Expertise Profiling in Evolving Knowledge-curation Platforms

Hasti Ziaimatin, Tudor Groza, Georgeta Bordea, Paul Buitelaar and Jane Hunter

***Abstract*—Expertise modeling has been the subject of extensive research in two main disciplines: Information Retrieval (IR) and Social Network Analysis (SNA). Both IR and SNA approaches build the expertise model through a document-centric approach providing a macro-perspective on the knowledge emerging from large corpus of static documents. With the emergence of the Web of Data there has been a significant shift from static to evolving documents, through micro-contributions. Thus, the existing macro-perspective is no longer sufficient to track the evolution of both knowledge and expertise. In this paper we present a comprehensive, domain-agnostic model for expertise profiling in the context of dynamic, living documents and evolving knowledge bases. We showcase its application in the biomedical domain and analyze its performance using two manually created datasets.**

***Index Terms*—Knowledge acquisition, Knowledge representation, Semantic Web, Text processing**

I. Introduction

ACQUIRING and managing expertise profiles represents a major challenge in any organization, as often, the successful completion of a task depends on finding the most appropriate individual to perform it. Furthermore, the use of expertise profiles to identify, acknowledge and recommend experts from within an online community, motivates additional participants to contribute to the community knowledge base. This collaborative input is vital to the capture and integration of diverse viewpoints and the efficient assembly of an extensive body of knowledge. In particular, many scientific research environments are increasingly dynamic and subject to rapid evolution of knowledge. Major scientific challenges such as global pandemics require teams of collaborators with expertise from a wide range of domains and disciplines. Better "expertise finders" would help identify the optimum set of researchers for a critical scientific challenge at any given time.

The topic of expertise modeling has been the subject of extensive research in two main disciplines: information retrieval (IR) and social network analysis (SNA). From the IR perspective, static documents authored by individuals (e.g. publications, reports) can be represented as bags-of-words (BOW) or as bags-of-concepts (BOC). The actual expertise identification is done by associating individual profiles to weighted BOWs or BOCs – either by ranking candidates based on their similarities to a given topic or by searching for co-occurrences of both the individual and the given topic, in the set of supporting documents. Such associations can then be used to compute semantic similarities between expertise profiles [1]. From the SNA perspective, expertise profiling is done by considering the graphs connecting individuals in different contexts, and inferring their expertise from the shared domain-specific topics [2]. Both IR and SNA techniques build the expertise model through a document-centric approach that provides only a macro-perspective on the knowledge emerging from the documents (due to their static, final nature, i.e. once written, the documents remain forever in the same form).

With the emergence of Web 2.0 [3] and then of the Semantic Web [4], there has been a significant shift from *static* documents to *evolving* documents. Wikis (starting with Wikipedia as a pioneering project) or collaborative knowledge bases, predominantly in the biomedical domain (e.g. AlzSWAN [5] or SKELETOME [6]) support this shift by enabling authors to *incrementally* and *collaboratively refine* the content of the embedded documents to reflect the latest advances in knowledge in the field. For example, AlzSWAN captures and manages hypotheses, arguments and counter-arguments in the Alzheimer's disease domain, while the Gene Wiki sub-project of Wikipedia supports discussions on genes.

This trend seems to emerge also in the scientific publishing process within some particular communities. Here, researchers try to shift from the traditional *document-centric* approach towards a *finer-grained contribution-oriented* approach in which hypotheses or domain-related innovations (in form of short statements) replace the publications. Examples include nano-publications [7] or liquid publications [8].

Regardless of the domain, the content of these living documents changes via *micro-contributions* made by individuals (e.g. *incremental updates* to Wikis or *contributions* in nano-publications), thus making the macro-perspective (provided by the document as a whole) no longer adequate for capturing the evolution of either the knowledge or the

Manuscript received August 28, 2012. This work was supported by the Australian Research Council (ARC) under the Linkage grant SKELETOME – LP100100156 and the Discovery Early Career Researcher Award (DECRA) – DE120100508.

H. Ziaimatin is with the School of ITEE, The University of Queensland, St. Lucia, QLD 4072, Australia (e-mail: h.ziaimatin@uq.edu.au).

T. Groza (corresponding author) is with the School of ITEE, The University of Queensland, St. Lucia, QLD 4072, Australia (phone: +61-7-3365-4539; e-mail: tudor.groza@uq.edu.au).

G. Bordea is with the Digital Enterprise Research Institute, The National University of Ireland, Galway, Ireland (georgeta.bordea@deri.org).

P. Buitelaar is with the Digital Enterprise Research Institute, The National University of Ireland, Galway, Ireland (paul.buitelaar@deri.org).

J. Hunter is with the School of ITEE, The University of Queensland, St. Lucia, QLD 4072, Australia (e-mail: jane@itee.uq.edu.au).

DOI: 10.5176/2010-3043_2.3.196

expertise.

Our goal is to advance the state of the art in expertise profiling by considering *living* documents; i.e. documents where *knowledge evolves* through *micro-contributions*. Most of the existing work has focused on the task of expert finding, i.e. given a set of documents and a set of expertise profiles, the task aims to find the best matches between the profiles and the documents ("*who's an expert in a particular topic?*"). Instead, we focus on *creating* expertise profiles in the context of evolving documents; i.e. given a series of micro-contributions, we aim to build an expertise profile for the author of those micro-contributions while taking into account the temporal aspect of contributions ("*what is the expertise of the person that has authored these contributions?*").

Our approach comprises two major elements: (i) a model, aimed at capturing micro-contributions in the macro-context of the host living documents, as well as the temporality of the expertise profiles; and (ii) a domain-agnostic methodology for extracting and building expertise profiles. In this paper, we showcase the methodology in the biomedical domain, mainly because of the existing tool support. However, the same methodology can be applied in other domains, by using a different set of tools, as we discuss later in the paper.

In order to analyze the efficiency of our methodology, and to understand its limitations, we have performed an evaluation on two datasets. However, the lack of a gold standard for expertise profiling, made it almost impossible to investigate the advantages that our methodology could bring to the task of expertise profiling in the context of micro-contributions, in comparison with traditional IR techniques. Nevertheless, we have performed similar experiments with two IR-based approaches, Saffron [9] and EARS[1], without conducting a direct comparison, due to the intrinsic differences among the approaches.

The remainder of the paper is structured as follows. In Section II, we provide a comprehensive overview of the related work. Section III describes the expertise-capturing model, while Section IV presents the generic methodology. Section V discusses its application in the biomedical domain and a series of experimental results, before concluding and outlining the future work in Section VI.

II. Related Work

Expertise profiling is an active research topic in a wide variety of applications and domains, including biomedical, scientific and education. In this section, we present a brief overview of the related efforts, with particular emphasis on the Information Retrieval and the Semantic Web domains. The two most popular and well performing approaches in the TREC[2] (Text Retrieval Conference) expert search task are profile-centric and document-centric approaches. These studies use the co-occurrence model and techniques such as Bag-of-Words or Bag-of-Concepts on documents that are typically large and rich in content. Often a weighted, multiple-sized, window-based approach in an information retrieval (IR) model is used for association discovery [10]. Alternatively the effectiveness of exploiting the dependencies between query terms for expert finding is demonstrated [11]. Other studies present solutions through effective use of ontologies and techniques such as *spreading* to link additional related terms to a user profile by referring to background knowledge [1].

Algorithms have been proposed to find experts in Wikipedia. One such study attempts to find experts in Wikipedia content or among Wikipedia users [12]. It uses semantics from Wordnet and Yago in order to disambiguate expertise topics and to improve the retrieval effectiveness. However, this study is unable to use the standards proposed for the evaluation of retrieval systems, as relevance assessments are required for representing the ground truth for a list of queries. Furthermore, none of the IR evaluation metrics can be used, since relevance judgments are not available on the Wikipedia collection or the list of queries to run. A relevant initiative to this task is the Web People Search task, which was organized as part of the SemEval-2007[3] evaluation exercise. This task consists of clustering a set of documents that mention an ambiguous person name according to the actual entities referred to using that name. However, the problem here is that the evaluated task is people name disambiguation and not expert finding. The Inex initiative [13], which provides an infrastructure for the evaluation of content-oriented retrieval of XML documents based on a set of topics, is also relevant but does not consider the expert finding task. To accomplish this task, the study aims to build a gold standard via manually and voluntary defined expertise profiles by Wikipedia users.

Such studies contribute to the task of expert finding and in the majority of cases, propose methods for finding experts, given a query or knowledge area in which experts are sought. Not only is expert finding a different task to expert profiling, but the methods applied in such studies rely on a large corpus of static documents (e.g. publications) and therefore are not suitable in the context of shorter text, such as micro-contributions in the context of living and evolving documents.

Another study, which introduces the task of *expert profiling,* also relies on queries for extracting expert profiles [14]; the first model uses traditional IR techniques to obtain a set of relevant documents for a given knowledge area (query) and aggregates the relevance of those documents that are associated with the given person. The second model represents both candidates and knowledge areas (queries) as a set of keywords and the skills of an individual are estimated based on the overlap between these sets.

The Entity and Association Retrieval System (EARS), is an open source toolkit for entity-oriented search and discovery in large test collections. EARS, implements a generative probabilistic modeling framework for capturing associations between entities and topics. Currently, EARS supports two main tasks: finding entities ("*which entities are associated with topic X?*") and profiling entities ("*what topics is an entity associated with?*"). EARS employs two main families of models, both based on generative language modeling techniques, for calculating the probability of a query topic (q) being associated with an entity (e), P(q|e). According to one

[1] http://code.google.com/p/ears/
[2] http://trec.nist.gov/
[3] http://nlp.cs.swarthmore.edu/semeval/

family of models (Model 1) it builds a textual representation (i.e. language model) for each entity, according to the documents associated with that entity. From this representation, it then estimates the probability of the query topic given the entity's language model. In the second group of models (Model 2), it first identifies important documents for a given topic, and then determines which entities are most closely associated with these documents. We have conducted experiments with EARS using our biomedical use cases and included the results in this paper; however, as mentioned above, this system also relies on a given set of queries. Furthermore, as with other studies that target expert finding, EARS relies on a large corpus of static publications, while we aim at building expert profiles from *micro-contributions*, without relying on any queries.

Finally, in the same category of expertise finding, we find *SubSift* (short for submission sifting), which is a family of RESTful Web services for profiling and matching text [15]. It was originally designed to match submitted conference or journal papers to potential peer reviewers, based on the similarity between the papers' abstracts and the reviewers' publications as found in online bibliographic databases. In this context, the software has already been used to support several major data mining conferences. *SubSift*, similar to the approaches discussed above, relies on significant amounts of data and uses traditional IR techniques such as TF-IDF, bag-of-words (BOW) and vector based modeling to profile and compare collections of documents.

The *ExpertFinder* framework uses and extends existing vocabularies that have attracted a considerable user community already such as FOAF, SIOC, SKOS and DublinCore [16]. Algorithms are also proposed for building expertise profiles using Wikipedia by searching for experts via the content of Wikipedia and its users, as well as techniques that use semantics for disambiguation and search extension [12]. We have leveraged these prior efforts to enable the integration of expertise profiles via a shared understanding based on widely adopted vocabularies and ontologies. This approach will also lead to a seamless aggregation of communities of experts.

WikiGenes combines a dynamic collaborative knowledge base for the life sciences with explicit authorship. Authorship tracking technology enables users to directly identify the source of every word. The rationale behind *WikiGenes* is to provide a platform for the scientific community to collect, communicate and evaluate knowledge about genes, chemicals, diseases and other biomedical concepts in a bottom-up approach. *WikiGenes* links every contribution to its author, as this link is essential to assess origin, authority and reliability of information. This is especially important in the Wiki model, with its dynamic content and large number of authors [17]. Although *WikiGenes* links every contribution to its author, it doesn't associate authors with profiles. More importantly, it doesn't perform semantic analysis on the content of contributions to extract expertise.

As more and more Web users participate in online discussions and micro-blogging, a number of studies have emerged, which focus on aspects such as content recommendation and discovery of users' topics of interest, especially in Twitter. Early results in discovering Twitter users' topics of interest are proposed by examining, disambiguating and categorizing entities mentioned in their tweets using a knowledge base. A topic profile is then developed, by discerning the categories that appear most frequently and that cover all of the entities [18].

The feasibility of linking individual tweets with news articles has also been analyzed for enriching and contextualizing the semantics of user activities on Twitter in order to generate valuable user profiles for the Social Web [19]. This analysis has revealed that the exploitation of tweet-news relations has significant impact on user modeling and allows for the construction of more meaningful representations of Twitter activities. As with other traditional IR methods, this study applies bags-of-words (BOW) and TF-IDF methods for establishing similarity between tweets and news articles and requires a large corpus. In addition, there are fundamental differences between micro-contributions in the context of evolving knowledge bases, contributions to forum discussions and Twitter messages; namely, online knowledge bases don't have to be tailored towards various characteristics of tweets such as presence of @, shortening of words, usage of slang, noisy postings, etc. Also, forum participations are a much richer medium for textual analysis as they are generally much longer than tweets and therefore provide a more meaningful context and usually conform better to the grammatical rules of written English. More importantly, twitter messages do not evolve, whilst we specifically aim to capture expertise in the context of evolving knowledge.

The *Saffron* system provides users with a personalized view of the most important expertise topics, researchers and publications, by combining structured data from various sources on the Web with information extracted from unstructured documents using Natural Language Processing techniques [9]. It uses the Semantic Web Dog Food (SWDF) [20] corpus to rank expertise and makes a distinction between the frequency of an expertise topic occurring in the context of a skill type and the overall occurrence of an expertise topic. *Saffron* also extends information about people by crawling Linked Open Data (LOD) [21] from seed URLs in SWDF. The semantics of the SWDF and crawled data represented using Semantic Web technologies is consolidated to build a holistic view represented via the social graph of an expert.

Existing social networks such as *BiomedExperts* (BME)[4] provide a source for inferring implicit relationships between concepts of the expertise profiles by analyzing relationships between researchers; i.e. co-authorship. BME is the world's first pre-populated scientific social network for life science researchers. It gathers data from *PubMed*[5] on authors' names and affiliations and uses that data to create publication and research profiles for each author. It builds conceptual profiles of text, called *Fingerprints*, from documents, Websites, emails and other digitized content and matches them with a comprehensive list of pre-defined *fingerprinted* concepts to make research results more relevant and efficient.

[4] http://www.biomedexperts.com/
[5] http://www.ncbi.nlm.nih.gov/pubmed/

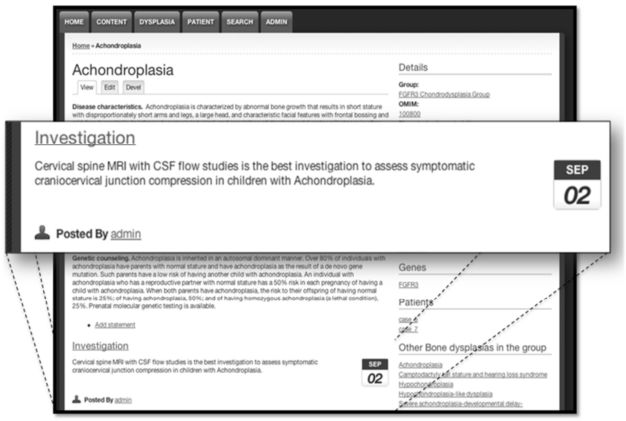

Fig. 1. Micro-contribution in the context of the SKELETOME platform.

III. An Ontology For Capturing Micro-contributions and Expertise Profiles

As mentioned in Sect. I, *micro-contributions,* represent incremental refinements by authors to an evolving body of knowledge. Examples of such micro-contributions are edits to a Wikipedia article or a Gene page in Gene Wiki, a statement in WikiGenes or OMIM[6], an argument in AlzSWAN or a statement in SKELETOME (Fig. 1). Regardless of the platform, we are interested in capturing the fine-grained provenance of these micro-contributions including the actions that lead to their creation, as well as the macro-context that hosts these contributions; i.e. paragraph or section of the document in which they appear. We have therefore created an ontology that combines coarse and fine-grained provenance modeling to capture such artefacts and their localization in the context of their host living documents.

Fig. 2 depicts the overall structure of our ontology. The objective has been to reuse and extend existing, established vocabularies from the Semantic Web that have attracted a considerable user community or are derived from de facto standards. This focus guarantees direct applicability and low entry barriers (compared to developing an entirely new ontology from scratch). We combine coarse and fine-grained provenance modeling using the SIOC ontology [22], with change management aspects captured by the SIOC-Actions module [23]. At the same time, we use the Annotation Ontology [24] to bridge the textual grounding and the ad-hoc domain knowledge, represented by concepts from domain-specific ontologies, and the Simple Knowledge Organization System (SKOS)[7] ontology to define the links to, and the relationships that occur between, these concepts.

As depicted in Fig. 2, our proposed ontology identifies four concepts and four relations illustrated with bold lines; it can be conceptually divided into two main parts: (i) a part modeling micro-contributions, and (ii) a part capturing expertise profiles. Both parts are discussed below.

The central concept of the first part is **Contribution** and we consider it to be a type of annotation (i.e. a subclass of **AO: Annotation**). We model the contributed text and its semantics at different conceptual levels. Therefore, a piece of text within a living document (modeled by **SIOC: Item**) is modified (*sioca: modifies*) by an action (e.g. add, delete, update) and can be clearly localized via pointer constructs – which in our case are represented by **AO: Selector** (s) on a **PAV: SourceDocument** (s). From a semantic perspective, the same action leads (*sioca: product*) to an annotation; i.e. the micro-contribution (**Contribution**) by the author to the living document. Hence, micro-contributions are in fact semantic annotations which define the body of knowledge within evolving documents. Domain specific aspects of these semantic annotations are represented by **SKOS: Concept** (s), connected to the annotation via *ao: hasTopic*. To get a better understanding of the modeling described above, we present the example depicted in Fig. 1 using the OWL Manchester syntax.

```
Individual: MicroContribution1
  Types: Contribution, ao:Annotation
  Facts:
    ao:context TextSelector1
    ao:hasTopic Concept1

Individual: Concept1
  Type: skos:Concept
  Facts:
    skos:prefLabel "Achondroplasia"
    skos:exactMatch radlex:Achondroplasia

Individual: TextSelector1
  Types:
    ao:Selector, aos:TextSelector,
    aos:OffsetRangeSelector
  Facts:
    aos:offset 0, aos:range: 356
    ao:onSourceDocument AchondroplasiaSource

Individual: AchondroplasiaSource
  Type: pav:SourceDocument
  Facts:
    pav:retrievedFrom AchondroplasiaPage
    pav:sourceAccessedOn "2012-08-01"

Individual: AchondroplasiaPage
  Type: sioc:Item
```

The second part of the ontology models expertise profiles as **SKOS: Collection** (s) of concepts. Although very lightweight, our proposed model introduces three novelties when compared to other expertise profiling approaches.

In order to capture the *temporal aspect of expertise*, we differentiate between **Short Term** and **Long Term** profiles. A **Short Term Profile** is a collection of concepts identified within a specific period of time (modeled via concepts introduced by the Time Ontology). A **Long Term Profile**, on the other hand, *aggregates* all the **Short Term Profile** (s) built for a particular expert. Intuitively, this enables a mechanism for tracking and analyzing the evolution of expertise. The actual method for creating these profiles is described in Section IV.

Expertise profiles are more than just collections / bags of concepts. Domain specific entities present in micro-contributions are captured in our model by **SKOS: Concept**

[6] http://omim.org/
[7] http://www.w3.org/TR/skos-reference

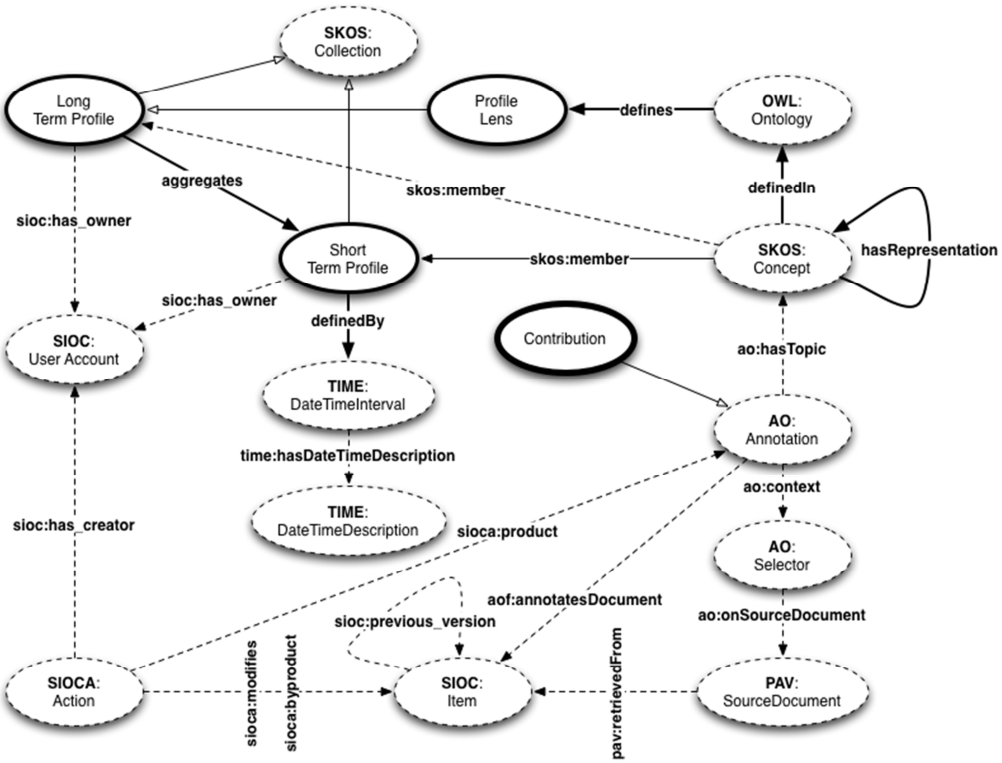

Fig. 2. An ontology for capturing micro-contributions and expertise profiles.

proxies[8]. By using the *hasRepresentation* relation between such proxies, we support the clustering of concepts in a manner similar to the semiotic triangle [25]. A particular entity, e.g. FGFR3, can be modeled as an abstract concept with multiple representations, each of which corresponds to a concept from a different ontology; e.g. Gene Ontology, Bone Dysplasia Ontology. This enables us to capture the semantics of micro-contributions by considering the best-suited concepts from one or multiple ontologies, while keeping track of the provenance of concepts (via *definedIn* **OWL: Ontology**). This will in turn result in creating a more accurate representation of expertise by avoiding duplication of the same concepts.

Maintaining the provenance of the domain specific concepts enables us to create multiple views over a **Long Term Profile** via lenses defined by particular ontologies. In our model, all **SKOS: Concept** (s) are *definedIn* an **OWL: Ontology**, which in turn may define (via the *defines* relation) a **Profile Lens** – a subclass of the **Long Term Profile**. This provides the opportunity to view a long-term profile from different ontological perspectives, each of which only considers concepts from a particular ontology. From an abstract perspective, since an ontology represents the conceptualization of a specific domain, profile lenses represent a domain-specific view over the expertise of an individual.

IV. Expertise Profiling

Our proposed methodology for creating expertise profiles is generic and can be applied to any domain, provided that appropriate tool support exists. Our goal is to provide a methodology for capturing micro-contributions and creating profiles, while ensuring that the methodology is not restricted to specific tools or frameworks within a domain. For a better understanding of the process, we exemplify the methodology in the context of the biomedical domain in Sect. V.

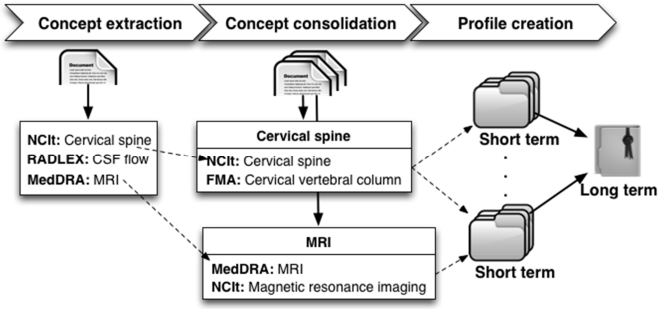

Fig. 3. Expertise profile creation methodology.

Our methodology consists of three main steps, as depicted and exemplified in Fig. 3; (i) Concept extraction; (ii) Concept consolidation; and (iii) Profile creation. We outline each step in the following sections.

A. Concept extraction

The concept extraction step aims to identify domain specific concepts in micro-contributions. From an ontological perspective, the goal is to populate the micro-contribution part of our ontology by creating appropriate annotations; i.e.

[8] This also enables the introduction and usage of concept-to-concept relationships at a later stage, e.g., *skos: broader*, *skos: narrower*, etc.

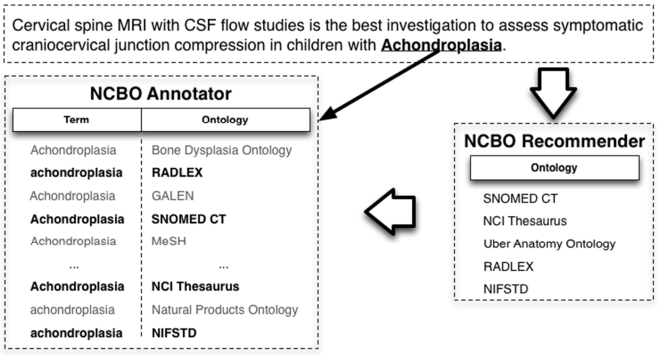

Fig. 4. Example of concept consolidation.

Contribution(s) that represent domain entities (**SKOS: Concept**(s)) captured within the text of the micro-contributions. Looking at the example presented in Fig. 1 – "*Cervical spine MRI with CSF flow studies is the best investigation to assess symptomatic craniocervical junction compression in children with Achondroplasia*" – the aim is to annotate those text chunks that represent domain concepts (e.g. *cervical spine*, *MRI*, *craniocervical junction compression* or *Achondroplasia*) and link them to an instance of a **Contribution,** representing the micro-contribution within which they have been identified.

This can be achieved by employing a typical information extraction or semantic annotation process, which is, in principle, domain dependent[9]. Hence, in order to provide a profile creation framework applicable to any domain, we don't restrict this step to the use of a particular concept extraction tool / technique.

B. Concept consolidation

Over the course of the last decade we have witnessed an increase in the adoption of ontologies as a domain conceptualization mechanism. While this has resulted in the formal conceptualization of a significant number of domains, it has also led to the creation of duplicated concepts; i.e. concepts defined in the context of multiple domains, and hence, ontologies. For example, the concept *Cervical spine* is now present in at least seven ontologies (cf. NCBO Bioportal[10]), while *MRI* is defined by at least 20 ontologies (cf. NCBO Bioportal). From a semiotic perspective, this can be seen as a symbol with multiple manifestations (or materializations), with each manifestation being appropriately defined by the underlying contextual domain.

Domain specific concepts captured within micro-contributions may also be defined in multiple ontologies. As a result, we have introduced the *concept consolidation* step that aims to cluster multiple representations of the same concept identified in one micro-contribution and across multiple micro-contributions. Fig. 3 depicts an example of consolidation output, where the concepts **NCIt: Cervical spine** and **MedDRA: MRI** which have resulted from concept extraction are consolidated under the abstract concepts *Cervical spine* and *MRI*, respectively, each of which has additional representations in **FMA: Cervical vertebral column** and **NCIt: Magnetic resonance imaging**.

As discussed in the previous section, our ontology is capable of capturing this semiotic perspective via the *hasRepresentation* relation between **SKOS: Concept**(s) and by keeping track of the provenance of concepts (*definedIn* **OWL: Ontology**). Below we present the example depicted in Fig. 4 using the Manchester syntax.

```
Individual: Concept1
  Type: skos:Concept
  Facts:
    skos:prefLabel "Achondroplasia"
    hasRepresentation C1, C2, C3

Individual: C1
  Type: skos:Concept
  Facts:
    skos:exactMatch radlex:achondroplasia
    definedIn http://radlex.org

Individual: C2
  Type: skos:Concept
  Facts:
    skos:exactMatch ncit:Achondroplasia
    definedIn http://nci-thesaurus.org

Individual: C3
  Type: skos:Concept
  Facts:
    skos:exactMatch snomed_ct:Achondroplasia
    definedIn http://snomed.org
```

Concept consolidation aggregates less prominent concepts with concepts that are manifestations of the same entities and appear more frequently; hence it provides a more accurate and coherent view over entities identified within micro-contributions. It is, however, an optional step and its realization usually depends on the concept extraction mechanism, in addition to an entity co-reference resolution technique.

As discussed in Sect. V, our experiments in the biomedical domain use the NCBO Annotator[11] for concept extraction and the results produced by the NCBO Recommender for concept consolidation.[12]. For example, if we consider the micro-contribution presented in Fig. 1, the NCBO Annotator annotates the concept **Achondroplasia** from 18 different ontologies; however, only the concepts that belong to the most suitable ontologies for annotating the micro-contribution, as recommended by the NCBO Recommender, are retained (Fig. 4). An abstract concept (**SKOS:Concept**) representing **Achondroplasia** is created, under which all retained concepts representing this entity from different ontologies are consolidated (through the *hasRepresentaton* relation).

C. Profile creation

The goal of this phase is to use the extracted and consolidated concepts to create two types of expertise profiles: (i) **Short Term Profile** (s); and (ii) a **Long Term Profile**. The expertise of an individual is dynamic and usually changes with time. Short-term profiles aim to capture periodic bursts of

[9] Generic IE / semantic annotation pipelines have been proposed, however, most research shows that there is always a trade-off between efficiency and domain independence.

[10] http://bioportal.bioontology.org/

[11] http://www.bioontology.org/annotator-service

[12] http://bioportal.bioontology.org/recommender
The use of the NCBO Recommender enables us to provide a more coherent view over the annotations provided by the NCBO Annotator.

expertise in specific topics, over a length of time. Long-term profiles, on the other hand, provide an overarching view of the expertise of an individual by taking into account all short term profiles (and hence all micro-contributions) of the expert. A long-term profile for an author consists of concepts that satisfy the *uniformity* and *persistency* criteria across all short term profiles for that author.

Short Term Profile creation. Using the provenance information captured by the ontology, we propose an approach for computing short-term profiles. Before discussing the actual computation, we need to re-iterate the concept consolidation phase and explain its role in building profiles.

As mentioned in the previous section, the consolidation step clusters domain specific entities that are manifestations of the same abstract concept. This is realized via the *hasRepresentation* relation between **SKOS: Concept**(s), as illustrated in the example presented in Section IV.B. We refer to a cluster representing an abstract concept as a v*irtual concept*. Virtual concepts represent an abstract entity and contain domain specific concepts from different ontologies, which are manifestations of the abstract entity. Virtual concepts are central to both short term and long term profile creation methods. The consolidation step is optional, and hence, instead of such *virtual concepts*, one may opt to directly process the results of the concept extraction phase. In this case, the virtual concept notation used in the profile creation formulae, should be replaced with a notation representing a domain specific concept.

A short-term profile represents a collection of concepts extracted from micro-contributions over a period of time. In order to compute a short term profile, we propose a ranking of all concepts identified within that time span based on an individual weight that takes into account the normalized frequency and the degree of co-occurrence of a concept with other concepts identified within the same period. The equation below lists the mathematical formulation of this weight. The intuition behind this ranking is that the expertise of an individual is more accurately represented by a set of co-occurring concepts forming an expertise context, rather than by individual concepts that occur frequently outside such a context.

$$W(V_c) = \frac{Freq(V_c)}{N_V} * \sum_{i=1}^{N_v - 1} PPMI(V_c, V_{ci})$$

The elements of the equation above are: V_c – the virtual concept for which a weight is calculated, N_v – total number of virtual concepts in the considered time window, and PPMI – the positive pointwise mutual information [26], as defined below:

$$PPMI(C_1, C_2) = log \frac{p(C_1, C_2)}{p(C_1) * p(C_2)} = log \frac{N_c * Freq(C_1, C_2)}{Freq(C_1) * Freq(C_2)}$$

N_c – the total number of concepts and Freq(C_1, C_2) – the joint frequency (or co-occurrence) of C_1 and C_2. PPMI is always positive, i.e. if PPMI(C_1, C_2) < 0 then PPMI(C_1, C_2) = 0.

Long Term Profile creation. The goal of the Long Term Profile is to capture the collection of concepts occurring both persistently and uniformly across all Short Term Profiles for an expert. Unlike other expertise profiling approaches, we consider uniformity as important as persistency; i.e. an individual is considered to be an expert in a topic if this topic is present persistently and its presence is distributed uniformly across all short term profiles for that expert. Consequently, in computing the ranking of the concepts in the Long Term Profile, the weight has two components, as listed in the equation below:

$$W(V_c) = \propto * (e^{-\Delta(V_c)} - \frac{\Delta(V_c)}{e}) + (1 - \propto) * \frac{Freq(V_c, S)}{N_s}$$

where N_s is the total number of Short Term Profiles, Freq(V_c, S) is the number of Short Term Profiles containing V_c, α is a tuning constant and $\Delta(V_c)$ is the standard deviation of V_c, computed using the equation below. The standard deviation of V_c shows the extent to which the appearance of the virtual concept in the Short Term Profiles deviates from a uniform distribution. A standard deviation of 0 represents a perfectly distributed appearance. Consequently, we've introduced a decreasing exponential that increases the value of the uniformity factor inversely proportional to the decrease of the standard deviation – i.e. the lower the standard deviation, the higher the uniformity factor.

$$\Delta(V_c) = \sqrt{\sigma(V_c)^2};\ \sigma(V_c)^2 = \frac{1}{N_s} * \sum_{i=1}^{N_s} [(ST_i - ST_{i-1}) - M_{ST}(V_c)]^2$$

$$M_{ST}(V_c) = \frac{1}{N_s} * \sum_{i=1}^{N_s} (ST_i - ST_{i-1})$$

Where (ST_i – ST_{i-1}) represents the window difference between Short Term Profiles in which a virtual concept appears, and $M_{ST}(V_c)$ is the mean of all window differences. In practice, we aim to detect uniformity by performing a linear regression over the differences between the short-term profiles that contain the virtual concept.

V. Experiences with Modeling Expertise Profiles in the Biomedical Domain

In order to exemplify the application of the methodology described in this paper and to get a better understanding of its strengths and limitations, we applied it to the biomedical domain. More specifically, we performed an experiment using data from the Molecular and Cellular Biology Wiki[13] (MCB) and the Genetics[14] Wiki projects (both sub-projects of Wikipedia), and a series of tools provided by NCBO[15]. In the following sections, we detail the characteristics of the datasets, the tools used for concept extraction and consolidation, the experimental results and the lessons learned.

[13] http://en.wikipedia.org/wiki/Wikipedia:MCB
[14] http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Genetics
[15] http://www.bioontology.org

It is important to note that performing a full-fledged comparative evaluation of our approach has not been possible, due to the lack of a gold standard. The experiments we have performed used manually created expertise profiles as baseline, which present a series of challenges, as discussed later in this section.

A. Datasets

The Molecular and Cellular Biology Wiki project aims to organize information in articles related to molecular and cell biology in Wikipedia. Similarly, the Genetics Wiki project aims to organize improvement and maintenance of genetics articles in Wikipedia. The underlying articles in both projects are constantly updated through expert contributions. Wikipedia allows authors to state opinions and raise issues in the discussion pages. These incremental additions to content, or micro-contributions, give the knowledge captured within the environment a dynamic character.

We have collected micro-contributions for 22 authors from the MCB project and 7 authors from the Genetics project over the course of the last 5 years. These contributions resulted in a total of ~4,000 updates, with an average of 270 tokens per micro-contribution and an average of 137 micro-contributions per author.

The 29 authors selected for the datasets were the only ones that had provided a personal perspective on their expertise when joining the corresponding project. Although a much larger number of participants are available, not all of them provide a sufficiently detailed description of their expertise. We were interested in expertise profiles that mention areas of expertise, rather than the position of the participant (e.g. "post doc" or "graduate student") or their interest in this project (e.g. "improving Wikipedia entries", "expanding stub articles"). Each of the 29 authors selected from all the participants provided an average of 4.5 expertise topics in their profiles. We used these topics to form corresponding long-term profiles for each author, which we have used as our baseline. An example of such a profile is the one for author "AaronM" that specifies: "cytoskeleton", "cilia", "flagella" and "motor proteins" as his expertise.

B. Tool support

As discussed in Sect. IV, our methodology may be implemented using domain-specific tools, which enable an accurate extraction of the concepts captured within micro-contributions. Since the datasets we had available were from the biomedical domain, we have chosen the NCBO Annotator [27] to perform the concept extraction phase and use the results produced by the NCBO Recommender [28] to perform concept consolidation.

The NCBO Annotator workflow is composed of two main steps; first the biomedical free text is given as input to the concept recognition tool used by the annotator along with a dictionary. The dictionary (or lexicon) is constructed using ontologies configured for use by the Annotator. As most concept recognizers take as input a resource and a dictionary to produce annotations, the only customization to the biomedical domain, would be the biomedical ontologies used by the Annotator. In other words, by using the Annotator, we aren't taking advantage of any specific functionality or feature that would otherwise be unavailable if other annotators or techniques were to be used in the context of fields other than the biomedical domain.

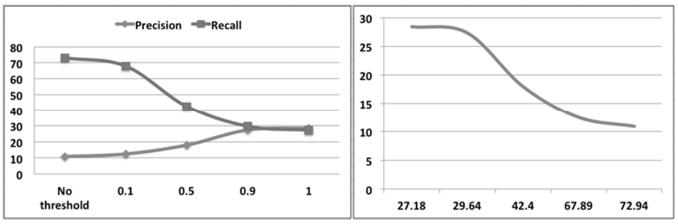

Fig. 5. Expertise creation efficiency. [A] Precision and recall subject to a weight threshold; [B] Precision-recall curve for different weight thresholds.

We would also like to emphasize that the Annotator can be configured to produce direct or semantically expanded annotations. In the latter case, the direct annotation is described along with the concept from which the annotation is derived; i.e. using the *is-a* relationships between concepts; however, we have specifically configured the annotator to perform direct annotations; i.e. annotations are performed directly on the underlying terms and *not* generalized to parent concepts. This configuration emulates entity recognition in traditional IR techniques, and thus removes any bias when comparing the performance of our methodology against such methods.

The NCBO Recommender identifies and ranks the most suitable ontologies for annotating a textual entry. As previously mentioned, the Annotator already helps with the concept consolidation, as it provides multiple concept candidates for the same text chunk. However, we've decided to introduce an additional consolidation step, via the Recommender, to create a more coherent view over the domain specific concepts derived from micro-contributions.

C. Experimental results

The main goal of the experiments discussed in this section has been to test the efficiency of the long-term profile generation. Using micro-contributions from the Molecular and Cellular Biology and Genetics Wiki projects, we've created long-term profiles for all 29 authors as part of our benchmark. As previously mentioned, the baseline consisted of the expertise profiles these authors have created when they joined the corresponding projects. In terms of efficiency measures, we have considered precision and recall as defined in the context of information retrieval.

Fig. 5 depicts the results achieved by our methodology. Part A tracks the values of Precision and Recall for different concept weight thresholds (see Sect. IV.C for the long-term profile creation), while part B provides a different perspective over the same results, by showing the evolution of precision for different recall cut-off points. From part A, we can observe that if we don't set any threshold on the weight of the concepts in the long term profiles, the achieved precision is 10.86% for a recall of 72.94%. Setting and subsequently increasing the threshold has positive effects on the precision, increasing from 12.44% at a 0.1 threshold to 28.47% at a 1.0 threshold, at the expense of the recall, which decreases from 67.89% to 27.18%.

In order to provide a more comprehensive interpretation of these results, we have performed the same experiment using Saffron and EARS, two systems that employ IR-based techniques. It is important to note that the results are not directly comparable because of two reasons: (i) the evaluation of Saffron is based on a dichotomous model, i.e. the terms resulting from the profile creation do not have weights attached; hence, when comparing them to the baseline, they are either present or not; (ii) the goal and workflow of the EARS system are different to those of Saffron and our methodology; in the context of our experiment, EARS requires as input both the micro-contributions dataset as well as the expected expertise profiles (profiles defined by the authors), the result being a ranked association of individual to expertise; hence, by default the recall will be high, as the evaluation of the expertise is performed on a closed, previously-known set of concepts. Nevertheless, from a technical perspective, it is interesting to analyze the challenges posed by using a different kind of dataset, on the performance of these systems (since most IR-based approaches rely on large corpus of data).

Table 1. Efficiency results of the Saffron and EARS systems

Saffron		EARS	
Prec.	**Recall**	**Prec.**	**Recall**
7.54%	9.63%	7.42%	83.43%

Table 1 summarizes the Precision and Recall values achieved by the two systems. It can be observed that Precision is fairly similar for all three approaches (including ours when no threshold is set), however Recall varies considerably. As already mentioned, in the case of EARS, a high Recall value was expected due to the experimental setup.

The above listed results shed a positive light onto the performance of our system. By setting an appropriate threshold, i.e. 0.5 for concept weight, our system is able to deliver a significantly improved precision (almost 20%), although at the price of a lower recall (around 40%). While these results can be further improved, they are encouraging as they illustrate that expertise profiling using micro-contributions in the context of evolving knowledge is significantly enhanced by implementing our proposed methodology, which combines concept consolidation and long-term profile generation based on uniformity and persistency.

D. Discussion

The experimental results presented in the previous section have been influenced by a series of factors. Firstly, choosing an appropriate set of tools for the concept extraction and concept consolidation phases is crucial. As already mentioned, we believe that these tools should be domain-specific, in order to achieve reasonable results. We opted for using the NCBO Annotator and Recommender; as a result, these tools had a massive influence on the final results. While the Annotator is used predominantly in the biomedical domain; i.e. the domain chosen for our experiments, its underlying technology is, in fact, domain agnostic, as the only customization to the domain is the biomedical ontologies configured for constructing the dictionary used by the annotator's concept recognizer. Its semantic annotation capability has been particularly beneficial for our approach, since it also supports the consolidation phase. However, its versatility comes at the price of extraction efficiency, as an exact match is required between the terms present in text and the labels of the ontological concepts, in order for annotations to be detected. For example, a simple usage of the plural of a noun (e.g. cilia) is enough to miss an ontological concept (such as Cilium); an issue that is usually resolved in most IR approaches by the use of lemmatization. We have also tried to alleviate this problem through concept consolidation by detecting the intersection of groups of concepts resulting from annotation of different, but semantically similar entities across micro-contributions and using their union to create virtual concepts. Although, this method has resulted in a significant improvement of the results produced by the annotator, we have observed few instances where such concepts haven't been integrated.

Secondly, the difference in abstraction between the content of the micro-contributions and the expertise profiles provided by the authors plays a crucial role in evaluation. Micro-contributions are generally very specific; i.e. the terminology describes specific domain aspects, while expertise profiles defined by experts and used as our baseline, consist of mostly general terms (e.g. genetics, bioinformatics, microbiology, etc). This makes direct comparison very challenging. The use of ontologies enables us to take into account more than just the actual concepts extracted from micro-contributions, by looking at their ontological parents or children. Consequently, we would be able to realize a comparison at a similar abstraction level, which could improve the evaluation results.

Finally, the weight assigned to each concept in the long term profile consists of both the uniformity and persistency of the concept across all short term profiles for an expert, in comparison to all other expertise profiling approaches that consider only a persistency factor. Hence, we provide the flexibility of computing expertise profiles that focus on uniformly behaving concepts or on concepts that are uniformly present throughout time.

VI. CONCLUSION

In this paper, we have presented an approach for modeling and creating expertise profiles from micro-contributions emerging from living documents. We proposed a domain-agnostic methodology for creating short-term and long-term profiles, while capturing the temporality in expertise. Our proposed ontology captures and stores micro-contributions, short term and long term profiles. Future work will focus on improving the concept consolidation phase and using the short term profiles to analyze the temporal aspects of expertise. In addition, we intend to facilitate domain-specific views over the expertise of an individual through implementing ontological lenses over long-term profiles.

REFERENCES

[1] R. Thiagarajan, G. Manjunath and M. Stumptner. "Finding experts by semantic matching of user profiles", Technical Report HPL-2008-172, HP Laboratories, 2008

[2] J. Zhang, J. Tang, J. Li, "Expert finding in a social network", Advances in Databases: Concepts, Systems and Applications, 2007, pp. 1066-1069.
[3] T. O'Reilly and J. Musser, "Web 2.0: Principles and best practices", O'Reilly Media, 2006.
[4] T. Berners-Lee, J. Hendler and O. Lassila, "The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities", Scientific American, 284(5), 2001, pp. 34-43.
[5] T. Clark and J. Kinoshita, "AlzForum and SWAN: The present and future of scientific Web communities", Briefings in Bioinformatics, 8(3), 2007, pp. 163-171.
[6] T. Groza, A. Zankl, Y-F. Li and J. Hunter, "Using Semantic Web Technologies to Build a Community-Driven Knowledge Curation Platform for the Skeletal Dysplasia Domain", In Proc. of the 10th International Semantic Web Conference, 2011, pp. 81-96.
[7] B. Mons and J. Velterop, "Nano-Publication in the e-science era", In Proc. of the Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009), Washington DC, USA, 2009.
[8] F. Casati, F. Giunchiglia and M. Marchese, "Liquid publications, Scientific Publications Meet the Web", Technical Rep. DIT-07-073, Informatica e Telecomunicazioni, University of Trento, 2007.
[9] F. Monaghan, G. Bordea, K. Samp and P. Buitelaar, "Exploring Your Research: Sprinkling some Saffron on Semantic Web Dog Food", In Proc. of the Semantic Web Challenge at the International Semantic Web Conference, Shanghai, China, 2010.
[10] J. Zhu, D. Song and S. Rueger, "Integrating multiple windows and document features for expert finding", Journal of the American Society for Information Science and Technology, 60(4), 2009, pp. 694-715.
[11] L. Yang and W. Zhang, "A study of the dependencies in expert finding", In Proc. of the 2010 Third International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 2010.
[12] G. Demartini. "Finding experts using Wikipedia". In Proc. of the ExpertFinder Workshop, co-located with ISWC 2007, Busan, Korea, 2007.
[13] N. Fuhr, N. Govert, G. Kazai and M. Lalmas, "INEX: INitiative for the Evaluation of XML retrieval". In Proc, of the SIGIR 2002 Workshop on XML and Information Retrieval, 2002.
[14] K. Balog and M. de Rijke, "Determining expert profiles (with an application to expert finding)", In Proc. of the 20th International Joint Conference on Artificial Intelligence, 2007, pp. 2657–2662.
[15] S. Price, S, P. A. Flach, S. Spiegler, C. Bailey and N. Rogers, "SubSift Web Services and workflows for profiling and comparing scientists and their published works", In Proc of the 2010 IEEE 6th International Conference on eScience, 2010.
[16] B. Aleman-Meza, U. Bojars, H. Boley, J. Breslin, M. Mochol, L. Nixon, A. Polleres and A. Zhdanova, "Combining RDF vocabularies for expert finding", In Proc. of the 4th European Semantic Web Conference, Innsbruck, Austria, 2007, pp. 235-250.
[17] R. Hoffmann, "A wiki for the life sciences where authorship matters", Nature genetics, 40(9), 2008, pp. 1047-1051.
[18] M. Michelson and S. Macskassy, "Discovering users' topics of interest on twitter: a first look", In Proc. of the 4th Workshop on Analytics for Noisy Unstructured, co-located with the 19th ACM CIKM Conference, 2010, pp. 73-80.
[19] F. Abel. Q. Gao, G. Houben and K. Tao, "Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web". In Proc. of the 8th Extended Semantic Web Conference, 2011, pp. 375-389.
[20] K. Moeller, T. Heath, S. Handschuh and J. Domingue, "Recipes for Semantic Web Dog Food – The ESWC and ISWC Metadata Projects", In Proc. of the 6th International Semantic Web Conference, 2007, pp. 802-815.
[21] C. Bizer, T. Heath and T. Berners-Lee, "Linked data – The story so far", International Journal on Semantic Web Inf. Syst, 5(3), 2009, pp. 1-22.
[22] J. Breslin, S. Decker, A. Harth and U. Bojars, "SIOC: An approach to connect web-based communities", The International Journal of Web-based Communities, 2(2), 2006, pp. 133-142.
[23] P.A. Champin and A. Passant, "SIOC in action representing the dynamics of online communities", In Proc. of the 6th International Conference on Semantic Systems, 2010, pp. 1-7.
[24] P. Ciccarese, M. Ocana, L. Castro, S. Das and T. Clark, "An open annotation ontology for science on Web 3.0". Journal of Biomedical Semantics, 2(Suppl 2), 2010, S4.
[25] C. Ogden and I.A. Richards, "The Meaning of Meaning: A study of the influence of language upon thought and of the science of symbolism, Magdalene College, University of Cambridge, 1923
[26] Y. Niwa and Y. Nitta, "Co-occurrence vectors from corpora vs. distance vectors from dictionaries", In Proc. of the 15th International Conference of the Association of Computational Linguistics, 1994, pp. 304-309.
[27] C. Jonquet, N. Shah and M. Musen, "The Open Biomedical Annotator", In Proc. of the Summit of Translational Bioinformatics, 2009, pp. 56-60.
[28] C. Jonquet, M. Musen and N. Shah, "Building a biomedical ontology recommender web service", Journal of Biomedical Semantics 1(Suppl 1), 2010, S1.

Hasti Ziaimatin is currently studying for a PhD in Computer Science at The University of Queensland, Australia. Her PhD thesis focuses on modeling and capturing expertise from fine-grained micro-contributions made to evolving knowledge. Hasti is interested in acquiring the provenance of knowledge contributed by the community of experts and modeling the expertise of the contributors by combining ontologies, Linked Data, information retrieval and collaboration aspects. Hasti received her Bachelor of Information Systems from The University of Auckland, New Zealand and has over ten years of extensive work experience as a Senior Software Engineer, specialising in Java, web application development and J2EE technologies, analysis, design, development and implementation of complex software systems.

Tudor Groza is a Postdoctoral Research Fellow in the e-Research group of the School of ITEE, at The University of Queensland. Here is works on building community-driven knowledge curation platforms using domain specific knowledge or knowledge emerging from argumentative discourse networks. Tudor has seven years experience in the R&D of Semantic Web and Linked Data technologies, with a focus on Healthcare and Life Sciences, scientific authoring and publishing and personal information management. He has published more than 25 papers on the above-mentioned topics and serves constantly as reviewer for the main Semantic Web journals and for conferences or workshops in the Semantic Web and knowledge capturing and acquisition areas.

Georgeta Bordea is currently preparing her PhD thesis at the Digital Enterprise Research Institute (DERI), National University of Ireland, Galway, in the Unit for Natural Language Processing (UNLP) under the supervision of Paul Buitelaar. She obtained a Master degree at the Polytechnic University of Bucharest, Romania, under the supervision of Tudor Groza, Siegfried Handschuh and Stefan Trausan-Matu, with a thesis concerning the extraction of claims from scientific publications. Her research interests are in ontology-based information extraction from text, ontology learning and population, domain adaptation of NLP applications and automatic terminology extraction.

Paul Buitelaar is a senior research fellow and head of the Unit for Natural Language Processing of DERI, a leading research institute in semantic technologies at the National University of Ireland, Galway. His main research interests are in language technology for semantic-based information access. He has been a researcher and/or project leader on a number of national and international funded projects, on concept-based and cross-lingual information retrieval, semantic navigation, ontology-based information extraction and ontology learning, semantic-based multimedia analysis. Currently, his work has a focus on linguistic analysis and NLP-based applications for expertise mining, i.e. the extraction of expertise topics from text collections, such as scientific literature, for identifying experts and emerging technological and scientific communities. Another line of current research is on 'lexicalized ontologies', i.e. on the linguistic grounding of knowledge representation in ontologies with multilingual terms and their linguistic (lexical) information.

Jane Hunter is Professor of eResearch within the School of ITEE at the University of Queensland. Her area of expertise is the application of Semantic Web technologies to the integration, analysis, modelling and visualization of scientific and humanities datasets. She has published over 100 peer-reviewed papers and is a member of the Scientific Committee of the ICSU World Data System.