding System" in: Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007), September 17-19, 2007, Irvine, California, USA

[10] Ackerman, M.S., Wulf, V. and Pipek, V.; "Sharing Expertise: Beyond Knowledge Man-agement"; MIT press, (2002).

[11] Krulwich, B. and Burkey, C.; "ContactFinder agent: answering bulletin board questions with referrals" in: Proceedings of the National Conference on Artificial Intelligence, pages 10-15, 1996

[12] Balog, K., Azzopardi L. and De. Rijke, M.; "Formal models for expert finding in enterprise corpora" in: Proceedings of the ACM SIGIR, pages. 43-50, 2006.

[13] Fang, H. and Zhai, C.; "Probabilistic models for expert finding" in: Proceedingsof the ECIR, pages 418-430, 2007

[14] Petkova, D. and Croft, W.; "Hierarchical language models for expert finding in enterprise corpora" in: Proceedings of the 18th IEEE International Conference on Tools with Artificial Intel-ligence, pages 599-608, 2006

[15] Serdyukov, P. and Hiemstra, D.; "Modeling documents as mixtures of persons for expert finding" in: Proceedings of the ECIR, pages 309-320, 2008.

[16] Gómez-Pérez, A., Ramírez, J., and Villazón-Terrazas, B., "An Ontology for Modelling Human Resources Management Based on Standards" in: B. Apolloni et al. (Eds.): KES 2007/WIRN 2007, Part I, LNAI 4692, pp. 534–541, 2007

[17] Dorn, J., Naz, T., and Pichlmair, M., "Ontology Development for Human Resource Management" in: "Proceedings of 4rd International Conference on Knowledge Management", Ch. Stary, F. Barachini, and S. Hawamdeh (Hrg.); Series on Information&Knowledge Management, 6 (2007), ISBN: 978-981-277-058-5; S. 109 - 120.

[18] Dorn, J. and Naz, T.; "Meta-search in Human Resource Development", in: Proceedings of 4th Int. Conference on Knowledge Systems, Bangkok, Thailand, 2007

[19] Dorn, J. and Pichlmair, M.; "A Competence Management System for Universities", in: European Conference on Information Systems, St. Gallen, 2007

[20] Aleman-Meza, B., Bojars, U., Boley, H., Breslin, J.G., Mochol, M., Nixon, L.JB., Polleres, A., and Zhdanova, A.V., "Combining RDF Vocabularies for Expert Finding"

[21] Dittmann, L.;"Towards Ontology-based Skill Management, Projektbericht zum Verbundprojekt KOWIEN", Universität Duisburg-Essen, 2003.

[22] Abramowicz, W., Wieloch, K.; "Raport podsumowujący wyniki prac przeprowadzonych w ramach zadań Z1.1, Z1.2 oraz Z2.1", Technical report of the eXtraSpec project, Department of Information Systems, Poznan University of Economics, 2009

[23] Abramowicz, W., Kaczmarek, T., Stolarski, P., Węcel, K., and Wieloch, K.; “Architektura systemu wyszukiwania ekspertów eXtraSpec”, in: Proceedings of “Technologie Wiedzy w Zarządzaniu Publicznym”, Hucisko, 19-21 September 2010

[24] http://extraspec.kie.ue.poznan.pl/, last access date: 20.01.2012

[25] http://www.bizwiz.com, last access date: 20.01.2012

[26] http://www.xing.com, last access date: 20.01.2012

[27] http://linkedin.com. last access date: 20.01.2012

[28] http://www.iso.org/iso/en/prods-services/popstds/currencycodeslist.html, last access date: 20.01.2012

[29] http://ec.europa.eu/eurostat/ramon/index.cfm?TargetUrl=DSP_PUB_WELC, last access date: 20.01.2012

[30] http://online.onetcenter.org/, last access date: 20.01.2012

[31] http://cs.yale.edu/homes/dvm/daml/time-page.html, last access date: 20.01.2012

[32] http://www.foaf-project.org/, last access date: 20.01.2012

[33] http://sioc-project.org/, last access date: 20.01.2012

[34] http://www.imc.org/pdi/, last access date: 20.01.2012

[35] http://dublincore.org/, last access date: 20.01.2012

[36] http://www.bls.gov/soc, last access date: 20.01.2012

[37] http://www.w3.org/TR/swbp-skos-core-spec, last access date: 20.01.2012

[38] http://www.europa-pages.com/jobs/europass.html, last access date: 20.01.2012

[39] http://www.nauka.gov.pl/szkolnictwo-wyzsze/system-szkolnictwa-wyzszego/uczelnie/, last access date: 20.01.2012

[40] http://www.praca.gov.pl/pages/klasyfikacja_zawodow2.php, last access date: 20.01.2012

[41] http://lucene.apache.org, last access date: 20.01.2012

[42] Berners-Lee, T., Hendler, J. & Lassila, O., “The semantic web”, Scientific American, May, 2001, pages 35-43.

[43] Shadbolt, N.; Berners-Lee, T.; Hall, W., “The Semantic Web Revisited”, IEEE Intelligent Systems Journal, Vol. 21, no. 3, 2006 page. 96-101.

[44] Navigli, R., Velardi, P. “An Analysis of Ontology-based Query Expansion Strategies”, Proceedings of Workshop on Adaptive Text Extraction and Mining at the 14th European Conference on Machine Learning, Cavtat-Dubrovnik, Croatia, 2003, pp. 42–49

[45] Michalski, R., Palus, S., Kazienko, P., “Matching Organizational Structure and Social Network Extracted from Email Communication”, Lecture Notes in Business Information Processing, 2011, Volume 87, Part 6, Part 6, 197-206

[46] http://www.w3.org/RDF/, last access date: 20.01.2012

[47] http://www.w3.org/TR/2004/REC-owl-features-20040210/, last access date: 20.01.2012

[48] Abramowicz, W., Bukowska. E., Dzikowski, J., Filipowska, A., Kaczmarek, M., “Web Interface for Semantically Enabled Experts Finding System”, in: ICEIS 2011, Proceedings of the 13th International Conference on Enterprise Information Systems , Beijing, SciTePress – Science and Technology Publications, 2011. pp. 291-296, ISBN 978-989-8425-56-0

[49] Distribution, G., & Lundvall, B. A. “The Knoweldge-based Economy”, Development (96), 115.,OECD, 1996. http://www.oecd.org/dataoecd/51/8/1913021.pdf, last access date:20.01.2012.

[50] Chasins, Jeff. “Social media, recruiting, and job boards: which way are we going?” ere.net. 14 Sept. 2010. Ere Media, Inc. 22 December 2010, http://community.ere.net/blogs/job-board-doctor/2010/09/social-media-recruiting-and-job-boards-whichway-are-we-going, last access date: 20.01.2012