Semantic Annotation of Documents Applied to e-recruitment

L. Yahiaoui, Z. Boufaida and Y. Prié

Abstract—**This paper presents an approach based on semantic annotation of CVs and job offers for automating recruitment on the web. The main idea consists on modeling formally the semantic content of these documents in term of their acquirements (case of a CV) or requirements (case of a job offer) using a shared ontology between recruiters and job seekers. The domain ontology built is inspired from the most significant parts of these documents (personal qualifications, diplomas and job experiences) and handle the competencies management. It describes a model of competency as well as hierarchies of topics that a competency can have. It allows the end user to explicitly enrich his document with metadata (annotations). Semantic matching between supply and demand, based on the computation of a coefficient, can be applied in a superficial or a deep way. Superficial matching deals with all acquirements/requirements mentioned explicitly by a job seeker/recruiter (a special diploma, a special job experience or a special personal qualification), whereas competency based matching deals with all competencies underlying the user's document.**

Index Terms—**Competency, Semantic annotation, Semantic matching, Semantic Web**

I. Introduction

THE evolution of job market has proven that traditional methods of recruitment are becoming inefficient. Internet has introduced a new way of managing human resources. Nowadays, job seekers can send their CVs directly to companies (email) or to dedicated servers on the Web. Recruiters, on the other side, can publish their job offers on the Web with a significant reduction in cost and time. In this context, electronic recruitment tends to automate matching between the published CVs and job offers. The major problem is that these resources are often badly used because available management techniques and tools are purely syntactic and remain limited in front of the increasing number of documents to process and the need for a more semantic interpretation of their content.

Automatic matching between supply and demand requires the use of new approaches based on semantic Web technologies. The idea consists on extending syntactic structures of documents with a semantic content in order to make them machine-understandable [8]. For that, two approaches are proposed: (i) semantic annotation of documents which consists on using a shared ontology to enrich documents with metadata [11] and (ii) semantic indexing of documents based on the construction of an index that will have a structure inspired from the used ontology.

In what follows, we propose a simple approach based on semantic annotation of documents to automate the e-recruitment process. The main idea consists on formally modeling the semantic content of these documents in term of their acquirements/requirements, in a simple and efficient way, by using a shared ontology between job seekers and recruiters. Concepts of this ontology are inspired from the most significant parts of these documents and competency is considered as the crucial element in the proposed modeling. This objective requires the description of the global architecture of our semantic annotation and matching system (presented in section II), the definition of a competency model (section III), the development of an ontology used to annotate documents with their semantic contents (section IV) and the definition of an efficient and simple semantic matching process between CVs and job offers (Section V). A pre-evaluation of our approach is shown in Section VI then a conclusion and perspectives for improving this work are given at the end of this paper.

II. The Annotation & Matching System Architecture

The architecture of the annotation and semantic matching system is illustrated in "Fig. 1". It is composed of:

A. The ER-ontology

An ontology framework composed by ontologies related to each other, dedicated to annotate CVs and job offers by its concepts instances. The metadata repository is used to store generated annotations.

Manuscript received October 23, 2006. This work was supported in part by the Laboratory "LIRE", Computer science Department, University of Constantine-Algeria.

L. Yahiaoui is with the computer science department, Laboratory "LIRE", University Mentouri Constantine, Constantine 25000 Algeria (phone: 00213-31818817; fax: 00213-31818817; e-mail: yahiaoui_lilapg@yahoo.fr).

Z. Boufaida is with the computer science department, Laboratory "LIRE", University Mentouri Constantine, Constantine 25000 Algeria (phone: 00213-31818817; fax: 00213-31818817; e-mail: boufaida@hotmail.com).

Y. prié belongs to laboratory "LIRIS", UMR 5205 CNRS, University Claude Bernard Lyon 1, F-69622 Villeurbanne Cedex, France (phone: 0033-472431636; fax: 0033-472431536; e-mail: yannick.prie@liris.cnrs.fr).

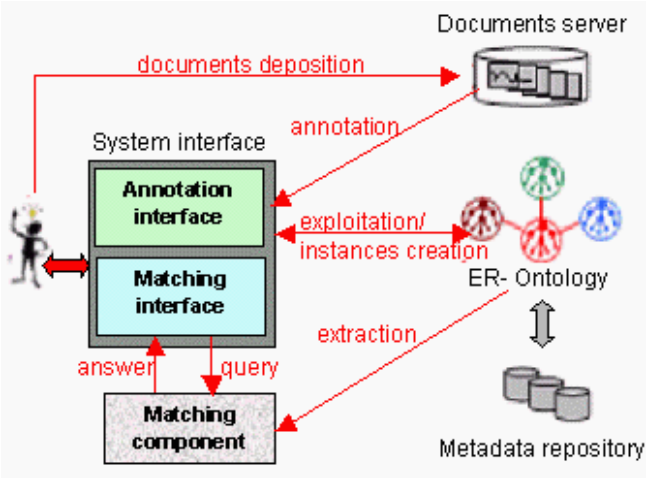

Fig. 1. Semantic annotation & matching system architecture.

B. *The XML/HTML documents server*

It allows the storage and the management of documents to be annotated (CVs and job offers).

C. *The system interface*

It offers two functions. The annotation interface gives the end-user the possibility to annotate his document by using the ER-ontology and generating metadata (annotations). The matching interface allows the end-user to submit queries to the matching component and presents the returned results. A recruiter can find the most qualified candidate according to his needs; whereas a job seeker can find the best job fitting to his qualifications.

D. *The matching component*

It interprets the user's query to get the URI of the user's document (CV/job offer) and the kind of semantic matching to apply, then it calculates coefficients of semantic matching (superficial or competency based coefficients) between the user's document and all available annotated documents (if the end-user is a job seeker, the matching process will use his CV and all available job offers). The result is a set of pairs (URI/ C_match), where URI is the identifier of the found document and C_match the associated coefficient (percentage) of semantic matching (superficial or competency based).

III. THE COMPETENCY MODEL

Human resources management is based on the knowledge of individuals and their competencies, as well as on the knowledge of the organization and its jobs. By mapping these competencies, it is possible to enhance recruitment [13]. This requires an explicit representation of competencies and thus a model for this concept. A competency can be identified as a set of knowledge used to accomplish a task [13]. It can appear as an aptitude (behaviour) or a scientific and technical competency (a knowledge or a know-how). The scientific and technical competency can be specific (related to a particular domain) or general. In this work, we are interested in "computer science and telecommunications" domain so our scientific and technical competency relates a competency object to a competency level [10]. The competency object can be a «technology topic» or a «software artefact». The competency level can have one of the following values: Basic (B or 20%), Application (A or 50%), Master ship (M or 70%) or Expert (E or 90%). Aptitudes, identified by their names, are inspired from CIGREF [5]. "Fig. 2" illustrates the competency model adopted.

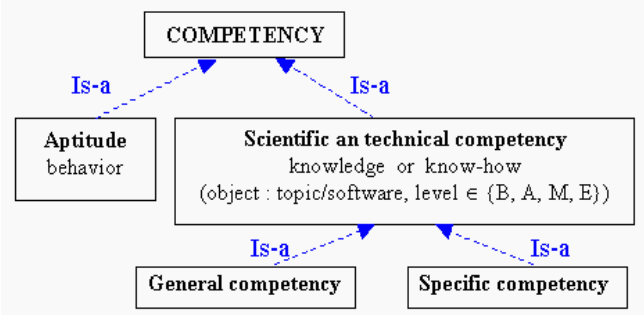

Fig. 2. The competency model.

This competency model seams to be simple compared with the one proposed in the project COMMONCV [4] because our objective is to find a compromise between simplicity and efficiency. Adding the context in the competency model will insure a better matching between CVs and job offers but it will also complicate this process.

IV. SEMANTIC CONTENT MODELING BASED ONTOLOGY

Ontology, considered as a formal and explicit specialization of a shared conceptualization, is the key element of the semantic web. Ontologies are crucial for e-recruitment because they allow recruiters and job seekers to share a common reference system to describe contents of their documents in a non-ambiguous, simple, semantic and a formal way. The importance of formalization is to allow an automatic matching between supply and demand.

A. *The ER-ontology architecture*

Elements of the ER-ontology (Electronic-Recruitment ontology) are inspired from the most significant and common parts between CVs and job offers. This includes personal information, diplomas, job experiences and explicit competencies acquired by a candidate (CV) or required by a job position (job offer). Furthermore, a job or a diploma mobilizes a subset of elementary competencies [4], what make the competency the crucial element in the proposed modeling. For the construction of the ER-ontology, some ideas are inspired from existing works [3] and [14]. We have chosen METHONTOLOGY [6] as a development method. "Fig. 3" illustrates the global architecture of the ER-ontology in term of linked ontologies (or sub-ontologies). These ontologies are

detailed in "Fig. 4" as a set of concepts hierarchies (shown as rectangles) with semantic relations between them. Competency model concepts are distinguished by doubled edges.

The domain of this ontology is "Computer Science and Telecommunications". It is considered as a framework composed by five ontologies:

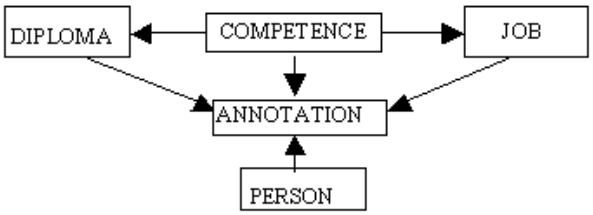

Fig. 3. Sub-ontologies of the ER-ontology.

1) The ontology "PERSON": composed only of one concept "Person" which describes the most important personal characteristics that a recruiter can require (and that a candidate can have). It includes: sex, maximum age, military service, residence (country and city), driving licence, familial state and nationality.

2) The ontology "DIPLOMA": Describes concepts related to diplomas and trainings. This includes Diploma families, valid domain diplomas and a diplomas reference system inspired here form the Algerian high education system [9]. It is related to the ontology "COMPETENCY" to attest competencies mobilized by a particular diploma.

3) The ontology "JOB": Describes concepts related to job experiences. This includes job families, existing domain jobs and a jobs reference system inspired form CIGREF [5]. It is related to the ontology "COMPETENCY" to attest competencies mobilized by a particular job.

4) The ontology "COMPETENCY": Describes the adopted competency model and hierarchies of objects ("TechnologyTopic" or "SoftwareArtefact") that can have the scientific and technical competency [10]. In the computer science domain, a topic can be general, mathematic or specific to this domain. The hierarchy of the general topic is inspired from general knowledge of CIGREF [5] and that of the mathematic topic inspired from the Algerian high education programs in computer science [9]; Whereas the hierarchy of the computer technology topic is inspired from information system competencies of CIGREF [5], the Algerian high education programs in computer science [9] and other modeling works related to computer science disciplines [1]. The hierarchy of topics is built in order to cover the majority of computer science disciplines including knowledge and know-how. Each topic is characterized by an attribute "weight" which represents the percentage of its contribution in its parent topic. This hierarchy will allow persons handling diplomas to bring their knowledge closer to competencies required by a particular job throw the computation of a semantic matching coefficient

5) The ontology "ANNOTATION": Allows associating a resource with all its corresponding acquirements/requirements (case of a CV/a job offer). The concept "Resource" describes the document to be annotated through its URI (Unified Resource identifier) and type (CV or a job position). The concept "AcquiRequi" is specialized in elements that a resource can be annotated with and links the "ANNOTATION" ontology with the other sub-ontologies. The concept "Annotation" relates the two former concepts in order to annotate a specific resource with a set of acquirements/requirements. The role of this ontology can be replaced by a semantic annotation tool.

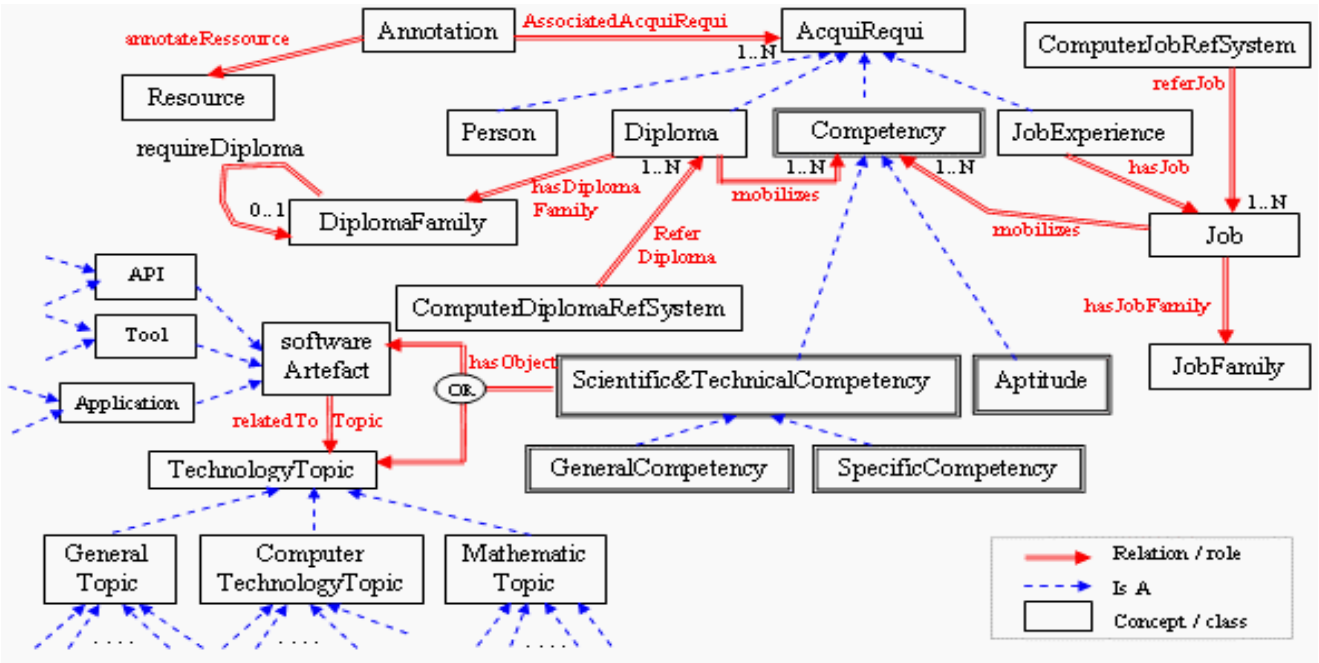

Fig. 4. The detailed architecture of the ER-ontology.

The ER-ontology contains 510 concepts : 109 concepts belong to the general topic hierarchy, 351 concepts belong to the computer technology topic hierarchy and 18 concepts belong to the software artefact hierarchy. These concepts are characterised by 20 attributes and 17 relations. This ontology is implemented as a single ontology in OWL (Ontology Web Language) [2] using Protégé_3.1[12].

B. Semantic annotation process

Concepts instances of Sub-ontologies "JOB", "DIPLOMA" and "COMPETENCY" are created by the system administrator (before any annotation by the end-user). The end-user can use these instances during the annotation of his document. An instance of the class "Competency" is created for each subclass of "TechnologyTopic" or "SoftwareArtefact" with the four possible competency levels {B, A, M, E). Instances of classes "Job" and "Diploma" are related to instances of the class "Competency" that they mobilize. The role of the end-user consists on:

--Creating an instance of the class "Resource" to describe the document to be annotated.

--Creating an instance of the class "Person" to describe personal information of a candidate (case of a CV) or the required personal information (case of a job offer).

--Creating instances of the class "JobExperience" to describe the candidate's job experiences or job experiences required by a recruiter. This description includes the name of a job and years of expertise.

--Creating instances of the class "AcquiRequi" with all the requirements of a job offer or the acquirements of a candidate, by using available instances.

--Creating instances of the class "Annotation" to link acquirements/requirements to the annotated resource.

In our current version, instantiating classes is a manually process using the interface of protégé-OWL[12].

V. SEMANTIC MATCHING BETWEEN DOCUMENTS

Once documents are annotated, a semantic matching algorithm can be applied between a particular CV (CV_1) and a job offer (P_1). This matching is based on the computation of a coefficient (percentage) which can be done according to two different but complementary techniques: (i) superficial semantic matching takes into account requirements or acquirements that annotate a document at a superficial level, whereas (ii) competency based semantic matching uses all competencies underlying the annotated document.

A. Competency based semantic matching

This kind of semantic matching is interested in competencies underlying the annotated documents. The main idea consists on searching each required competency (by a job offer) in the set of competencies acquired by a candidate (CV). If this competency exists, a weight will be cumulated; otherwise the topic's hierarchy of this competency (the case of a scientific and technical competence) will be exploited to calculate the level of the candidate in this topic. A weight is associated to each type of competency. For example, we can assign a coefficient of 2 to both "GeneralCompetency" and "Aptitude", and a weight of 6 to "SpecificCompetency". We have chosen these initial values for coefficients according to the importance of each type of competency in the recruitment process, but they can be adjusted according to tests results.

Competency based matching (R_{OF} , R_{CV})

```
tot_weight ← 0  /*requirements coefficients accumulation*/
calc_weight ←0 /*acquirements coefficients accumulation*/
Extraction_Competencies (C_CV , C_OF)
For each   c_r ∈ C_OF  Repeat /*a required competency */
  If Aptitude (c_r) then  /*case of a behavior */
    tot_weight ← tot_weight + Coef_type (Aptitude)
    If  c_r ∈ C_CV  then  /*the Competency is acquired*/
      calc_weight ← calc_weight + Coef_type (Aptitude)
  Else    /*Scientific &technical competency */
    If GeneralCompetency (c_r) then
      C1← Coef-type (GeneralCompetency)
    Else C1 ← Coef-type (SpecificCompetency)
    tot_weight ← tot_weight + C1
    X←{c∈ C_CV / c.hasObject = c_r .hasObject) /*same topic*/
    If X =∅  then
       lev ← Evaluate _subtopics  (class (c_r.hasObject), 1)
    ELSE  c_A← c∈ X / (∀y ∈ X, y.level ≤ c_A.level)
       lev ← c_A.level  /*choosing the best level*/
    If  (lev ≥ c_r. level)  then  /*the level acquired greater*/
       calc_weight ← calc_weight + C1
    Else  K := integer ((c_r . level - lev) / 20) /*level rounding*/
       calc_weight ←  calc_weight  + (C1* (1-K / 4))
 End repeat
   C_match ← (calc_weight / tot_weight) * 100.
End
```

Evaluate _subtopics (T , coef)

```
F ← {F_i , i≥ 0/  F_i → T} /* sub-concepts of the topic T */
lev ← 0 /* the level of the condidat in the topic T */
If  F = ∅ then
  return (lev*coef) /* coef = % of participation of T in his parent topic */
Else
  For each F_i ∈ F  Repeat /* for each sub-concept of T*/
    C←{c∈ C_CV / F_i(c.hasObject)} /*set of acquired competencies
    that have an instance of F_i as a topic*/
    If C ≠∅  then /*choosing the best acquired competency*/
       x ← c∈ C/ (∀y ∈ C, y. level ≤ x. level)
       lev ← lev + (x. level* (x. hasObject . coef))
    Else   /*go deeper in the hierarchy of topics*/
      lev ← lev + Evaluate _subtopics (F_i, coef (F_i ,T)) /*recursively*/
  End Repeat
 return (lev*coef)
End if
End
```

Fig. 5. Semantic matching algorithm.

Note : the parameter Coef used in the function *Evaluate_Subtopics* reflect the percentage of participation of the topic T in its parent topic. For example we can estimate that the percentage of participation of the topic "Software design" in its parent topic "software engineering" is 25% (or 0.25).

The scientific and technical competence level is evaluated as (B ≅20%) if level<25%, as (A ≅50%) if 25% ≤level < 60%, as (M ≅70%) if 60% ≤level ≤75% and as (E ≅90%) if level >75 %. "Fig. 5" illustrates the competency based semantic matching algorithm, witch interests as most, between a CV (R_{CV}) and a job offer (R_{OF}). The following Conventions are used:

--C(I): I is an individual concept/class C (so class(I)=C).

--I.atrName: the value of the attribute "atrName" of the individual I or all individuals related to I by the role "atrName".

--A → C: class A is a sub-class of class C.

The function "Extraction_Competencies" extracts all competencies underlying the CV in the set C_{CV} and those of the job offer in the set C_{OF}. Theses competencies can be explicit (explicit annotations) or implicit (mobilized by a particular diploma or a job experience).

In addition to the power of expression offered by OWL, used to implement the ER-ontology, powerful inference services are offered by the a reasoner called RACER [7]. This reasoner is a knowledge representation system that implements a highly optimized tableau calculus for a very expressive description logic. It can interpret OWL documents and offers reasoning services for multiple T-Boxes and for multiple A-Boxes as well.

At the terminological level, various types of queries can be applied. For instance: to check the consistency of a concept or to control relations between concepts (descendants or parents). The first functionality was used for the validation of the ER-ontology, while the second one can be used to exploit the topics hierarchy of the scientific and technical competency in the implementation of the competency based matching algorithm (for example : to implement $F \leftarrow \{Fi , i \geq 0 / Fi \rightarrow T\}$).

At the A-Boxes level, other queries are possible. The most interesting for us are : calculating the direct type (class) of an individual, which can be used in the IF-statements (for example: If GeneralCompetency(Ci)) and extracting instances of a particular class, even according to various criteria, based on analysing roles and attributes of these instances. RQL (Racer Query Language), which is an extended query language for RACER, makes it possible to use complex queries on OWL documents that can be useful in the implementation of extraction functions mentioned in the proposed matching algorithms (for example: Extraction_Competencies(Ccv , Cp)).

B. Superficial semantic matching

Acquirements or requirements that can explicitly annotate a document (CV/job offer) have four types: a competency, a diploma, a job experience (job + years of expertise) or personal information. In superficial matching, researching a particular job offer requirement in the candidate's acquirements set is done with exactitude (exists or not). A weight is associated to each type of acquirements/requirements to reflect its importance in the computation of the matching coefficient. For example, we can assign the coefficient 8 to the type "Person" (1 for each personal qualification), 10 for the type "Diploma", 20 for the type "JobExperience" and 5 for the type "Competency". These weights can be adjusted according to tests results. The matching coefficient will depend on the difference between the some of weights of all job requirements (tot_weight) and the some of weights of requirements satisfied by the candidate (calc_weight). It is calculated as :

C_match = (calc_weight/ tot_weight)*100.

VI. DIPLOMAS AND JOBS MATCHING

The test of our approach on a set of documents has given satisfying results. From the simplicity and efficiency view points, the two proposed techniques of semantic matching offer to the recruiter a deep vision of the received CVs (satisfaction of superficial and deep requirements). Furthermore, a job seeker have the possibility to make closer his competencies with those are required by a particular job position. "Table.I" shows the results of the competency based semantic matching between five job offers (with different job positions) and four CVs (candidates having distinct diplomas).

TABLE I
RESOLTS OF THE SEMANTIC MATCHING BASED COMPETENCY (%)

Diplomas \ Jobs	TNT	AM	D	DBA	EOS
AB	60,50	61,36	59,32	54,38	42,60
BSE	55,63	67,74	80,79	55,21	37,00
BIS	57,60	90,66	72,60	73,18	38,80
BSTIC	76,71	66,41	70,48	65,33	48,40

Jobs: TNT= Technician in network and telecommunication, AM= applications manager, D= developer, DBA= data base administrator, EOS= expert in operation system. Bachelor

Diplomas: AB = Academic Bachelor, SEB= Bachelor in Software Engineering, ISB = Bachelor in Information System, STICB = Bachelor in Science and Technology of Information and Communication.

It is clear that these coefficients reflect the relation between jobs and diplomas. For Instance a candidate having a professional Bachelor in Science and Technology of Information and Communication (BSTIC) is the most qualified to get the position of a Network and Communications Technician among the other candidates.

VII. CONCLUSION

In this paper, a simple and efficient approach based on semantic annotation for automating electronic recruitment was proposed. It is characterized by modeling the semantic content of CVs and job offers using a shared ontology between recruiters and job seekers. Elements of this ontology are inspired from the most significant parts of these documents. It

allows also competencies management via a competency formal modeling. The ER-ontology is implemented in OWL, by using the powerful inference services of the RACER reasoner, all acquirements/requirements related to a particular document, including competencies, can be deduced (inferred) and used by the two original algorithms of semantic matching proposed (superficial and competency based).

Future work aims to validate this approach on real data (a site of CVs and job offers) and enhance the competency model as well as the semantic matching process. We tend also to implement an interface for annotating documents easier to use then the Protégé-OWL interface that we use actually, as well as to generalize the ER-ontology to other domains. Furthermore, the different conceptual models used in the different countries, especially for describing diplomas and jobs, should be handled.

REFERENCES

[1] A. Abran, J. W. Moore, P. Bourque, R. Dupuis et L. L. Tripp, "A guide to the Software Engineering Body of Knowledge-SWEBOK", *IEEE Computer Society Professional project, 2004.* Available: http://www.swebok.org

[2] S. Bechhofer, F.-V. Harmelen, J. Hendler, I. Horrocks, "OWL Web Ontology Language Reference", 2004. Available: http://www.w3.org/TR/2004/REC-owl-ref-20040210/

[3] C. Bizer, R. Heese, M. Mocho, R. Oldakowski, R. Tolksdorf, R. Eckstein, "The Impact of Semantic Web Technologies on Job Recruitment Processes", in *In International Conference workshop on computer science (WI'05),* 2005.

[4] M. Bourse, M. Harzallah, M. Leclère, F. Trichet, "COMMONCV: modeling the competencies underlying a Curriculum Vita", IRIN research report N° 2, 2002.

[5] GIGREF, "Nomenclature 2005, les emplois-métiers du système d'information dans les grandes entreprises", 2005. Available: http://www.cigref.fr/cigref/livelink.exe/Nomenclature_RH_2005.pdf?func=doc.Fetch&nodeId=401472&docTitle=Nomenclature_RH_2005%2Epdf

[6] M. Fernandez, P.-A. Gomez, N. Juristo, "Methontology: from ontological art toward ontological engineering", *In Spring symposium series on ontological engineering AAAI97, USA*, 1997.

[7] V. Haarslev, R. Moller, M. Wessel, "RACER User's Guide and Reference Manual (Version 1.7.19)", 2004. Available: http://coli.lili.uni-bielefeld.de/~felix/lehre/ws04_05/ontologischeRessourcen/addLiterature/haarslev-undmoeller04.pdf

[8] P. Laublet, C. Reynaud, J. Charlet, "Sur quelques aspects du Web sémantique", 2002. Available: sis.univ-tln.fr/gdri3/fichiers/assises2002/papers/03-WebSemantique.pdf

[9] M. E. S, "Reforme LMD de l'enseignement supérieur", university of Constantine, department of computer science, 2004.

[10] ONTOlogen Group (DFKI 's Knowledge Management Department), Competency Ontology : a project for modelling competency using protègé-2000), 2002. Available: www.dfki.uni-kl.de/~elst/ONTOlogen/doc/ONTOlogen-148.htm

[11] Y. Prié, S. Garlatti, " Annotations et Méta-données dans le Web sémantique", *in Revue I3 Information – Interaction – Intelligence, numéro Hors-série Web Sémantique*, 24 pp, 2003.

[12] Protègè OWL : http//protégé/standford/edu/plugins/owl

[13] C. Rieu, C.-M. Rousseau, C. Roche, "Gestion des compétences: un modèle opérationnel à base d'ontologie", *(du E-Management à la E-RH) Colloque*, Paris-Dauphine Univ., Paris, France, 2005.

[14] F. Trichet, AnnotatingWithCigref : a project for annotation CVs, 2002. Available : www.sciences.univ-nantes.fr/irin/commoncv/ production.