THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Tracking expertise profiles in community-driven and evolving knowledge curation platforms

Hasti Ziaimatin

BCom (Information Systems)

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2014

School of Information Technology and Electrical Engineering

Abstract

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. User profiling has been extensively utilised as a basis for recommendation, personalisation and matchmaking systems. Accurate user profile generators can improve interaction and collaboration between researchers working in similar domains but in different locations or organizations. They can also assist with identifying the optimum set of researchers with complementary skills for cross-disciplinary research teams at a given time. The topic of expertise modelling has been the subject of extensive research in two main disciplines: Information Retrieval (IR) and Social Network Analysis (SNA). Traditional IR and SNA expertise profiling techniques rely on large corpora of static documents authored by an expert, such as publications, reports or grants, the content of which remains unchanged due to the static and final nature of such resources. Consequently, such techniques build the expertise model through a document-centric approach that provides only a macro-perspective of the knowledge emerging from such documents.

With the emergence of Web 2.0, there has been a significant increase in online collaboration, giving rise to vast amounts of accessible and searchable knowledge in platforms where content evolves through individuals' contributions. This increase in participation provides vast sources of information, from which knowledge and intelligence can be derived for modelling the expertise of contributors. However, with the proliferation of collaboration platforms, there has been a significant shift from static to evolving documents. Wikis or collaborative knowledge bases, predominantly in the biomedical domain, 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. Regardless of the domain, the content of these living documents changes via *micro-contributions* made by individuals, thus making the macro-perspective, provided by the document as a whole, no longer adequate for capturing the evolution of knowledge or expertise. Hence, expertise profiling is presented with major challenges in the context of dynamic and evolving knowledge. Thus, the shift from static documents to living documents requires a shift in the way in which expertise profiling is performed.

This thesis examines methods for advancing the state of the art in expertise modelling by considering *dynamic* content; i.e., platforms in which, knowledge *evolves* through *micro-contributions*. Towards this goal, a novel expertise profiling framework is introduced that provides solutions for expertise modelling in the context of platforms where knowledge is subject to continuous *evolution* through experts' micro-contributions; i.e., given a series of micro-

contributions, the aim is to build an expertise profile for the author of those micro-contributions. Furthermore, as the expertise of an individual is dynamic and usually changes with time, the proposed framework aims at capturing the *temporality* of expertise, in order to facilitate *tracking* and analysis of changes in interests and expertise over time.

The proposed framework comprises three major elements: (i) a model, aimed at capturing the fine-grained provenance of micro-contributions and evolving content in the macro-context of the host living documents, as well as the temporality of micro-contributions; (ii) a domain-independent methodology for building expertise profiles by capturing expertise topics in micro-contributions and consolidating them to weighted concepts from domain ontologies, and (iii) a profile refinement mechanism for complementing expertise profiles by integrating contextual factors in existing social expert networks.

Furthermore, the proposed expertise profiling framework creates profiles containing ontological concepts, each of which represents an area of expertise. This provides the flexibility of using the structure of domain ontologies to represent the expertise topics embedded in the micro-contributions of an expert, at different levels of granularity. In addition, using ontological concepts to represent expertise topics facilitates the use of semantic similarity for comparing profiles that describe expertise at different levels of abstraction. This in turn facilitates the semantic evaluation of expertise profiles, rather than evaluation based on the exact matching of concepts or terms. Moreover, using the structure of ontologies allows experts to customise the granularity of their profiles in order to complement their existing profiles with fine-grained domain concepts representing knowledge embedded in their micro-contributions to evolving knowledge-curation platforms.

Finally, this thesis presents the Profile Explorer visualization tool, which serves as a paradigm for exploring and analysing *time-aware* expertise profiles in knowledge bases where content evolves over time. Profile Explorer facilitates browsing, search and comparative analysis of evolving expertise, independent of the domain and the methodology used in creating profiles.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with The University Library and, subject to the General Award Rules of The University of Queensland, immediately made available for research and study in accordance with the *Copyright Act 1968*.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Publications during candidature

- Ziaimatin, H., Groza, T., & Hunter, J. (2011). Expertise Modelling in Community-driven Knowledge Curation Platforms. *ADVANCES IN ONTOLOGIES*.

- Zankl, A., Groza, T., Li, Y. F., Ziaimatin, H., Paul, R., & Hunter, J. (2011). The SKELETOME Project: Towards a community-driven knowledge curation platform for Skeletal Dysplasias. In *10th Biennal Meeting of the International Skeletal Dysplasia Society*.

- Ziaimatin, H., Groza, T., Bordea, G., Buitelaar, P., & Hunter, J. (2012). Expertise profiling in evolving knowledge-curation platforms. *Global Science and Technology Forum Journal on Computing*, 2(3), pp. 118-127.

- Ziaimatin, H., Groza, T., & Hunter, J. (2013). Semantic and Time-Dependent Expertise Profiling Models in Community-Driven Knowledge Curation Platforms. *Future Internet*, 5(4), pp. 490-514.

- Ziaimatin, H., Groza, T., Tudorache, T. & Hunter, J. (2014). Modelling expertise at different levels of granularity using semantic similarity measures in the context of collaborative knowledge-curation platforms. Manuscript submitted for publication.

- Ziaimatin, H., Groza, T. & Hunter, J. (2014). Building expertise profiles from micro-contributions and social collaboration factors. Manuscript submitted for publication.

Publications included in this thesis

- Ziaimatin, H., Groza, T., Bordea, G., Buitelaar, P., & Hunter, J. (2012). Expertise profiling in evolving knowledge-curation platforms. *Global Science and Technology Forum Journal on Computing*, 2(3), pp. 118-127.

  This publication is mainly incorporated as Chapter 3 and partially as Chapter 4. The statement of contribution is listed in the following table:

Contributor	Statement of contribution
Hasti Ziaimatin (Candidate)	Designed experiments (80%) Wrote the paper (70%)
Dr. Tudor Groza	Designed experiments (20%) Wrote the paper (20%)
Prof. Jane Hunter	Wrote and edited paper (10%)

- Ziaimatin, H., Groza, T., & Hunter, J. (2013). Semantic and Time-Dependent Expertise Profiling Models in Community-Driven Knowledge Curation Platforms. *Future Internet*, 5(4), pp. 490-514.

  This publication is mainly incorporated as Chapter 4 and partially as Chapter 8. The statement of contribution is listed in the following table:

Contributor	Statement of contribution
Hasti Ziaimatin (Candidate)	Designed experiments (80%) Wrote the paper (80%)
Dr. Tudor Groza	Designed experiments (20%) Wrote the paper (10%)
Prof. Jane Hunter	Wrote and edited paper (10%)

- Ziaimatin, H., Groza, T., Tudorache, T. & Hunter, J. (2014). Modelling expertise at different levels of granularity using semantic similarity measures in the context of collaborative knowledge-curation platforms. Manuscript submitted for publication.

  This manuscript is mainly incorporated as Chapter 6. The statement of contribution is listed in the following table:

Contributor	Statement of contribution
Hasti Ziaimatin (Candidate)	Designed experiments (80%) Wrote the paper (60%)
Dr. Tudor Groza	Designed experiments (20%) Wrote the paper (30%)
Prof. Jane Hunter	Wrote and edited paper (10%)

- Ziaimatin, H., Groza, T. & Hunter, J. (2014). Building expertise profiles from micro-contributions and social collaboration factors. Manuscript submitted for publication.

  This manuscript is mainly incorporated as Chapter 7. The statement of contribution is listed in the following table:

Contributor	Statement of contribution
Hasti Ziaimatin (Candidate)	Designed experiments (80%) Wrote the paper (60%)
Dr. Tudor Groza	Designed experiments (20%) Wrote the paper (30%)
Prof. Jane Hunter	Wrote and edited paper (10%)

Contributions by others to the thesis

Prof. Jane Hunter and Dr. Tudor Groza, played an advisory role to the author of this thesis. They provided guidance, constructive criticisms and helped generate ideas throughout the work presented in this thesis.

Statement of parts of the thesis submitted to qualify for the award of another degree

None.

Acknowledgements

This doctoral dissertation was accomplished with the enormous support of several great people. I would like to express my warmest appreciation to those who have been an essential part of my achievement. First and foremost, I thank the Almighty God for the numerous blessings He has bestowed upon me throughout my dissertation journey and for providing me with strength and resources to complete my thesis, despite the difficult times I faced in the past couple of years.

I cannot begin to express my unfailing gratitude and love to my husband, Mohammad Ali, for providing me with continuous support and encouragement throughout my years of study. I am truly blessed and thankful for having you in my life.

My utmost gratitude goes to Prof. Jane Hunter, my principal supervisor, for granting me the opportunity to pursue a PhD in the excellent environment provided by the eResearch Lab at the University of Queensland. I would also like to thank her for her high-quality supervision, scholarly guidance, motivation and constructive criticism throughout the PhD program.

I owe my sincere gratitude to my co-supervisor, Dr. Tudor Groza, for his patience, enthusiastic support and guidance through every step of the PhD program, and for all he has taught me. His expertise and patience has been remarkable and added considerably to my graduate experience.

I would like to thank all of the staff in eResearch Lab, especially, Mrs. Carol Owen, for her assistance, encouragements and most of all, for being a lovely friend and companion in difficult times; Dr. Nigel Ward, who has been the chair of my thesis committee, and has helped to coordinate several milestones and offered me constructive feedback to progress my thesis. I would also like to thank my classmates, Hamed Hassanzadeh, Suleiman Odat, Juana Gao, David Yu and Razan Paul for all of their support and friendship throughout my PhD journey.

My deepest gratitude goes to my loving father and mother, Sam and Soudabeh, for their unconditional support, encouragements and their tremendous sacrifices to ensure that I had an excellent education. I would also like to acknowledge the unconditional love, support, guidance and tremendous sacrifices of my grandparents, Reza and Irandokht Bassiri.

I am blessed with a number of wonderful friends and family who have been there for me, through thick and thin, who have listened, counselled, commiserated and celebrated with me. People who deserve special mention are my lovely aunt, Cherrie Bassiri, who constantly encouraged me to follow my dream of pursuing a PhD, my lovely mother-in-law, Parvaneh, my dear friends, Chris Strom, Susan Rahimi and John Rahimi.

I would like to dedicate this thesis to my brother, Hootan and my grandmother, Irandokht, who sadly passed away during the completion of my PhD. There isn't a day that goes by that I don't think of you and wish you were healthy, happy and here sharing this life with us. Love always.

Keywords

Expertise profiling; Knowledge-curation platforms; Micro-contributions; Annotation, Ontologies; knowledge acquisition; knowledge representation; Semantic Web; Text processing; Expertise visualization; Social expert networks; Contextual factors

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 080107, Natural Language Processing, 30%
ANZSRC code: 080607, Information Engineering and Theory, 50%
ANZSRC code: 080603, Conceptual Modelling, 20%

Fields of Research (FoR) Classification

FoR code: 0806, Information Systems, 70%
FoR code: 0801, Artificial Intelligence, 30%

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AO	Annotation Ontology
API	Application Programming Interface
BME	BiomedExperts
EARS	Entity and Association Retrieval System
IC	Information Content
iCAT	ICD Collaborative Authoring Tool
ICD-11	International Classification of Diseases ontology, revision11
IDF	Inverse Document Frequency
IR	Information Retrieval
L&C	Leacock and Chodorow
LCS	Least Common Subsumer
MCB	Molecular and Cellular Biology
NCBO	National Centre for Biomedical Ontology
NLM	National Library of Medicine
NLP	Natural Language Processing
OPM	Open Provenance Model
OWL	Web Ontology Language
Profiles RNS	Profiles Research Networking Software
RDF	Resource Description Framework
SIOC	Semantically-Interlinked Online Communities
SKOS	Simple Knowledge Organization System
SNA	Social Network Analysis
SOAP	Simple Object Access protocol
STEP	Semantic and Time-dependent Expertise Profiling
TF	Term Frequency
TREC	Text Retrieval Conference
TWFG	Topical and Weighted Factor Graph
UMLS	Unified Medical Language System
VRE	Virtual Research Environment
W&P	Wu and Palmer
W3C	World Wide Web Consortium
WHO	World Health Organisation

Chapter 1 Introduction

1.1 Background

Organizations are constantly seeking individuals with expertise in specific topics and therefore require extensive profiling systems to enable them to efficiently locate experts with the required knowledge. Moreover, there is growing recognition that enabling timely access to relevant expertise within organizations is critical to the efficient running of enterprise operations. For example, the employees of geographically dispersed organizations typically have difficulty in determining what others are doing and which resources can best address their problems. Failure to foster exchange within the knowledge community leads to duplication of effort and an overall reduction in productivity levels [1]. In addition, many scientific research environments are increasingly dynamic and subject to rapid evolution of knowledge. Global scientific challenges, such as 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. Furthermore, nomination of experts in a scientific community, through current and comprehensive expertise profiles, motivates a potentially larger number of authors to contribute to the community which is vital to the integration of diverse viewpoints and the efficient assembly of an extensive body of knowledge [2].

However, expertise is not easily identified and is even more difficult to manage on an ongoing basis which leaves vast resources of tacit knowledge and experience untapped. Filling out comprehensive profiling systems and keeping them up to date requires extensive manual effort and has proven to be impractical. Research into expertise profiles at IBM has found that after 10 years of repetitive and consistent pressure from the executives, including periodic emails sent to experts to remind them to update their profiles, only 60% of all IBM profiles are kept up-to-date [3]. This clearly indicates that a manual approach to profiling expertise is not sufficient and an automated solution is required to create and maintain expertise profiles.

Expertise Retrieval is an active research topic in a wide variety of applications and domains, including biomedical, scientific and education [4, 5, 6]. Most of the existing research has focused on the task of *expert finding*, i.e., given a set of documents and a set of expertise profiles, the aim is to find the best matches between the profiles and the topics that emerge from those documents ("*who are the experts in a particular topic?*"). The associated research topic of *expert profiling* focuses on identifying a list of expertise topics in which a person is knowledgeable ("*what topics does this person know about?*") [4].

Expert finding (identifying a list of people who are knowledgeable about a given topic) has attracted significant attention from both research and industry communities. Contrary to traditional Information Retrieval (IR) systems, the target of expert finding is individual people (named entity) rather than documents. This task is usually addressed by uncovering associations between people and topics. The strength of association between a person and a topic, determines the person's level of competence in that topic. Co-occurrences of a person's name with topic terms in the same context are often assumed to be evidence of expertise [7].

A number of models have been developed to capture the association between topics and experts. For example, Generative Probabilistic Models estimate associations between topics and people as the likelihood that the particular topic was generated by a given candidate (topic generation models) or that a probabilistic model based on the topic generated the candidate (candidate generation models) [8]. Discriminative models capture associations between topics and people by directly estimating the binary conditional probability that a given pair of a topic and a candidate expert is relevant [9].

Determining an association between candidate experts and expertise that emerges from their publications has proven to be a complex task. Current approaches to expertise finding and expertise profiling associate an expert with the "tacit knowledge" that emerges from the explicit knowledge (e.g., documents and publications) associated with that expert. Thus, such approaches must overcome both the challenges of document retrieval, in addition to the challenges associated with the task of expertise retrieval. Generally, experts are identified by analysing documents associated with them, through authorship, mentions or citations. Such associations are not always an accurate indication of expertise in topics that emerge from the documents. Furthermore, heterogeneous sources used as evidence of expertise are assumed to be of equal importance, while in practice, some sources provide a much stronger evidence of expertise than others. Other limitations of current approaches include the inability to determine changes in a person's expertise over time or to extract expertise from non-traditional publications (such as online blogs, wikis, twitter etc).

The Text Retrieval Conference (TREC) [10] enterprise track has been the major forum for empirically comparing Expertise Retrieval techniques. Essentially, the two most popular and well-performing types of approaches in TREC expert search task are profile-centric and document-centric approaches. Profile-centric approaches create a textual representation of a person's knowledge according to the documents with which he/she is associated [11]. These "pseudo documents" can then be ranked using standard document retrieval techniques. Document-centric approaches can be generalised as a two-stage model. First, a document relevance model finds documents relevant to a topic. Second, an association discovery model, which is typically a

window-based, co-occurrence model, ranks candidates mentioned in these documents based on a combination of the document's relevance score and the degree to which the person is associated with that document [12]. Such traditional approaches rely on analysing a *large corpus* of *static* documents for expertise retrieval.

The field of Social Network Analysis (SNA) considers the graphs connecting individuals in different contexts, and infers their expertise from the shared domain-specific topics [5]; e.g., researchers co-authoring publications with other researchers inherit part of their co-authors' expertise, extracted from non-co-authored publications. In a study which addresses the task of expert profiling, the profile of an individual is defined to be a record of the types and areas of skills of that individual (topical profile) plus a description of his/her collaboration network (social profile). Experts' social profiles are described through a graph representation of the collaboration network, where nodes represent people and weighted, directed edges (based on co-authored documents) reflect the level of collaboration [13]. A recent study, which addresses the task of finding similar experts, demonstrates that models which combine content-based and contextual factors (social information) can significantly outperform existing content-based models [14].

Finally, in the Semantic Web [15] domain, expertise is captured using ontologies and then inferred from axioms and rules defined over instances of these ontologies. A recent study has investigated an ontological approach to expertise profiling by developing a formal ontology for representing and reasoning about skills and competencies in a dynamic environment [16]. Another study introduces an ontology for competency management and considers expertise to be a level of competency characterised by performance. According to this study, criteria such as frequency, scope, autonomy, complexity and context can be used as performance indicators for evaluating expertise [17].

1.2 Collaboration Platforms

The World Wide Web (WWW) has changed dramatically over the recent years, moving from a static one-way medium toward a more *dynamic* platform, transforming the mechanisms and workflows of collaboration. More specifically, with the emergence of Web 2.0 [18], there has been a significant increase in online collaboration, through Web-based communities of users such as Wikis, blogs and social networks. People are no longer merely consumers of content and applications; they are participants, creating content and interacting with different services and users. More and more people are sharing and exchanging knowledge through collaborative online communities; e.g., contributing to knowledge bases such as Wikipedia and using peer-to-peer (P2P) technologies, where experts share their knowledge and expertise through *micro-contributions* to the

underlying knowledge base. This increase in community participation and content creation presents new opportunities for mining expertise from the tacit knowledge embedded in such platforms.

Micro-contributions or incremental refinements to the *structured* or *unstructured* content of collaboration platforms provide a dynamic environment where knowledge is subject to ongoing evolution. Examples of *unstructured* contributions, in natural language form, can be seen in platforms such as Wikis (starting with Wikipedia as a pioneering project) or collaborative knowledge bases, predominantly in the biomedical domain, e.g., AlzSWAN [19]. These platforms enable authors to incrementally and collaboratively refine the content of 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 [20] (a sub-project of Wikipedia) supports discussions on genes. Figure 1-1 illustrates an example of a micro-contribution extracted from the *Skeletome* knowledge base [21], a discussion and collaboration platform on skeletal dysplasias. The background image depicts a page containing general information about Achondroplasia, a type of skeletal dysplasia. The overlaid image illustrates a *micro-contribution*, created by an expert by adding an investigation item to the definition of Achondroplasia.

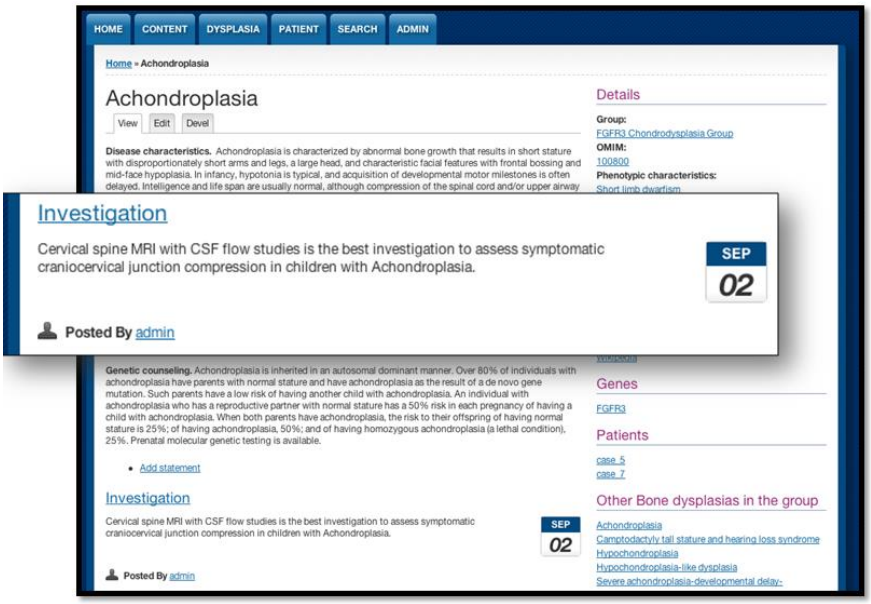

Figure 1-1: Example of a micro-contribution in the Skeletome Knowledgebase

The *WikiProject Medicine* [22] is a platform where people interested in medical and health content on Wikipedia can discuss, collaborate or debate related issues. *Stack Overflow* [23] uses a similar approach to knowledge sharing, however with a focus on programming and code development / deployment.

Examples of *structured* contributions are evident in the context of collaborative ontology engineering projects, where changes contributed by experts target ontological concepts. For

example, the 11th revision of the *International Classification of Diseases ontology, ICD-11* [24], is currently under active development by the World Health Organization [25], involving over 270 experts from around the world. In this context, knowledge evolves through experts' contributions to structured content, i.e., ontological concepts. Building ontologies in a collaborative and increasingly community-driven manner has become a central paradigm of modern ontology engineering. This understanding of ontologies and ontology engineering processes is the result of intensive theoretical and empirical research within the Semantic Web community, supported by technology developments such as Web 2.0 [18]. With increasing adoption and relevance, ontologies have significantly increased in size, resulting in an evolution in the way ontologies are engineered. Because no single individual has the expertise to develop such large-scale ontologies, ontology-engineering projects have evolved from small-scale efforts involving just a few domain experts to large-scale projects that require input from and collaboration between, dozens or even hundreds of experts and other stakeholders [26].

In addition, collaboration platforms enable researchers and scientists to connect, network, communicate and collaborate. Thus, *contextual factors*, such as the collaboration structure and experts' relationships in scientific social networks provide an additional source of tacit and implicit knowledge for modelling expertise. A representative example in this category is the *ResearchGate* network of scientists and researchers [27], where knowledge continuously evolves through the addition and sharing of new publications, contributions to Q&A forums and qualitative assessment of collaborators' contributions (i.e., voting system).

Regardless of the various types of knowledge embedded in collaboration platforms, i.e., *structured contributions*, *unstructured contributions* and *social factors*, experts' *micro-contributions* provide vast resources of implicit knowledge and experience, while giving the knowledge captured within the environment a *dynamic* character. Traditional expertise profiling approaches (that typically rely on large corpora of static documents) have limited applicability in the context of such dynamic knowledge environments.

Hence, this thesis proposes an Expertise Modelling Framework, which advances the state of the art in expertise profiling by considering *living* documents; i.e., documents where knowledge evolves through micro-contributions. This work is motivated by: the emergence of Web 2.0, resulting in an increasing trend in online participation and knowledge sharing; the increasing importance of online profiles in generating reputations and visibility in particular communities; and the increasing use of online profiles by head hunters and employment agencies.

1.3 Challenges

Regardless of the domain, traditional approaches to expertise profiling raise a number of challenges, when applied to micro-contributions in the context of evolving knowledge bases. Such approaches associate "person mentions" with "query words" in the same context as evidence of expertise [7]. In other words, they measure the frequency of topics and the co-occurrence of topics and experts in documents and therefore, rely on analysing *large corpora* of *static* documents; e.g., publications, grants and reports. However, the content of collaboration platforms is *dynamic* and continuously changes through experts' micro-contributions. The *short* and *sparse* content of micro-contributions does not provide sufficient context for modelling expertise using traditional techniques. Thus, in the context of collaborative knowledge bases, where content evolves over time, a model is required that can derive expertise by performing *semantic* analysis of the short and sparse content of micro-contributions.

Furthermore, incremental refinements to collaborative knowledge bases, give rise to *dynamic* content, which can be analysed to *track changes* in experts' skills and interests over time. However, as traditional techniques analyse static documents, they are unable to model ongoing changes in peoples' expertise and interests over time.

Traditional expertise profiling methods adopt a *document-centric* approach, by associating an expert with expertise topics that emerge from the documents associated with the expert. Thus, in modelling expertise, such approaches do not consider the expert's specific contributions to these document/s (in terms of quantity, quality or topic). In the context of knowledge curation platforms, where content evolves through collaborative efforts of many experts, a model is required to profile the expertise of every contributing author based on his/her contributions.

Current expertise retrieval methods adopt a *macro-perspective* of documents authored by individuals, associating experts with expertise topics embedded in a document as a whole. However, such techniques do not provide sufficient evidence of expertise. For example, an individual may be considered an expert in a particular topic because he/she has authored or co-authored documents in the topic, but their actual contributions to the authored documents cannot be established merely by considering co-authorship. In order to provide *evidence* of an individual's expertise, a model is required that captures a *fine-grained* representation of the individual's contributions and their provenance.

Finally, traditional expertise profiling approaches make extensive use of unstructured data and therefore, have very limited inference capabilities. An approach is required for capturing and representing the *semantics* of knowledge contributed by experts, using structured and widely adopted vocabularies and ontologies, such as ontologies published in the Linked Data Cloud [28].

This in turn facilitates integration of profiles into the Linked Data Cloud and provides the foundations for creating overarching views of expertise, by complementing published profiles using the structured and interlinked data. Figure 1-2 provides a comparison of traditional expertise modelling techniques with expertise profiling in the context of collaboration platforms.

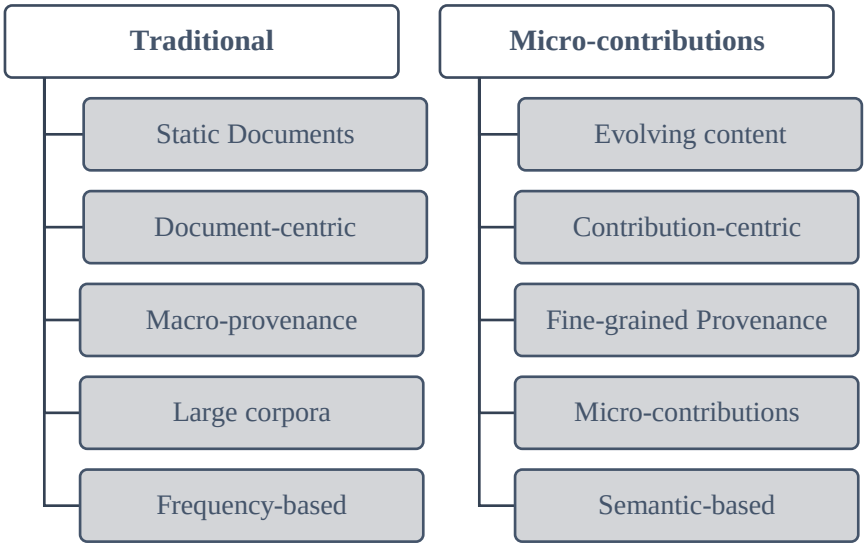

Figure 1-2: Comparison of traditional expertise modelling and expertise profiling in collaboration platforms

1.4 Motivation and Significance

To date, most expertise profiling approaches aim to associate experts with the tacit knowledge embedded in explicit sources associated with those experts (i.e., large corpora of static documents authored by the experts). However, with the advent of Semantic Web and Web 2.0 and the associated significant increase in social networking, online collaborations, and community-generated content, an alternative source of explicit knowledge has emerged, from which the expertise of contributors can be mined. This alternative source of expertise, i.e., *micro-contributions*, consists of short and sparse content contributed to Web-based community fora such as Wikis, blogs and social networks, where knowledge continuously *evolves* over time. Thus, traditional approaches to expertise profiling, which rely on analysing large corpora of static documents, are inadequate when applied to micro-contributions. This thesis aims to overcome the limitations and challenges associated with traditional expertise profiling - when applied in the context of *dynamic micro-contributions*. Towards this goal, an Expertise Profiling Framework is proposed, which creates *fine-grained* and *time-aware* expertise profiles by tapping into the knowledge contributed by experts to collaboration platforms; i.e., *micro-contributions*. The framework incorporates a model, which refines expertise profiles by integrating *contextual factors*, such as the *implicit* and *explicit relationships* between *experts* in *social networks*, with content-based factors (i.e., the topics that arise within micro-contributions).

The Expertise Profiling Framework captures the *temporal* and *dynamic* characteristics of expertise, enabling one to monitor not only the activity performed by individuals, but also the change in personal interests and the progression of an expert's knowledge over time. This is analogous to the progression of scientific hypotheses, from simple ideas to scientifically proven facts.

The Expertise Profiling Framework proposed in this thesis, represents the implicit knowledge embedded in micro-contributions using terms from machine-processible domain ontologies. This in turn facilitates the application of reasoning techniques developed by the Semantic Web community. From a technical perspective, building expertise profiles from concepts defined in widely adopted ontologies enables individuals to publish and integrate their profiles as *structured data* on the *Web*. This provides "*expertise seekers*" and "*web crawlers*" with access to expertise profiles and facilitates better consolidation of the profiles, which in turn occasions a seamless aggregation of communities of experts. Furthermore, the links between ontological concepts in expertise profiles and concepts in the Linked Data Cloud [28], can be discovered and used to complement the published profiles, providing access to richer, more accurate and more up-to-date expertise profiles.

In some communities, there has also been lobbying for a change in scientific publishing from the current document-centric approach (e.g., full journal or conference papers) to a micro-contribution approach in which hypotheses or domain-related assertions are published in the form of short statements in online knowledge bases or in which multiple contributors work on a document collaboratively. Examples of this new trend can be seen in recent initiatives promoting the adoption of *nano-publications* [29] and *liquid publications* [30]. In this new environment, mapping such micro-contributions to expertise will be essential in order to support the development of reputation metrics. The research presented in this thesis, focuses only on building expertise profiles from micro-contributions. However the resulting expertise profiles provide a robust foundation upon which novel trust and reputation models can be applied.

Furthermore, the proposed model complements authorship recognition. In addition to identifying authorship, it attaches semantics to authored content and builds profiles based on authored contributions. An essential element of being a scientist is recognition of expertise by others in the community, which translates into jobs, grants, publications and collaborators. Expertise profiles will therefore provide authors with due recognition for their contributions, which will in turn motivate further contribution to and collaboration in community-driven knowledge curation platforms.

1.5 Scenarios

The realization of the goal of creating *fine-grained* and *time-aware* expertise profiles, using *micro-contributions* to collaboration platforms, where knowledge *evolves* over time, can be pragmatically described as a series of scenarios or requirements:

- Mining expertise from implicit knowledge embedded in experts' micro-contributions to collaborative, knowledge-curation platforms;
- Facilitating individual attribution and evidence of expertise;
- Facilitating greater visibility of expertise on the Web;
- Enabling experts to describe their expertise at various levels of granularity;
- Facilitating the tracking and analysis of changes in expertise over time;
- Enabling experts to complement existing profiles with knowledge embedded in social networks.

The following describes each of the abovementioned requirements in more detail:

Mining expertise from implicit knowledge embedded in experts' micro-contributions to collaborative, knowledge-curation platforms

With the proliferation of the Web of Data [28], characterised by the increasing use of ontologies, via Semantic Web [15] and Web 2.0 [18], there has been a significant increase in online collaboration. This has in turn given rise to knowledge bases in which content continuously evolves through individuals' contributions. Thus, experts' micro-contributions to evolving knowledge-curation platforms, provides a rich source for mining the expertise and knowledge of contributors. Figure 1-3 depicts an example of a micro-contribution (highlighted in red) and its encapsulating context.

> The most recognizable and most common form of Dwarfism in humans is Achondroplasia, which accounts for 70% of dwarfism cases and produces rhizomelic short limbs, increased spinal curvature, and distortion of skull growth. *With Achondroplasia, the body's limbs are proportionately shorter than the trunk (abdominal area), with a larger head than average and characteristic facial features. Achondroplasia is an autosomal dominant disorder caused by the presence of a faulty allele in the genome. If a pair of Achondroplasia alleles are present, the result is fatal.* Achondroplasia is a mutation in the fibroblast growth factor receptor 3. In the context of Achondroplasia, this mutation causes FGFR3 to become constitutively active, inhibiting bone growth. Research by urologist Harry Fisch of the Male Reproductive Center at Columbia Presbyterian Hospital indicates that in humans this defect may be exclusively inherited from the father and becomes increasingly probable with paternal age: specifically males reproducing after 35. This condition occurs in 4 to 15 out of 100,000 live births.

Figure 1-3: Example of a micro-contribution and its encapsulating context

In collaborative knowledge bases, documents are neither static (as they are continuously and incrementally refined) nor lead to large corpora authored by individual experts (usually authors edit a fraction of a document, which is closer to their expertise/interest). Therefore, in order to facilitate

expertise profiling using micro-contributions, in the context of collaboration platforms, this thesis proposes a framework, which supports the paradigm shift from *static* knowledge embedded in documents to *evolving* knowledge brought by *micro-contributions* to the content of *dynamic* collaboration platforms. Furthermore, the proposed framework adopts a *contribution-centric* view of the platform, and captures and analyses the "*semantics*" of the short and sparse content of micro-contributions.

Facilitating individual attribution and evidence of expertise

Traditional approaches to expertise profiling, associate an expert with expertise and knowledge embedded in the document/s which the expert has authored/co-authored. The Expertise Profiling Framework presented in this thesis, facilitates *individual attribution* by associating an individual with expertise topics embedded in his/her *micro-contributions*, rather than topics that emerge from the documents that host those micro-contributions. Towards this goal, the proposed framework captures the coarse and fine-grained provenance of micro-contributions and their localization in the context of their host living documents; e.g., the sentence, paragraph, subsection and section of the document in which they appear. This in turn enables the analysis of micro-contributions, using the broader context within which they are made. In addition, capturing the fine-grained provenance of micro-contributions provides *evidence* for the expertise topics associated with an individual. In other words, the proposed model links expertise topics associated with an expert, to the content of the expert's micro-contributions, rather than the entire content of the documents to which he/she has contributed.

Facilitating greater visibility of expertise on the Web

Many scientific research domains are subject to rapid evolution of knowledge, leading to the proliferation of special-purpose knowledge bases for keeping up with the most recent advances in the field. Consequently, experts often contribute to various collaboration platforms and social networks, resulting in contributions across multiple knowledge bases.

The Expertise Profiling Framework proposed in this thesis, models expertise using concepts from widely adopted *ontologies* and vocabularies in the Semantic Web. This facilitates the integration of an author's expertise, emerging from his/her contributions to each of these isolated silos, providing a comprehensive view of the expert's skills and experience, using a shared understanding. Furthermore, *structured* expertise profiles, i.e., profiles containing ontological concepts, can be integrated into the Web, making them visible and accessible to Web crawlers and Web 2.0 enabled applications. In addition, publishing profiles containing structured data to the Web, facilitates detection of links and relationships between ontological concepts in profiles and

concepts in the Linked Data Cloud [28], which can be used to complement the published profiles, providing a more comprehensive, accurate and up-to-date view of experts' skills and experiences.

Enabling experts to describe their expertise at various levels of granularity

The Expertise Profiling Framework proposed in this thesis, represents expertise topics embedded in micro-contributions, using concepts from domain ontologies. The use of ontologies enables one to take into account more than just the actual domain concepts, by looking at their ontological parents and children. Thus, the proposed framework, provides the flexibility to customise the granularity of domain concepts representing expertise topics in profiles, by using expertise centroids - i.e., ontological concepts that act as representatives for an area of the ontology by accumulating high similarity values against all micro-contributions located in that area. This in turn enables experts to complement their existing online profiles with *fine-grained* domain concepts that represent the *implicit knowledge* embedded in their *micro-contributions* to *collaboration platforms*.

Furthermore, the ability to represent expertise at various levels of granularity facilitates comparison of profiles, which describe expertise at different levels of abstraction. This has in turn facilitated the evaluation of the Expertise Profiling Framework proposed in this thesis. This framework uses experts' micro-contributions to create profiles, which represent the knowledge and expertise of contributing authors. As micro-contributions are generally very specific, (i.e., the terminology describes specific domain aspects (e.g., insulin, hypoglycaemia, beta cells, and pancreas)) the generated profiles will define expertise at a correspondingly fine-grained level. However, experts often describe their expertise using very generic topics (e.g., Chemistry, Biology, Cell and Genetics). Thus, profiles created from experts' micro-contributions and profiles described by experts, contain concepts at different levels of abstraction. Therefore, in order to facilitate comparison and evaluation, the proposed framework should provide the ability to describe expertise at a level of granularity that is comparable to the profiles defined by the experts.

Facilitating the tracking and analysis of changes in expertise over time

The expertise of an individual is dynamic and typically changes over time. The proposed framework captures and tracks the *temporality* of expertise, by tracking the *evolution* of *micro-contributions over time*. Temporal analysis of expertise enables one to determine the level of activity in particular topics over time, detect the timeframes where an expert demonstrates "peak activity" in particular topics and identify the most/least active experts in particular topic/s. From a project management perspective, this provides the ability to determine if participants' activities are

in line with the focus of the project or to ascertain the level of collaboration among subject matter experts.

Enabling experts to complement existing profiles with knowledge embedded in social networks

Scientific social networks not only support knowledge evolution through experts' continuous contributions to the underlying knowledge, but also provide a paradigm for experts to communicate, collaborate, network and share knowledge. The Expertise Profiling Framework proposed in this thesis, adopts a "*network perspective*" of collaboration platforms and analyses experts and their contributions as embedded in a network of relations. The collaboration context and social collaboration relationships among experts in these networks provide an additional source of implicit knowledge for modelling expertise. For example, in Q&A forums, the context is formed by a question and its associated answers, while social factors can be captured implicitly via relationships formed by participating in the same Q&A forums and discussions, votes on questions and answers, or explicitly via "Following" / "Co-author" relationships between experts. The framework proposed in this thesis, combines content-based factors, i.e., experts' micro-contributions with contextual factors, i.e., collaboration context and social collaboration relationships among experts, to refine expertise profiles. Furthermore, expertise profiles are refined using the semantic relationships between ontological concepts in collaborators' micro-contributions and profiles.

1.6 Hypothesis, Aims and Objectives

The hypothesis that underpins the research described in this thesis is that:

A comprehensive, fine-grained provenance model, that is able to capture and consolidate structured and unstructured micro-contributions made within the context of multiple host documents, will improve expertise profiling in evolving, dynamic knowledge bases.

This hypothesis raises a series of research questions:

- How can expertise be modelled using the fine-grained provenance and the evolution of micro-contributions in the context of evolving knowledge?
- How can the temporal and dynamic characteristics of expertise be captured in order to create profiles, which enable the tracking of changes in expertise and interests over time?
- How can different perspectives (requirements and performance) of the proposed expertise profiling methodology be obtained and investigated in the context of both structured and unstructured micro-contributions in different knowledge domains?

- How can the granularity of expertise profiles be customised in order to accurately represent the knowledge and skills of contributing experts, and facilitate comparison and evaluation of profiles that describe expertise at different levels of abstraction?
- How can expertise profiles be refined and enriched using the contextual factors that exist within social expert networks?

The research questions listed above can be mapped to the following objectives:

O1. Development of a *comprehensive and fine-grained Provenance Model* for capturing *structured* and *unstructured* micro-contributions, by combining coarse and fine-grained provenance, change management and concepts from domain-specific ontologies.

O2. Development of a *Semantic and Time-dependent Expertise Profiling methodology* by linking the textual representation of expertise topics in micro-contributions to weighted concepts from domain ontologies, whilst capturing the *temporality* of expertise.

O3. Application of the Semantic and Time-dependent Expertise Profiling methodology to different types of community-driven, dynamic knowledge-curation platforms; i.e., both *unstructured* and *structured* micro-contributions in the context of a range of knowledge domains.

O4. Development of a mechanism for *customising* the *granularity* of ontological concepts in expertise profiles in order to: (i) describe expertise with a level of specificity that accurately represents the knowledge embedded in micro-contributions, and; (ii) facilitate the *comparison* and *evaluation* of profiles which describe expertise at different levels of abstraction.

O5. Development of a *Profile Refinement Model* by integrating contextual factors from social expert networks, with the Semantic and Time-dependent Expertise Profiling methodology, in order to improve the accuracy of expertise profiles.

O6. Development of a Profile Visualization paradigm to facilitate analysis and tracking of evolving expertise and interests over time

1.7 General Overview of the Research Framework

This thesis proposes an Expertise Modelling Framework for capturing and representing the expertise of individuals who contribute to the evolution of knowledge in collaboration and knowledge curation platforms. The framework is domain-agnostic and aims to support the hypothesis, aims and objectives outlined in Section 1.5 and scenarios presented in Section 1.4. Figure 1-4 depicts a high-level overview of the proposed framework. The following describes the main constituents of this framework.

1.7.1 The Fine-grained Provenance Model

The Fine-grained Provenance Model captures and represents micro-contributions in the context of their host documents. It documents the change management aspects of the platform, i.e., activities including *update, add, delete,* that result in incremental refinements to content. Furthermore, it keeps track of the revisions to host documents, generated by miro-contributions.

At the centre of the model is the *Fine-grained Provenance Ontology,* which combines coarse-grained and fine-grained provenance modelling to capture *micro-contributions* and their *localization* in the context of their *host living* documents. Figure 1-3 depicts a micro-contribution within its context. This in turn, facilitates semantic analysis of *short* and *sparse* contributions, by identifying the broader context which encapsulates every micro-contribution. The Fine-grained Provenance Ontology adopts a "*contribution-oriented*" approach to expertise modelling by capturing the domain concepts which represent expertise topics emerging from every micro-contribution. It bridges the gap between the textual grounding of expertise topics in micro-contributions and the domain knowledge (i.e., ontological concepts). Representing expertise topics using ontological concepts, enables us to use the structure of domain ontologies to determine the relationships between concepts that describe a micro-contribution and concepts that describe the broader context within which the contribution is made. These relationships, e.g., superclass/subclass, can be used to enhance the set of concepts representing the contribution, resulting in a more comprehensive view of the expertise and skills of the contributing author. The conceptual representation of an expert's micro-contributions is then used to create *semantic* and *time-aware* expertise profiles.

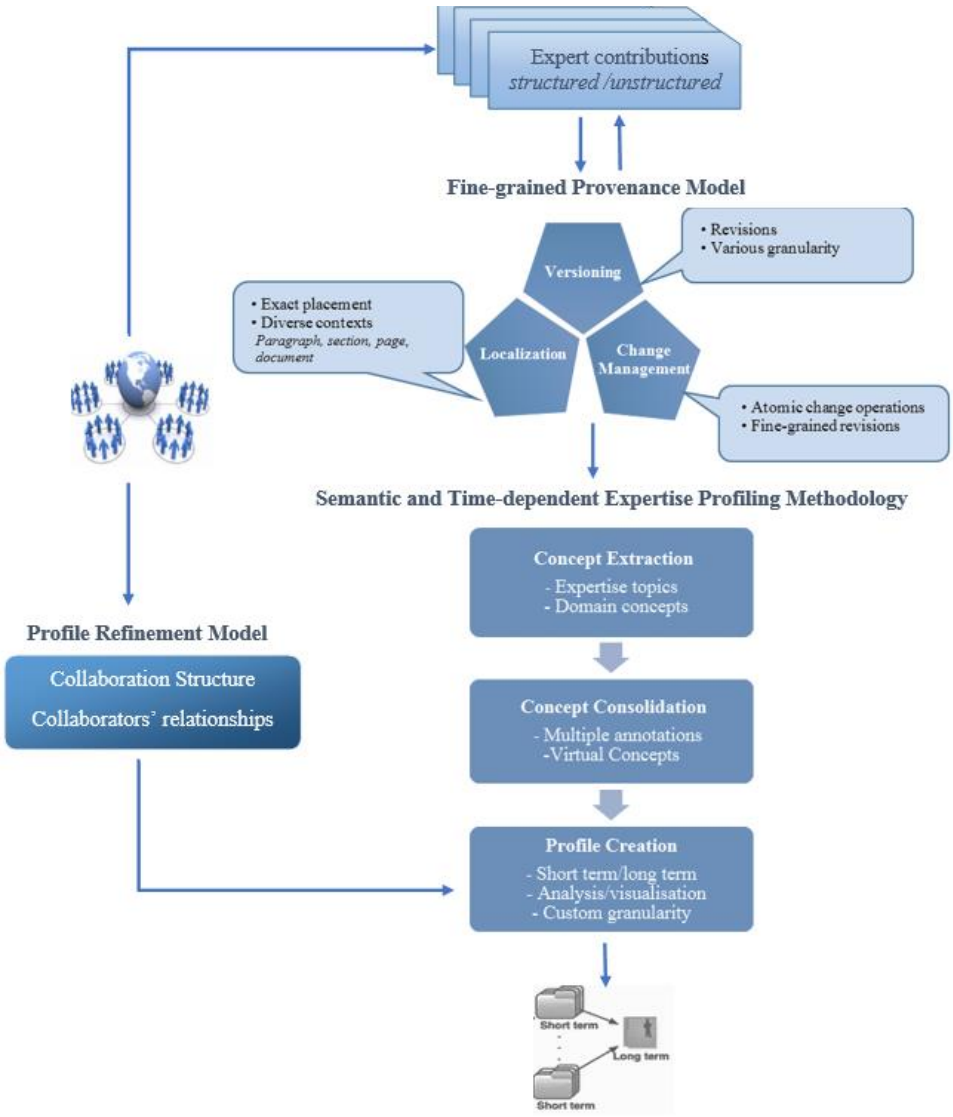

Figure 1-4: High-level overview of the Expertise Profiling Framework

1.7.2 The Semantic and Time-dependent Expertise Profiling Methodology

The Semantic and Time-dependent Expertise Profiling (STEP) methodology uses experts' contributions (whose coarse and fine-grained provenance is captured and represented by the Fine-grained Provenance Model), to build expertise profiles containing domain concepts, while tracking changes in the experts' areas of expertise and interests over time. STEP comprises three main phases, Concept Extraction; Concept Consolidation; Profile Creation; as depicted in Figure 1-4 and described in the following three sections.

Concept Extraction

This phase aims at capturing experts' micro-contributions to a collaboration platform. It annotates micro-contributions and represents expertise topics using concepts from domain ontologies. This is achieved by employing a typical information extraction or semantic annotation

process, which is, in principle, domain dependent[1]. Hence, in order to provide a profile creation framework applicable to any domain, this step is not restricted to the use of a particular concept extraction tool or technique. Using domain-specific annotation tools as the only method for annotating micro-contributions with ontological concepts would render the profiles dependent on the accuracy of annotations performed by the tools. Therefore, the pluggable architecture of STEP is used to integrate Language Modelling techniques [31], i.e., Topic Modelling [32] and N-gram Modelling [33], with the Concept Extraction phase. This approach demonstrates that combining language modelling techniques, (which are applicable to any domain) with the STEP methodology, improves the accuracy of expertise profiles, by reducing the effects of domain-specific concept extraction tools and techniques.

In addition to micro-contributions (content-based information), this phase also captures contextual factors (e.g., patterns of communication) from the network. More specifically, it captures the relationships among experts in existing social networks; e.g., co-authorship, follower/following, and ad hoc relationships formed through participation in discussions and Q&A forums. In addition to the relationships, additional collaboration attributes, such as the rankings or number of positive and negative votes (e.g., the thumbs-up/thumbs-down system used by platforms such as StackOverflow [23] to quantify the quality of questions and answers) associated with "questions and answers" contributed by experts, are also captured.

Concept Consolidation

Over the course of the last decade there has been an increase in the adoption of ontologies in order to provide machine-processible conceptualization of a domain. While this has resulted in the formal conceptualization of a significant number of domains, it has also led to the creation of duplicate concepts; i.e., the same concepts defined in the context of multiple domains, and hence, within multiple ontologies – and having slightly different definitions in each. For example, in the biomedical domain, the concept "Viral Gastroenteritis" is now present in at least seven ontologies (cf. NCBO Bioportal [34]), while "Alagille Syndrome" is defined by at least 22 ontologies (cf. NCBO Bioportal [34]). 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. Consequently, expertise topics identified in an expert's micro-contributions are annotated with concepts from multiple ontologies. This phase consolidates concepts resulting from annotation of *lexically different*, but *semantically similar* entities across

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

micro-contributions and uses their union to create "*Virtual Concepts*". A "Virtual Concept" represents an abstract entity and contains domain-specific concepts from different ontologies, which are manifestations of the abstract entity. For example, concepts from various ontologies that represent the topics "Gene", "RNA", "DNA" and "Gene Sequencing", are manifestations of the virtual concept "Genetics". Hence, virtual concepts provide comprehensive and coherent views over entities identified in an expert's micro-contributions and serve as the building blocks for generating expertise profiles using STEP.

Furthermore, in the context of contributions to structured content, i.e., collaborative ontology engineering projects where experts' contributions target ontology concepts, *semantic similarity* is used, in order to: (i) determine the level of profile abstraction that accurately represents the knowledge of contributing authors, and (ii) customise the level of granularity of concepts representing an expert's knowledge and expertise. As mentioned in Section 1.4, this enables experts to complement their existing profiles with a fine-grained representation of the knowledge which they have contributed to collaboration platforms. In addition, it facilitates the comparison and evaluation of profiles that describe expertise at different levels of abstraction.

Profile Creation

This phase uses the extracted and consolidated concepts to create *time-aware* expertise profiles. Capturing the temporal characteristics of expertise is extremely valuable as it enables the changes in an expert's interests and expertise to be tracked and analysed over time. In order to facilitate tracking and analysis of changes in expertise, two types of profiles are created for every expert; (i) *short term* profiles and (ii) *long term* profile.

A *short term* profile represents a collection of concepts extracted from micro-contributions of an expert, over a specific window of time. Short term profiles aim to capture periodic bursts of expertise in specific topics, over a length of time. This phase involves the development and application of methods aimed at capturing time-windows in which an expert demonstrates high levels of activity in particular topics of expertise.

A *long term* profile, on the other hand, provides 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. The long term profile of an expert consists of concepts that appear *persistently* and spread *uniformly* across all short term profiles of the expert. Unlike traditional approaches, the expertise profiling model proposed in this thesis considers uniformity as important as persistency; i.e., an individual is considered to be an expert in a topic if this topic is detected in his/her contributions over a long period of time (persistency) and its presence is distributed uniformly across the majority of short term profiles for that expert.

Furthermore, the "*Profile Explorer*" [35] visualization tool is proposed, in order to provide a friendly and intuitive framework for visualization and analysis of evolving interests and expertise over time. Profile Explorer facilitates visualization of short term and long term profiles and provides a framework for conducting comparative analysis of experts and expertise by linking an expert's long term profile with short term profiles and underlying contributions. Profile Explorer creates a domain-independent paradigm that facilitates visualization, search and comparative analysis of expertise profiles.

1.7.3 The Profile Refinement Model

The STEP methodology described above creates profiles that represent expertise using *content-based* factors, i.e., experts' micro-contributions to collaboration platforms. The *Profile Refinement Model* captures and analyses *contextual factors* (i.e., social network information embedded in collaboration platforms), in order to provide additional evidence of expertise. While in previous phases, the focus was only on experts' attributes and contributions to the underlying knowledge base, in this step a "*network perspective*" of the platform is adopted, viewing experts and their contributions as part of a network of relations.

The Profile Refinement Model uses the implicit knowledge embedded in the *context* within which every micro-contribution is made, to refine expertise profiles of contributors. Furthermore, it identifies and analyses implicit relationships (e.g., relationships formed between experts by *participating* in Q&A discussions) and explicit relationships (e.g., "*following*" and "*co-author*") between experts in the network, to refine the expertise profiles of collaborators.

In addition, the model uses the structure of domain ontologies to determine semantic relationships (e.g., superclass/subclass) between domain concepts in collaborators' contributions and profiles. Collaborators' profiles are subsequently refined using the type and strength of their relationships and the semantic associations between concepts in their profiles and contributions.

1.8 Original Contributions

This thesis presents a series of contributions to the current state of the art, as listed below:

1.8.1 Expertise profiling using the fine-grained provenance of micro-contributions

The Expertise Profiling Framework proposed in this thesis, combines coarse and fine-grained provenance modelling to capture micro-contributions and their localisation in the context of the living documents that host them (The Fine-grained Provenance Model described in Chapter 3). The model adopts a "*contribution-oriented*" view of the platform, thereby facilitating fine-grained expertise profiling, by analysing the contexts which encapsulate micro-contributions. Such contexts provide sufficient content for semantic analysis of short fragments of micro-contributions, while

limiting the analysis to content modified by micro-contributions (as opposed to the whole document). Capturing the provenance of micro-contributions enables us to perform comparisons between those concepts emerging from micro-contributions and those concepts embedded in the broader contexts. Furthermore, capturing the fine-grained representation and provenance of an expert's micro-contributions is used as evidence of expertise associated with the expert.

1.8.2 Creating semantic and time-aware expertise profiles

The framework proposed in this thesis analyses experts' *micro-contributions* to dynamic, collaborative knowledge-curation platforms, in order to generate semantic and time-aware expertise profiles (The Semantic and Time-dependent Expertise Profiling (STEP) methodology described in Chapter 4). STEP captures the *temporal* aspect of expertise and differentiates between *short term* and *long term* profiles, facilitating analysis and tracking of changes in expertise and interests over time.

While a number of research efforts analyse large corpus of static documents authored by an expert to determine the changes in expertise over time [36], the research proposed in this thesis is the first attempt at determining the *temporality* of expertise by analysing *micro-contributions* to *evolving* knowledge. In addition, prior research efforts assume regular and set time intervals for creating expertise profiles [36]. This research, on the other hand, generates both short term profiles based on regular time-intervals, but also presents a method for identifying *time-windows*, where an expert exhibits "*peak activity*" in specific topics of expertise. These time-windows, which are of variable lengths, emerge as experts focus on specific activities and areas of interests. Thus, the time intervals depend on the temporal distribution of an expert's contributions, rather than on pre-configured timeframes.

Furthermore, most expertise profiling approaches consider *persistency* of a concept/topic to be an indication of its significance. However, this research considers *uniformity*, to be just 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.

1.8.3 Expertise profiling using micro-contributions in a range of knowledge domains

The Expertise Profiling Framework proposed in this thesis is domain-agnostic, i.e., applicable to all domains. Therefore, its applicability has been investigated in the context of different dynamic, knowledge-curation platforms. More specifically, the proposed STEP methodology has been studied in the context of: (i) *unstructured micro-contributions*, i.e., experts' micro-contributions target knowledge bases in natural language form; e.g., Wiki projects (Chapter 5) and (ii) *structured micro-contributions*, i.e., experts' micro-contributions target ontological concepts; e.g., micro-

contributions in the context of collaborative ontology engineering projects (Chapter 6). The different types of micro-contributions in the context of various knowledge domains provide different perspectives of STEP, which is used to design a framework that is applicable to all domains, i.e., domain-agnostic.

1.8.4 Creating expertise profiles at various levels of granularity

This thesis proposes methods for creating expertise profiles that describe the expertise and knowledge of individuals at various levels of granularity (Chapter 6), in order to: (i) represent expertise with a level of specificity that reflects the knowledge embedded in micro-contributions; (ii) facilitate comparison and evaluation of profiles that describe expertise at different levels of abstraction; (iii) customise the granularity of ontological concepts in expertise profiles; and (iv) complement experts' existing profiles with fine-grained domain concepts, representing the implicit knowledge embedded in their micro-contributions to evolving knowledge-curation platforms.

1.8.5 Combining contextual and content-based factors for expertise profiling

The Expertise Profiling Framework proposed in this thesis, identifies and analyses contextual factors embedded in existing social networks, to refine expertise profiles created from content-based factors (i.e., micro-contributions). Existing scientific and professional networks, such as *BiomedExperts* [37], provide a source for inferring implicit relationships between concepts in experts' profiles by analysing co-authorship relationships between experts. However, co-authorship reflects collaboration on static publications and resources. Other types of implicit relationships that exist between experts in a social network are often not taken into account. The profile refinement approach proposed in this thesis recognises both explicit relationships (e.g., *co-authorship* and *following*) and implicit relationships (e.g., relationships formed through participating in Q&A discussions and forums). The assumption is that experts contributing to the same topics have similar or related expertise. Furthermore, the context within which every micro-contribution is made, in addition to the experts who contribute to these contexts are identified. Expertise profiles of experts contributing to the same contexts, i.e., collaborators, are subsequently refined using the semantic relationships between concepts in their profiles and micro-contributions (Chapter 7).

1.8.6 Visualising time-aware expertise profiles

This thesis introduces the "*Profile Explorer*" visualization tool [35], which serves as a customizable interface to facilitate visualization, search and comparative analysis of expertise profiles. Profile Explorer enables the visualization of short term and long term profiles and provides a framework for conducting comparative analysis of experts and expertise by linking an expert's long term profile with his/her short term profiles and micro-contributions (Chapter 8).

1.9 Thesis Outline

This section provides a brief description of the remaining chapters of this thesis.

Chapter 2 discusses background research in the areas of collaboration platforms, ontologies, text analytics, expertise modelling and expertise profiling. It then discusses the application of these concepts to expertise modelling through the analysis of micro-contributions to evolving, collaborative knowledge-curation platforms.

Chapter 3 introduces the *Fine-grained Provenance Model for Micro-contributions*, which captures micro-contributions (including the actions that lead to their creation, as well as the context that hosts these contributions; i.e., sentence, paragraph or section of the document in which they appear). The model introduces an ontology that combines coarse and fine-grained provenance modelling to capture such artefacts and their localization in the context of their host living documents.

Chapter 4 introduces the *Semantic and Time-dependent Expertise Profiling*, (*STEP*), methodology for creating expertise profiles using *micro-contributions* to collaboration platforms, whilst also capturing the *dynamic* and *temporal* characteristics of expertise.

Chapter 5 discusses the application of the *Semantic and Time-dependent Expertise Profiling* (STEP) methodology to *unstructured* micro-contributions. Furthermore, this chapter discusses the integration of Language Models into STEP, in order to minimise the effects of domain-specific tools on the accuracy of resulting profiles. Experiments are performed on designated experts' micro-contributions to the *Molecular and Cellular Biology (MCB)* [38] and *Genetics* [39] Wiki projects (sub-projects of Wikipedia). In order to evaluate the original STEP methodology and the STEP methodology enhanced by integrating Language Models, experimental results are compared with the results generated both manually and by expertise profiling systems that use traditional IR techniques to analyse large corpora of static publications. This chapter evaluates the STEP profiles by: firstly comparing them against profiles manually generated by the authors when they first join these projects; and secondly comparing them with the results generated by the two traditional expertise profiling systems.

Chapter 6 discusses the application of the *Semantic and Time-dependent Expertise Profiling* (STEP) methodology to *structured* micro-contributions that have been generated during collaborative authoring of the *International Classification of Diseases revision 11 ontology*, (*ICD-11*) [24]. In addition, it demonstrates the use of *ontology structures* and *semantic similarity* for describing expertise at various levels of *granularity*. Furthermore, it showcases two major aspects:

(i) a novel semantic similarity metric, in addition to an approach for creating bottom-up baseline expertise profiles using expertise centroids; and (ii) the application of STEP in this new environment combined with the use of the same semantic similarity measure to both compare STEP against baseline profiles, as well as investigate the coverage of these baseline profiles by STEP.

Chapter 7 discusses the application of STEP in the *ResearchGate* [27] social expert platform and demonstrates how micro-contribution contexts and intrinsic and extrinsic contextual factors can be leveraged to improve the resulting profiles. In addition, it presents manual evaluation results computed with the assistance of nine ResearchGate experts.

Chapter 8 presents the *Profile Explorer* visualization tool, which serves as an extensible/customizable framework for exploring and analysing time-aware expertise profiles in knowledge bases where content evolves over time. Furthermore, it proposes a method, which uses the temporal aspect captured by the STEP model, to identify time-windows where an expert demonstrates peak activity in particular topics of expertise. Finally, it presents the results of a useability testing performed on Profile Explorer, in addition to identified strengths, limitations and future research directions.

Chapter 9 concludes the thesis by summarising the presented work and discussing its main original contributions, while presenting a series of insights gained from this research. Finally, the outstanding challenges and areas that require further investigation, improvement and development, are described.

Chapter 2 Foundational Aspects

This chapter provides a high-level overview of the key concepts upon which the work presented in this thesis is built. Section 2.1 provides an overview of Web and Web 2.0, traditional collaboration platforms and social networks. Section 2.2 provides an overview of ontologies, particularly in the expertise modelling and biomedical domains as well as semantic similarity techniques. Section 2.3 describes different approaches to text analytics. Section 2.4 describes expertise modelling in the context of Information Retrieval, Social Networks and the Semantic Web. Section 2.5 discusses various sources of knowledge in collaboration platforms, used in the expertise modelling framework proposed in this thesis. Section 2.6 discusses how the key concepts described in Sections 2.1-2.5 are applied to expertise modelling in community-driven and knowledge-curation platforms. Section 2.6 also highlights the limitations of existing approaches in the context of micro-contributions and identifies the major unresolved issues that provide the motivation for this thesis.

2.1 Social Collaboration platforms

2.1.1 From Web to Web 2.0

In recent years, there has been a transition from static HTML Web pages to a more dynamic Web that involves community-generated content and a greater focus on collaboration and sharing of information. Unlike the initial version of the Web, where the users were mainly "passive consumers" of content, users are now offered easy-to-use services that enable anyone to produce content and publish it on the Web. Mashups, blogs, wikis, feeds and social networking/tagging systems are all examples of such services. The Social Web is represented by a class of Web sites and applications in which user participation is the primary driver. The characteristics of such systems are well described by Tim O'Reilly under the banner of Web 2.0 [18]. In particular, Web 2.0 focuses on creating knowledge through collaboration and social interactions among individuals (e.g., Wikis) [40]. This increase in participation and content creation has given rise to large online volumes of information, from which knowledge and intelligence can be derived through the application of useful reasoning and data mining techniques.

2.1.2 Traditional Web Collaboration Platforms

Web 2.0 technologies have demonstrated the value of "crowdsourcing", i.e., harnessing users across the Internet to acquire information, expertise and ideas, help solve problems, accomplish objectives and foster innovation. Furthermore, collaboration platforms have emerged as a category of business software that adds broad social networking capabilities to work processes. The goal of

a collaboration software application is to foster innovation by incorporating knowledge management into business processes so employees can share information and solve business problems more efficiently.

With the emergence of Web 2.0, there has been a significant increase in online collaboration, giving rise to vast amounts of accessible and searchable knowledge in the context of platforms where content evolves through individuals' contributions. Blogs and Wikis are prime examples of collaboration through the Internet, a feature of the group interaction that characterizes the social Web [41]. Blogs and Wikis are used by individuals who contribute to the content as well as those who reference the content as resources. Blogs allow members to share ideas and other members to comment on those ideas, while Wikis facilitate group collaboration. Both of these tools open a gateway of communication in which social interaction leads to the ongoing development of the Web [42]. For example, the *RNA Wiki Project* [43] aims to better organise information in articles related to RNA on Wikipedia and AstraZeneca's science-focused blog, *LabTalk* [44] enables scientists, researchers and academics to discuss novel ideas, research and innovation. The knowledge in these platforms continuously evolves through experts' *unstructured* micro-contributions, i.e., micro-contributions in natural language form.

Discussions about the Social Web often use the phrase "collective intelligence" or "wisdom of crowds" to refer to the added value created by the collective contributions of all collaborators writing articles for Wikipedia, sharing tagged photos on Flickr, sharing bookmarks on Del.icio.us or streaming their personal blogs into the blogosphere [41].

The goal of the research in this thesis is to use content-based factors (i.e., micro-contributions) and contextual factors (i.e., collaborators' relationships) to profile the expertise of individuals, who contribute to the evolution of knowledge in collaboration platforms.

2.1.3 Social Expert Platforms

Collaborative Platforms on the Web can be investigated not only by considering the resulting knowledge, but also by looking at the social ties that connect the contributing members – or more concretely, by analysing the underlying social network. A social network is a social structure made up of a set of social actors (such as individuals or organizations) and a set of relationships between these actors. Social network analysis provides methods for analysing the structure of whole social entities as well as theories explaining the patterns observed in these structures. The study of these structures uses social network analysis to identify local and global patterns, locate influential entities, and examine network dynamics [45]. In particular, social networking in scientific communities enables experts and scientific groups to expand their knowledge base and share ideas. In addition, researchers and experts use social networks to maintain and develop professional

relationships, share knowledge and information and establish collaborations in common fields of expertise and interest. Below, four examples of different types of social expert platforms are described.

ResearchGate is a social networking site with more than 3 million scientists and researchers, who share papers and exchange domain-specific knowledge [27]. The site offers tools and applications for researchers to interact and collaborate. *Topics*, ResearchGate's Q&A forum, enables members to ask questions, get answers and share interesting content with one another about specific topics. ResearchGate has reported that approximately 12,342 questions were answered within their 4,000 topics in 2011 alone [46].

The *myExperiment* Virtual Research Environment (VRE) [47] is a joint effort from the universities of Southampton, Manchester and Oxford in the UK. It provides a social networking site for scientists, enabling researchers to share digital items associated with their research. In particular, it enables experts to share and execute scientific workflows and supports the individual scientist on their personal projects, forming a distributed community with scientists elsewhere who would otherwise be disconnected. *myExperiment* enables scientists to share, re-use and repurpose experiments, in order to reduce time-to-experiment, share expertise and avoid reinvention — and it does this in the context of the scholarly knowledge lifecycle. Hence myExperiment is a community social network, a market place, a platform for launching workflows and a gateway to other publishing environments. The myExperiment VRE has successfully adopted a Web 2.0 approach in delivering a social website where scientists can discover, publish and curate scientific workflows and other artefacts. It shares many characteristics of other Web 2.0 sites, such as providing users with a profile. However, features that distinguish myExperiment from other social networking sites, such as Facebook and Myspace, especially with respect to meeting the needs of its research user base include support for credit, attributions and licensing, fine control over privacy, a federation model and the ability to execute workflows.

Quora [48] is a question-and-answer website where questions are created, answered, edited and organized by its community of users. Quora aggregates questions and answers to topics. Users can collaborate by editing questions and suggesting edits to other users' answers. One thing that differentiates Quora from other question & answer platforms is how they incorporate the aspect of gamification into their platform. Quora users can easily earn credits by preforming the platforms' norms & prescriptions. For example, a user would be rewarded credits for providing a quality answer. With these credits, users are able to individually ask and compensate experts to answer a certain question. The aspect of being able to ask experts question in exchange for credits is extremely unique to Quora's platform.

The World Health Organization [25] is using Social and Semantic Web technologies to enable the collaborative development of the 11th revision of the International Classification of Diseases ontology (ICD-11) [24]. Health officials use ICD in all United Nations member countries to compile basic health statistics, monitor health-related spending, and to inform policy makers [49]. A large community of medical experts around the world is involved in the authoring of ICD-11 using a collaborative Web-based platform, called iCAT (ICD Collaborative Authoring Tool), a customisation of the generic Web-based ontology editor, WebProtégé [50]. To date, more than 270 domain experts around the world have used iCAT to author 45,000 classes, to perform more than 260,000 changes and to create more than 17,000 links to external medical terminologies [49].

2.2 Ontologies

An ontology is defined as a formal, explicit specification of a shared conceptualization [51]. In computer science and information science, ontologies are used to formally represent knowledge within a domain. Ontologies are the structural frameworks for organizing information and are used in artificial intelligence, the Semantic Web, systems engineering, software engineering, biomedical informatics, library science, enterprise bookmarking, and information architecture to formally represent knowledge about the world or some part of it. An ontology provides a common machine processible vocabulary to denote the *types*, *properties* and *relationships* of *concepts* in a domain [52]. In the Semantic Web domain, ontologies are represented using the Web Ontology Language (OWL) [53] and the Resource Description Framework (RDF) [54]. OWL is a family of knowledge representation languages or ontology languages for authoring ontologies or knowledge bases and RDF is a family of World Wide Web Consortium (W3C) [55] specifications used as a general method for defining concepts or modelling information about Web resources.

2.2.1 Ontologies for Expertise Modelling

Competence management is an important research topic in the more general area of knowledge management. Competence management can play a critical role at both an organizational and personal level, as it identifies the key knowledge that an employee or an organization should possess in order to achieve his/its targets [56]. Research has shown that competence and skills management can directly empower a company's workforce leading to an increase in the company's competitive advantage, innovation, and effectiveness [57]. Subsequently, Web data mining techniques (named entity recognition and co-occurrence data) have been employed to link the individuals in an organisation with expertise and associates [58]. Automatic topic extraction techniques have also been applied to scientific publications to streamline searches for competency management and expertise [59].

More specifically, a number of research efforts have focused on *ontology-based competency management*. In 2000, Sure et al proposed an approach that performs competency management by matching people to positions, providing more comprehensive knowledge about individuals' skills, using background knowledge from an ontology and secondary information such as project documents [60]. In 2007, Paquette proposed an ontology for designing competency-based learning and knowledge management applications and a software framework for ontology-driven e-learning systems [61]. This work identifies several performance indicators such as frequency, scope, autonomy, complexity and context for evaluating expertise [61]. In addition, in 2008 Heath & Motta [62], developed the Hoonoh ontology for describing trust relationships in the context of word of mouth information seeking. While the Hoonoh ontology is not specific to describing individuals' expertise, it does enable these relationships to be expressed, thereby making it suitable for use in expert-finding applications. It also provides the means to model a number of other relationships, which are highly relevant to applications and services in this domain

Unlike previous efforts, the Expertise Modelling Framework proposed in this thesis uses an ontology for capturing and representing the fine-grained provenance of micro-contributions in the living documents that host them. This ontology captures the exact placement of contributions in the underlying content at different levels of granularity, e.g., paragraph, section, sub-section, page, document. It also captures the actions that lead to the creation of micro-contributions, e.g., *update*, *delete* and *add* as well as document revisions resulting from such actions. It thus captures and represents the evolution of knowledge, which in turn facilitates capturing and tracking the changes in individuals' expertise and interests over time.

2.2.2 Biomedical Ontologies

Ontologies have grown to be one of the great enabling technologies of modern bioinformatics. They are used both as terminological resources and as resources that provide important semantic constraints on biological entities and processes [63]. Ontologies provide conceptual representations of the terms used within biomedical literature. The conceptual representation of the content of documents in turn enables development of sophisticated information retrieval tools for organising documents based on categories of information in the content [64, 65].

Over the past years, there has been an exponential growth in amount of biomedical and health information available in digital form. In addition to the 23 million references to biomedical literature currently available in PubMed [66], other sources of information are becoming more readily available. For example, digitisation efforts have resulted in the availability of large volumes of historical material and there is a wealth of information available in clinical records, whilst the growing popularity of social media channels has resulted in the creation of various specialised

groups. With such a deluge of information at their fingertips, domain experts and health professionals have an ever-increasing need for tools that can help them isolate relevant information in a timely and efficient manner. Consequently, enormous effort has been invested and progress has been made, in developing tools, methods and resources in the biomedical domain [67].

The Unified Medical Language System (UMLS) [68] is a compendium of controlled vocabularies maintained by the U.S. National Library of Medicine (NLM) [69], unifying over 100 dictionaries, terminologies, and ontologies in its Metathesaurus. Overall, NLM provides over 200 knowledge sources and tools that can be used for text mining. Other sets of ontologies that are maintained through collaborative effort include the OBO Foundry [70] and the National Centre for Biomedical Ontology (NCBO) [71].

The *International Classification of Diseases, revision 11, (ICD-11) ontology* [24], is currently under active development. International Classification of Diseases is the standard diagnostic classification developed by the World Health Organisation (WHO) [25] to encode information relevant for epidemiology, health management and clinical use [72]. The knowledge-curation process of the ICD-11 ontology is done in a collaborative manner by experts from diverse institutions around the world. Each expert contributes to this process by authoring (i.e., creating, modifying, removing) ontological concepts.

The proposed Expertise Modelling Framework that is the focus of this thesis, is applied and evaluated using *structured* micro-contributions generated within the context of collaborative authoring of the *ICD-11* ontology [24].

Moreover, the expertise modelling framework proposed in this thesis employs ontologies in multiple ways: (i) ontologies are used to annotate the text chunk or context that encapsulates a micro-contribution in order to map expertise topics to domain concepts; (ii) ontologies provide the means to identify and group lexically different, but semantically similar terms and represent them using domain concepts, e.g., “diabetes” and “high blood sugar” are both manifestations of the concept “diabetes mellitus” from the Human Disease Ontology; (iii) representing expertise topics using ontological concepts facilitates the refinement of expertise profiles based on the semantic relationships between concepts that represent the expertise of collaborating experts; (iv) expertise profiles containing ontological concepts can be published and integrated as structured data on the Web, making them more visible to “*expertise seekers*” and “*Web crawlers*”; and (v) analysis and comparison of concepts in expertise profiles with concepts in the Linked Data Cloud [28] provides access to a richer, more accurate and more up-to-date set of concepts representing the expertise of individuals.

2.2.3 Semantic Similarity

Measuring semantic similarity is a critical step when trying to align documents that are described using ontological concepts. For example, an assessment of concept alikeness improves the understanding of textual resources and increases the accuracy of knowledge-based applications [73]. The adoption of ontologies during annotation provides a means to compare entities on aspects that would otherwise be difficult to compare. For instance, if two gene products are annotated using the same schema, they can be compared by comparing the terms with which they are annotated. While this comparison is often done implicitly (for instance, by finding the common terms in a set of interacting gene products), it is possible to perform an explicit comparison using semantic similarity measures [74]. In general, a semantic similarity measure is a function that, given two ontology terms or two sets of terms annotating two entities, returns a numerical value reflecting the closeness in meaning between them. Several approaches have been defined for quantifying semantic similarity, the two most prominent ones being: (i) node-based, in which the main data sources are the ontological concepts and their properties; and (ii) edge-based, which uses the edges between the ontological concepts and the edge *types* as the data source. Note that there are other approaches for comparing terms that don't use semantic similarity; for example, systems that select a group of terms, which best summarise or classify a given subject based on the discrete mathematics of finite partially ordered sets [73].

Node-based approaches rely on comparing the properties of the terms involved, which can be related to the terms themselves, their ancestors, or their descendants. One concept commonly used in these approaches is *Information Content* (IC) [75], which provides a measure of how specific and informative a term is. Information Content-based (IC) approaches assess the similarity between concepts as a function of the Information Content shared between the concepts. The amount of shared information is represented by the IC of their Least Common Subsumer (LCS) - i.e., the most specific taxonomical ancestor of the two concepts in a given ontology [75]. IC quantifies the semantic content of a concept and incorporates taxonomical evidence explicitly modelled in ontologies (such as the number of leaves/hyponyms (specialisations) and ancestors/subsumers). The IC of a concept can be either computed from its probability of occurrence in a corpus (i.e., frequently appearing concepts have lower IC), or from its degree of taxonomical specialisation in the background ontology (i.e., the larger the number of hyponyms (subclasses) of a concept, the more general its meaning and the lower its IC). Pure ontology-based approaches, like the latter one, are preferred to corpora-based ones due to their higher scalability.

Edge-based approaches rely on the structural model defined by the taxonomical relationships in the ontology. These approaches base the similarity assessment on the length of the shortest path

separating two concepts, defined by going through taxonomical generalisations modelled in the ontology [76]. The shortest taxonomical path between two concepts is the one that goes through their Least Common Subsumer (LCS), which also represents their commonality. The following lists some of the well-known edge-based similarity measures. Rada [76], is a simple edge-counting measure, which quantifies the semantic distance between two concepts C1 and C2 as the sum of the number of links from C1 and C2 to their LCS; i.e., their minimum taxonomical path (Eq. 2-1).

$$Dis_{Rada}(C_1, C_2) = N_1 + N_2 \quad \text{(Eq. 2-1)}$$

Where N_1 and N_2 represent the number of links from C1 and C2 to their LCS, respectively.

Leacock and Chodorow [77] normalise the value by the maximum depth of the taxonomy (D), evaluating the path length in a non-linear fashion (Eq. 2-2).

$$Sim_{L\&C} = -log\left(\frac{N_1 + N_2 + 1}{2D}\right) \quad \text{(Eq. 2-2)}$$

Where N_1 and N_2 represent the number of links from C1 and C2 to their LCS, respectively and D represents the maximum depth of the taxonomy.

Wu and Palmer [78] consider the relative depth of the LCS of concept pairs in the taxonomy as an indication of similarity (Eq. 2-3).

$$Sim_{W\&P} = -log\left(\frac{2N_3}{N_1 + N_2 + 2N_3}\right) \quad \text{(Eq. 2-3)}$$

Where N_1 and N_2 represent the number of links from C1 and C2 to their LCS, respectively and N_3 represents the relative depth of the LCS of concept pairs in the taxonomy.

Other approaches also use path length in addition to other structural characteristics of a taxonomy, such as the relative depth of concepts, and local densities of taxonomical branches. Because several heterogeneous features must be evaluated, these approaches assign weights to balance the contribution of each feature in the final similarity value. These measures, also considered to be hybrid approaches, depend on the empirical tuning of weights according to background ontology and input terms, resulting in ad hoc solutions that cannot be easily generalised [73]. The main advantage of edge-counting measures is their simplicity. However, edge-based approaches are based on two assumptions that are seldom true in ontologies: (i) nodes and edges are uniformly distributed; and (ii) edges at the same level in the ontology correspond to the same

semantic distance between terms. Furthermore, terms at the same depth do not necessarily have the same specificity or semantics, and edges at the same level do not necessarily represent the same semantic distance [74].

In the context of this thesis, semantic similarity measures are used to customise the granularity of expertise profiles generated by the proposed framework, in order to: (i) represent expertise with a level of specificity which accurately represents the knowledge conveyed in micro-contributions; (ii) facilitate comparison of profiles describing expertise at different levels of abstraction; and (iii) investigate the coverage and alignment between expertise embedded in micro-contributions and expertise profiles created by the proposed framework.

2.3 Text Analytics

Text analytics, refers to the process of deriving high-quality information from text (relevant, novel and interesting), by detecting patterns and trends using methods such as statistical pattern learning [79]. Text mining typically involves the process of: structuring the input text (by parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database); deriving patterns within the structured data; and finally evaluation and interpretation of the output [80]. The following provides a high level overview of methods used in text mining.

2.3.1 Natural Language Processing in the Biomedical Domain

Text analysis involves a wide range of technologies including: information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization, and predictive analytics. The overarching goal is to turn text into data for analysis, via the application of natural language processing (NLP) and analytical methods [80].

Within the biomedical domain, the widespread application of high-throughput techniques, such as gene and protein analysis, has generated massive volumes of data. This growth is accompanied by a corresponding increase in associated biomedical literature, in the form of articles, books and technical reports. In order to organize and manage this data, manual curation efforts have been established e.g., to identify entities (e.g., genes and proteins) [81] and their interactions (e.g., protein-protein) [82]. However, manual annotation of large quantities of data is a very demanding and expensive task, making it difficult to maintain the annotation of these databases. These factors have naturally led to increasing interest in the application of text mining systems to help perform those tasks [83]. One major focus has been on Named Entity Recognition (NER), the task of identifying words and phrases in free text that belong to certain classes of interest [84]. The

development of NER and normalization solutions requires the application of multiple techniques, which can be conceptualized as a simple processing pipeline [85]. This design improves system robustness, i.e., one could replace one module with another (possibly superior) module, with minimal changes to the rest of the system [86]. This is the intention behind pipelined NLP frameworks, such as GATE [87], IBM (now Apache) Unstructured Information Management Architecture (UIMA) [88] and the Natural Language Toolkit (NLTK) [198].

In the context of this thesis, as described in Chapter 1, micro-contributions don't offer sufficient context (due to their short and sparse content) for the analysis performed by NLP techniques. Rather, a model is required to complement the content of micro-contributions, without considering the whole content of their host documents (Chapter 3). Moreover, a methodology is required that can extract the semantics conveyed by micro-contributions (Chapter 4).

2.3.2 Concept Recognition

In the domain of biomedical informatics, the task of concept recognition involves mapping biomedical text to a representation of biomedical knowledge consisting of inter-related concepts, usually codified as an ontology or a thesaurus [89]. Despite the ever-increasing amount of biomedical literature and resources and the availability of biomedical ontologies through BioPortal [90], manual ontology-based annotations are unlikely to scale. This is mainly due to the large number of biomedical ontologies, which are often subject to ongoing changes and frequently contain overlapping concepts.

The National Centre for Biomedical Ontology (NCBO) [71] is a leading scientific organization that is applying semantic technologies to biomedicine. One of the main objectives of NCBO is to build tools and Web services to enable the use of ontologies and terminologies. The centrepiece of NCBO is the BioPortal – a Web-based resource that makes more than 270 biomedical ontologies and terminologies available for research. In addition to providing a comprehensive library of biomedical ontologies and terminologies, the NCBO develops tools and services that use those ontologies to aid biomedical investigators in their work. These tools are all available through a Web-browser interface, as well as programmatically via Web services [91]. In particular, NCBO has developed the Open Biomedical Annotator (NCBO Annotator) Web Service [92], enabling end users to utilise ontologies (from *UMLS* [68] and *BioPortal* [90]) for annotation of biomedical resources with minimal effort [89].

Within this thesis, the NCBO Annotator web service is used to map arbitrary keywords and natural language text occurring in micro-contributions to standardized ontological terms. Figure 2-1 depicts an example of annotations derived from an expert's micro-contribution.

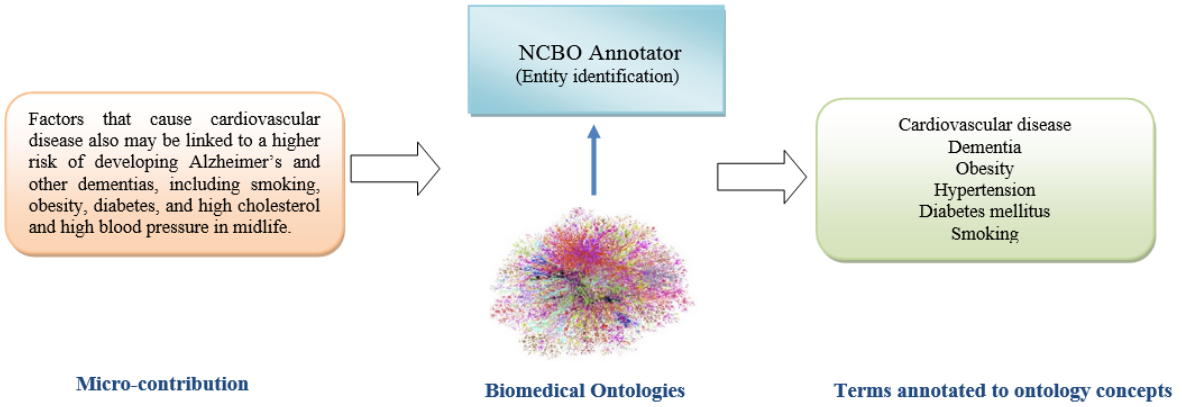

Figure 2-1: Example of annotations derived from a micro-contribution

The NCBO Annotator takes as input some specified text and generates as output a set of terms derived from BioPortal-stored ontologies, such that the terms refer to concepts that the NCBO Annotator identifies in the text. It provides a mechanism to determine what the text is 'about' in terms of standardized, ontological entities. The structure of the ontologies in BioPortal [90] permits the NCBO Annotator to associate the text not only with particular terms (e.g., *Coronary Heart Disease* from the *Experimental Factor Ontology* [93]), but also with more general terms (e.g., *Cardiovascular Disease*). This provides access to a rich set of descriptors representing the semantics of micro-contributions at different levels of granularity and generality.

Micro-contributions are annotated using the ontologies stored in BioPortal. The NCBO Ontology Recommender Service [94] is used to determine the ontologies that provide the best coverage for capturing the entities in a micro-contribution. This service takes as input the micro-contribution text and returns as output an ordered list of ontologies available in BioPortal, the terms of which would be most appropriate for annotating the corresponding text. In all experiments performed in this thesis, the five most highly ranked ontologies identified by the recommender service are used to generate annotations. Thus, terms identified in micro-contributions are often mapped to domain concepts from different ontologies. Figure 2-2 illustrates an example of multiple annotations for a single term, i.e., the term "Cardiovascular Disease" is mapped to related concepts in three different ontologies.

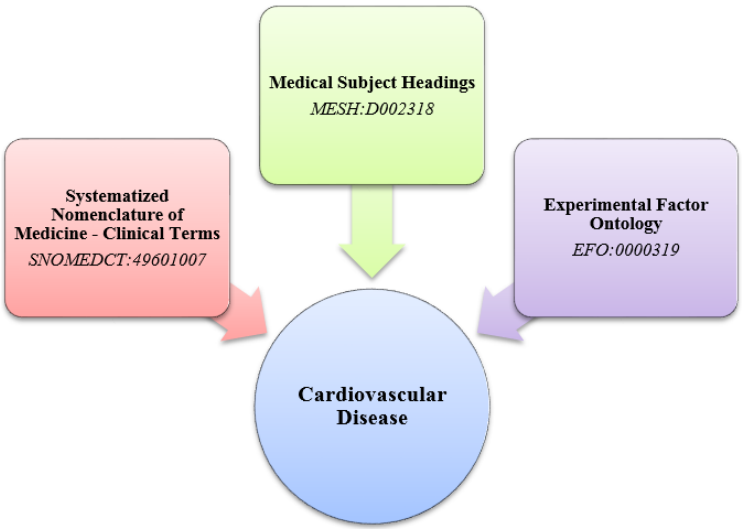

Figure 2-2: Example of annotations from multiple ontologies

2.3.3 Statistical Language Modelling

The goal is to provide a methodology for capturing micro-contributions and creating profiles, whilst ensuring that the methodology is not restricted to specific tools or frameworks of a particular domain. Therefore, the concept extraction process should not be limited to a particular tool or technique. The NCBO Annotator's underlying technology is similar to most concept recognizers, however, it's predominantly used in the biomedical domain and therefore, using it as the only means of extracting concepts from micro-contributions, will result in expertise profiles, which are heavily dependent on the accuracy of annotations produced by the NCBO Annotator. Consequently, in order to reduce the effects of domain-specific annotation tools on the accuracy of the generated profiles, *Language Models* [31] are incorporated into the expertise profiling methodology. The terms generated by applying language models to micro-contributions are subsequently combined with terms annotated by the NCBO Annotator [92], for modelling the expertise of contributing experts.

A statistical language model assigns a probability to a sequence of words by means of a probability distribution. The experiments described in this thesis use Topic Modelling [32] and N-gram Modelling [33] techniques. Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents. Topic modelling algorithms can be adapted to many kinds of data. Among other applications, they have been used to find patterns in genetic data, images and social networks [95]. N-gram models are analogous to placing a small window over a sentence or a text, so that only *n* words are visible at a time. The simplest n-gram model is therefore a so-called unigram model. This is a model which only looks at one word at a time [33].

In the context of this thesis, Topic Modelling and N-gram Modelling have been integrated with the proposed expertise profiling methodology, in order to minimise the effects of domain-specific annotation and concept extraction tools and techniques on the resulting profiles. More specifically, as outlined in Chapter 5, micro-contributions are lemmatised followed by (i) Topic Modelling and (ii) N-gram Modelling, in two separate experiments. The resulting topics and n-grams are then mapped to concepts from domain-specific ontologies, using the NCBO Annotator. The aim of this experiment is to use domain-independent methods for identifying expertise topics, thus reducing the effects of concept recognition inefficiencies that may exist in domain-specific tools. Chapter 5 includes a detailed discussion on some of the extraction inefficiencies associated with the NCBO Annotator.

2.4 Expertise Modelling

Traditional Expertise Retrieval techniques model the associations between query topics and people and rank topics based on the strength of their association with an individual. The two most popular and well performing approaches in the TREC (Text Retrieval Conference) expert search task [10] are profile-centric and document-centric approaches. Profile-based methods create a textual representation of a person's knowledge according to the documents with which they are associated [96]. These representations i.e., "pseudo documents" can then be ranked using standard document retrieval techniques. These representations are built irrespective of queries, therefore these models are also referred to as query-independent approaches [97]. Document-based methods, also referred to as query-dependent approaches [97], do not directly model the knowledge of a person. They first find documents relevant to the query and then rank candidates mentioned in these documents based on a combination of the document's relevance score and the degree to which the person is associated with that document. A person, therefore, is represented by a weighted set of documents. There are also hybrid methods that build candidate profiles in a query-dependent way – such as the previous research that models documents as mixtures of persons [98, 99].

Traditional approaches rely on associations between people and documents. For example, a person who is associated with a document on a given topic is more likely to be an expert on the topic than a person who is not associated with documents on that topic. Document-candidate associations are represented in different ways; however, in general, these associations are established in two steps: (i) for every document in a collection, the set of candidates that are associated with that document, are identified (e.g., authors or people mentioned in the content), and (ii) for each of the document-candidate pairs identified, the strength of the association is estimated (e.g., by considering other documents associated with the candidate). Other approaches consider co-occurrence information of person mentions and query words in the same context as evidence of

expertise [97]. A number of studies use the co-occurrence model and techniques such as Bag-of-Words [100] or Bag-of-Concepts [101] on documents that are typically large and rich in content. A common method is to apply a weighted, multiple-sized, window-based approach in an information retrieval (IR) model to association discovery [102]. The effectiveness of exploiting the dependencies between query terms for expert finding has also previously been demonstrated [103]. Other studies present solutions that combine the use of ontologies and techniques such as *spreading*, to link additional related terms to a user profile by referring to background knowledge [104].

The following describes previous approaches to expertise modelling.

2.4.1 Expertise Retrieval using Content-based Features

A number of studies have focused on the automatic generation of expertise profiles using publications and static documents. In particular, the Ulm Rare Disease Centre is developing an automated system which employs bibliometric analyses to discover, retrieve and continuously update information on rare disease experts [105]. Another study has implemented a researcher network knowledge base by integrating publications from the Digital Bibliography & Library Project (DBLP) computer science bibliography as well as researcher Web pages [106]. Furthermore, the agent-based approach for finding experts within knowledge intensive organisations [107] and the semantic repository approach for locating academic experts [108] both partly rely on publication analysis. The central premise of these approaches is that if a person has (co)authored a significant number of publications on a specific subject, then this person can be seen as a potential expert in that subject [106, 109]. However, bibliometric analysis can only provide insights on experts who actively publish. Experts with no publishing activity are unlikely to be discovered. Additionally, not every author may be an actual expert on the research topic, e.g., in the case of honorary authorships [105].

Another study focuses on expertise retrieval within a bounded organizational setting (intranet) that differs from the W3C [110] setting—one in which relatively small amounts of clean, multilingual data are available, that cover a broad range of expertise areas, as can be found on the intranets of universities and other knowledge-intensive organizations [111]. Typically, this setting features several additional types of structure: topical structure (e.g., topic hierarchies as employed by the organization), organizational structure (faculty, department), as well as multiple types of documents (research and course descriptions, publications and academic homepages). The study focuses on a number of research questions: Does the relatively small amount of data available on an intranet affect the quality of the topic-person associations that lie at the heart of expertise retrieval

algorithms? How do state-of-the-art algorithms developed on the W3C data set perform in the alternative scenario? Do the lessons from the Expert Finding task at TREC carry over to this setting? How does the inclusion or exclusion of different documents affect expertise retrieval tasks? How can the topical and organizational structure be used for retrieval purposes?

2.4.2 Expertise Retrieval using Online Discussions

Algorithms have also been 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. These prior efforts have been leveraged 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. [112]. A related past initiative is the Web People Search task, which was organized as part of the SemEval-2007 [113] 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 focus of this effort is on *people name disambiguation* and not *expert finding*. The INEX initiative [114, 200], 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 their objective, INEX aims to build a gold standard via manually- and voluntarily-defined expertise profiles generated by Wikipedia users.

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 [120]. The feasibility of linking individual tweets with news articles has also been analysed for enriching and contextualizing the semantics of user activities on Twitter in order to generate valuable user profiles for the Social Web [121]. This analysis has revealed that the exploitation of tweet-news relations has significant impact on user modelling and allows for the construction of more meaningful representations of Twitter activities.

As with other traditional IR methods, this study [121] applies bags-of-words (BOW) [100] and TF-IDF [117] 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 the 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 (max. 140 characters) 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 the Expertise Profiling Framework proposed in this thesis specifically aims to capture expertise in the context of *evolving* knowledge.

Another study [123] leverages the appearance of user traces in the form of linked data for expert finding. It examines how Linked Data metrics, which reveal the constitution of a linked dataset (or set of datasets), could help to detect a good type of user trace to use for expert finding, and thus help the user prioritize those expertise hypotheses that rely on this particular type of trace.

2.4.3 Expertise Retrieval Software

Previous research that falls within the same category of expertise finding as this thesis, is *SubSift* (short for submission sifting) [115]. Subsift is a family of *RESTful* Web services [116] for profiling and matching text. 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 relies on significant volumes of data and uses traditional IR techniques such as *Term Frequency (TF) – Inverse Document Frequency* (*IDF*) [117], *Bag-of-Words* (BOW) [100] and *Vector-based Modelling* [118] to profile and compare collections of documents.

The Entity and Association Retrieval System (EARS) [122], is an open source toolkit for entity-oriented search and discovery in large test collections. EARS, implements a generative probabilistic modelling framework for capturing associations between entities and topics. Currently, EARS supports two main tasks: (i) finding entities ("which entities are associated with topic X?") and; (ii) profiling entities ("what topics is an entity associated with?"). EARS employs two main families of models, both based on generative language modelling techniques, for calculating the probability of a query topic (q) being associated with an entity (e), P(q|e). According to one 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.

The *ExpertFinder* framework uses and extends existing vocabularies that have attracted a considerable user community already such as FOAF, SIOC, SKOS and DublinCore [119].

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 [2]. 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.

2.4.4 Expertise Retrieval using Contextual Factors

While current Expertise Retrieval efforts focus on the task of expertise mining using content-based factors, a number of recent research efforts have emerged which consider the problem of expertise mining from several other perspectives, including contextual factors. As a result, content-based, expert finding approaches have been extended with contextual factors that have been found to influence human expert finding. In particular, one study [14] analyses a community of science communicators in a knowledge-intensive environment. Given an example expert, the aim is to find similar experts, by combining expertise-seeking and retrieval research. First, a user study is conducted to identify contextual factors that may play a role in the specific goal and environment. Then, expert retrieval models are designed to capture these factors, combined with content-based retrieval models and evaluated in a retrieval experiment. The main finding is that while content-based features are the most important, human participants also take contextual factors into account, such as media experience and organizational structure. Experiments demonstrate that models combining content-based and contextual factors can significantly outperform content-based models.

Similarly, *SmallBlue,* a social-context-aware expertise search system, mines an organisation's electronic communication to provide expert profiling and expertise retrieval. Both textual content of messages and social network information (patterns of communication) are used [98, 124].

Another study [199] proposes a novel approach to expert finding in large enterprises or intranets by modelling candidate experts (persons), organizational documents and various relations among them with so-called expertise graphs. As distinct from the state-of-the-art approaches estimating personal expertise through one-step propagation of relevance probability from documents to the related candidates, this method is based on the principle of multi-step relevance propagation in topic-specific expertise graphs.

2.4.5 Expertise Retrieval using Social Factors

Several research efforts have focused on expertise modelling using Social Network Analysis techniques [125]. The majority of such research efforts analyse each person's local information and relationships separately and combine them in an ad-hoc approach. For example, the issue of expert finding has been investigated in an email network [126]. This study utilizes the link between authors and receivers of emails to improve the expert finding result. In addition, link-based algorithms, e.g., PageRank [127] and HITS [128], can be used to analyse the relationships in a social network, which might improve the performance of expert finding. However, a problem common to both PageRank and HITS is topic drift. Because they give the same weight to all edges, the pages with the most in-links in the network being considered tend to dominate, whether or not they are the most relevant to the query.

Existing social networks such as *BiomedExperts* (BME) [37] provide a source for inferring implicit relationships between concepts within expertise profiles by analysing 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 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.

SciVal Experts [129] is a resource for finding experts and fostering collaboration. It creates researcher profiles with automatically updated publication and grant information and faculty-generated curriculum vitae, capturing a more comprehensive view of a researcher's body of work. Powered by the Elsevier Fingerprint Engine [130], SciVal Experts scans and analyses every publication in the Scopus database [131], creating Fingerprints of individual researcher's expertise and exposing connections among authors. Similar to BiomedExperts [37], it analyses large corpora of static documents and connects researchers based on co-authored publications.

Profiles Research Networking Software (RNS) [132] is an open source tool which aims to speed the process of finding researchers with specific areas of expertise for collaboration and professional networking. Profiles RNS analyses publication data to define a researcher's professional interests with a set of prioritized keywords. In addition, it automatically creates networks based on current or past co-authorship history, organizational relationships and geographic proximity and extends these networks by discovering new connections, such as identifying "similar people" who share related keywords. Furthermore, users can manually create active networks by identifying advisor, mentor and collaborator relationships to colleagues.

Within [133], a hybrid approach has been proposed for integrating topic identification and community detection techniques, recognising that communities and topics are interwoven and co-evolving. While most scientometric evaluations of topics and communities have been conducted independently and synchronically, this study examines the dynamic relationship between topics and communities. The hybrid approach demonstrates the interactive nature of topics and communities, confirming that topics can be used to understand the dynamics of community structures, leading to an enhanced understanding of a particular domain.

Another approach to expert finding in a social network takes into consideration not only each person's local information but also relationships between persons [5]. This study consists of two steps: (i) *Initialization* and (ii) *Propagation*. In *Initialization*, each person's local information is used to calculate an initial expert score for each person. The basic idea in this stage is that if a person has authored many documents on a topic or if the person's name co-occurs many times with the topic, then it is likely that he/she is a candidate expert on the topic. The strategy for calculating the initial expert scores is based on the probabilistic information retrieval model. For each person, a 'document', d, is first created by combining all his/her person local information. It estimates a probabilistic model for each 'document' and uses the model to calculate the relevance score of the 'document' to a topic. The score is then viewed as the initial expert score of the person. In *Propagation*, it makes use of relationships between persons to improve the accuracy of expert finding. The basic idea here is that if a person knows many experts on a topic or if the person's name co-occurs many times with another expert, then it is likely that he/she is an expert on the topic. This research proposes a propagation-based approach based on propagation theory [134]. It views the social network as a graph. In the graph, a weight is assigned to each edge to indicate how well the expert score of a person propagates to its neighbours. These so-called propagation coefficients range from 0 to 1 inclusively and can be computed in many different ways. Experimental results show that the proposed approach outperforms the baseline, which only considers each person's local information.

Another investigation [135] studies the problem of topic-level expert finding within a citation network. This study proposes a topical and weighted factor graph (TWFG) model to combine all the candidates' personal information (i.e., topic relevance and expert authority) and the scholarly network information (i.e., citation relationships) in a unified way.

2.4.6 Expertise Retrieval in the Semantic Web

In the Semantic Web domain, expertise modelling involves capturing expertise using ontologies or inferring it via axioms and rules defined over instances of these ontologies [15]. In particular, the *Saffron* system [6], provides insights into a research community by analysing their

main topics of investigation and the individuals associated with these topics. The Saffron system performs expert finding and profiling by extracting terms from text, at a level of specificity, which describes areas of expertise accurately. A graph-based algorithm is employed to construct topical hierarchies using only domain corpora. The knowledge of an expert is estimated using topical hierarchies, based on how well they cover subordinate expertise topics [136].

ourSpaces [137] is a Virtual Research Environment that makes use of Semantic Web technologies to create a platform for supporting multi-disciplinary research groups. The main semantic components of the system are a framework for capturing the provenance of the research process, a collection of services to create and visualise metadata and a policy reasoning service. The ontological support in *ourSpaces* facilitates capturing entities such as artefacts, people and processes and the links between them. This 'linked data' approach makes additional aspects of information discovery and presentation possible within ourSpaces.

eagle-i [138] is an ontology-driven framework for biomedical resource curation and discovery, focusing on resources that are commonly generated but rarely shared, e.g., reagents, protocols, instruments, expertise, organisms, and biological specimens. The framework aims at collecting information about "invisible" research resources and adding value to resource data by identifying and documenting meaningful semantic relationships between them. *eagle-i* aims to enhance resource discovery and interoperability by adopting existing biomedical vocabularies and ontologies and linking content in public repositories.

VIVO [139] is an open source *Semantic Web* platform that enables the discovery of research and scholarship across disciplinary and administrative boundaries through interlinked profiles of people and other research-related information. VIVO is populated with information about researchers, allowing them to highlight areas of expertise, display academic credentials, and visualize academic and social networks and display information such as publications, grants, teaching, service, and more. VIVO and other compatible applications produce a rich network of information across institutions, organizations, and agencies that can be searched to foster collaboration and enable open research discovery. VIVO provides network analysis and visualization tools to maximize the benefits afforded by the data available in VIVO.

Recently the eagle-i [138] and VIVO [139] projects have been coordinating efforts in order to address overlapping areas of interest. The Clinical and Translational Science Awards (CTSA) [140] program managed through the National Centre for Advancing Translational Sciences (NCATS) [] is dedicated to improving the sharing of resources and clinical expertise in support of translational science. To this end, they have recently funded *CTSAconnect* [141], a project that will integrate information about research resources (captured by eagle-i) and researcher profiles (captured by

VIVO) into one single ontology suite, i.e., the *Integrated Semantic Framework* (ISF). This new framework will extend coverage to include representation of clinical encounters and develop a data model and algorithms for computing practitioner expertise and publishing it as Linked Data [28].

2.5 Knowledge Sources in Collaboration Platforms

The expertise profiling framework proposed in this thesis uses various sources of implicit knowledge embedded in collaboration platforms, each of which provides a different perspective of the model, enabling the design of an abstraction layer that will render the final model into a domain-agnostic form. The following sub-sections describe various sources of implicit knowledge analysed by the expertise profiling model proposed in this thesis.

2.5.1 Unstructured Micro-contributions

The content of collaborative knowledge bases is dynamic and subject to continuous evolution through experts' micro-contributions. In the context of collaboration platforms such as the Molecular and Cellular Biology (MCB) [38] and Genetics [39] Wiki projects, the underlying content evolves through experts' *unstructured* micro-contributions, comprising short fragments of text in natural language form (Figure 2-3). In this thesis, ontology-based annotation of unstructured micro-contributions is performed to derive the semantics of knowledge contributed by experts. Furthermore, language modelling techniques are applied to extract topics and terms from unstructured contributions, in order to complement and reduce the impact of domain-specific annotation tools.

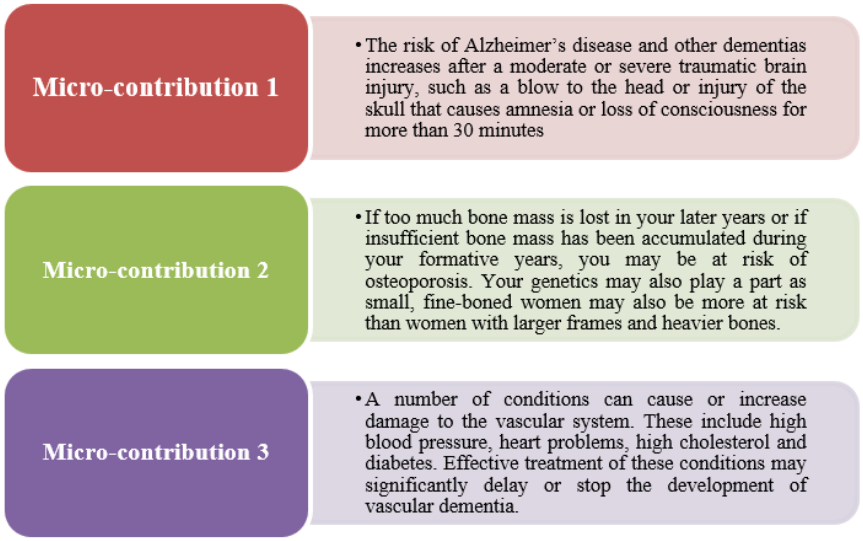

Figure 2-3: Examples of unstructured micro-contributions

2.5.2 Structured Micro-contributions

Building ontologies in a collaborative and increasingly community-driven fashion has become a central paradigm of modern ontology engineering. This collaborative approach to ontology

engineering is the result of intensive theoretical and empirical research within the Semantic Web community, supported by technology developments such as Web 2.0 [142]. In this context, experts contribute and author ontological concepts, which given their intrinsic nature, could be viewed as *structured* micro-contributions to the development of the underlying ontology.

Within this thesis, the proposed expertise profiling model is applied to and evaluated using structured micro-contributions made to the International Classification of Diseases ontology, revision11 (*ICD-11*) [24]. The International Classification of Diseases (ICD) is the foundation for the identification of health trends and statistics globally. It is the international standard for defining and reporting diseases and health conditions. The 11th version, ICD-11, is in development phase and due to be finalized in 2017 [72]. Figure 2-4 depicts a snapshot of the ICD-11 ontology.

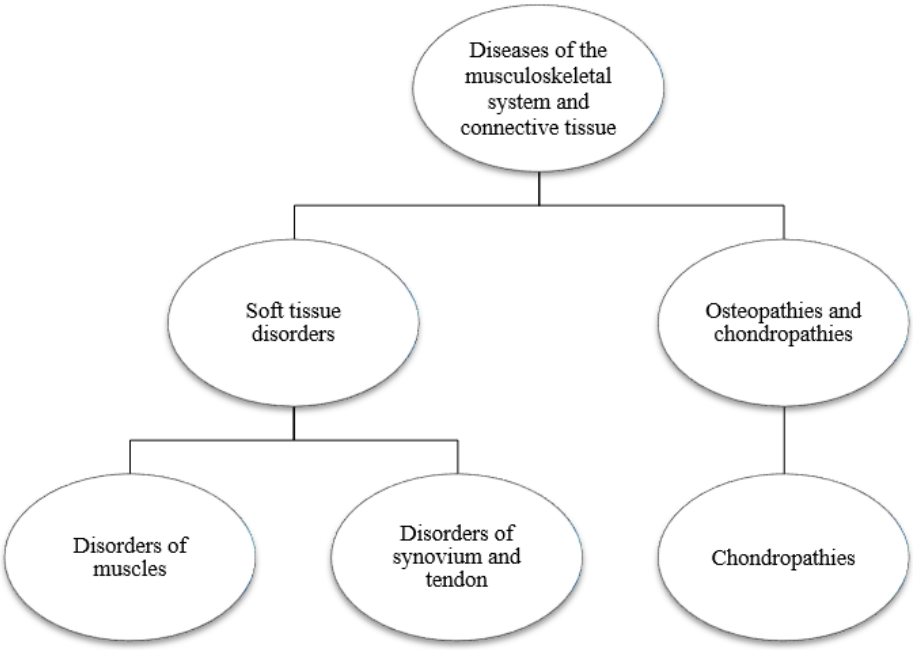

Figure 2-4: A Snapshot of the ICD-11 Ontology

2.5.3 Micro-contribution Contexts

The expertise profiling model proposed in this thesis analyses existing collaboration networks and processes micro-contributions – taking into account the *context* in which they occur, e.g., an answer provided by an expert as a contribution towards a question raised by another expert, is processed taking into account both the original question and all of the other answers to the question. Furthermore, expertise profiles are refined using the expertise and the strength of relationships among collaborating experts. For example, *co-authorship*, *following/follower* and *context collaborator* relationships are all recognized. Context collaboration refers to ad hoc relationships formed during discussions on common topics, i.e., Q&A discussions. Figure 2-5 depicts an example of profile refinement. In Figure 2-5, the expertise profile of Expert1 is refined based on the

expertise of his/her collaborators. Profile refinement is also performed by analysing semantic relationships between concepts in collaborators' profiles. For example, the hierarchical structure of the Bone Dysplasia ontology [143], determines that "*Short-rib Dysplasia*" in the profile of Expert1 and "*Acromelic Dysplasias*" in the profile of Expert3 share a common superclass, i.e., "*Bone Dysplasia*", which is added to the refined profile of Expert1. The concept "*Bowed Legs*" from the Human Phenotype Ontology [144] has been assigned a higher weight, as it is a topic of expertise shared by Expert1 and all his/her collaborators. A detailed description and discussion of algorithms and mechanisms used in profile refinement is outlined in Chapter 7 of this thesis.

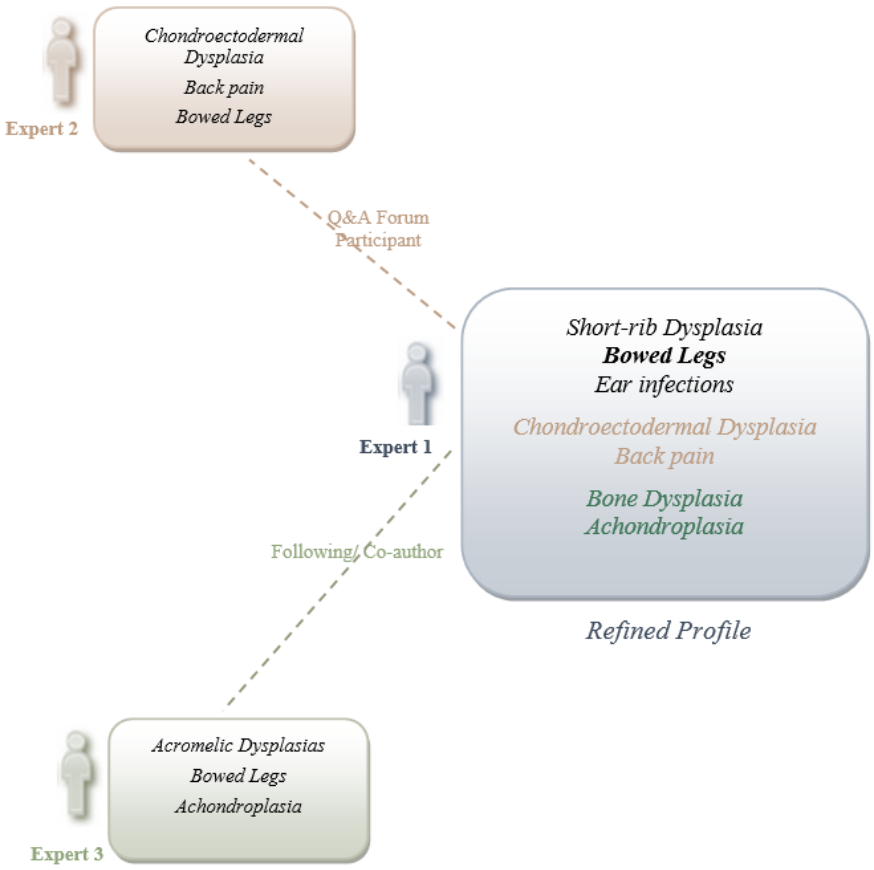

Figure 2-5: Example of Profile Refinement using Social Collaboration Factors

2.6 Discussion

Despite significant previous research focusing on expertise finding and expertise profiling, modelling expertise in the context of collaboration platforms still presents a range of unresolved issues and challenges. Current research efforts primarily take a document-centric view of static documents authored or co-authored by an expert such as publications, grants and reports. Such techniques adopt a macro-perspective of documents and associate individuals with expertise topics that emerge from the entire content of these documents. The macro-perspective is unable to associate individuals to their micro-contributions in the content of documents and thus, cannot

provide detailed evidence of expertise. In addition, current approaches to expertise modelling rely on analysing large corpora of static documents. Additionally, traditional expertise retrieval relies on analysing static documents, i.e., documents where content does not evolve, as once written, the documents remain fixed forever in the same form. Consequently, such techniques are unable to capture and track the changes in evolving knowledge or changes in the expertise and interests of contributors.

This thesis proposes an innovative framework for modelling expertise in *collaborative, knowledge curation platforms*, where knowledge is *dynamic* and continuously *evolves* through incremental refinements to content or *micro-contributions.* The proposed model facilitates *individual attribution*, i.e., the expertise of individuals is modelled, based on the knowledge contributed by the individual, rather than the knowledge that emerges from the document/s or the knowledge base as a whole. Thus, the *fine-grained provenance* of micro-contributions and their *localisation* in the context of the *living documents* that host them, is captured/documented. This fine-grained perspective facilitates expertise modelling using experts' contributions, while providing the means to view every contribution within a broader context, i.e., its encapsulating content, e.g., paragraph, section, sub-section, etc. This in turn, provides adequate context for processing the *short* and *sparse* content of micro-contributions.

Furthermore, the proposed Expertise Modelling Framework uses experts' micro-contributions to collaboration platforms to create *structured* expertise profiles, i.e., expertise profiles containing concepts from domain ontologies, each of which represents a topic of expertise. This in turn facilitates greater visibility of expertise, as profiles can be published and integrated into the *Web of Data* [28]. Experts often contribute to multiple scientific networks. Thus, profiles that represent the knowledge contributed by an expert to each of these networks can be integrated to create an overarching view of the expert's skills and experiences. In addition, semantic associations among concepts representing expertise profiles and concepts in the *Linked Open Data* [28], can be used to complement expertise profiles, identify the optimum set of collaborators for critical scientific challenges or accelerate scientific discoveries by recognising connections across domains.

Moreover, semantic similarity measures are leveraged to create profiles at *different levels of abstraction*, thus facilitating *comparison* and *evaluation* of profiles describing expertise with different granularity. In addition, semantic similarity is used to create *fine-grained* representations of contributed knowledge and to investigate the extent to which the profiles created by the framework, cover the expertise embedded in micro-contributions.

Additionally, the proposed framework captures the *temporal aspect of expertise*, by capturing *micro-contributions*, the *actions* that lead to their creation, e.g., update, delete and add operations on

the host documents and the *revisions* of the host documents that result from such operations. This information is used subsequently to devise algorithms and models for analysing and tracking *expertise* and *interests over time*.

In addition to experts' contributions, the proposed framework uses *social factors* to *refine* expertise profiles. In particular, the *collaboration structure* of experts in existing *social networks*, (i.e., collaborators' *relationships* and the *strength* of those relationships) is leveraged. Collaborators' profiles are refined by taking into account the *semantic associations* between concepts representing the expertise and contributions of collaborators.

This is the first approach that combines social relationship analyses, semantic similarity measures and dynamic semantic analysis of micro-contributions and their context, to generate more precise expertise profiles that can be tracked over time and compared across domains.

The next Chapter (3) describes the Fine-grained Provenance Model and Ontology that underpins the innovative methods that have been developed to extract fine-grained expertise profiles from micro-contributions, as described in Chapters 4-8.

Chapter 3 A Fine-grained Provenance Model for Micro-contributions

3.1 Introduction

The framework proposed in this thesis, aims to profile the expertise of an author, using the implicit knowledge embedded in his/her micro-contributions, rather than the knowledge embedded in the document/s that host those micro-contributions. More specifically, the aim is to model expertise using the contributed content, thus, facilitating *individual attribution*. Towards this goal, this chapter describes the *Fine-grained Provenance Model*, developed for capturing the *fine-grained provenance* of micro-contributions in the context of platforms, where knowledge *evolves* over time (Objective 1 (O1) in Section 1.5 of Chapter 1). The Fine-grained Provenance Model facilitates expertise modelling using micro-contributions and the encapsulating contexts which host them; e.g., paragraph, section or page in which a contribution is made. Therefore, the model complements the short and sparse content of micro-contributions with their encapsulating content; thus, facilitating *semantic analysis* of the contributed content. In addition, the fine-grained provenance of micro-contributions can be used as evidence for the addition or removal of expertise topics within an expert's profile.

The model combines *coarse* and *fine-grained* provenance modelling to capture and represent *micro-contributions* and their *localisation* in the context of their *host living* documents. In particular, the model facilitates a *contribution-oriented* view of a platform, by representing micro-contributions and their context, at different levels of *granularity*; e.g., paragraph, sub-section, section, page and document (in increasing order of coarseness). In other words, a micro-contribution can be viewed as a complete entity (e.g., paragraph, sub-section, section, page, document) or as a constituent of the paragraph, subsection, section or page in which it is made. The model also captures and represents *revisions* resulting from such incremental refinements. The fine-grained provenance and the localisation of micro-contributions, in addition to the change management aspects of the platform such as *actions* (that lead to the creation of micro-contributions) and document *revisions*, are used by the proposed expertise profiling methodology, described in Chapter 4, to create *semantic* and *time-aware* expertise profiles.

Section 3.2 outlines the requirements that underpin the design of the model. Section 3.3 describes the Fine-grained Provenance Ontology, developed for capturing micro-contributions and expertise profiles in collaboration platforms. Section 3.4 concludes this chapter with a discussion of the results. The work described in this chapter is published in [145].

3.2 Requirements

The emergence of different types of collaborative environments, such as Wikis, content management systems, and collaborative ontology editors, enables novel ways of curating knowledge, hence transforming the workflow from being curator-centred to being community-driven. Such systems provide the means for communities of experts in different fields, to create, share and reuse knowledge collaboratively. The goal of such systems is to foster long term expansion and maximisation of knowledge curation, extraction and reasoning, by creating live knowledge bases within their specific domains [146].

A typical workflow within such platforms involves the evolution of knowledge through contributions from multiple collaborating experts. Figure 3-1 depicts an example of evolving knowledge, where the contribution of one expert modifies/complements the contribution made by another expert. In this example, Expert1 has made a contribution by updating the content in the document describing "Achondroplasia" and Expert2 has made a micro-contribution by deleting some of the content contributed by Expert1. The incremental refinements, such as add, delete and update, performed by experts on content hosted by collaboration platforms, result in micro-contributions and revisions to the underlying documents. From an expertise profiling perspective, the collection of an expert's micro-contributions provides a valuable resource from which the expertise and interests of the expert can be inferred.

An analysis of micro-contributions, information flows and typical interactions among experts in collaborative knowledge-curation platforms highlighted a number of key requirements, which have been accommodated into the design of the model. In particular, the change management aspects of the platform such as the *actions* that lead to the creation of micro-contributions, e.g., *updates*, *additions*, *deletions* and *revisions* to host documents, must be captured. In addition, because the goal is to map expertise topics embedded in micro-contributions to concepts from ontologies in any domain, modularisation has also been identified as a key requirement. The following sub-sections describe in greater detail, the specific requirements that have been identified and how they have been accommodated into the model.

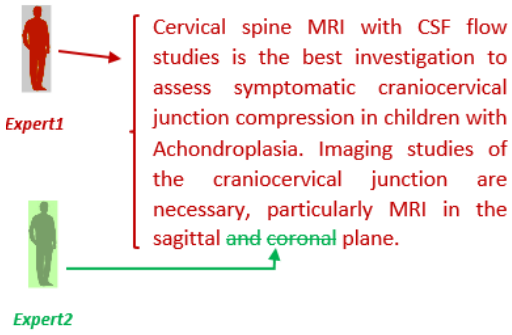

Figure 3-1: Example of two micro-contributions within the same context

3.2.1 Identification and Revision

Figures 1-1 and 3-1 illustrate two examples of micro-contributions to collaboration platforms. Individual contributions are uniquely identified as chunks of text, that encapsulate the contribution semantics and that are constituents of the larger host living documents. Micro-contributions represent incremental refinements to the content of collaboration platforms, whereby knowledge evolves over time. As outlined in Chapter 1, the short and sparse nature of micro-contributions and the evolving content of collaboration platforms, present challenges for current expertise modelling approaches, which rely on analysing large corpora of static documents.

The Fine-grained Provenance Model proposed in this chapter, identifies and represents micro-contributions at various levels of granularity, e.g., as a complete entity (paragraph, subsection, section or page), or as a constituent of an encapsulating paragraph, subsection, section or page. This approach overcomes the inadequate content of micro-contributions without having to consider the entire content of host documents. Moreover, by documenting the revisions made on these elements (chunks of text), the evolution of micro-contributions and the associated individual's activities can be tracked. This in turn, provides a way to monitor not only the change in personal interests over time, but also the maturation (or regression) of an individual's expertise [147].

3.2.2 Support for Domain Knowledge and Specific Complementary Models

The proposed model is designed to be extensible – enabling domain-specific knowledge/concepts to be easily incorporated. Ontologies from a variety of domains can be plugged into the model dynamically and used to link the textual representation of expertise topics to domain concepts. Furthermore, the model is complemented with specific modules for capturing coarse and fine-grained provenance and change management aspects of evolving knowledge [147].

3.2.3 Modularisation

Modularization represents a key requirement for ontologies in order to achieve re-use and evolution [148]. With this aim in mind, domain knowledge and processes from the proposed fine-grained provenance model are decoupled. This leads to a model that supports evolution, extensibility and integration with ontologies from a variety of domains [147].

More specifically, in order to achieve high modularisation, the Fine-grained Provenance Model comprises two layers: the Contribution layer and the Expertise Profile layer. Furthermore, the model builds on existing widely adopted upper level ontologies (the Open Provenance Model and SKOS). This approach facilitates modularization and extensibility across domains and enables the model to be used for knowledge acquisition and reasoning purposes.

3.3 An Ontology for Capturing Micro-contributions and Expertise Profiles

As mentioned in Chapter1, micro-contributions represent incremental refinements by authors to an evolving body of knowledge. Examples of such micro-contributions include: edits to a Wikipedia article or a Gene page in Gene Wiki [149]; a statement in *WikiGenes* [2] or *OMIM* [150]; an argument in *AlzSWAN* [19]; or a statement in *SKELETOME* [21] (Figure 1-1). Regardless of the platform, the aim is to capture 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., the sentence, paragraph or section of the document in which they appear. Therefore an ontology is created to capture such artefacts and their localization in the context of their host living documents.

The objective is 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 goal guarantees direct applicability, greater re-use and low entry barriers (compared to developing an entirely new ontology from scratch). Coarse and fine-grained provenance modelling are combined using the *SIOC* ontology [151], with change management aspects captured by the *SIOC-Actions* module [152]. The *Annotation Ontology* [153] is used to bridge the textual grounding and the ad-hoc domain knowledge, represented by concepts from domain-specific ontologies. The *Simple Knowledge Organization System* (*SKOS*) [154] ontology is used to define the links to, and the relationships that occur between, these concepts. Figure 3-2 depicts the overall structure of the ontology.

Furthermore, ontology mappings are defined between the *Open Provenance Model Ontology* [155] and the fine-grained provenance model using the SKOS vocabulary. The W3C Provenance Incubator Group [156] has used the OPM as a reference for mapping the most widely used provenance ontologies. OPM is a general and broad model that encompasses many aspects of provenance and already represents an ongoing community effort that spans several years, benefiting from many discussions, practical use, and several versions. Many groups are currently mapping their vocabularies to OPM.

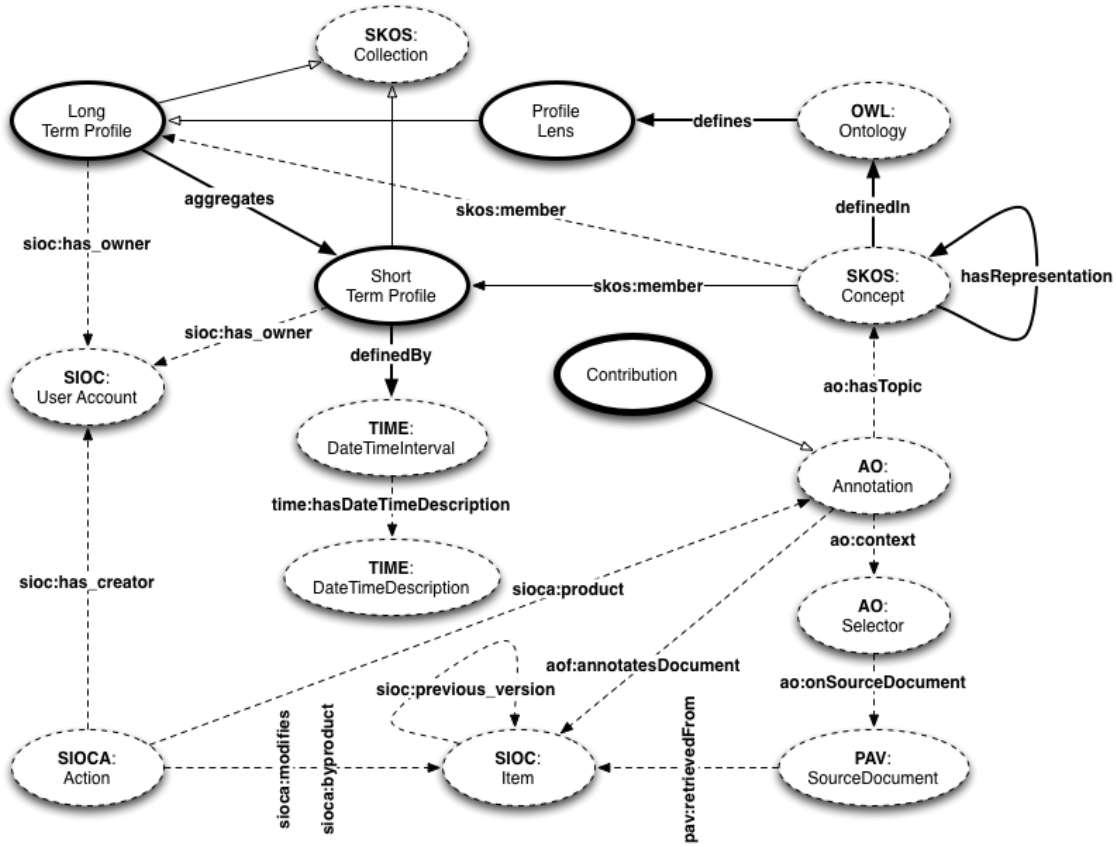

Figure 3-2: An ontology for capturing micro-contributions and expertise

As depicted in Figure 3-2, the proposed ontology identifies four concepts and four relations illustrated with bold lines. It can be conceptually divided into two main parts: (i) Part 1 that models micro-contributions, and (ii) Part 2 that captures expertise profiles. Both parts are discussed below.

The central concept of Part 1 is **Contribution,** which is considered to be a type of annotation (i.e., a subclass of **AO: Annotation**). The contributed text and its semantics are modelled at different conceptual levels. Therefore, a piece of text within a living document (modelled by **SIOC: Item**) is modified (*sioca: modifies*) by an action (e.g., add, delete, update) and can be clearly localized via pointer constructs – which 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*.

Figure 3-3 illustrates the example depicted in Figure 3-1 for Expert1 (topic: Achondroplasia) and Expert2 (topic: coronal plane) using the OWL Manchester syntax. As depicted in the following example, an expertise topic annotated in a micro-contribution, is mapped to a domain concept and

its exact placement in the context of the host document is captured (using the offset and range attributes of the Selector class in the Annotation Ontology). Thus, evidence of domain concepts representing expertise topics in an expert's profile, can be identified by linking the concepts to their textual representation in the expert's contributions. Furthermore, when profiling the expertise of an individual, only the concepts that emerge from the individual's contributions and their encapsulating contexts (which can be identified using the Selector class in the Annotation Ontology) are taken into account, rather than the entire content of the document. Our model facilitates this *contribution-oriented* approach to expertise profiling, by capturing and representing the fine-grained provenance of micro-contributions.

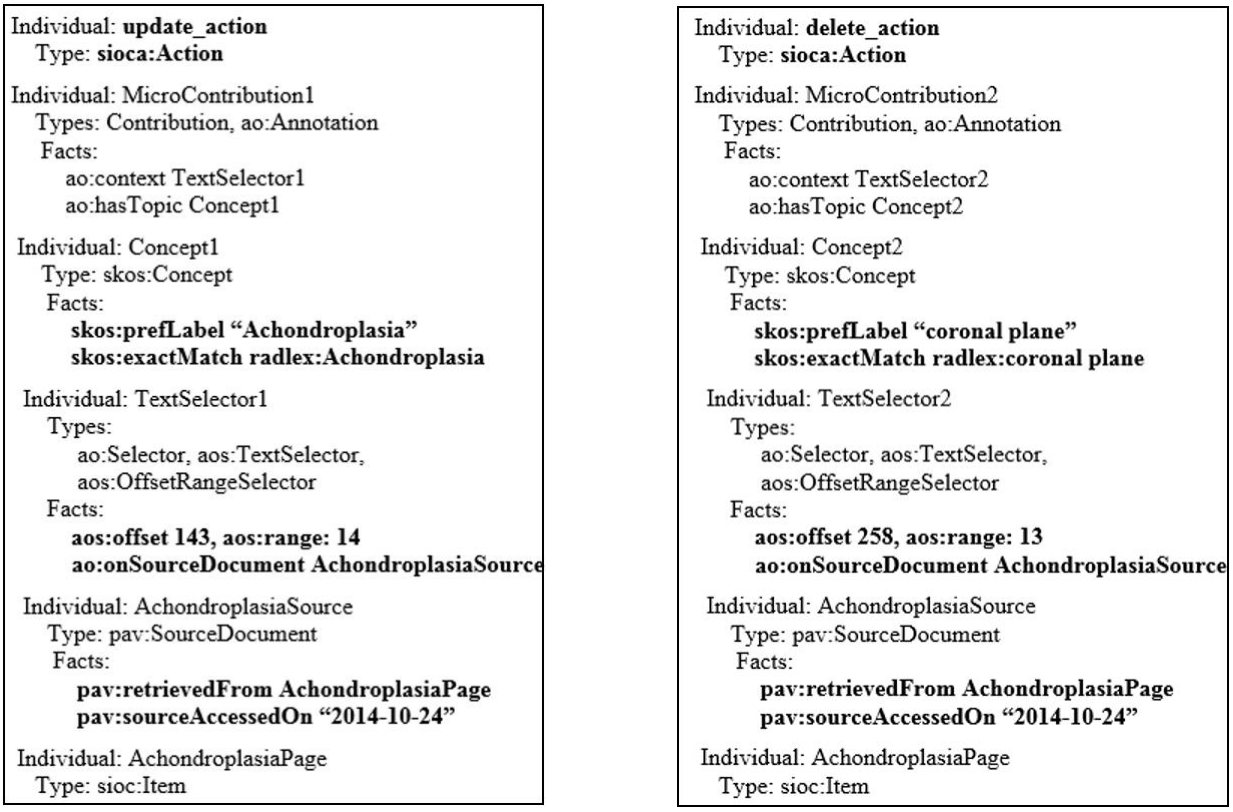

Figure 3-3: Example for Expert1 (topic: Achondroplasia) and Expert2 (topic: coronal plane) using the OWL Manchester syntax

Part 2 of the ontology models expertise profiles as **SKOS: Collection** (s) of concepts. Although very lightweight, the proposed model introduces three novel aspects when compared to other expertise profiling approaches.

In order to capture the *temporal aspect of expertise*, the proposed model differentiates between **Short Term** and **Long Term** profiles. A **Short Term Profile** is a collection of concepts identified within a specific period of time (modelled via concepts introduced by the Time Ontology). A **Long Term Profile**, on the other hand, *aggregates* all the **Short Term Profile** (s) generated for a particular expert. Intuitively, this provides a mechanism for tracking and analysing the evolution of

an individual's expertise over both the short and long term. The actual method for creating these profiles is described in Chapter 4.

Expertise profiles are more than just collections / bags of concepts. Domain-specific entities present in micro-contributions are captured in the model by the use of **SKOS: Concept** proxies[2]. By using the *hasRepresentation* relation between such proxies, the clustering of concepts is performed in a manner similar to the semiotic triangle [157]. A particular entity, e.g., FGFR3, can be modelled as an abstract concept with multiple representations, each of which corresponds to a concept from a different ontology; e.g., Gene Ontology or the Bone Dysplasia Ontology. This facilitates capturing the semantics of micro-contributions by considering the best-suited concepts from one or more ontologies, while keeping track of the provenance of concepts (via *definedIn* **OWL: Ontology**). In other words, an abstract entity represented by an instance of SKOS: Concept, can be defined by several concepts, each of which is also an instance of SKOS: Concept and belongs to an ontology (represented by **OWL: Ontology).** For example, the abstract entity "*MRI*" can be represented by the concept "*magnetic resonance imaging*" from the SNOMED-CT ontology as well as by the concept "*MRI imaging protocol*" from the Biomedical Informatics Research Network Project Lexicon. This approach will result in creating a more accurate representation of expertise by linking expertise to related concepts from multiple ontologies.

Maintaining the provenance of domain-specific concepts enables the creation of multiple views over a **Long Term Profile** via lenses defined by particular ontologies. In the proposed 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.

3.4 Conclusion and Future Work

This chapter introduces the Fine-grained Provenance Model for Micro-contributions, an important step towards meeting the principle objective of this thesis – fine-grained expertise profiling by analysing micro-contributions to evolving knowledge-bases (Objective 1 (O1) in Section 1.5 of Chapter 1). An ontology is developed for capturing and representing the fine-grained provenance of micro-contributions in the living documents that host them. The ontology captures

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

the exact placement of contributions in the underlying content at different levels of granularity, e.g., sentence, paragraph, sub-section, section, page, document. It also captures the actions that lead to the creation of micro-contributions, e.g., update, delete and add as well as document revisions resulting from such actions. The model captures and represents the evolution of knowledge over time, which in turn facilitates capturing and tracking the changes in individuals' expertise and interests over time.

Fine-grained provenance modelling facilitates analysis of micro-contributions using the encapsulating content, thus providing adequate context for semantic analysis of the short and sparse content of contributions. In addition, the fine-grained provenance of micro-contributions can be used as evidence of expertise in topics represented by domain concepts in individuals' profiles.

The main contribution of the model is that it facilitates *individual attribution*, by providing a *contribution-oriented* view of the platform. This in turn facilitates expertise profiling by analysing the *contributed content*. As outlined in Chapters 1 and 2, this is in contrast to traditional approaches, which profile expertise by associating individuals with expertise topics that emerge from the entire content of the authored or co-authored documents. Finally, instances of the model are not only useful for expertise profiling, but can also act as a personal repository of micro-contributions, to be published, reused or integrated within multiple evolving knowledge bases.

The aim is to create a comprehensive model for capturing and representing the fine-grained provenance of micro-contributions to evolving knowledge platforms. Thus, the SIOC-Actions module [152] is used to capture the actions that lead to the creation of micro-contributions, e.g., add, delete, update. Future work will focus on leveraging this information, in order to determine the quality of micro-contributions and adjust the weight of concepts in expertise profiles, accordingly. For example, an expert could modify a document by making a series of micro-contributions. All or some of these micro-contributions may subsequently be rolled back by another expert. This would then result in a lower ranking of concepts that emerge from those contributions in the expert's profile.

The next chapter, Chapter 4, describes the Semantic and Time-dependent Expertise Profiling Methodology and the way in which the Fine-grained Provenance Ontology is populated as micro-contributions are processed and expertise profiles are created.

Chapter 4 The Semantic and Time-dependent Expertise Profiling Methodology

4.1 Introduction

The previous chapter presented the Fine-grained Provenance Model for Micro-contributions – which captures and represents micro-contributions in the context of the evolving documents that host them. This chapter proposes the *Semantic and Time-dependent Expertise Profiling (STEP) methodology*, which analyses the fine-grained provenance of micro-contributions to represent the textual grounding of expertise topics, using weighted *concepts* from domain ontologies. In addition, the STEP methodology uses the change management aspects captured by the Fine-grained Provenance Model (i.e., update, delete and add actions resulting in micro-contributions and document revisions), to create *time-aware* expertise profiles, which facilitate tracking and analysis of changes in expertise and interests over time. The STEP methodology is developed to satisfy the objective of creating *time-aware* expertise profiles, while representing the knowledge embedded in *micro-contributions* using *weighted concepts* from domain ontologies (i.e., Objective 2 (O2) in Section 1.5 of Chapter 1).

Section 4.2 describes in detail, the three main phases of the STEP methodology (Concept Extraction, Concept Consolidation and Profile Creation). Section 4.3 provides a discussion outlining the pros and cons of this approach and Section 4.4 concludes with a summary of the outcomes of this chapter. (The work presented in this chapter is published in [145] and is one of the main foundations of the Expertise Modelling Framework proposed in this thesis.)

4.2 Expertise Profiling

Semantic and *Time-dependent Expertise Profiling* (*STEP*) provides a generic methodology for modelling expertise in the context of evolving knowledge. It consists of three main modules, as depicted in Figure 4-1; (i) Concept Extraction; (ii) Concept Consolidation; and (iii) Profile Creation.

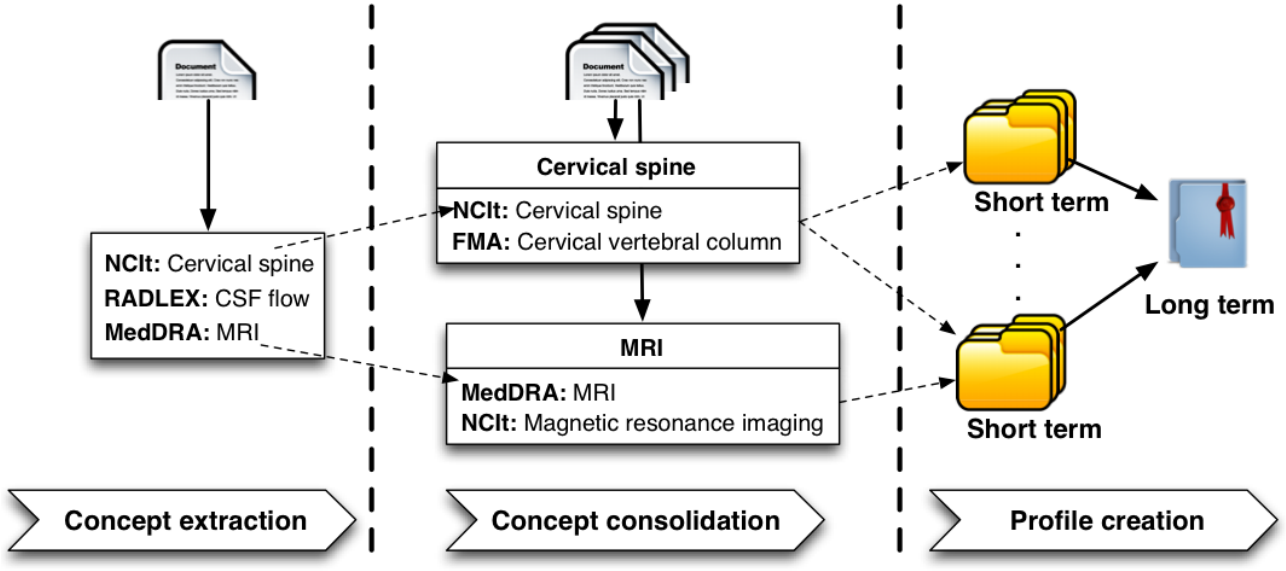

Figure 4-1: Semantic and Time-dependent Expertise Profiling Methodology

4.2.1 Concept Extraction

The concept extraction step aims to identify domain-specific concepts within micro-contributions. From an ontological perspective, the goal is to populate the micro-contribution part of the Fine-grained Provenance Ontology by creating appropriate annotations; i.e., **Contribution**(s) that represent domain entities (**SKOS: Concept**(s)) captured within the text of the micro-contributions. Consider the example presented in Figure 3-1 – "*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,** that represents 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[3]. Hence, in order to provide a profile creation framework applicable to any domain, this step is not restricted to the use of a particular concept extraction tool / technique.

4.2.2 Concept Consolidation

Over the course of the last decade there has been 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, in the NCBO

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

Bioportal [34] – i.e., the largest repository of biomedical ontologies – the concept *Cervical spine* is present in at least seven ontologies, while *MRI* is defined by at least 20 ontologies. From a semiotic perspective, this can be seen as a symbol with multiple manifestations (or materializations) [157], with each manifestation being appropriately defined by the underlying contextual domain. Figure 4-2 depicts an example of concept consolidation.

Domain-specific concepts captured within micro-contributions may also be defined in multiple ontologies. As a result, the *concept consolidation* step is introduced, which aims to cluster multiple representations of the same concept identified in one micro-contribution and across multiple micro-contributions. Figure 4-1 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**.

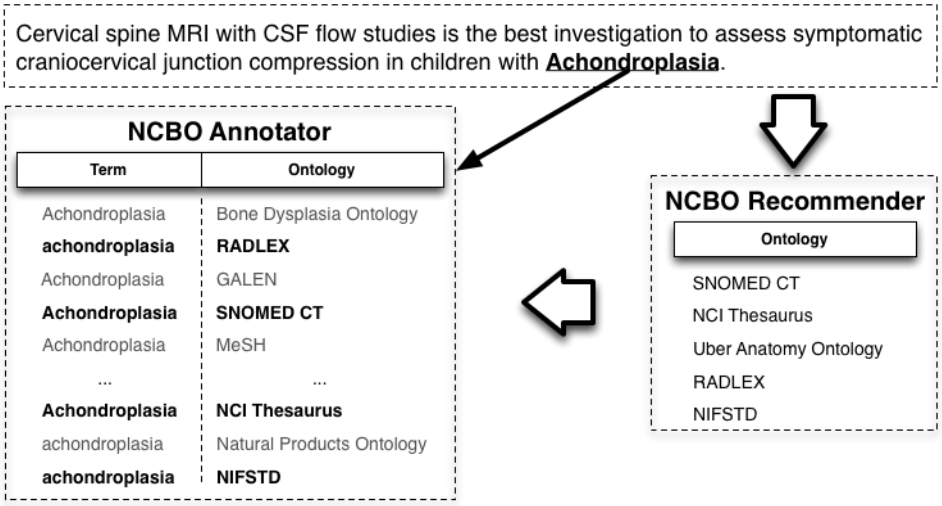

Figure 4-2: Example of concept consolidation

As discussed in Chapter 3, the Fine-grained Provenance Ontology for Micro-contributions 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**). The following figure represents the example depicted in Figure 4-2 using the Manchester syntax.

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.

```
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
```

Figure 4-3: Multiple annotations for "Achondroplasia" presented in Manchester syntax

As discussed in Chapter 2, the expertise profiling model proposed in this thesis is applied and evaluated in the context of collaboration platforms in the biomedical domain, due to the widespread availability of both resources and tools. However, the proposed methodology for creating expertise profiles is generic and can be applied to any domain, provided that appropriate tool support exists. The experiments presented in this thesis are conducted in the biomedical domain and use the NCBO Annotator [92] for concept extraction and the results produced by the Biomedical Ontology Recommender Web service [94] for concept consolidation. For example, consider the micro-contribution presented in Figure 4-2. The NCBO Annotator annotates the term **Achondroplasia** with concepts from 18 different ontologies; however, only the concepts that belong to the most suitable ontologies for annotating the micro-contribution, as recommended by the Biomedical Ontology Recommender, are retained (Figure 4-2). 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).

4.2.3 Profile Creation

The goal of this phase is to use the extracted and consolidated concepts to create time-aware expertise profiles by differentiating between ***Short term*** and ***long term*** profiles. The expertise of an individual is dynamic and typically changes over time. *Short term* profiles aim to capture periodic bursts of expertise in specific topics, over *contiguous*, *non-overlapping* intervals of time; e.g., the STEP methodology may be configured to create short term profiles representing the expertise of

individuals within arbitrary regular time windows, such as two weeks or one month. *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. In other words, the *long term* profile of an expert, is created by analysing the distribution of expertise concepts across all of his/her *short term* profiles.

Short Term Profile creation. Using the provenance information captured by the Fine-grained Provenance Ontology for Micro-contributions, an approach is proposed for computing short term profiles. Before discussing the actual computation, the following re-iterates the concept consolidation phase and explains 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 4.2.2. A cluster representing an abstract concept is referred to 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 specific period of time. In order to compute a short term profile for an expert, the concepts identified in the expert's micro-contributions within a specified time-window (e.g., two weeks) are ranked 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. Eq. 4-1 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.

$$W(V_c) = \frac{Freq(V_c)}{N_V} * \sum_{i=1}^{N_v - 1} PPMI(V_c, V_{ci}) \qquad \text{(Eq. 4-1)}$$

Where $c \neq c_i$ and V_c is the virtual concept, for which a weight is calculated, N_V is the total number of virtual concepts in the considered time window, and $PPMI$, is the positive pointwise mutual information [158], as defined in Eq. 4-2:

$$PPMI(C_1, C_2) = log \frac{p(C_1,C_2)}{p(C_1)*p(C_2)} = log \frac{N_c*Freq(C_1,C_2)}{Freq(C_1)*Freq(C_2)} \quad \text{(Eq. 4-2)}$$

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 represent an overarching view of an individual's expertise. The method aims to capture the collection of concepts occurring both *persistently* and *uniformly* across all short term profiles for an expert, by considering 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 Eq. 4-3:

$$W(V_c) = \alpha * (e^{-\Delta(V_c)} - \frac{\Delta(V_c)}{e}) + (1 - \alpha) * \frac{Freq(V_c, S)}{N_s} \quad \text{(Eq. 4-3)}$$

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, a decreasing exponential is introduced, which 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 (Eq. 4-4).

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

$$M_{ST}(V_c) = \frac{1}{N_s} * \sum_{i=1}^{N_s} (ST_i - ST_{i-1}) \quad \text{(Eq. 4-4)}$$

Where ST_i represents a short term profile window in which V_c appears and ST_{i-1} represents the previous short term profile window in which V_c appears, $(ST_i - ST_{i-1})$ 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, the aim is to detect uniformity by performing a linear

regression over the differences between the time-windows representing the short term profiles that contain the virtual concept.

Figure 4-4 depicts an example of short term profiles created for an expert. In this example, every short term profile represents expertise topics that emerge from the expert's contributions in a particular month. A more detailed discussion of the time-windows represented by short term profiles is outlined in Chapter 8. Pre-configured monthly durations are used in this example for simplicity sake only. In reality, experts are more likely to make micro-contributions about a specific topic/concept over irregular intervals (a few days, weeks or months). This would be reflected in variable time intervals/windows associated with that individual's short term profiles.

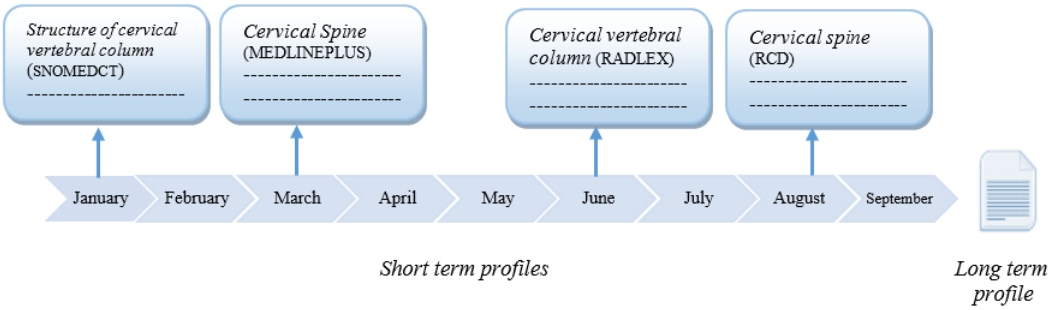

Figure 4-4: Example of short term profiles of an expert

Consider short term profiles where the concept "Cervical Spine" has been identified as an area of expertise. As illustrated in Figure 4-4, "Cervical Spine" can be viewed as an abstract entity or a virtual concept, with multiple manifestations represented by concepts from different ontologies (discussed in Section 4.2.2). In other words, the abstract entity, "Cervical spine", has been represented by the "Structure of cervical vertebral column" concept from the SNOMED CT ontology [159], "Cervical Spine" from the MEDLINEPLUS ontology [160], "Cervical vertebral column" from the RADLEX ontology [161] and "Cervical spine" from the RCD ontology [162], in the short term profiles created from contributions made in the months of January, March, June and August, respectively. Furthermore, while the concept "Cervical spine" (and its multiple representations), don't appear in all the short term profiles created for the expert, their appearance is persistent and more or less uniformly distributed across the short term profiles.

4.3 Discussion

The STEP methodology provides a domain-agnostic method for creating *semantic* and *time-aware* expertise profiles and serves as the cornerstone of the proposed expertise profiling framework and the foundation upon which the work presented in other chapters is built. STEP is applied to various knowledge domains, each of which provides a different perspective of the

methodology, facilitating the design of an abstraction layer that renders the final expertise profiling framework into a domain-agnostic form.

Unlike traditional expertise retrieval techniques, the STEP methodology creates expertise profiles by analysing the *short* and *sparse* content of *micro-contributions* in the context of *evolving* knowledge. However, as micro-contributions don't offer sufficient context for analysis, the content of every contribution is analysed in the context of its encapsulating content. The encapsulating context is captured at different levels of granularity through the Fine-grained Provenance Model proposed in Chapter 3. Thus, STEP facilitates *individual attribution*, by profiling the expertise of an individual using his/her micro-contributions, rather than traditional techniques, which rely on analysing the entire content of the documents to which one or more experts contribute.

Furthermore, as discussed in Chapter 3, the Fine-grained Provenance Model is designed to be extensible – enabling the plugging-in of relevant domain-specific ontologies. This in turn facilitates the extraction, capture and representation of topics that occur in micro-contributions, using ontological concepts.

In addition, unlike traditional approaches, the expertise profiling methods presented in this chapter, consider *uniformity* as important as *persistency*. To be precise, the long term profile of an expert, is generated by extracting the concepts that occur both *persistently* and *uniformly* across all the short term profiles for that expert.

Furthermore, Statistical Language Modelling techniques are integrated with STEP (Chapter 5) in order to minimise the effects of domain-specific concept extraction/recognition tools and techniques on the resulting profiles.

Finally, contextual factors embedded in social networks are integrated with STEP (Chapter 7) in order to refine the expertise profiles created. Contextual factors include the *context* within which every micro-contribution is made, as well as the *intrinsic* and *extrinsic relationships* that exist among experts who contribute to these contexts.

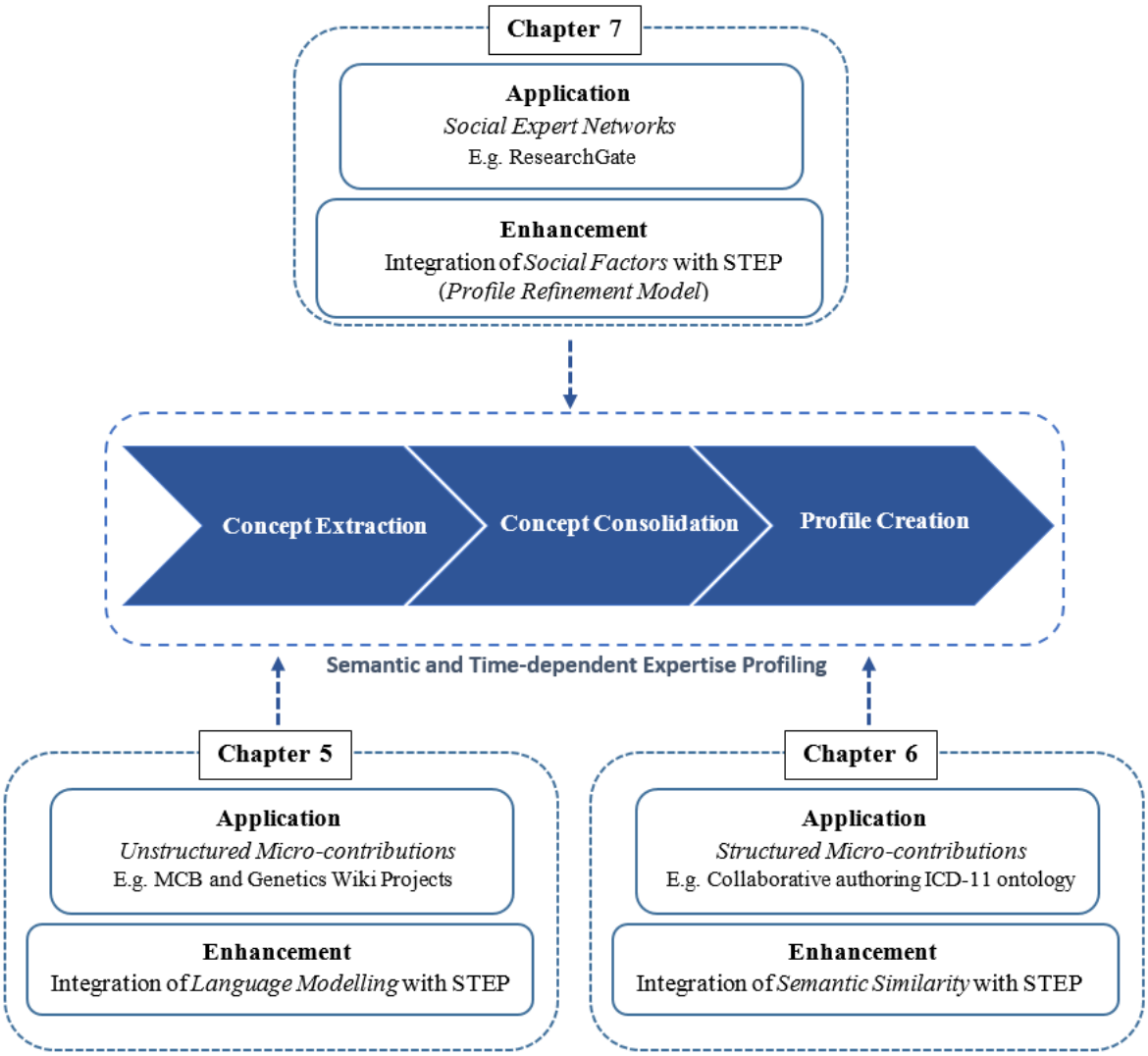

Figure 4-5: Applications and enhancements to the STEP methodology

Figure 4-5 depicts the relationship between STEP and the other constituents of the proposed framework. STEP is applied to *unstructured* micro-contributions in Chapter 5. More specifically, STEP is applied to contributions in natural language form, in the context of the *Molecular and Cellular Biology (MCB)* [38] and *Genetics* [39] Wiki projects (sub-projects of Wikipedia). Furthermore, Language Models are integrated with STEP in order to minimise the effect of domain-specific annotation and concept extraction tools on the resulting profiles. Evaluation results of applying the generic STEP methodology and STEP integrated with Language Models, to unstructured contributions, are also presented in Chapter 5.

The application of STEP to *structured* micro-contributions is investigated in Chapter 6. In particular, STEP is applied to micro-contributions made during collaborative authoring of the International Classification of Diseases ontology – Revision 11 (*ICD-11*) [24]. This chapter also presents the use of semantic similarity measures for creating profiles that represent expertise at a

level of specificity, which corresponds to topics embedded in micro-contributions. The use of semantic similarity is also investigated for creating profiles at different levels of granularity.

The *Profile Refinement Model* is described in Chapter 7. This model aims at integrating *social factors* with STEP, in order to refine expertise profiles using contextual factors embedded in scientific social networks. The Profile Refinement Model is applied to micro-contributions in the context of the *ResearchGate* social networking site for scientists and researchers [27]. This chapter demonstrates the use of ontological structures for complementing expertise profiles of collaborators, based on the relationships between experts, i.e., the types and strength of relationships among collaborating experts. The results of manual evaluations performed by nine ResearchGate experts is also presented and discussed.

4.4 Conclusion and Future Work

This chapter presents the methodology for creating *semantic* and *time-aware* expertise profiles by analysing micro-contributions made to evolving knowledge bases (e.g., knowledge curation platforms in the biomedical domain). STEP serves as the foundation upon which the expertise modelling framework proposed in this thesis is built and provides a critical role in meeting the objective (O2) of creating *semantic* and *time-dependent* expertise profiles, while capturing the *temporality* of expertise, as outlined in Section 1.5 of Chapter 1.

The STEP methodology creates profiles representing expertise using concepts from domain ontologies, by tapping into the *semantics* conveyed by micro-contributions. Previous chapters, highlighted the fact that semantic analysis of micro-contributions is essential, as such contributions don't offer sufficient content for applying methods used by traditional approaches, which rely on analysing large corpora. Furthermore, the semantic analysis performed on micro-contributions, provides a more comprehensive and accurate view of expertise, through the use of ontologies. As described in Section 4.2.2, the Concept consolidation phase of the STEP methodology, creates a consolidated view of abstract entities in micro-contributions that have been defined using concepts from different ontologies. Moreover, the weight attached to these concepts takes into account all the manifestations of the same entity, and therefore represents the true significance of topics in expertise profiles. This is in contrast to traditional text-based approaches, which treat every manifestation of the same entity, as a separate topic on its own, and hence are unable to represent and accurately rank the collective view of semantically similar expertise topics in profiles.

Furthermore, STEP creates profiles that capture the temporality of expertise. This, in turn, facilitates tracking and analysing changes in expertise and interests over time. While some existing research efforts have focused on temporal expert profiling [36], they rely on analysis of large corpora of static documents and representing expertise during specific regular or non-regular

intervals. For simplicity sake, the example provided in this chapter (Figure 4-4) generates short term profiles using regular time intervals (calendar months). However, Chapter 8 presents a method for identifying *time-windows*, where an expert exhibits "*peak activity*" in specific topics of expertise. These time-windows are of different lengths and emerge as experts focus on various activities and adopt different perspectives and interests, thus, allowing the time intervals to be determined based on an expert's contributing activity, rather than pre-configured timeframes. A detailed discussion of the temporal aspect of expertise is presented in Chapter 8.

The next chapter, Chapter 5, presents the application of STEP to unstructured micro-contributions in the context of two different Wiki projects and demonstrates the integration of Language Models with the STEP methodology. It also presents experiments and evaluation of the generic STEP methodology and STEP integrated with Language Models on unstructured micro-contributions.

Chapter 5 Application of STEP to Unstructured Micro-contributions

5.1 Introduction

The previous chapter introduced the Semantic and Time-dependent Expertise Profiling (STEP) methodology, for creating expertise profiles by analysing micro-contributions to collaborative knowledge curation platforms. STEP links the textual representation of expertise topics in micro-contributions to weighted concepts from domain ontologies, whilst capturing the temporality of expertise.

The STEP methodology is the foundation upon which the expertise profiling framework proposed in this thesis is built. As outlined in Chapter 1, one of the main objectives of this research (O3, Section 1.5), is to determine STEP's applicability to different types of community-driven, dynamic knowledge-curation platforms, in the context of a range of knowledge domains. Each of these knowledge domains provides a different perspective of STEP, which is used to design a framework that is applicable to all domains, i.e., domain-agnostic. Towards this goal, this chapter investigates the application of STEP to *unstructured* (natural language) micro-contributions from two case studies. Moreover, enhancements to STEP that involve integrating it with *Language Models* are implemented and evaluated. These enhancements aim to improve the accuracy of expertise profiles and minimise the impact of domain-specific concept extraction tools and techniques.

Section 5.2 describes the two biomedical Wiki projects that are used to evaluate the STEP methodology when applied to unstructured micro-contributions. Section 5.3 describes the tools employed for the Concept Extraction and Consolidation steps. Section 5.4 describes how the Language Models are integrated within the STEP methodology to implement two enhanced methodologies: the *topic modelling approach* and *n-gram approach*. Sections 5.5 and 5.6 describe the experiments and experimental results produced by applying the *original*, *topic modelling* and *n-gram* methodologies to *unstructured* micro-contributions. Section 5.7 compares the experimental results with traditional IR techniques. Section 5.8 provides a discussion of the results. Finally, Section 5.9 concludes with a summary of the research outcomes described in this chapter. The work presented in this chapter is published in [145, 163].

5.2 Use Cases

The STEP methodology was implemented and evaluated using contributions extracted from the *Molecular and Cellular Biology* (*MCB*) [38] and the *Genetics* [39] Wiki projects (both sub-projects of Wikipedia). Wikipedia allows authors to state opinions and raise issues in the discussion pages. MCB aims at organizing information in articles related to molecular and cell biology in Wikipedia. Similarly, the Genetics Wiki project involves the collaborative improvement and maintenance of genetics articles in Wikipedia. The underlying articles in both projects are constantly updated through expert contributions.

The following presents two examples of micro-contributions to existing articles in the MCB project (in this case by author *Jpkamil*) on different dates:

- 4 February 2008—Lipase article: *Lipoprotein lipase functions in the blood to act on triacylglycerides carried on VLDL (very low density lipoprotein) so that cells can take up the freed fatty acids. Lipoprotein lipase deficiency is caused by mutations in the gene encoding lipoprotein lipase.*

- 15 February 2008—Lipase article: *Pancreatic lipase related protein 1 is very similar to PLRP2 and HPL by amino acid sequence (all three genes probably arose via gene duplication of a single ancestral pancreatic lipase gene). However, PLRP1 is devoid of detectable lipase activity and its function remains unknown, even though it is conserved in other mammals.*

The Fine-grained Provenance Model introduced and discussed in Chapter 3, captures the localisation of micro-contributions within the host documents. This enables the STEP methodology to analyse micro-contributions at different levels of contextual granularity; e.g., using the paragraph, subsection, section or host document in which they appear. The experiments presented in this chapter use only the micro-contributions for expertise modelling, as the aim is to demonstrate the performance of STEP in facilitating individual attribution. In other words, the aim is to evaluate the extent to which expertise profiles created by STEP represent the knowledge contributed by experts, rather than the knowledge that emerges from host documents. In Chapter 7, the context in which each micro-contribution is made is taken into account. More specifically, experiments are initially conducted that apply STEP to micro-contributions – these results are then compared with experiments that apply STEP to micro-contributions taking into account both the context in which they are made as well as the intrinsic and extrinsic relationships that exist between experts who contribute to these contexts. These experiments are designed to quantify the effects of combining contextual and content-based factors.

Short term and long term profiles are created, using experts' *unstructured* micro-contributions, i.e., micro-contributions in natural language form. These profiles are created using the methods and algorithms described in Chapter 4. Short term profiles, i.e., expertise profiles created over

contiguous, non-overlapping intervals (e.g., two-week or one month time-windows)—allow one to determine bursts of activities related to particular topics within the corresponding intervals; e.g., the level of participation of an individual in a project. In the example provided above, one could infer that *Jpkamil* has been active within this period in the area of *Lipase genes*.

Long term profiles, i.e., expertise profiles created over the entire history of an individual—can be used to determine how long individuals were experts in a specific topic, or how recently they demonstrated expertise in this topic.

Micro-contributions for 22 authors from the MCB project [38] and 7 authors from the Genetics project [39] over the course of the last 5 years were collected. These contributions resulted in a total of 4,000 updates, with an average of 270 words per micro-contribution and an average of 137 micro-contributions per author. Each of the 29 authors selected from all the participants provided an average of 4.5 expertise topics in their profiles. These topics were used to create long term profiles for each author, representing the baseline. An example of such a profile is the one for author "*AaronM*" that specifies: "*cytoskeleton*", "*cilia*", "*flagella*" and "*motor proteins*" as his expertise.

The 29 designated authors, whose micro-contributions were collected, were those who provided a personal view of their expertise when they registered/joined each project. While a much larger number of participants were available, the vast majority did not provide a sufficiently detailed description of their expertise. Experiments were performed using only those experts whose personal profiles listed topics of expertise, rather than simply their role (e.g., "post doc" or "graduate student") or interest in the project (e.g., "improving Wikipedia entries", "expanding stub articles").

5.3 Tool Support for Concept Extraction and Consolidation

The STEP methodology can be implemented using domain-specific tools, which enable an accurate extraction of the concepts embedded in micro-contributions (Chapter 4). Within this thesis, the biomedical domain is chosen for application and evaluation purposes because of the ready availability of existing tools that can be employed. To evaluate the STEP methodology, the NCBO Annotator [92] is used as an underlying concept extraction technique and the Biomedical Ontology Recommender Web service [94] is used to perform concept consolidation (Chapter 4).

The National Centre for Biomedical Ontology Annotator, NCBO Annotator, is an ontology-based Web service for annotating biomedical textual content with biomedical ontology concepts. The biomedical community uses the Annotator service to tag textual datasets automatically with concepts from more than 200 ontologies (sourced from the two most important set of biomedical ontology & terminology repositories: the UMLS Meta thesaurus [68] and NCBO BioPortal [34]). The annotation (or tagging) of *unstructured* free-text data with ontological concepts transforms it

into structured and standardized data and enables it to become part of the biomedical Semantic Web – expanding the knowledge base that leads to translational scientific discoveries [164].

The workflow of the NCBO Annotator's Web service is composed of two main steps. Firstly, the biomedical free text is provided as input to the concept recognition tool used by the Annotator, together with a dictionary. The dictionary (or lexicon) is constructed using ontologies configured for use by the NCBO Annotator. The Web service uses *Mgrep* [165], a concept recognizer with a high degree of accuracy (>95%) in recognizing disease names [166] developed by the National Centre for Integrative Biomedical Informatics (NCIBI) at the University of Michigan [167]. Mgrep implements a novel radix-tree-based data-structure that enables fast and efficient matching of text against a set of dictionary terms. In the second step of the workflow, the biomedical annotator uses an *is_a* transitive-closure component and leverages UMLS Meta thesaurus CUI-based (Concept Unique Identifier-based) mappings in order to expand the annotations created by Mgrep. The NCBO Annotator is publicly available and deployed as a SOAP (Simple Object Access Protocol) [168] and RESTful (REpresentational State Transfer) Web service [116].

The NCBO 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 relationship* between concepts. However, in the experiments described here, the Annotator is configured to perform *direct* annotations only i.e., annotations were 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 the methodology against such methods (Section 5.7).

Although the NCBO Annotator is predominantly used in the biomedical domain, its underlying technology is domain-agnostic. Like most concept recognizers, it takes as input a textual resource to be annotated and a dictionary to produce annotations. Hence, the only customization to the biomedical domain is the specification of the biomedical ontologies used by the Annotator. In other words, by using the NCBO Annotator, the experiments 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.

However this versatility comes at the price of extraction efficiency, as an exact match is required between the terms present in the text and the labels of ontological concepts, in order for annotations to be detected. For example, a simple usage of the plural of a noun (e.g., Flagella) is enough to miss an ontological concept (such as Flagellum); furthermore, in some cases, only constituents of a phrase are annotated (e.g., "tibial shaft"); aggregating partial annotations does not

accurately convey the semantics of the whole term (e.g., consolidating concepts representing "shaft" and "tibial" does not convey the same semantics as concepts representing "tibial shaft").

The impact of this problem is minimised by consolidating and representing semantically similar concepts using *virtual concepts* in the Concept Consolidation phase of the STEP methodology. Concept consolidation is realized with the help of the Biomedical Ontology Recommender Web service [94], which identifies and ranks the most suitable ontologies for annotating a textual entry. While the NCBO Annotator assists with concept consolidation by providing multiple concept candidates for the same text chunk, an additional consolidation phase is introduced, via the Biomedical Ontology Recommender Web service, or Recommender. The consolidation phase creates a more coherent view over the domain-specific concepts derived from micro-contributions. Given textual metadata or a set of keywords describing a domain of interest, the Recommender suggests ontologies appropriate for annotating or representing the data. Appropriateness is evaluated according to three main criteria; coverage, or the ontologies that provide most terms covering the input text; connectivity, or the ontologies that are most often mapped to by other ontologies; and size, or the number of concepts in the ontologies.

While concept consolidation results in a significant improvement in the expertise topics produced by the Annotator, in some cases, domain concepts representing expected expertise topics were either not included in the results, or were ranked inaccurately. Exhaustive analysis and resolution of sub-optimal performance by the NCBO Annotator in the context of these use cases, is outside the scope of this research. However, experiments clearly indicate that the accuracy of resulting profiles is directly influenced by the quality of annotations produced by the annotator. Since the STEP methodology provides a pluggable architecture, in order to reduce the effects of domain-specific concept extraction tools on the accuracy of the generated profiles, an approach is proposed, which integrates Language Models [31] with the Concept Extraction phase of the STEP methodology. The following section outlines the proposed methods, which are domain-agnostic, in order to ensure that the overall architecture remains *domain-independent*.

5.4 Integrating Language Models with STEP

This section describes the integration of Language Models with the STEP methodology. The aim is to complement the concept extraction phase of the STEP methodology by using domain-agnostic methods to identify expertise topics embedded in micro-contributions. This in turn reduces reliance on domain-specific concept extraction tools and techniques, minimising their effects on the resulting profiles. More specifically, two sets of experiments were performed for enhancing the STEP methodology. These experiments involve firstly applying *lemmatisation* to *unstructured micro-contributions* followed by: either (i) *topic modelling* [32]; or (ii) *n-gram modelling* [33].

Figure 5-1 illustrates the three different approaches to the concept extraction phase which were implemented and compared (i.e., *original*, *topic modelling* and *n-gram modelling* approaches).

5.4.1 Lemmatization

As outlined above, the NCBO Annotator will not generate annotations for terms that vary from their base or dictionary form (*lemma*). Therefore *Lemmatization* [169] was performed on micro-contributions prior to extracting concepts using the NCBO Annotator. Lemmatization, which is the algorithmic process of determining the *lemma (base or dictionary form)* for a given word, improves the accuracy of information extraction tasks. Morphological analysis of biomedical text is most effective when performed by a specialized lemmatization program for biomedicine [170]. Hence, the experiments described in this chapter used *BioLemmatizer* [171] to conduct morphological processing/lemmatization of micro-contributions, prior to concept extraction. It is important to note the distinction between *stemming* and *lemmatization*; a stemmer operates on a single word, removing the end without knowledge of the context, and therefore cannot discriminate between words that have different meanings depending on part of speech. Lemmatization, on the other hand, uses a vocabulary and morphological analysis of words, to remove inflectional endings only and to return the base form of a word, known as the lemma. Therefore, in this research project, micro-contributions were lemmatized as a pre-processing step, in order to facilitate understanding of *context* and to determine the part of speech of a word in a sentence.

As described in Chapter 4, topics identified in micro-contributions are annotated using concepts from multiple ontologies; thus, a term is often represented by a cluster of concepts, each of which belongs to a different ontology. Topics that are lexically different but semantically similar (e.g., diabetes and high blood sugar), are therefore represented using clusters of concepts, which often contain common annotations. The *Concept Consolidation* phase of STEP detects these common annotations and combines them to create *virtual concepts*, which represent an abstract entity (e.g., diabetes) and contain concepts that are manifestations of the abstract entity (e.g., high blood sugar, hyperglycaemia).

As described in Section 5.3, an exact match is required between the terms present in micro-contributions and the labels of ontological concepts (which are often in lemma form), in order for annotations to be detected. Thus, lemmatisation, which determines the lemma (base or dictionary form) for terms, increases the number of matches between terms in micro-contributions and the label of ontological concepts, leading to an increase in the number of annotated concepts representing a given topic. As virtual concepts are created by detecting and aggregating common annotations among clusters of concepts representing topics in micro-contributions, this leads to an

increase in the number of detected virtual concepts, leading to enhanced capture of the semantics of micro-contributions and more accurate ranking of concepts in expertise profiles.

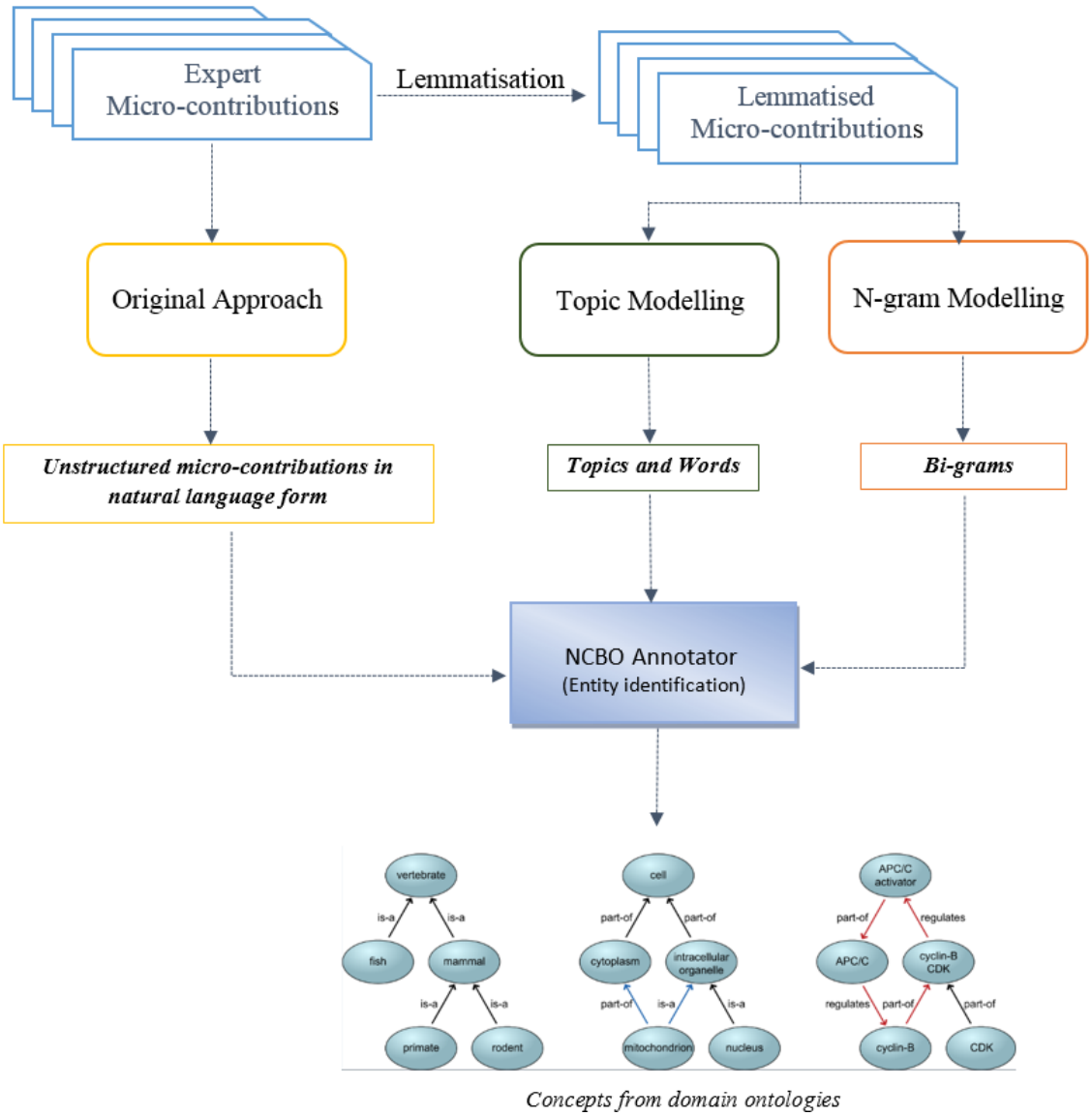

Figure 5-1: Overview of the original, topic modelling and n-gram modelling approaches to Concept Extraction

5.4.2 Topic Modelling

Topic models discover the *main themes* that pervade a large and otherwise unstructured collection of documents. Topic models can organize the collection according to the discovered themes and can be applied to massive collections of documents and adapted to many kinds of data. Among other applications, they have been used to find patterns in genetic data, images and social networks [95].

Incremental and collaborative refinements to content in collaborative knowledge platforms, including micro-contributions in the biomedical domain, usually contain discussions on a variety of

topics. In order to discover the *abstract topics* and the *hidden thematic structure* of micro-contributions, *topic modelling* was performed on all contributions made by an author. More specifically, the *Latent Dirichlet Allocation (LDA)* topic model [172] was chosen, which allows documents to encapsulate a mixture of topics. The intuition behind LDA is that documents exhibit multiple topics; e.g., a discussion regarding *Achondroplasia*, a disorder of bone growth, in *SKELETOME* [173], will most likely include information about the possible causes of the disease (such as inheritance and genetic mutation, genes and chromosomes), diagnosis methods, treatment options, medications, etc.

LDA is a statistical model of document collections that defines a topic to be a distribution over a fixed vocabulary. For example, the "genetics" topic has words about genetics (such as FGFR3 gene, chromosome, *etc*.) with high probability. Each micro-contribution made by an expert is seen as an *exhibition of these topics in different proportion*. Defined topics and words included in those topics were annotated to derive domain-specific concepts from domain ontologies.

As depicted in Figure 5-1, *Lemmatization* followed by *Topic Modelling* was integrated within the *Concept Extraction* phase of the STEP methodology. In particular, an expert's micro-contributions were lemmatised in order to retrieve terms in their lemma form. The lemmatised micro-contributions were then fed to the MALLET package [174], which implements LDA [172] as the topic model. In order to address the inefficiencies of topic modelling using sparse content, for a given expert, MALLET was configured to train a topic model by aggregating the expert's micro-contributions. This model was subsequently used to obtain a higher quality of *terms* learned from the expert's individual contributions. *Terms* identified by this process were then mapped to domain concepts, using the NCBO Annotator. The annotated concepts were then used to create short term and long term expertise profiles according to the profile creation methods described in Chapter 4.

5.4.3 N-gram Modelling

Latent Dirichlet Allocation [172] is based on the "bag-of-words" [100] assumption, in that the order of words in a document does not matter. However, *word order* and *phrases* are often critical to capturing the meaning of text. N-gram models are analogous to placing a small window over a sentence or text, in which only *n* words are visible at the same time. Therefore the experiments were performed using the N-gram modelling technique presented in [33], where every sequence of two adjacent entities (*bi-gram model*) in micro-contributions from an expert are identified and annotated with concepts from domain ontologies.

As depicted in Figure 5-1, *Lemmatization* followed by *N-gram* modelling was integrated with the *Concept Extraction* phase of the STEP methodology. In particular, an expert's micro-contributions were processed to remove stop words and then lemmatised in order to retrieve terms

in their lemma form. The lemmatised micro-contributions were further processed to extract all *bi-grams*. A Markov Chain [176] was subsequently constructed with one state per word, and with a special state reserved for end of text. The probability of one word appearing after another was estimated from the relative bigram frequencies in the collection of an expert's micro-contributions. All bi-grams identified by this process were subsequently mapped to domain concepts, using the NCBO Annotator. The annotated concepts were then used to create short term expertise profiles. Long term profiles were subsequently created from the short term profiles. Short term and long term profiles were created according to the profile creation methods described in Chapter 4.

5.5 Experimental Setup

The main goal of the experiments discussed in this section was to test and compare the efficiency and accuracy of different methods for generating *long term* expertise profiles. The experimental process involved extracting and comparing expertise profiles generated via the following methods:

(i) The original STEP methodology, where concepts are extracted from micro-contributions using the NCBO Annotator tool; *i.e.*, the *original approach*;

(ii) The enhanced STEP methodology, where Topic Modelling is integrated with the Concept Extraction phase in STEP; *i.e.*, the *topic modelling approach* and

(iii) The enhanced STEP methodology, where N-gram Modelling is integrated with the Concept Extraction phase in STEP; *i.e.*, the *n-gram modelling approach*.

Short term profiles of an individual represent the expertise inferred from his/her contributions within contiguous, non-overlapping intervals in time. Furthermore, the long term profile for the individual is created by analysing the distribution of expertise concepts across all of his/her short term profiles. Two sets of experiments were performed with each of the approaches described above. The first set of experiments used two-week intervals to create short term profiles; i.e., every short term profile represented a two-week time-window. Corresponding long term profiles were created from these short term profiles, as per the methodology described in Chapter 4. The second set of experiments used one-month intervals to create short term profiles, followed by the compilation of the corresponding long term profiles. Comparisons and analysis of the two sets of profiles confirmed that the long term profiles generated from short term profiles representing two-week intervals, described expertise with higher accuracy than long term profiles created from short term profiles representing one-month intervals. Therefore, Section 5.6 presents experimental results and evaluations performed on long term profiles created from short term profiles that represent expertise in contiguous, non-overlapping two-week time-windows.

It is important to note that the STEP methodology can be configured to create short term profiles using either regular or non-regular intervals. Different time intervals can be used to detect specific patterns in experts' contributing activities. Larger time windows provide an indication of an individual's topics of expertise over an extended period of time, while shorter time windows facilitate analysis of the *changes* in topics and interests over time. Thus, shorter time windows (e.g., 2 weeks cf. 4 weeks) generate a more accurate representation of expertise in the corresponding long term profile, as they facilitate more fine-grained detection of topics' occurrence, uniformity and persistency over time. Chapter 8 proposes a method for determining time windows of variable length, in which an expert exhibits high activity in particular topics of expertise for short bursts of time.

All methods were performed on micro-contributions for the 29 authors selected from the Molecular and Cellular Biology [38] and Genetics [39] Wiki projects. These authors were chosen because of the availability of manually input personal profiles, which were used to provide baseline/benchmark long term profiles for testing. As discussed and illustrated in Section 5.2, the baseline profiles were created by experts when they registered/joined each project and represent their own personal views of their knowledge and experience. Baseline profiles typically list topics of expertise at high levels of abstraction, such as *Genetics*, *Chemistry, Cell* and *Biology*.

It is important to note that the baseline profiles (created by the experts) and profiles created by the original, topic and n-gram modelling approaches, describe the expertise of individuals at different levels of abstraction. Micro-contributions tend to be very specific, i.e., terms identified in micro-contributions describe very specific domain aspects. Thus, STEP profiles describe expertise at a low level, while baseline profiles, created by the experts, provide a high level, more abstract description of expertise. For example, an expert might specify *Cardiology* as one of his/her areas of expertise, while the expert's STEP profile is more likely to identify his/her expertise using more precise domain concepts *Pulmonary Stenosis* and *Balloon Valvuloplasty*. The difference in abstraction between the baseline and STEP profiles plays a crucial role in evaluation, as it makes direct comparison very challenging. As discussed in Section 5.9, the generation of profiles that represent expertise at a level comparable with the baseline is investigated in Chapter 6.

In terms of efficiency measures, *F-score*, *Precision* and *Recall* were used, as defined in the context of Information Retrieval. In the context of these experiments, the value of *F-score* provides a measure of the accuracy of profiles generated by each approach, by considering both Precision and Recall. F-score is the harmonic mean of precision and recall, with its best value at 1 and worst value at 0. For a given expertise profile, *Precision* is the number of correct concepts (concepts matching the baseline) divided by the number of all returned concepts (total number of concepts in

the generated profile) and *Recall* is the number of correct concepts (concepts matching the baseline) divided by the number of concepts that should have been returned (total number of concepts in the baseline).

The process of matching STEP-extracted expertise concepts to gold standard entries in the baseline profiles (i.e., expertise profiles described by authors when joining the MCB and Genetics projects) was done in an exact manner. A correct match was recorded when a gold standard entry textually matched any of the labels or synonyms of resulting STEP virtual concepts (or concepts which were manifestations of the virtual concepts). The Concept Consolidation Phase plays an important role in this setting, by aggregating semantically similar expertise topics.

5.6 Experimental Results

5.6.1 Experiments with the Original STEP Methodology

This section depicts the results achieved by the original STEP methodology, i.e., the *original approach*. Figure 5-2 tracks the values of Precision and Recall for different concept weight thresholds (see Chapter 4 for long term profile creation), while Figure 5-3 provides a different perspective over the same results, by showing the precision and recall for different weight thresholds (labelled on the graph). From Figure 5-2, it can be observed that if a threshold is not set on the weight of the concepts in the long term profiles (i.e., concept weight threshold is 0), 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%.

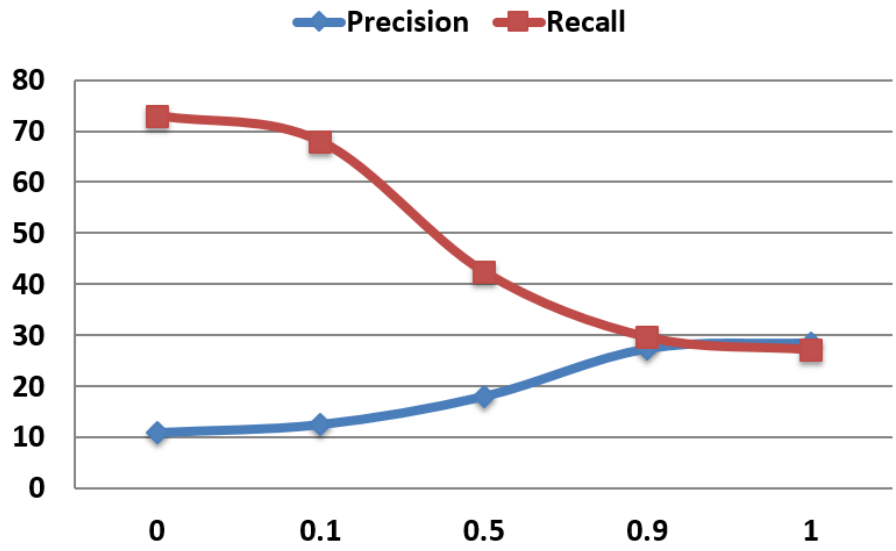

Figure 5-2: Precision and recall subject to a weight threshold

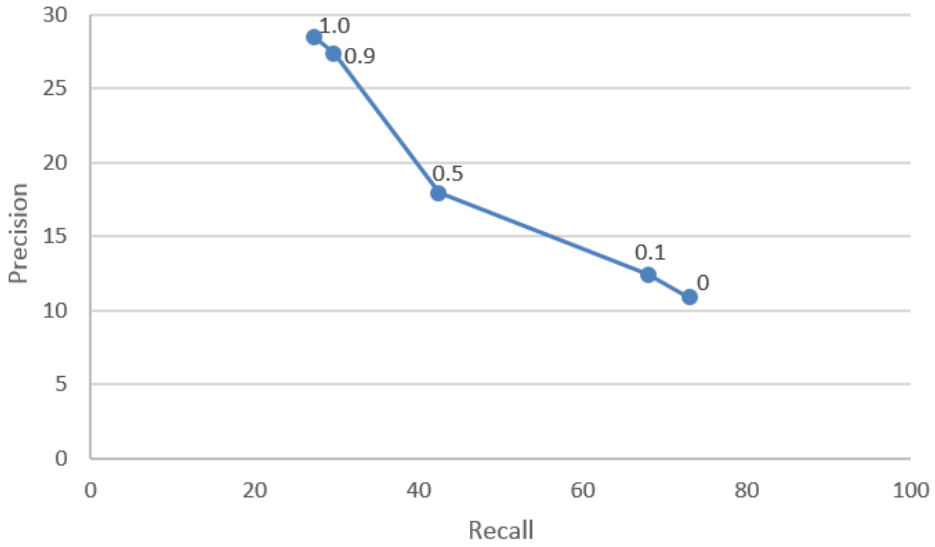

Figure 5-3: Precision-recall curve at different weight thresholds

5.6.2 Experiments with the enhanced STEP Methodology

Figure 5-4 depicts and compares the results achieved by the *original approach*, the *topic modelling approach* and the *n-gram modelling approach* and tracks the values of F-Score for different concept weight thresholds. If a concept weight threshold is not set (all virtual concepts are included in generated profiles) or the thresholds are set to < 0.5 (virtual concepts with weight < 0.5 are included in generated profiles), the *original approach* achieves the highest F-score. This is due to the fact that profiles generated by topic modelling and n-gram modelling approaches contain more noise as they include additional concepts representing the topics and n-grams derived from experts' micro-contributions.

Subsequently increasing the concept weight threshold from 0.5 to 1.2 results in consistently higher F-scores being achieved by topic modelling and n-gram modelling approaches, with the highest F-score achieved by n-gram modelling at concept weight threshold of 1 (31.94%). The enhanced F-Score of profiles generated by topic and n-gram modelling approaches, is partly due to the presence of concepts representing the topics and n-grams derived by these approaches, as well as a reduction of noise in the profiles as a result of increasing the concept weight threshold.

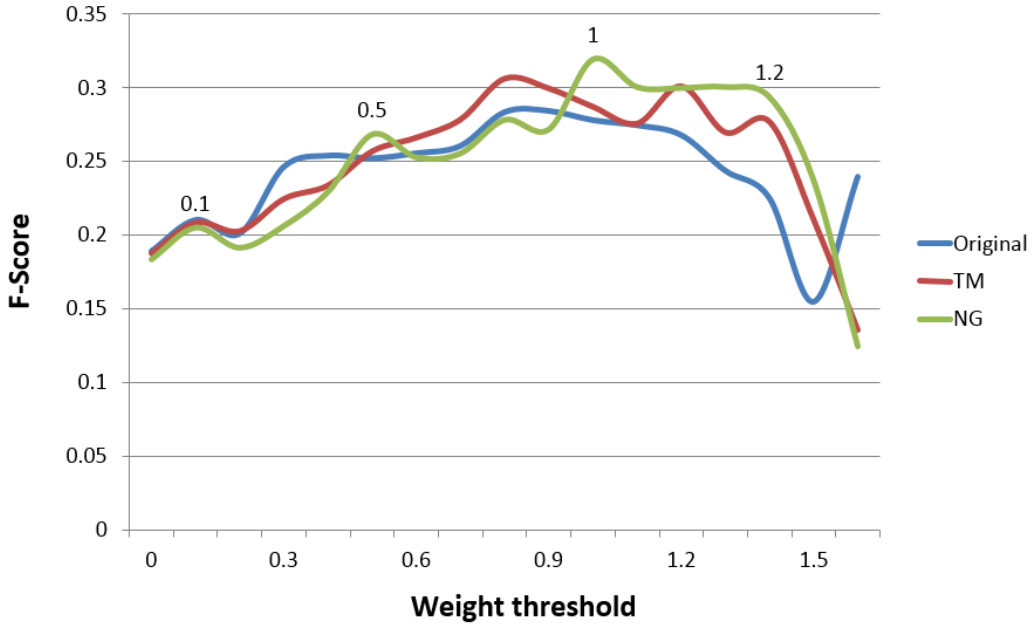

Figure 5-4: F-Score at different concept weight thresholds achieved by the original approach, topic modelling (TM) and n-gram modelling (NG)

Increasing the concept weight threshold from 1.2 to 1.5, results in a decline in the value of F-score for all approaches, but with n-gram modelling maintaining the highest F-Score of the three approaches (at all thresholds in this range). The decline in the value of F-Score is due to the exclusion of a large number of concepts from generated profiles as a result of higher thresholds. This in turn results in the decline of F-score for the topic modelling and n-gram modelling approaches up to the weight threshold of 1.6.

It is important to note that while topic modelling and n-gram modelling approaches use *meaning*, *context* and *themes* for identifying terms, the original approach analyses micro-contributions as bag-of words. Consequently, expertise profiles created by the original approach describe expertise using additional terms, some of which constitute noise. As the concept weight threshold is increased to 1.6 and above, the F-Score of profiles created by language modelling methods decreases due to the exclusion of a large number of concepts from the profiles. However, for profiles generated by the original approach, the same high weight thresholds result in the exclusion of concepts representing noise. This in turn increases *precision* (as profiles contain fewer concepts that represent noise) and decreases *recall* (due to the exclusion of a large number of concepts from profiles), leading to an increase in the value of F-Score for profiles generated by the original approach.

Figure 5-5 depicts the relationship between precision and recall for different concept weight thresholds (these thresholds are labelled). If a threshold is not set on the weight of the concepts in the long term profiles, the *original approach* achieves the best precision; *i.e.*, precision is 10.86%

for a recall of 72.94%, followed by topic modelling (precision: 10.79% and recall: 72.94%) and n-gram modelling (precision: 10.42% and recall: 76.82%). Overall, at concept weight thresholds of less than 0.5, topic modelling and n-gram modelling approaches have resulted in lower precision, as additional concepts included in the profiles (*i.e.*, concepts representing topics and n-grams derived from experts' micro-contributions) have contributed to more noise in the profiles (demonstrated by a higher recall).

Increasing the concept weight threshold to 0.5 results in an increase in the precision achieved by all methods, with n-gram modelling achieving the highest precision (18.35%), albeit at the expense of the recall (49.96%). Subsequently increasing the threshold results in further improvements to the precision achieved by all approaches, however at the expense of lower recall values. The best precision is achieved by the *topic modelling approach* at the concept weight threshold of 1.2 (34.08%) followed by the *n-gram modelling approach* (32.47%) and the *original approach* (27.68%). The results indicate that a higher accuracy is achieved by *topic modelling* and *n-gram modelling approaches* by setting a concept weight threshold, which minimizes the noise. Increasing the concept weight threshold above 1.2, results in a significant decrease in both the precision and recall values achieved by all methods; this is due to the exclusion of a large number of concepts with weights below such high thresholds.

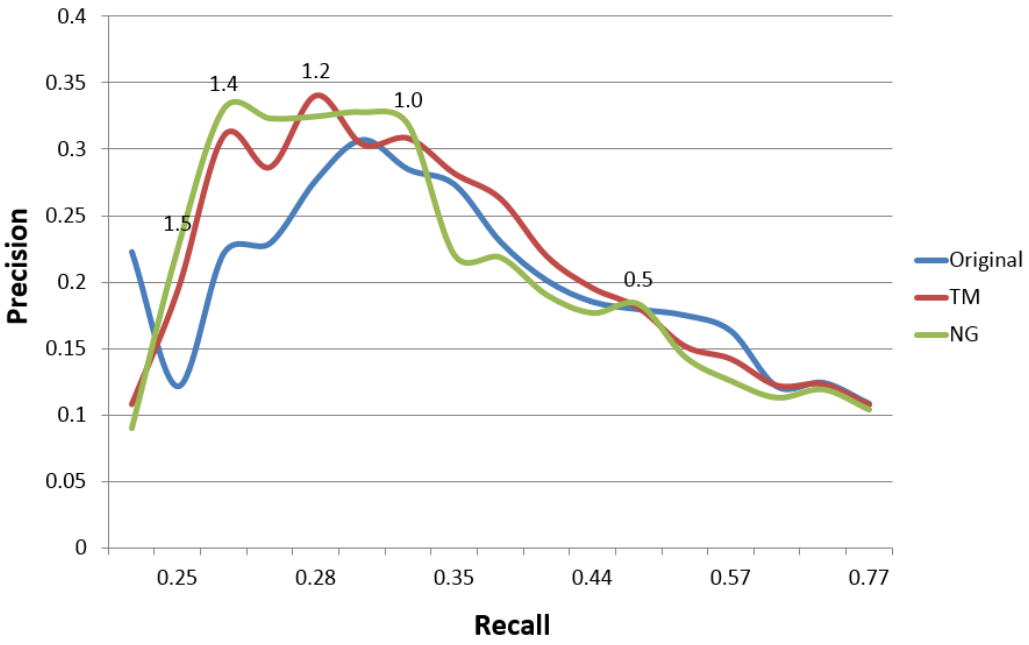

Figure 5-5: Precision-recall curve at different concept weight thresholds

Experimental results indicate that at concept weight thresholds greater than 0.4, *topic modelling* and *n-gram modelling* approaches consistently achieve higher accuracy in comparison to the *original approach*. The *topic modelling approach* demonstrates the highest precision at the

threshold of 1.2, although at the expense of recall. Overall, the *n-gram modelling approach* achieves the highest accuracy (F-score: 31.94%) at the concept weight threshold of 1. The enhanced accuracy is due to the fact that the *n-gram modelling approach* derives n-grams by taking into account *word order* and *context* (*higher precision*) and *multiple words* and *phrases* (*higher recall*).

The following tables demonstrate examples of concepts in the gold standard for two participants in the MCB Wiki project and the weights in the profiles generated by the three approaches presented in this chapter. N/A denotes concepts in the gold standard that are not present in the profiles generated by each approach. These tables are provided for illustrative purposes but individually are not statistically representative of the overall results.

Table 5-1: Comparison of profiles (concepts and weights) generated by the Original, Topic and N-gram Modelling approaches for author "Jpkamil"

	Gold Standard Concepts							
	Virology	Virus	Herpes	Molecular	Biology	Enzyme	Biochemistry	Lipase
Original	N/A	0.64	N/A	0.16	0.09	0.82	N/A	0.92
Topic Modelling	N/A	0.65	N/A	0.17	0.09	0.88	N/A	0.91
N-gram Modelling	0.46	0.46	N/A	0.13	0.14	0.88	N/A	0.92

Table 5-2: Comparison of profiles (concepts and weights) generated by the Original, Topic and N-gram Modelling approaches for author "pez2"

	Gold Standard Concepts			
	Enzyme	Vitamin K	Serine Protease	Natural Killer Cells
Original	0.7	0.2	N/A	N/A
Topic Modelling	0.75	0.4	N/A	N/A
N-gram Modelling	0.95	0.4	N/A	N/A

5.7 Comparative Analysis with Traditional IR Systems

In order to provide a more comprehensive interpretation of these results, the same experiment was performed using *Saffron* [175] and *EARS* [122], two systems that employ IR-based techniques. It is important to note that the results are not directly comparable for 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 the STEP methodology. In the context of these experiments, 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 analyse the performance of these systems when applied to a different type of dataset (micro-contributions as opposed to large corpuses of data that most IR-approaches rely on). Default configurations were used for both Saffron and EARS to create profiles for the experts designated as use cases for this study.

Table 5-3 summarizes the results achieved by Saffron and EARS systems, in comparison to the original, topic and n-gram modelling approaches of creating long term profiles. The results for the latter three are reported based on a concept weight threshold of 1.0, where the n-gram modelling approach achieved the highest accuracy (i.e., F-Score of 31.94%).

Table 5-3: Efficiency results of Saffron, EARS, Original STEP and Enhanced STEP approaches

Saffron			EARS			Original			Topic Modelling			N-gram Modelling		
Prec.	Recall	F-Score	Prec.	Recall	F-Score	Prec.	Recall	F-Score	Prec.	Recall	F-Score	Prec.	Recall	F-Score
7.54	9.63	8.46	7.42	83.43	13.63	28.47	27.18	*27.81*	30.85	26.91	*28.75*	31.84	32.04	**31.94**

As illustrated by the results, the best accuracy is achieved by the n-gram modelling approach followed by topic modelling and the original approaches. Furthermore, even the original approach (i.e., the approach with the lowest accuracy among the approaches based on the STEP methodology), achieves a higher accuracy in comparison to the Saffron and EARS systems, although at the expense of a lower recall (i.e., 27.18%) compared to the EARS system. However, as already mentioned, in the case of EARS, a high Recall value was expected due to the experimental setup. This reflects positively on the performance of the STEP methodology, in comparison with these two traditional IR systems. 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 the STEP methodology.

5.8 Discussion

The original STEP methodology (i.e., *original approach*) profiles expertise using *unstructured* micro-contributions and a domain-specific concept extraction tool, i.e., the NCBO Annotator. Experiments using micro-contributions from the MCB and Genetics Wiki Projects have demonstrated that the *original approach* produces profiles with higher accuracy than traditional IR approaches, which perform expertise profiling using large corpus of static documents such as publications and reports (Section 5.7). Moreover, STEP captures the *temporal* aspect of expertise by creating short term and long term profiles.

The results of experiments using the enhanced STEP methodologies, i.e., STEP integrated with topic modelling (*topic modelling approach*) and STEP integrated with n-gram modelling (*n-gram modelling approach*), indicated the potential for further improvements in performance. By setting an appropriate threshold, i.e., concept weight threshold of 1.0, the *n-gram modelling approach* delivers a significantly improved accuracy (F-score: 31.94%).

The experimental results achieved by approaches based on the STEP methodology (i.e., original, topic and n-gram modelling approaches) are encouraging as they illustrate that even the original STEP methodology generates profiles with statistically significant higher accuracy than traditional expertise retrieval systems. While the reasons for statistically significant differences in performance of the original STEP, topic and n-gram modelling approaches can be investigated further, the focus of these experiments was to demonstrate that the concept extraction phase of the STEP methodology is not restricted to specific tools or techniques. In other words, not only can *domain-independent* methods be successfully integrated with STEP for concept extraction, they also lead to an improvement in performance. The statistically significant extent of this improvement wasn't the focus of these experiments, rather the focus was to establish the feasibility of using domain-independent methods for concept extraction in order to minimise the influence of domain-specific tools on the resulting profiles.

It is important to note that the results discussed in this chapter are directly dependent on the underlying concept extraction phase – i.e., the NCBO Annotator, which has been used for annotating terms that result from topic and n-gram modelling. However, the way in which terms are identified by the enhanced STEP approaches are domain-agnostic and differ from the method used by the Annotator for identifying terms. Therefore, the proposed approaches aim at *complementing term/topic extraction* given the *context of micro-contributions*.

5.9 Conclusions and Future Work

This chapter demonstrated the application of the STEP methodology (and enhanced methodologies) to *unstructured* micro-contributions to generate expertise profiles. The objective was to evaluate STEP as an expertise profiling methodology, in the context of evolving community-generated knowledge platforms containing unstructured micro-contributions (O3 in Section 1.5 of Chapter 1).

The experimental process and results of applying the original STEP methodology (i.e., *original approach*) to *unstructured* micro-contributions in the context of the MCB and Genetics Wiki projects were also presented and discussed. Evaluation results confirm that the STEP methodology creates expertise profiles with higher accuracy than two systems (Saffron and EARS), which use traditional IR methods and rely on the analysis of large corpora of static documents.

Furthermore, this chapter proposed and demonstrated the integration of two Language Modelling techniques with the STEP methodology. The pluggable architecture of STEP enabled the Concept Extraction phase to be enhanced - with *Lemmatization* as a pre-processing step, followed by either *topic modelling* or *n-gram modelling*. Evaluation results demonstrate a significant improvement in the accuracy of profiles generated by incorporating Language Models into STEP, as these approaches facilitate a domain-independent method for identifying entities in micro-contributions and therefore reduce reliance on domain-specific concept extraction tools and techniques.

Traditional approaches to expertise profiling associate an individual with expertise topics that emerge from a collection of static documents, such as publications, reports and grants, etc. Therefore, such approaches only take into account the presence of expertise topics in the documents associated with a person; i.e., persistency. However, as described in Chapter 4, the STEP methodology, ranks domain concepts in the long term profile of an expert by incorporating both the *uniformity* and *persistency* of the concepts across all the short term profiles of the expert. Hence, it provides the flexibility of computing expertise profiles that focus on *uniformly* behaving concepts or on concepts that are *uniformly* present throughout time.

The experimental setup for evaluating the three expertise profiling approaches used exclusively generated long term profiles. Future work will focus on overcoming the challenges of evaluating short term profiles. Assessing the validity and accuracy of expertise profiles is, by default, a subjective process. The complexity of performing such an assessment increases significantly in the case of short term profiles because of their intrinsic temporal nature. Consequently, novel, incremental ways of evaluating expertise profiles are required, in order to enable an appropriate tracking of the temporal aspect.

As described in Section 5.2, the baseline/benchmark data consists of expertise profiles defined and created by experts when they registered with the MCB and Genetics projects. Direct comparison of the expertise profiles generated by STEP and the baseline profiles, proved to be challenging, as the two sets of profiles represent expertise topics at different levels of abstraction. Expertise profiles generated by STEP using micro-contributions are typically very specific; i.e., the terminology describes specific domain aspects, while expertise profiles defined by experts when they register, mostly consist of general terms (e.g., *genetics*, *bioinformatics*, *microbiology*, etc.). The use of ontologies provides a means to compare not just the actual concepts extracted from micro-contributions, but also their ontological parents or children. This is investigated in Chapter 6, where methods are proposed for tailoring the expertise profiles, in order to achieve a level of abstraction comparable to the baseline.

The next chapter, chapter 6, presents the application of the STEP methodology to *structured* micro-contributions in the context of the collaborative authoring of the International Classification of Diseases, revision 11, ontology (ICD-11) [24]. It also demonstrates the use of semantic similarity measures for creating expertise profiles at different levels of granularity, thereby facilitating comparison of profiles, which represent expertise at different levels of abstraction.

Chapter 6 Application of STEP to Structured Micro-contributions

6.1 Introduction

Chapter 5 investigated the application of the Semantic and Time-dependent Expertise Profiling (STEP) methodology to unstructured micro-contributions (i.e., micro-contributions in natural language form), in the context of two Wiki projects in the biomedical domain. The research aimed at determining the applicability of the STEP methodology to community-driven, dynamic knowledge-curation platforms, in the context of knowledge domains containing *unstructured* micro-contributions (O3 in Section 1.5 of Chapter 1).

Towards this same objective, this chapter extends the investigation to *structured* micro-contributions, in order to analyse and evaluate the proposed methodology in a different context. This study presents the application of STEP to *structured* micro-contributions associated with the *International Classification of Diseases* (*ICD*) *revision 11*, (*ICD-11*) ontology [24]. ICD is the standard diagnostic classification developed by the World Health Organisation (WHO) [25] to encode information relevant for epidemiology, health management and clinical use [177]. Experts from diverse institutions around the world collaboratively curate the knowledge associated with the ICD-11 ontology. Each expert contributes to this process by authoring (i.e., creating / modifying / removing) ontological concepts.

Ontologies have become key elements in the design and development of intelligent decision-support systems, information retrieval systems and knowledge discovery applications and have been increasingly widely adopted, in particular by the biomedical community [146]. As a result, ontology engineering has evolved into a community-driven process, where experts focus on a particular domain to perform collaborative knowledge-curation. In this context, instead of contributing and authoring text, experts contribute and author ontological concepts, which due to their intrinsic nature, can be regarded as *structured* micro-contributions to the underlying ontology.

One of the major lessons learned from the application of STEP to unstructured micro-contributions, presented in Chapter 5, was the need for creating baseline profiles at a level of abstraction as close as possible to the actual micro-contributions. For example, an author of the MCB project described his expertise using very high-level concepts, such as *Genetics*, *Chemistry*, *Cell* and *Biology*, while the bottom-up profiles (generated by STEP) included topics such as *Metabolic pathways* and *Lipoprotein lipase*. Although both profiles may be accurate, direct comparison will yield very few common terms. This gives rise to another major objective of the

expertise profiling framework proposed in this thesis (Objective 4 (O4) in Section 1.5 of Chapter 1); i.e., development of a mechanism for *customising* the *granularity* of ontological concepts in expertise profiles in order to describe expertise at a level of specificity that accurately represents the knowledge embedded in micro-contributions, and that facilitates *comparison* and *evaluation* of profiles which describe expertise at different levels of abstraction. To achieve this aim, the research in this chapter proposes and evaluates a novel approach that takes advantage of the semantically related nature of *structured* micro-contributions. Three aspects are considered and discussed below.

Firstly, a novel method for creating bottom-up baseline expertise profiles using expertise centroids (that are identified using a semantic similarity metric) is proposed. Expertise centroids are ontological concepts that act as representatives for an area of the ontology by accumulating high similarity values against all micro-contributions located in that area. Such centroids not only streamline the evaluation of the methodology, but also provide a more accurate representation of the actual expertise.

Secondly, the results of applying STEP to the ICD-11 ontology engineering environment are described and the benefits of using semantic similarity for comparing the generated expertise profiles with baseline profiles, is demonstrated.

Thirdly and finally, a method for selecting the level of abstraction of STEP expertise profiles by analysing the coverage of the baseline profiles is described and discussed. The work presented in this chapter is in submission [178].

6.2 Materials and Methods

6.2.1 Experimental Data

The structured micro-contributions used in this research, have been compiled from the collaborative engineering process associated with the development of the ICD-11 ontology. The development of ICD-11 is a large-scale project with high visibility and impact for healthcare around the world. The ontology is currently being curated via a shared Web-based process, where many experts contribute, improve and review the domain-specific concepts. The collaborative authoring process is similar to other community curated knowledge bases – e.g., the Molecular and Cellular Biology (MCB) Wikipedia project – the difference being the resulting content. Within the ICD-11 project, experts provide incremental changes to ontological concepts, as opposed to free text contributions to existing articles. Hence, this general scenario appears to be appropriate for applying and evaluating the STEP methodology in the context of structured micro-contributions. In order to have a better understanding of the process, a brief description of the ICD-11 workflow and datasets is provided below.

A large community of medical experts around the world is involved in the authoring of ICD-11 via a collaborative Web-based platform, called *iCAT*. iCAT is a customisation of the generic Web-based ontology editor, Web Protégé [50]. To date, more than 270 domain experts around the world have used iCAT to author 45,000 classes, perform more than 260,000 changes and create more than 17,000 links to external medical terminologies [49]. iCAT uses the *Change and Annotation Ontology* (*ChAO*) [179] to represent changes and therefore provides a semantic log of changes and annotations. Change types are ontology classes in ChAO and *changes* to the ICD-11 ontology are instances of these classes. Similarly, *notes* (or *annotations*) that users attach to classes or threaded user discussions are also stored in ChAO. Every *change* and *annotation* provides information about the user who performed it, the involved concept, a timestamp and a short description of the changed or annotated concepts/properties.

Two main types of data were used for the analysis:

1. The semantic log of changes and annotations to the ICD-11 ontology (extracted from a snapshot of ChAO on 18th March 2013; and
2. The structure of the ICD-11 ontology.

The following illustrates two examples of contributions to ICD-11, extracted from iCAT.

URI: http://who.int/icd#2255_ea0b2e17_d398_4474_8ed0_c2b2ced85d96
Label: Proliferative Diabetic Retinopathy
Type: Composite_Change
Date: 05/05/2010 14:53:19
Description: Added a new definition. Prefilled to BC9.2 Proliferative Diabetic Retinopathy
Apply to: http://who.int/icd#2255_ea0b2e17_d398_4474_8ed0_c2b2ced85d96

URI: http://who.int/icd#1727_ea0b2e17_d398_4474_8ed0_c2b2ced85d96
Label: Combined arterial and vein occlusion
Type: Composite_Change
Date: 05/05/2010 13:17:17
Description: Create class with name: H34.81 Combined arterial and vein occlusion,
parents: H34.8 Other retinal vascular occlusions

ICD-11 has a very large change log; however, the majority of users perform a very small number of changes on a very small number of concepts – i.e., up to five ontological concepts. The large majority of changes are made by a minority of users that perform bulk operations on a large number of concepts – due to their position in the project; e.g., administrators or group leaders committing or approving a large number of changes. Changes such as commit and approve operations, involve a large number of concepts, but do not necessarily reflect the expertise of users who perform them. For example, a Working Group may manage an entire branch of ICD-11, such as *Infectious diseases.* However, all changes to the structure and content of the *Infectious diseases*

branch may be committed by a single user (e.g., the chair of the Working Group) at regular intervals. This leads to this particular user being associated with, for example, 6000+ concepts representing the expertise of the entire Working Group, which is not an accurate reflection of his/her individual expertise. Thus, maintenance changes were excluded from the analysis.

Following the filtering process (i.e., removal of maintenance changes), the resulting dataset comprised a total of 19,888 changes by 22 authors involving 737 unique concepts over a period of four years. The focus is on the number of unique concepts to which an expert contributed (because *concepts* represent *expertise*), rather than on the total number of changes made by the expert.

The hierarchical structure of the ICD-11 ontology (Figure 6-1 illustrates a small sub-set of the entire ontological structure) facilitates access to concepts and their ontological parent-child relationships. Expertise is described at different levels of abstraction by applying semantic similarity measures to this structure.

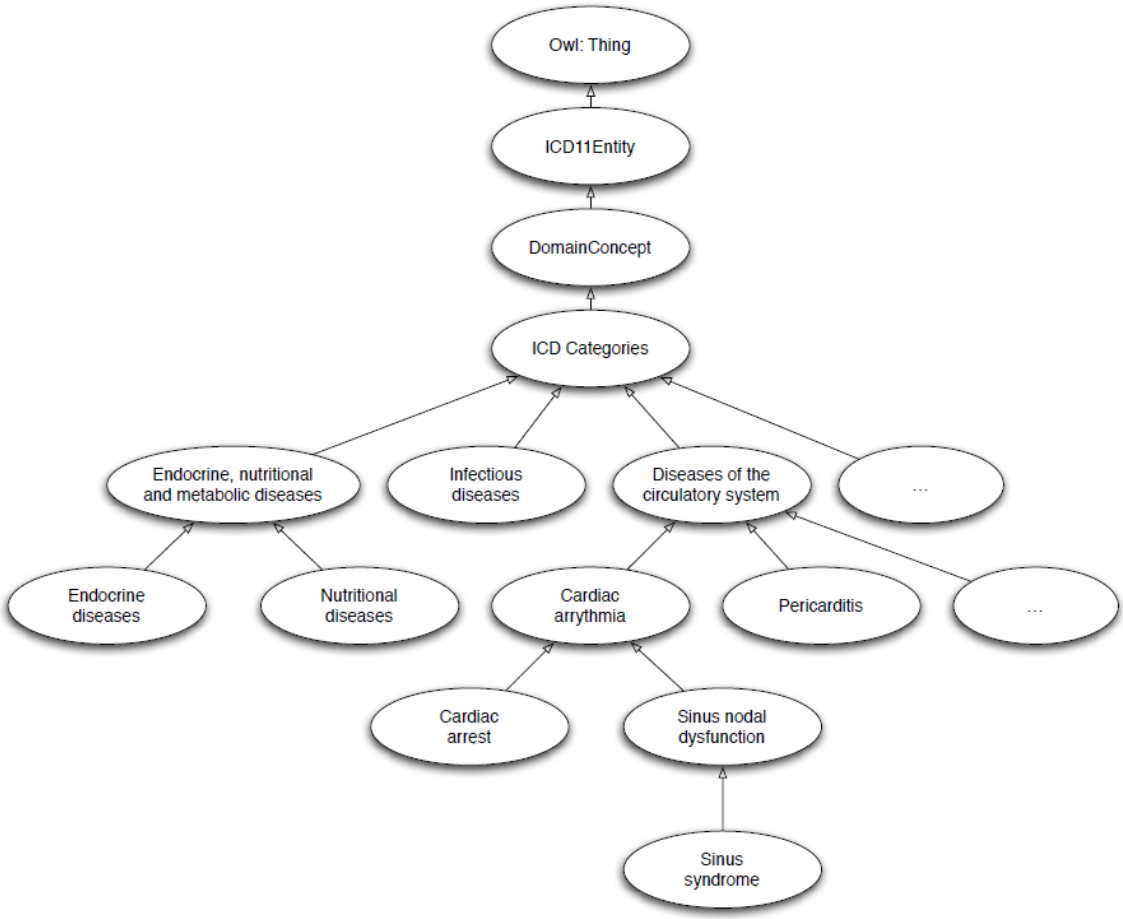

Figure 6-1: Excerpt from the ICD-11 Ontology showing its high-level structure. The classification of diseases in ICD-11 starts with a set of well-defined branches (e.g., Infectious diseases, Diseases of the circulatory system, etc.) and is refined into sub-categories, with the leaves of the ontology representing instances of actual diseases.

6.2.2 Semantic similarity measure for creating expertise centroids

A good semantic similarity measure needs to take into account the specific characteristics of the underlying ontology. The first goal is to define a semantic similarity that accurately reflects the

semantics of micro-contributions in close proximity, in addition to their degree of specificity in the larger context of the ontology. Hence, a hybrid approach was adopted, which is based on the work of Sanchez et al. [73]. More concretely, an edge-based measure (which measures the path between concepts) is redefined in terms of the Information Content (IC) of concepts, which is then expressed via its specificity in the ontology. Contrary to the classical method that computes IC using term appearance probabilities in a given corpora [75], here, IC is computed using the taxonomic structure of the ontology.

Following the method proposed in [73], the length of the minimum path separating two micro-contributions (concepts) – i.e., $min\_path(C_1, C_2)$ is considered. This path evaluates the differential semantic features of both concepts as a function of the amount of non-common ancestors found through the shortest link connecting them. In terms of IC, the minimum path length can be approximated as the sum of the amount of differential information between two concepts, as outlined in Eq. 6-1:

$$[IC(C_1) - IC(C_2)] + [IC(C_2) - IC(C_1)]$$

(Eq. 6-1)

The differential information of one concept compared to another can be quantified by subtracting their common information (i.e., the IC of the least common subsumer (LCS) of both concepts), from the IC of the concept alone. Formally, this is expressed in Eq. 6-2:

$$\begin{aligned} |\min\_path\,(C_1, C_2)| &= [IC(C_1) - IC(C_2)] + [IC(C_2) - IC(C_1)] \\ &\cong \left[IC(C_1) - IC\left(LCS\,(C_1, C_2)\right)\right] + \left[IC(C_2) - IC\left(LCS\,(C_1, C_2)\right)\right] \\ &= IC\,(C_1) + IC\,(C_2) - 2 * IC\left(LCS\,(C_1, C_2)\right) \end{aligned}$$

(Eq. 6-2)

Subsequently, the depth of a concept *C*, i.e., the *min_path* between *C* and the root node, is also redefined in terms of IC, as shown in Eq. 6-3:

$$depth(C) \cong IC(C) + IC(root) - 2 * IC\big(LCS(C, root)\big)$$

(Eq. 6-3)

As the root node is general enough and can potentially subsume any other concepts, its IC can be considered zero; therefore, the depth of a concept *C* can be approximated as in Eq. 6-4:

$$depth(C) \cong IC(C) - IC(root) \cong IC(C)$$

(Eq. 6-4)

Using the definitions above, the *Wu and Palmer similarity measure* (Eq. 2-3) is redefined in terms of IC. Wu and Palmer consider the relative depth of the LCS of concepts in the ontology as an indication of similarity. In other words, in addition to the shortest path separating two concepts, it takes into account the degree of taxonomical specialisation of their LCS. *Rada* [76] only considers the length of the minimum path connecting the concepts (Eq. 2-1). *Leacock and Chodorow* [77], considers the maximum depth of the ontology (Eq. 2-2); however, in the context of the experiments presented here, this isn't a differentiating factor in determining the similarity between concepts, as all concepts come from the ICD-11 ontology.

Table 6-1 provides examples of similarity values calculated for two pairs of concepts using the similarity algorithms described above. The concepts *Enteroviral gastritis* and *Cytomegaloviral gastritis* are both specific types of *Viral gastritis*, a common infection of the stomach and intestines. However, the concepts *Diseases of pancreas* and *Diseases of stomach* represent two different classes of diseases of the digestive system – hence their common LCS, *Diseases of the digestive system*.

Table 6-1: An example of concept similarity calculated for two pairs of concepts using various algorithms

Concept pair	LCS	LCS depth	Path	Rada	L&C	W&P
Enteroviral gastritis *Cytomegaloviral gastritis*	*Viral gastritis*	8	2	2	2.30	**0.889**
Diseases of pancreas *Diseases of stomach*	*Diseases of the digestive system*	4	2	2	2.30	**0.8**

From a medical perspective, the semantic similarity of concepts in the first pair is higher than the semantic similarity of concepts in the second pair. However, as shown in Table 6-1, the shortest path between the concepts in both pairs is the same. Furthermore, both the Rada and L&C algorithms calculate the same similarity value for concepts in both pairs, despite the fact that the taxonomical specialisation of the LCS of the concept pairs is significantly different, as highlighted by the difference in their depth in the ontology. *Viral gastritis* (depth=8) is more specialised than *Diseases of the digestive system* (depth=4). W&P, on the other hand, calculates a higher similarity between *Enteroviral gastritis* and *Cytomegaloviral gastritis* (similarity=0.889, LCS depth=8) compared to *Diseases of pancreas* and *Diseases of stomach* (similarity=0.8, LCS depth=4). Consequently, the Wu and Palmer similarity algorithm is redefined in terms of IC (Eq. 6-5) in order to calculate the pairwise similarity of concepts.

$$Sim_{w\&p}(C_1, C_2) = \frac{2 * depth(LCS(C_1, C_2))}{|\min\_path(C_1, C_2)| + 2 * depth(LCS(C_1, C_2))} \cong \frac{2 * IC(LCS(C_1, C_2))}{IC(C_1) + IC(C_2) - 2 * IC(LCS(C_1, C_2)) + 2 * IC(LCS(C_1, C_2))} = \frac{2 * IC(LCS(C_1, C_2))}{IC(C_1) + IC(C_2)}$$

(Eq. 6-5)

This framework for estimating edge-based similarity measures based on the IC of concepts relies heavily on accurate estimation of IC. In order to calculate the IC of concepts, a number of approaches were analysed, which only considered the subclasses of a concept relative to the maximum number of concepts in the taxonomy, e.g., [180] and [181]. None of these approaches consider the depth of a concept as expressed by its subsumers. Consequently, they are unable to differentiate between concepts with the same number of hyponyms/leaves but different depths in the taxonomy. Therefore, the approach proposed by Sánchez et al. [73] is used, which estimates IC intrinsically as the ratio between the number of leaves of *C*, as a measure of its generality, and the number of taxonomical subsumers, as a measure of its depth in the ontology (Eq. 6-6).

$$IC(C) = -logp(C) \cong -log\left(\frac{\frac{|leaves(C)|}{|subsumers(C)|} + 1}{\max\_leaves + 1}\right)$$

(Eq. 6-6)

where *leaves(C)* is the set of concepts found at the end of the taxonomical tree under concept *C* and *subsumers(C)* is the complete set of taxonomical ancestors of *C* including itself. It is important to note that in case of multiple-inheritance all the ancestors are considered. The ratio is normalised by the least informative concept (i.e., the root of the taxonomy), for which the number of leaves is the total number of leaves in the taxonomy (*max_leaves*) and the number of subsumers of the root including itself is 1. In order to produce values in the range of 0 and 1 and avoid log (0), 1 is added to both expressions. This approach also prevents dependence on the specificity and detail of the inner taxonomical structure by relying on taxonomical leaves rather than the complete set of subsumers.

6.2.3 Creating baseline expertise profiles from expertise centroids

Using the similarity measure defined above, baseline expertise profiles were created by selecting so-called *expertise centroids*. These are concepts associated with micro-contributions that have a high aggregated similarity value across all micro-contributions found in close proximity to

them in the ontology. In order to find *expertise centroids*, a matrix of the pair-wise similarity of all concepts to which an expert had contributed, is computed using the measure defined in Eq. 6-5 and described in the previous section. Subsequently, for every concept, the total pair-wise similarity is calculated by iterating over all pair-wise similarities computed in conjunction with all other concepts – as per Eq. 6-7:

$$TSim(c, C) = \sum_{i=1}^{n-1} sim(c, c_i)$$

(Eq. 6-7)

Expertise centroids, and the resulting baseline profiles, are then identified by using the median as a threshold over the set of all total pair-wise similarities – as shown in Eq. 6-8:

$$Baseline(C) = \{c | sim(c, C) \geq median[TSim(c_1, C), TSim(c_2, C), \dots, TSim(c_n, C)]\}$$

(Eq. 6-8)

When calculating the pair-wise similarity of concepts, the IC of the LCS of concept pairs is used; i.e., the most taxonomically specific ancestor of concept pairs is used to create baseline profiles. However, if the structure of the ICD-11 ontology is traversed to identify ancestors with lower taxonomical specification (lower information content), baseline profiles can be created that contain concepts describing expertise at higher levels of abstraction. In other words, more taxonomically specific ancestors result in finer-grained profiles, while ancestors which are less specific and therefore have a lower IC result in profiles containing concepts which represent expertise at higher levels of abstraction.

6.3 Experimental setup

The second goal of this study is to use the baseline expertise profiles for evaluating the application of the STEP methodology to *structured* micro-contributions. In addition, the aim is to demonstrate how expertise profiles, at different levels of abstraction, can be generated - by looking at the coverage of the STEP profiles over the baseline. The following describes the experimental setup for achieving these goals/tasks.

6.3.1 Evaluating STEP profiles against the baseline expertise profiles

Given two sets of concepts, one representing the STEP profile *SC* = $\{SC_1, SC_2, \dots, SC_n\}$ and one the baseline *BC* = $\{BC_1, BC_2, \dots, BC_n\}$, the aim of this task is to find the maximal subset of baseline concepts $\{BC_i \in BC\}$ or the maximal subset of STEP concepts $\{SC_i \in SC\}$ (depending on which initial set is larger) that maximises the overall similarity of *SC* against *BC*. To some extent,

the underlying principle is the same as in a standard experimental setup in which one requires the computation of Precision / Recall and F-Score, without relying on exact matching of the candidates against the gold standard. The most important constraint in this setting is that each concept from *SC* or *BS* can only be accounted for once – in order to avoid an artificial increase in similarity via multiple counts. The final similarity score is computed as the normalised sum of the pairwise $SC_i - BC_i$ similarity values (based on Eq. 6-5 and Eq. 6-6), as shown in Eq. 6-9:

$$Similarity(SC, BC) = argmax \left\{ \frac{p}{n} * \frac{q}{m} * \sum_{i=1}^{q} sim(c_i, c_k) \right\}$$

(Eq. 6-9)

where p is the number of concepts matched in *BC*, $n = |BC|$, $m = |SC|$, q is the number of concepts matched in *SC*, $c_i \in SC$, $c_k \in BC$ – such that the overall sum is maximised.

In the following example, the assumption is that for a given author, the baseline profile contains concepts C_1, C_2 and C_3 and the corresponding STEP profile contains concepts C_4 and C_5. Table 6-2 illustrates the concept similarity matrix for the profiles, while Eq. 6-10 explains the resulting values.

Table 6-2: Example of the similarity matrix computed for comparing a STEP profile to a baseline profile

	Baseline Concepts		
STEP Concepts	C_1	C_2	C_3
C_4	0.73	**0.52**	0.31
C_5	**0.89**	0.01	0.24

$$Similarity(STEP, BC) = \frac{2}{3} * \frac{sim(C_1, C_5) + sim(C_2, C_4)}{2} = \frac{2}{3} * \frac{0.89 + 0.52}{2} = 0.47$$

(Eq. 6-10)

As illustrated above, due to the single inclusion and maximality constraints, the final matching includes only the concepts C_1 and C_2 from the baseline (because the pair $C_2 - C_4$ has a higher similarity than any of the pairs formed by C_3). Furthermore, it can be observed that the maximum overall similarity is achieved by including the pairs $C_2 - C_4$ and $C_1 - C_5$ in the computation, rather than $C_1 - C_4$, even though the similarity of $C_1 - C_4$ (0.73) is higher than that of $C_2 - C_4$ (0.52). Finally, including (C_3, C_4) or (C_1, C_4) in the overall similarity, would lead to overrepresentation of similarity between the profiles, as the similarity of a single concept in the STEP profile, i.e., C_4 would be considered with multiple concepts in the baseline profile. In this example, $p = n = 2$ (since

all STEP concepts are included in the computation) and $q = 2$, $m = 3$, since only 2 of the 3 baseline concepts have been used.

To reiterate, the goal is to identify concepts in the compared profiles, which represent similar topics. The most extreme/rigid comparison, as described in Chapter 5, involves only counting exact matches between concepts listed in the profiles. This leads to underrepresentation of similarity, as different concepts in the two profiles may be semantically similar and represent similar topics. The other extreme is to include similarity between all concept pairs from the two profiles. This leads to overrepresentation of similarity, as any two concepts will have a similarity value associated with them represented in the matrix. Therefore, only the maximum pair-wise similarity of concepts is included in the overall result, ensuring that the components of every pair are only considered in one pair wise similarity in order to prevent overrepresentation of similarity of the corresponding profiles.

6.3.2 Investigating the coverage of STEP profiles over the baseline expertise profiles

The second aim is to investigate the generation of expertise profiles at different levels of abstraction. The method devised for performing experiments in this context relies on two aspects:

1. Defining and compiling the subset of baseline concepts that provide a target level of abstraction;
2. Defining and computing the coverage of the STEP concepts given the above-defined subset of baseline concepts at a specified target level of abstraction.

The first of the above aspects was studied via a clustering approach. More concretely, the 1:n relationship is observed between each concept in the STEP profile and all combinations of clusters that can be formed from the concepts in the baseline profile. This relationship was quantified by means of the centrality of the STEP concept in the context of a given cluster – as shown in Eq. 6-11. Centrality is calculated as the normalised sum of pair-wise similarity between a STEP concept and a baseline cluster of concepts, normalised by computing the proportion of baseline concepts in the cluster to the total number of concepts in the baseline profile. Again, the pair-wise similarity calculation uses the formulations defined in Eq. 6-5 and Eq. 6-6. Measuring this centrality, provides an understanding of the extent to which a STEP concept is able to cover (or represent) a set of baseline concepts. And since this relies on a semantic similarity measure that takes into account the path between the concepts, as well as the information content, a high centrality score will ensure an appropriate (mid) level of abstraction for the resulting expertise profile.

$$Centrality(c, Cl_n) = \frac{1}{n} * \frac{|Cl_n| - n}{|Cl_n|} * \left[\sum_{i=1}^{n} sim(c, t_i)\right]$$

(Eq. 6-11)

Where $Cl_n \subseteq Cl$, Cl – the set of all concepts in a baseline profile, $|Cl_n| = n$, $1 \leq n \leq |Cl|$, and $t_i \in Cl_n$.

The final overall coverage is then computed by finding the set of centrality measures that lead to the highest average – Eq. 6-12:

$$Similarity(STEP, Cl) = \frac{1}{|STEP|} * \sum_{i=1}^{|STEP|} max\|Centrality(c_i, Cl)\|$$

$$where\ c_i \in STEP$$

(Eq. 6-12)

As in the case of the previous section, the following presents an example illustrating the computation of this final coverage, where *BC = {C3, C4, C5}* and *SC = {C1, C2}*. The first step is to create all possible clustering combinations from *BC* – as shown in the second column of Table 6-3.

Table 6-3: Example of the similarity matrix for a STEP and its corresponding baseline profile

Cluster	STEP Concept	Centrality
$\boldsymbol{Cl_1}$	C_3	$Centrality(C_1) = sim_{w\&p}(C_1, C_3)$ $Centrality(C_2) = sim_{w\&p}(C_2, C_3)$
$\boldsymbol{Cl_2}$	C_4	$Centrality(C_1) = sim_{w\&p}(C_1, C_4)$ $Centrality(C_2) = sim_{w\&p}(C_2, C_4)$
$\boldsymbol{Cl_3}$	C_5	$Centrality(C_1) = sim_{w\&p}(C_1, C_5)$ $Centrality(C_2) = sim_{w\&p}(C_2, C_5)$
$\boldsymbol{Cl_4}$	C_3, C_4	$Centrality(C_1) = {}^{2}/_{3} * {}^{1}/_{2} * [sim_{w\&p}(C_1, C_3) + sim_{w\&p}(C_1, C_4)]$ $Centrality(C_2) = {}^{2}/_{3} * {}^{1}/_{2} * [sim_{w\&p}(C_2, C_3) + sim_{w\&p}(C_2, C_4)]$
$\boldsymbol{Cl_5}$	C_3, C_5	$Centrality(C_1) = {}^{2}/_{3} * {}^{1}/_{2} * [sim_{w\&p}(C_1, C_3) + sim_{w\&p}(C_1, C_5)]$ $Centrality(C_2) = {}^{2}/_{3} * {}^{1}/_{2} * [sim_{w\&p}(C_2, C_3) + sim_{w\&p}(C_2, C_5)]$
$\boldsymbol{Cl_6}$	C_4, C_5	$Centrality(C_1) = {}^{2}/_{3} * {}^{1}/_{2} * [sim_{w\&p}(C_1, C_4) + sim_{w\&p}(C_1, C_5)]$ $Centrality(C_2) = {}^{2}/_{3} * {}^{1}/_{2} * [sim_{w\&p}(C_2, C_4) + sim_{w\&p}(C_2, C_5)]$
$\boldsymbol{Cl_7}$	C_3, C_4, C_5	$Centrality(C_1) = {}^{3}/_{3} * {}^{1}/_{3} * [sim_{w\&p}(C_1, C_3) + sim_{w\&p}(C_1, C_4) + sim_{w\&p}(C_1, C_5)]$ $Centrality(C_2) = {}^{3}/_{3} * {}^{1}/_{3} * [sim_{w\&p}(C_2, C_3) + sim_{w\&p}(C_2, C_4) + sim_{w\&p}(C_2, C_5)]$

In the second step, the centrality of each concept in *SC* is computed against every possible clustering option. Assuming that C_1 yields the highest centrality relative to Cl_5 (cluster 5) and C_2 yields the highest centrality relative to Cl_7 (cluster 7), the overall similarity of the STEP and its corresponding baseline profile is obtained by computing the normalised sum of maximum centrality values for the concepts in the STEP profile (Eq. 6-13).

$$Similarity(STEP, BS) = \frac{1}{2} * \left(Centrality(C_1, Cl_5) + Centrality(C_2, Cl_7)\right)$$

(Eq. 6-13)

The best coverage of the baseline profile is determined by considering concepts in the clusters that yield the highest centrality values for concepts in the STEP profile; i.e., cluster 5 and cluster 7. This is achieved by taking the union of concepts in the clusters that result in the highest centrality for concepts in the STEP profile. In this example, the union of concepts in clusters 5 and 7 is C_3, C_4 and C_5.

Figure 6-2 depicts a concrete view over the centrality between a STEP concept (C_1 – *Iron deficiency anaemia due to decreased duodenal absorption*) and a cluster of baseline concepts consisting of C_3 (*Hereditary iron deficiency anaemia*) and C_5 (*Iron deficiency anaemia secondary to blood loss*). As discussed above, the centrality of C_1 is computed in the context of C_3 and C_5 using the semantic similarity measure defined in Eq. 6-5 and Eq. 6-6. In this example, the pair-wise similarity values $sim_{W\&P}$ (C_1, C_3) and $sim_{W\&P}$ (C_1, C_5) are calculated using the same LCS concept, i.e., *Iron Deficiency anaemia*. Furthermore, all concepts have very close IC (Information Content) (Eq. 6-6), as they all have the same number of subsumers, and relatively similar number of leaves (C_3 and C_5 have 2 leaves and C_1 has one leaf). This will lead to C_1 having a high centrality value when considered in conjunction with the $C_3 - C_5$ cluster.

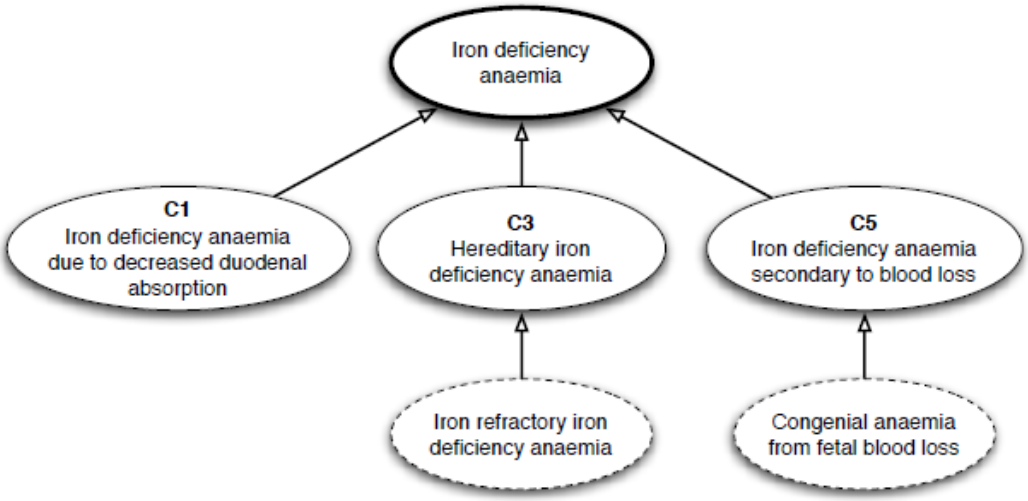

Figure 6-2: Excerpt from ICD-11 Ontology used to exemplify computation of the coverage of STEP profiles

6.4 Experimental Results

Structured micro-contributions were collected for the 22 experts from the iCAT system, each of which had contributed to an average of 33.5 ontological concepts. In the initial setup, baseline expertise profiles were created using the proposed semantic similarity measure, and STEP profiles were created by applying STEP to the experts' micro-contributions.

The following outlines the analysis performed to evaluate the proposed bottom-up method of creating baseline expertise profiles using expertise centroids and semantic similarity measures. Concepts included in the baseline profile of an author were selected based on their similarity with other concepts to which the author had contributed. More concretely, a concept is included in the baseline if, its total pair-wise similarity with other concepts, is greater than the median of total pair-wise similarity of all concepts.

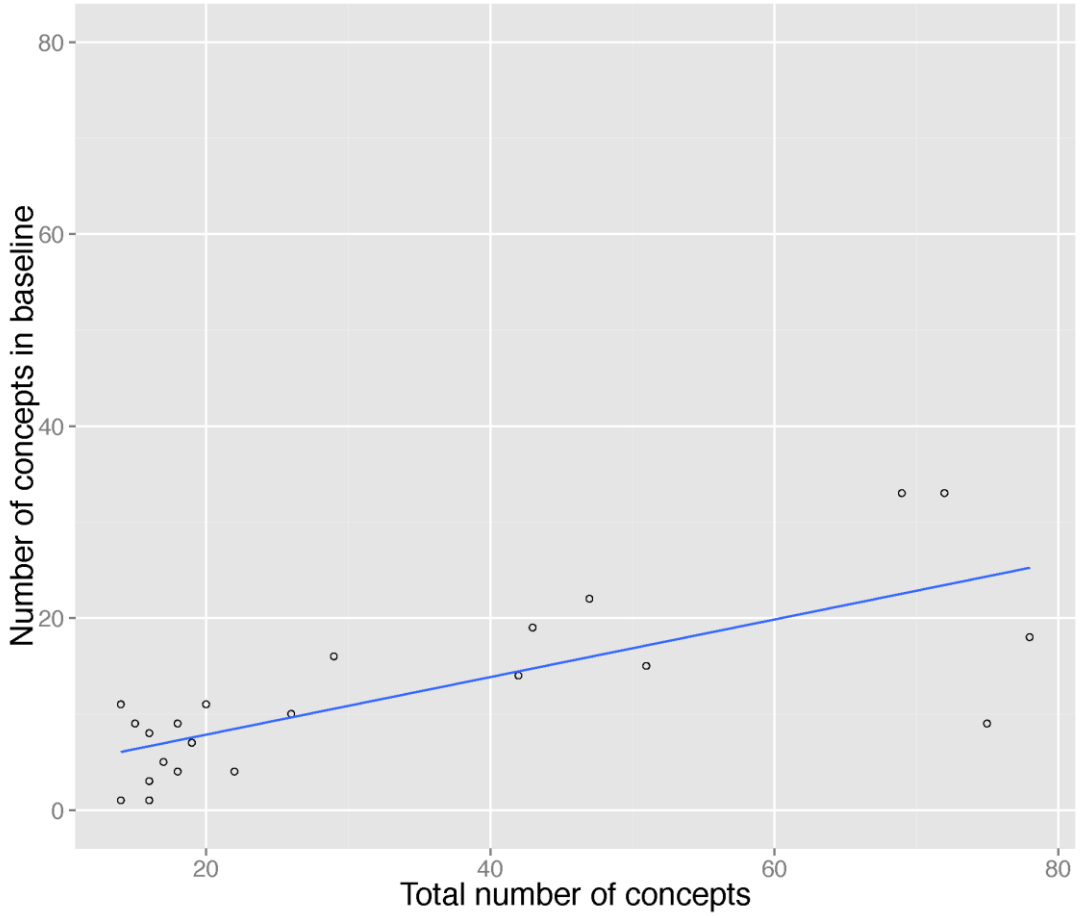

Figure 6-3: The creation of baseline expertise profiles from the total number of concepts authored by each of the 22 experts leads to a 64.45% decrease in the number of concepts, from an average of 33.5 concepts to 11.91 concepts per author.

The results of the baseline profile creation process are presented in Figure 6-3. As depicted in Figure 6-3, the process resulted in a 64.45% decrease in the number of concepts included in the baseline profiles, from an average of 33.5 concepts to 11.91 concepts per author. Qualitatively, the

expertise centroids were located, as expected, at a fairly uniform distance (both from a breadth, as well as from a depth perspective) from all concepts that were in close proximity.

A similar reduction effect can also be observed in the creation of STEP profiles, when generating expertise snapshots by increasing the threshold of the weights associated with the concepts contained within the profiles. As depicted in Figure 6-4, application of the STEP methodology to all contributions of an author leads to the inclusion of almost all concepts in the STEP profiles, when no threshold is specified (on average 32.95 concepts per author, compared to the initial 33.5 average concepts per author). However a filtering effect is seen when increasing the threshold – for an initial threshold of 0.05, there is a large reduction of 77.38% in the number of concepts (7.45 concepts/profile), followed by a quasi-linear behaviour for thresholds between 0.05 and 0.15 (falling to 2.71 concepts/profile at 0.15). Note that STEP profiles are built using a combination of uniformity and persistency measures. Hence, the increase in threshold leads to retaining only those concepts that are persistent and uniformly distributed throughout the entire time the author has contributed to the project – which is a normal expectation from an expertise profile.

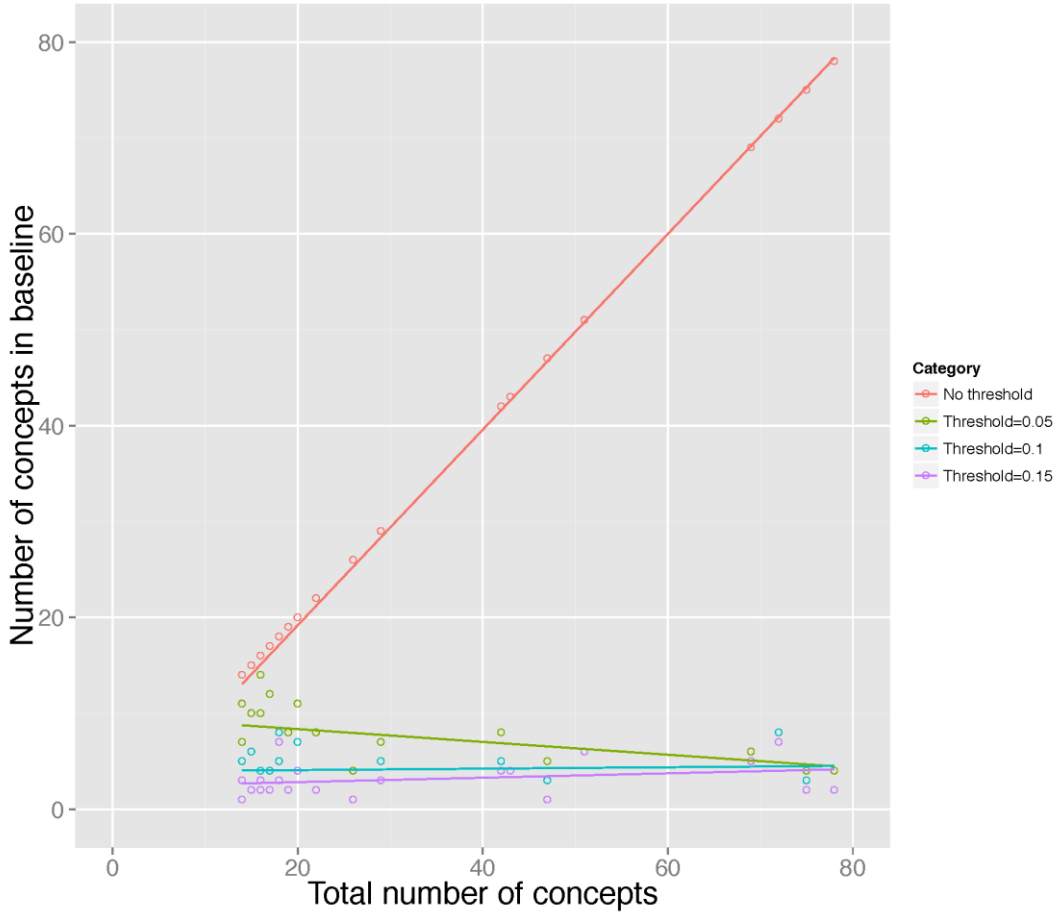

Figure 6-4: The effect of varying the weight threshold over STEP profiles. When no threshold is specified, the resulting profile mirrors the set of initial micro-contributions – on average 32.95 concepts/profile, compared to the initial average of 33.5 concepts/profile. A filtering effect occurs as the threshold is increased –an average of 7.45 concepts / profile at threshold of 0.05 falls to an average of 2.71 concepts / profile at threshold of 0.15.

Using the baseline expertise profiles, the STEP profiles were evaluated at different thresholds, using the same similarity measure. Experimental results are summarised in Figure 6-5 and detailed results are provided in Figure 6-6; i.e., expanded for all 22 experts.

As can be observed in Figure 6-5, when no threshold is imposed, i.e., all concepts in the STEP profiles are included in the comparison, an almost exact match is achieved between the STEP and baseline profiles (99.32%). Increasing the threshold on STEP to 0.05, 0.1 and 0.15, results in similarities of 49.12%, 26.17% and 22.91%, respectively. The results indicate that when the weight threshold = 0.05, there is an overall decrease of 77.38% in the number of concepts and 49.12% similarity is achieved against the baseline. While at the highest STEP threshold (0.15), only 8.24% of concepts are included in the STEP profiles, an almost 23% similarity is achieved. In practice, this shows that even the most restrictive STEP profile is still similar and able to match 23% of the baseline expertise.

These results were compared to those discussed in Chapter 5, where at weight thresholds of 0, 0.1 and 0.2, F-score values of 18.91%, 21.03% and 20.31% were achieved, respectively. While the results cannot be compared directly (since the previous results were generated using unstructured contributions and achieved via exact matching), the conclusion can be drawn that comparing profiles using semantic similarity methods and ontological relationships, results in more accurate comparisons than simply identifying exact matches between the content of profiles. The methods proposed here take into account different concepts that represent semantically similar topics. While the exact matching method used earlier considers such concepts to be completely different and therefore, results in a less accurate (and lower) measurement of similarity between profiles.

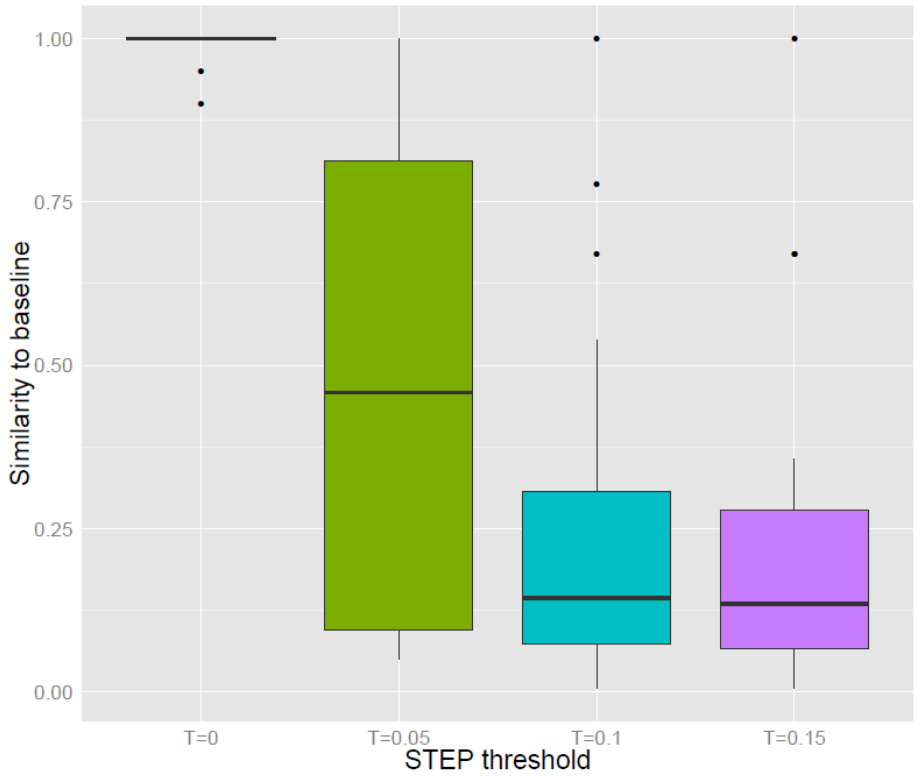

Figure 6-5: Summarised representation of the evaluation of STEP profiles using the baseline expertise profiles

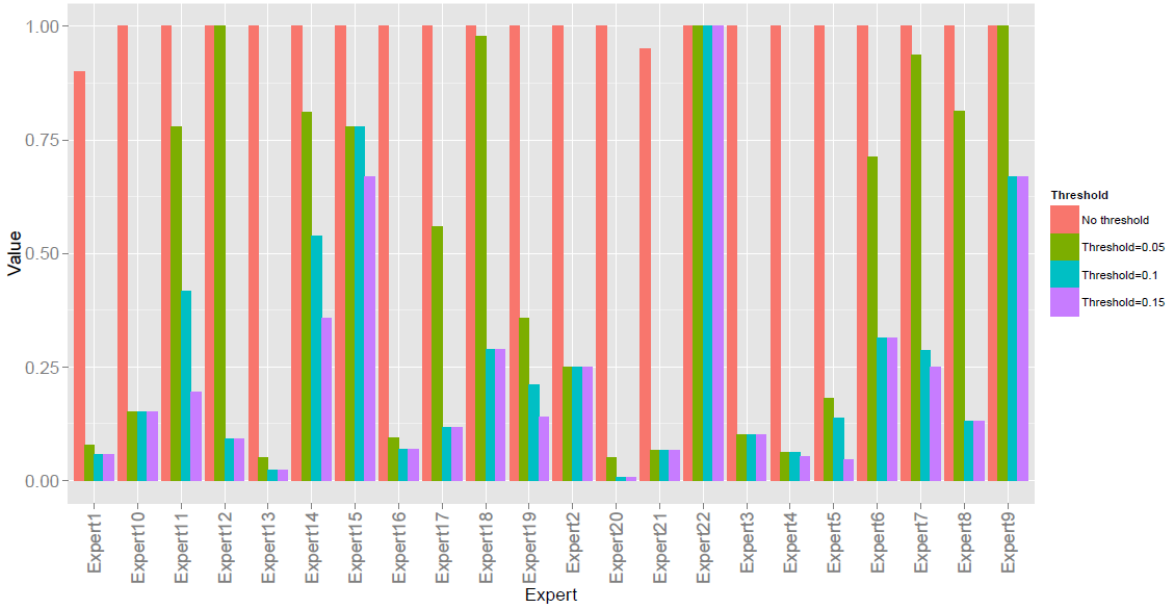

Figure 6-6: Expanded representation of the evaluation of STEP profiles using the baseline expertise profiles

Finally, an investigation was performed on the coverage of STEP profiles over the baseline profiles at different levels of abstraction. As shown in the results depicted in Figures 6-7 and 6-8, STEP profiles exhibit an almost constant behaviour in terms of coverage of the baseline profiles, independently of the imposed threshold. Increasing the threshold does lead to a small decrease in the centrality of concepts in the STEP profile relative to the optimal subset of baseline concepts. However, this decrease is minimal (on average 2% per threshold step) and is associated with

eliminating the concepts that contribute to noise, rather than excluding concepts in which an author has considerable expertise. This in turn suggests that weights associated with concepts in a STEP profile, represent the true level of an author's expertise in the topics represented by those concepts.

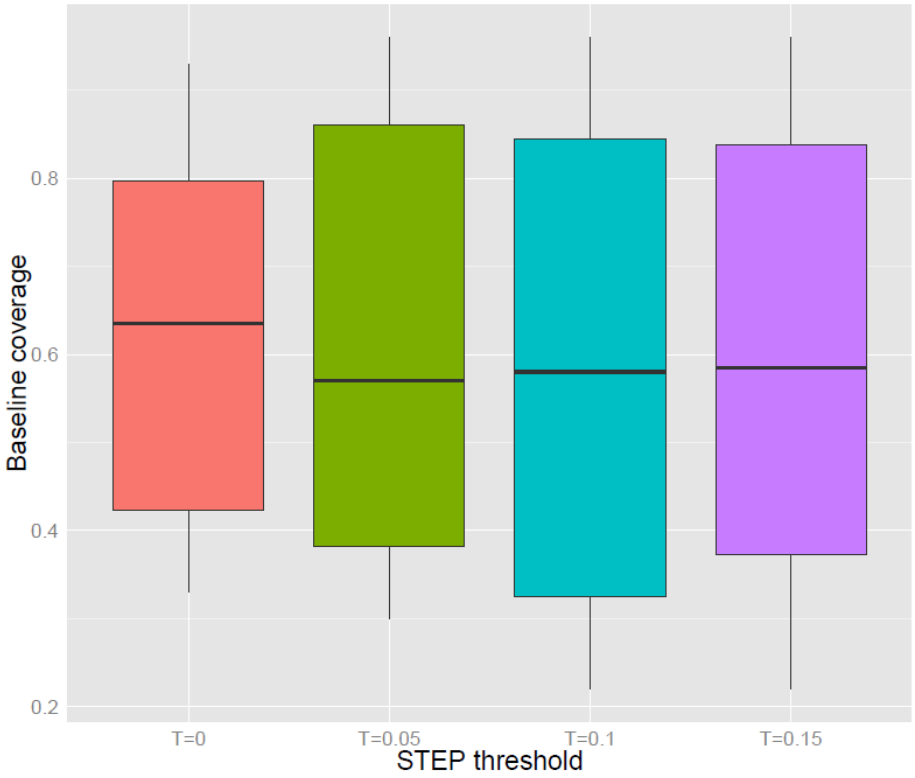

Figure 6-7: Summary illustrating the coverage of STEP profiles over the baseline profiles

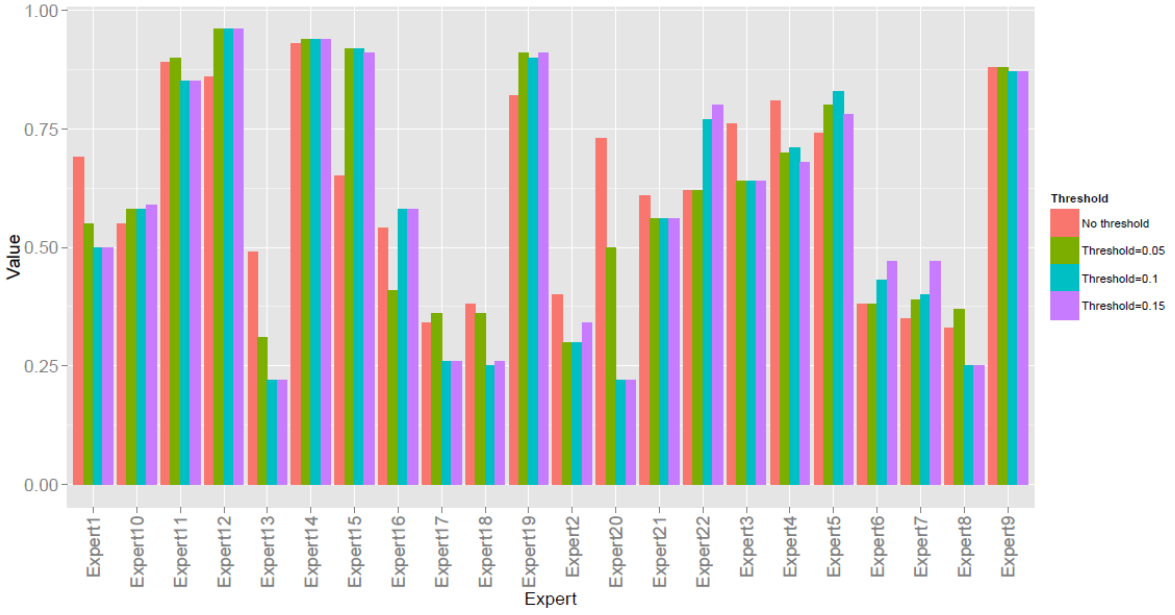

Figure 6-8: Detailed representation showing the coverage of STEP profiles over the baseline profiles

6.5 Discussion

One of the conclusions drawn from the experiments presented in Chapter 5, was that the comparison of baseline profiles created by experts (containing coarse-grained description of expertise) and expertise profiles created by STEP from micro-contributions (containing fine-grained description of expertise), proved to be challenging, as the two sets of profiles described expertise at different levels of abstraction. This in turn highlighted the importance of facilitating comparison and evaluation of expertise profiles that represent expertise topics at different levels of abstraction. This chapter proposed semantic similarity methods for creating profiles at different levels of granularity. Experiments were performed using these methods for creating baseline profiles at a level of abstraction, comparable with the STEP profiles, in order to facilitate comparison and evaluation. Experimental results presented above highlight the significance of *semantic similarity* methods in profiling expertise at diverse levels of abstraction and in assessing and comparing expertise profiles. However, this study has identified a number of limitations, as discussed below:

– The results above are generated using *structured* contributions in the context of collaborative authoring of the ICD-11 ontology and therefore, should be verified using unstructured contributions, where ontological concepts are derived through annotating experts' micro-contributions, as described in Chapter 5.

– Baseline expertise profiles were created and used to evaluate profiles created by the STEP methodology, however, a comprehensive evaluation can only be performed using a gold standard, which represents the true expertise of an individual with absolute confidence. Such gold standards would need to be developed manually and maintained/updated over time and even then would contain some subjectivity.

– The results presented here are based on a snapshot of the ICD-11 ontology, as the ontology is still under development. Hence, in order to create profiles that represent a comprehensive view of experts' expertise, the STEP methodology and the methods proposed in this chapter for fine-tuning the granularity of profiles, would need to be applied to the complete set of expert contributions to the collaborative authoring of ICD-11.

6.6 Conclusion and Future Work

This chapter demonstrated the application of STEP to *structured* micro-contributions in the context of collaborative authoring of the ICD-11 ontology [24]. The objective was to evaluate STEP as an expertise profiling methodology, in the context of evolving community-driven dynamic knowledge-curation platforms, containing *structured* micro-contributions (O3 in Section 1.5 of Chapter 1).

Furthermore, the investigations were also aimed at achieving another one of the major objectives of this thesis; i.e., *customising* the *granularity* of ontological concepts in expertise profiles, and facilitating *comparison* and *evaluation* of profiles which describe expertise at different levels of abstraction (O4 in Section 1.5 of Chapter 1). To this end, *semantic similarity* measures were proposed for creating fine-grained baseline profiles at a level of abstraction comparable with the STEP profiles. In addition, the alignment and coverage between STEP and baseline profiles was investigated.

In a number of studies, expertise is captured using ontologies and then inferred from axioms and rules defined over instances of these ontologies. In particular, the Saffron system [6], provides insights into a research community by analysing their main topics of investigation and the individuals associated with them. Saffron is based on research that performs expert finding and profiling by extracting terms from text, at a level of specificity, which describes areas of expertise accurately. A graph-based algorithm is employed to construct topical hierarchies using only domain corpora. The knowledge of an expert is estimated using topical hierarchies, based on how well they cover subordinate expertise topics [136].

Existing social networks such as BiomedExperts (BME) [37] provide a source for inferring implicit relationships between concepts of the expertise profiles by analysing relationships between researchers; i.e., co-authorship. BME gathers data from PubMed 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.

As opposed to the above-listed approaches, the methods proposed in this chapter create expertise profiles containing concepts from domain ontologies, by analysing *structured micro-contributions* (rather than large corpora of static documents). Furthermore, the proposed methods provide a means to *customise* the *granularity* of concepts that represent expertise topics in the resulting profiles, while capturing the *temporality* of expertise.

As outlined in the previous section, ideally the experimental results presented in this chapter should be verified using unstructured contributions. In this study, every structured contribution identifies the concept that has been the target of the change. Furthermore, all contributions target concepts which belong to the same ontology; i.e., ICD-11. This is in contrast to unstructured contributions in the context of collaboration platforms such as MCB [38] or Genetics [39] Wiki projects, where contributions in natural language form are extracted and annotated in order to map identified expertise topics to ontological concepts. As multiple domain ontologies are used for

annotating contributions, identified expertise topics can often be mapped to concepts from different ontologies.

In order to apply these methods to unstructured contributions, future work should focus on creating *ontological lenses*. An *ontology lens* provides a domain-specific view over the expertise of an individual by considering concepts that emerge from the annotation of the expert's contributions using a given ontology; e.g., all concepts from the SNOMED-CT ontology, that have emerged from annotating an expert's contributions, will constitute a SNOMED-CT lens; the GO lens will contain all concepts that emerge from annotating an expert's contributions using the Gene Ontology (GO). The ontology lens that best describes the expertise of the expert will be subsequently identified – i.e., the one that contains the highest number of concepts. The structure of the corresponding ontology will then be used to apply the semantic similarity methods proposed in this chapter for customising the granularity of the profile.

Furthermore, the application of methods proposed in this chapter to unstructured micro-contributions, will facilitate comparative analysis of the effects of *virtual concepts* (proposed in Chapter 5) and *semantic similarity* and *structure of ontologies*, proposed in this chapter, on the accuracy of expertise profiles created by the STEP methodology.

Future work will also focus on conducting *comprehensive evaluation* by using a gold standard, which represents the expertise of contributors with absolute confidence, rather than baseline profiles created as part of the experiments (e.g., fine-grained expertise profiles created by the experts). Moreover, the intention is to create expertise profiles that provide a *comprehensive view* of contributors' expertise, by applying STEP to the complete set of micro-contributions to collaborative authoring of ICD-11, rather than a snapshot of contributions used in this study (due to ongoing development of the ontology).

The next chapter, Chapter 7, investigates the impact of social factors on expertise profiles, by integrating contextual social factors acquired from social expert platforms with the STEP methodology.

Chapter 7 Integration of STEP with Social Factors

7.1 Introduction

The previous chapter (Chapter 6) demonstrated the application of STEP to *structured* micro-contributions, in the context of collaborative authoring of the ICD-11 ontology. In addition, it investigated and demonstrated the benefits of semantic similarity measures and ontological relationships for creating profiles at various levels of abstraction. The ability to generate expertise profiles at different levels of abstraction is essential to enable the evaluation and comparison of profiles describing expertise at different levels of granularity. Chapter 5, on the other hand, demonstrated the application of STEP to *unstructured* micro-contributions for creating semantic and time-aware expertise profiles. In both previous chapters (5 and 6), experts' micro-contributions were the only source of knowledge used for inferring the expertise of contributors.

This chapter investigates the effects of *social factors* on expertise profiling. Towards this goal, this study presents the *Profile Refinement Model*, which integrates *contextual factors* embedded in social networks, with the STEP methodology (O5 in Section 1.5 of Chapter 1). Therefore, in addition to experts' micro-contributions (i.e., content-based factor), this study takes into account the *context* within which every micro-contribution is made as well as the *intrinsic* and *extrinsic relationships* that exist among experts who contribute to these contexts.

Existing scientific and professional networks, such as BiomedExperts [37], provide a source for inferring implicit relationships between concepts in experts' profiles by analysing co-authorship relationships among experts. However, co-authorship reflects collaboration on static publications and resources. Moreover, co-authorship provides little to no information about the actual authored contributions. (Apart from certain assumptions about contributions based on the order of authorship e.g., the first author is the main contributor) [184]. Furthermore, other types of relationships among experts in a social network are often not taken into account; e.g., *following (explicit)* or *forum participations (implicit).*

The *Profile Refinement Model* proposed in this chapter, hypothesizes that relationships formed through *participation* in discussions and Q&A forums within existing social networks, can potentially provide valuable information for expertise profiling, based on the assumption that experts who contribute to the same topics have similar or related expertise. In order to evaluate the proposed model, this study uses the social mechanisms provided by the *ResearchGate* network [27]; in particular, it uses social factors embedded in the ResearchGate Q&A forums. Here, the *context* is represented by a question, and its associated answers, while

social factors can be captured implicitly via the number of votes on questions and answers, or explicitly via "*Following*"/"*Co-author*" relationships between experts.

Section 7.2 describes the ResearchGate use case whose data is used to evaluate the proposed model. Section 7.3 describes how the STEP process is augmented with social parameters to enhance the expertise profiles. Sections 7.4 and 7.5 describe the experimental set-up and experimental results. Section 7.6 provides a discussion of the pros and cons of this approach and Section 7.7 concludes the chapter with a summary of outcomes and a list of areas requiring future work. The work presented in this chapter is in submission [182].

7.2 Use case

ResearchGate is a social networking site with more than 3 million members (scientists and researchers) who share papers and exchange domain-specific knowledge. The site offers tools and applications for researchers to interact and collaborate. Topics, ResearchGate's Q&A forum, enables members to ask questions, get answers and share interesting content with one another. ResearchGate has reported that approximately 12,342 questions were answered about their 4,000 topics in 2011 alone [46]. ResearchGate provides experts with a social networking platform that goes beyond the standard professional profile creation and linking, by enabling members to increase their visibility via participation in discussions that take place in Q&A forums associated with diverse topics. Figure 7-1 depicts an example of several experts asking and answering questions in the context of such a Q&A forum.

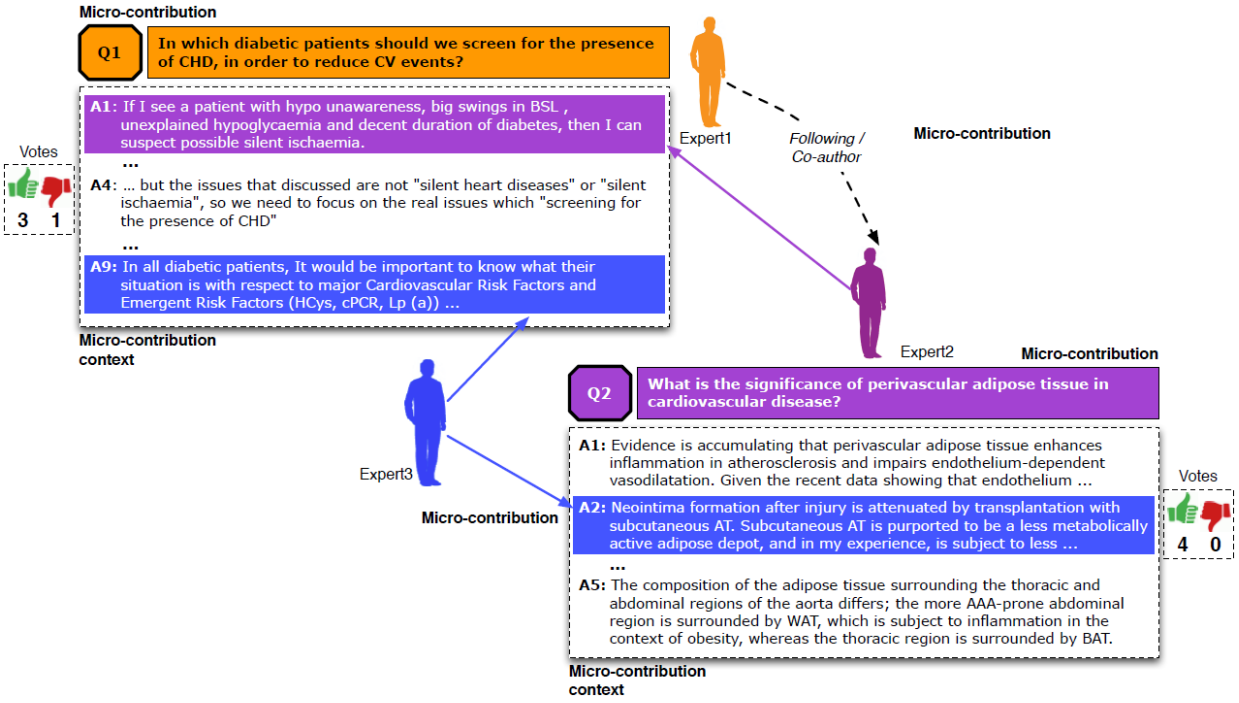

Figure 7-1: Example of Q&A forum in ResearchGate – micro-contributions via questions and answers

The proposed *Profile Refinement* model uses these micro-contributions to create expertise profiles in two settings:

- Individually – i.e., by only taking into account the micro-contributions of individual experts (e.g., *Q2* and *A1* for *Expert2*, or *A2* and *A9* for *Expert3)*.
- Context-driven – i.e., by taking into account the entire context of the micro-contributions (e.g., the answers for a given question, or the question and the entire set of answers for a given answer); for example, in the case of *Expert1*, this setting considers *Q1* and all its answers, while for *Expert3*, it considers *A9* together with *Q1* and all its other answers and *A2* together with *Q2* and all its other answers. This is based on the assumption that questions and answers are intrinsically related, and hence the topics emerging from the context can be used to enrich the expertise profile of the corresponding expert.

Social factors are embedded in the platform at diverse levels. On the one hand, the simple participation in a Q&A exchange can be regarded as a social factor – i.e., it creates an implicit relation between the experts asking and answering questions. On the other hand, such relations can also be expressed in an explicit manner by creating a *Following* link (i.e., when an expert *follows* the activity of another expert), capturing a *Co-author* link (i.e., when several experts are co-authors on a publication), or by voting positively or negatively on the existing micro-contributions (see Figure 7-1). The *Profile Refinement Model* uses this entire set of social factors to refine the weight of expertise topics in the profiles.

Data was gathered by collecting the publicly available micro-contributions of 39 experts in ResearchGate – with a focus on the biomedical domain. This resulted in a set of 3,412 micro-contributions (i.e., questions and answers), with an average of 87.5 micro-contributions per expert. From a contextual perspective, these micro- contributions were associated with 2,077 contexts (i.e., a question and its associated answers) – similar to the example depicted in Figure 7-1. On average, each such context had 8.8 experts contributing to it (including experts who were not one of the selected 39) and 12 answers (in addition to the question). The total number of votes associated with a context (i.e., a question and its answers) ranged between 0 and 119, with an average of 10.92.

7.3 Augmenting STEP with social factors

As outlined in Chapter 4, the STEP methodology consists of three steps: concept extraction, concept consolidation and profile creation. The following discusses the implementation and augmentation of these three steps in the context of ResearchGate. It is worth noting that, while this methodology is applied and evaluated using ResearchGate, the actual steps can be implemented in a

similar manner within other social expert platforms, by simply defining appropriate micro-contribution contexts and identifying the explicit relations that can be created or exist between experts.

7.3.1 Concept extraction

Concept extraction in STEP is delegated to tools or systems that are able to efficiently recognise domain-specific entities in free text. This enables the methodology to be abstracted above the underlying domain characteristics and create expertise profiles in a domain-agnostic manner. On the other hand, relying on external concept recognisers may be seen as a limitation, since the quality of the resulting profiles will be directly affected by the quality of the concept recognition tool.

Experimental results described in previous Chapters were conducted in the biomedical domain, because of the ready availability and maturity of both domain ontologies and high accuracy concept recognition tools e.g., the NCBO Annotator. As demonstrated in Chapter 5, the results achieved using the NCBO Annotator were satisfactory, with the proposed method outperforming existing approaches on Precision by over 20 percent (Section 5.7 in Chapter 5).

Consequently, this study focuses on the same domain and applies the same concept extraction pipeline but within the context of ResearchGate. Hence, the 3,412 micro-contributions were annotated with concepts from domain ontologies, using the NCBO Annotator. This resulted in summarising and representing every micro-contribution via a set of biomedical concepts. Furthermore, as with the experiments presented in Chapter 5, the methodology only considers concepts from the 5 most highly-ranked ontologies, as identified by the Biomedical Ontology Recommender. This filtering step is necessary in order to produce cohesive expertise profiles.

7.3.2 Concept consolidation

The ontological concepts that annotate experts' micro-contributions are consolidated by taking advantage of both the context of the micro-contributions, as well as of the intrinsic semantic relations that exist between them. As mentioned above, the context of a micro-contribution is provided by the question and answers directly associated with the micro-contribution. For example, the context of answer *A1* from *Expert2* in Figure 7-1 is provided by *Q1* and all its other answers. The context of question *Q2* includes *Q2* and all of its answers.

Consequently, given an expertise concept *C* in a particular context, *Context*, its initial weight W is computed as in Eq. 7-1, which denotes the frequency of *C* in *Context*. It is important to note that expertise concepts, such as *C*, are concepts emerging from direct expert micro-contributions – e.g., *hypoglycaemia* in *A1* of *Expert2*. The population of the final expertise profile is derived from the frequency of these concepts.

$$W_{Context}(C) = \frac{Count(C, N_{Context})}{N_{Context}}$$

(Eq. 7-1)

Where $N_{Context}$ represents the number of micro-contributions in *Context* – e.g., 10 for *Q1* in Figure 7-1 (1 question and 9 answers) – and *Count(C, $N_{Context}$)* denotes the count of concept *C* in these micro-contributions.

The above weight assumes exact matching – i.e., the expertise concept under scrutiny is found in the exact same form in diverse micro-contributions in the context. However, the use of ontologies enables one to also consider, and account for, semantically similar concepts, or more concretely, to employ the structure of the underlying ontologies to refine the weight associated with expertise concepts and to enrich the expertise profile with additional concepts that are not explicitly present in the expert micro-contributions. Hence, given an expertise concept *C*, the model takes advantage of the sub-sumption/hierarchical relationships between *C* and other concepts annotated in micro-contributions within a context using the following two scenarios – both of which are depicted in Figure 7-2.

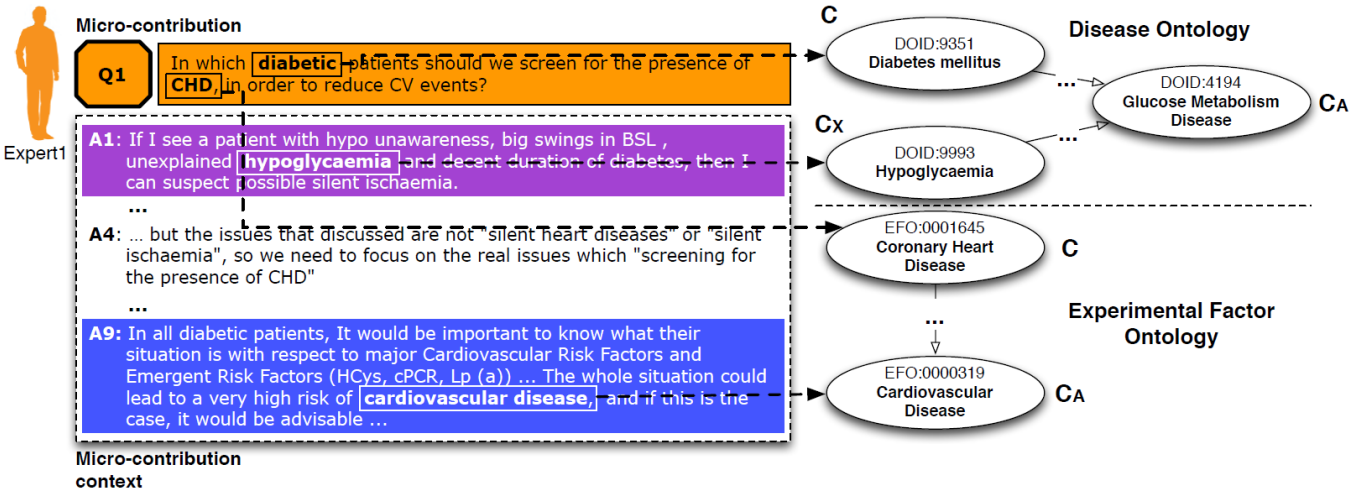

Figure 7-2: Example of concept consolidation using hierarchical relationships in the underlying ontology

1. *C* is a descendant of a concept C_A, case in which C_A is added to the list of expertise concepts with a weight defined by Eq. 7-2:

$$W(C_A) = \frac{W_{Context}\,(C_A)}{distance(C, C_A)}$$

(Eq. 7-2)

Where $W_{Context}(C_A)$ denotes the frequency of C_A in *Context* as per Eq. 7-1 and *distance(C, C_A)* is the hierarchical distance between *C* and C_A in the ontology.

2. *C* shares a common ancestor C_A with another concept C_x, case in which C_A is added to the list of expertise concepts with a weight defined by Eq. 7-3:

$$W(C_A) = \frac{W_{Context}(C_x)}{distance(C_x, C_A)} + \frac{W_{Context}(C)}{distance(C, C_A)} \quad \text{(Eq. 7-3)}$$

Where $W_{Context}$ denotes the frequency of C_x or *C* in *Context* as per Eq. 7-1 and *distance* is the hierarchical distance between C_x and C_A, and *C* and C_A respectively.

7.3.3 Profile Creation

Expertise profiles are created from weighted ontological concepts in two steps: Firstly, short term profiles are built by ranking the ontological concepts based on their aggregated context and social weight, as well as their pairwise mutual information. The role of the short term profiles is to capture the temporal aspect of expertise, or "bursts" of expertise over a restricted period of time. Secondly, the concepts are re-ranked by aggregating their weight across all short term profiles and by introducing additional factors that leverage their uniformity and persistency. The following paragraphs describe each of these steps.

A short term profile represents a collection of concepts identified and extracted from micro-contributions over a specific period of time. Consequently, micro-contribution contexts are grouped into contiguous, non-overlapping intervals of two-weeks and an interval-specific weight is computed for all ontological concepts in the corresponding micro-contributions. This weight takes into account both the expertise concepts (i.e., concepts emerging from an expert's micro-contributions), as well as the concepts resulting from the concept consolidation phase, described in the previous section.

Figure 7-3 presents an example of time interval groupings, where the expert micro-contributions are highlighted. The concepts emerging directly from these micro-contributions, or via concept consolidation (described in the previous section), are weighted and used to create short term profiles. As with the experiments presented in Chapter 5, in this chapter two sets of experiments were performed using (i) two-week and (ii) one-month time-windows for creating short term profiles. However, similar to the results achieved in Chapter 5, the long term profiles created from the short term profiles covering two-week intervals, represented expertise with higher accuracy, because they enable more fine-grained analysis of the periodicity of expertise concepts. Therefore, in all experiments, two-week time intervals are used for grouping micro-contribution contexts.

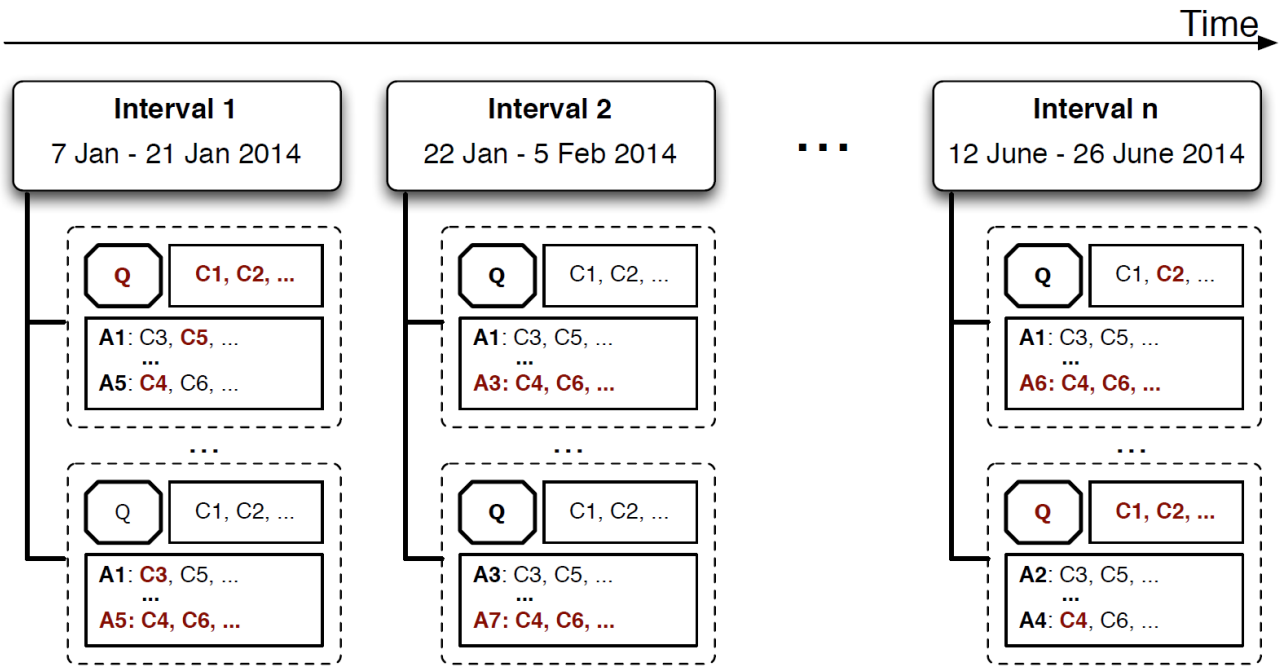

Figure 7-3: Example of time interval groupings for short term profile creation. Micro-contributions of the expert under scrutiny, as well as the direct and semantically similar expertise concepts, are represented in bold

The aim of this study is to investigate whether social factors can be used to build better expertise profiles. So far, the context weight of an expertise concept (Eq. 7-1) only captures the implicit social collaboration – i.e., the contribution made by multiple experts to a single question - answer set. However, as depicted in Figure 7-1, such an environment also provides access to additional explicit social factors that can also be used to refine the weight of the expertise concepts: (i) positive and negative votes and (ii) expert relationships (*Following* or *Co-authorship*). Hence, two additional factors are defined that take these social indicators into consideration:

- Quality factor (Q_F) – this factor aggregates the votes associated with the micro-contributions in a particular context. Q_F , is defined in Eq. 7-4 and denotes a normalised difference in positive and negative votes.

$$Q_F(C, Context) = \frac{Votes_{up} - Votes_{down}}{Votes_{Total}} \quad \text{(Eq. 7-4)}$$

- Social network factor (SN_F) – this factor aggregates the number of explicit social relationships that exist between the experts. SN_F is defined in Eq. 7-5, where $Expert \neq Expert_i$ and Rel_{SF} denotes the relationship strength factor and is: (i) 1/3, if only implicit collaboration exists; (ii) 2/3, if the implicit collaboration is augmented with one of the two types of explicit relations: *Following* or *Co-author*; and (iii) 1, if the implicit collaboration is augmented with both types of explicit relations.

$$SN_F(C, Context) = \frac{1}{N_{Experts}} * \sum_{i=1}^{N_{Experts}-1} Rel_{SF}(Expert, Expert_i)$$

(Eq. 7-5)

The final weight of a concept in a short term profile is computed using Eq. 7-6. This represents an adaptation of the original short term profile creation method (described in Chapter 4) to include these two social factors (Q_F and SN_F). In practice, the first component of the method (initially, denoting the frequency of the concept in the given time period) is replaced with an average over the implicit and explicit social factors.

$$W_{ShortTerm}(C) = \frac{W(C) * Q_F(C) + SN_F(C)}{2} * \sum_{i=1}^{N_C-1} PPMI(C, C_i)$$

(Eq. 7-6)

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

(Eq. 7-7)

Long term expertise profiles aim to provide a comprehensive and ranked view over the entire set of expertise concepts. The computation of the long term profile (Eq. 7-8) follows the original method (described in Chapter 4) by combining an aggregated perspective over the short term profiles with two indicators that reflect the *persistency* and *uniformity* of the expertise concepts. Persistency captures the overall frequency of a concept across all short term profiles, while uniformity models its occurrence patterns.

$$W_{LongTerm}(V_c) = \alpha * (e^{-\Delta(V_c)} - \frac{\Delta(V_c)}{e}) + (1 - \alpha) * \frac{Freq(V_c, S)}{N_s} + \frac{1}{N_s} \sum_{i=1}^{N_s} W_{ShortTerm}^{i}(C)$$ (Eq. 7-8)

where α is a tuning factor, N_S is the total number of short term profiles, *Freq(C, S)* is the number of short term profiles containing concept *C* and $\Delta(C)$ denotes the standard deviation of *C* computed in terms of windows of short term profiles in which the concept is present.

7.4 Experimental Setup

The expertise profiles created by STEP and augmented with social factors were evaluated with the help of ResearchGate experts. The publicly available data collected on 39 experts was used to create corresponding long term expertise profiles. Each expert was then invited to assess his/her own profile by means of a questionnaire, which listed the expertise concepts in a descending order according to their ranking in the profile. In order to reduce the complexity and duration of the task,

evaluation was performed on the top 50 ranked expertise concepts in each profile. The actual assessment was done using a 3-point Likert scale: (i) *Expert*; (ii) *Competent*; and (iii) *Novice*. More concretely, experts were asked to judge their own level of expertise in the processed concepts according to these three categories. It is worth noting that the social factors used to augment STEP, introduced concepts that may not have been explicitly present in micro-contributions, but were inferred from the contexts in which micro-contributions were made (Figure 7-2). The difference in expertise introduced by the Likert scale (from Novice to Expert) aims to accurately capture and reflect the concepts that were inferred from micro-contribution contexts in the evaluation results.

Nine experts out of the initial 39 assisted with the evaluation. The statistics associated with these nine experts are as follows: (i) total number of micro-contributions: 952; (ii) average micro-contributions per expert: 105.7; (iii) total number of micro-contribution contexts: 603; (iv) average micro-contributions per context: 11.6; (v) average contributing experts per context: 8.2.

The experiment computed Precision, as the percentage of concepts at different levels of expertise represented by the Likert scale – e.g., the percentage of concepts associated with the Expert level. Furthermore, it analysed the percentage of these concepts at different ranking cut-offs – e.g., top 10%, 15%, 20%, etc of the evaluated concepts. Finally, in order to investigate the value contributed by the additional contextual and social factors in building expertise profiles, the resulting profiles were compared against a baseline computed by applying the original STEP methodology to experts' micro-contributions – i.e., only using concepts that emerge from micro-contributions, without taking into account the context or the social factors.

7.5 Experimental Results

Figure 7-4 depicts the distribution of expertise judgement results across the nine experts. Overall, the percentage of concepts in the profiles was split almost uniformly across all three expertise categories – i.e., 32.79% Novice (ranging between 13.75% and 55.69%), 34.02% Competent (ranging between 18.13% and 60%) and 33.18% Expert (ranging between 5.06% and 55%). Hence, considering the Expert category as the single correct target class, results in a Precision of 33.18%. Similarly, merging the Competent and Expert categories into a single target class, will increase the precision to 67.20% (34.02% + 33.18%).

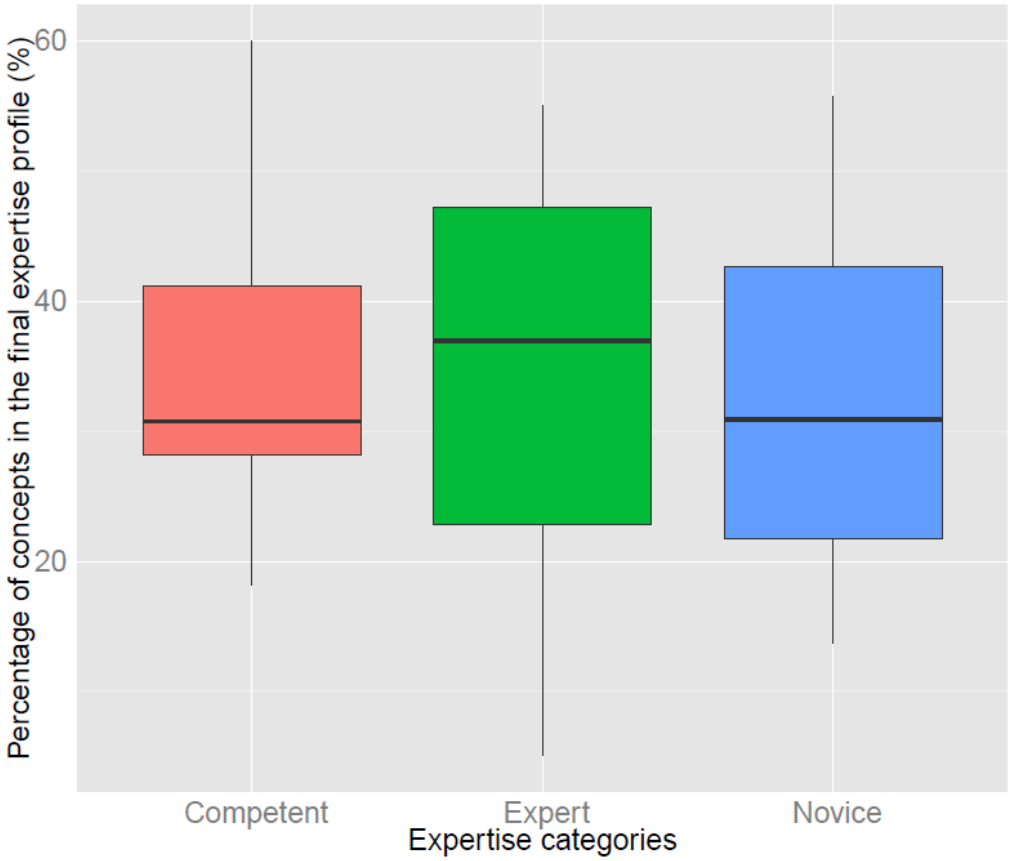

Figure 7-4: Distribution of the evaluated expertise concepts mapped to three expertise categories: Novice, Competent and Expert

In order to determine the relation between the ranking of expertise concepts in profiles and associated expert evaluations (category), the study investigated the percentage of top-ranked concepts above an N% cut-off. Figure 7-5 depicts these percentages for each individual category, with the analysis run at increasing 5% cut-offs. Figure 7-5 clearly demonstrates that Expert concepts are ranked at the top of the expertise profile for any cut-off above 75%. More concretely, if the first 20% of the 50 expertise concepts evaluated per expert (i.e., top 10 concepts) were selected, half of them (50%) would be in the *Expert* category, approx. 33% would be in the *Competent* category and the rest (approx. 17%) would be in the *Novice* category. In conclusion, the method is able to rank, with a high precision, those concepts that reflect the user's true expertise.

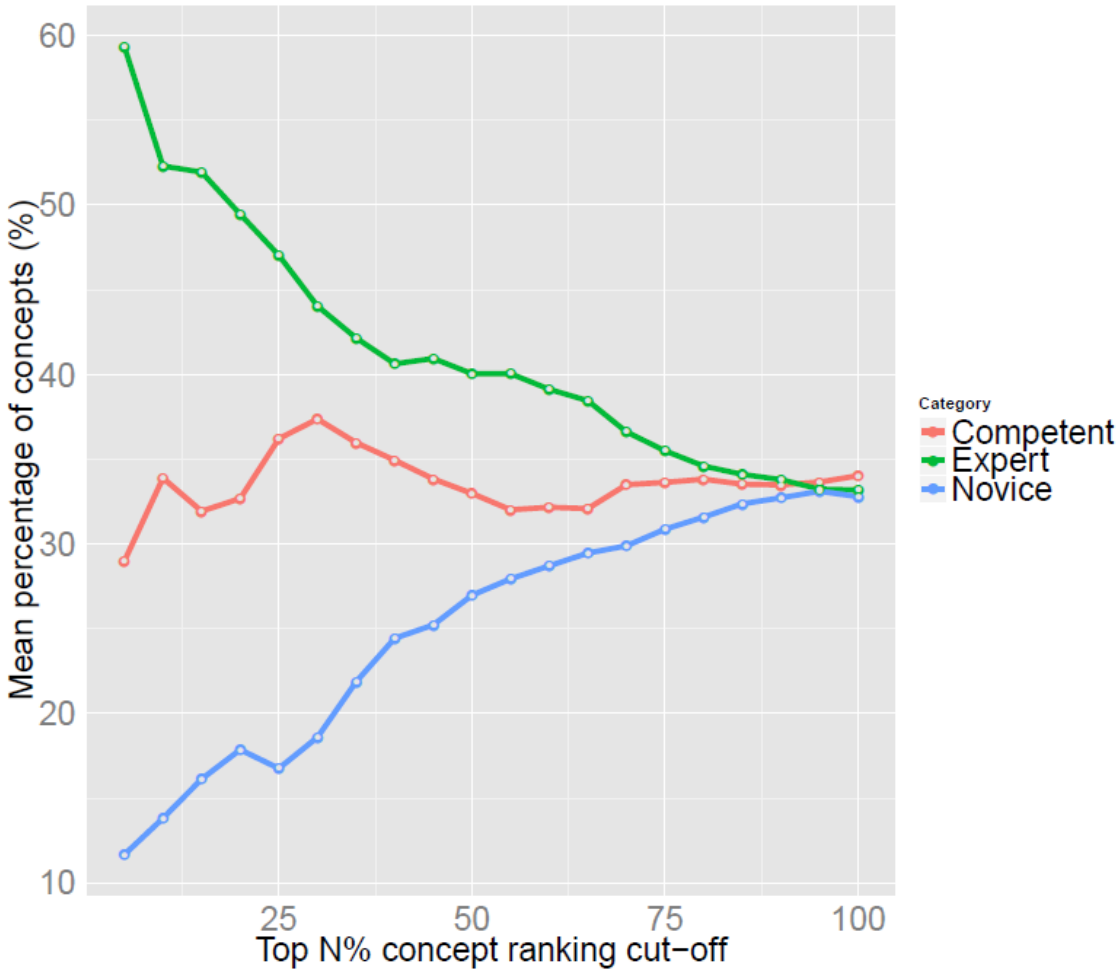

Figure 7-5: Coverage of expertise concepts mapped to the three expertise categories when introducing increasing ranking cut-offs

A second objective of this research was to investigate the potential benefit of including contextual and social factors in the building of expertise profiles. To evaluate the benefits, baseline expertise profiles were first created using the original STEP methodology – i.e., without taking into account micro-contribution contexts or explicit social relationships. Figure 7-6 depicts this comparison according to the three categories of expertise.

On average, around 65% of the *Expert* and 75% of the *Competent* profiles (where *Expert* represents 34% of the total number of concepts, and *Competent* 33%) emerged from the social context, while in the case of the *Novice* category, the percentage increases to around 85%. These results demonstrate the value added by using the social context and relationships when creating expertise profiles – in particular, when the underlying raw data has a fine granularity. The results also demonstrate the expected behaviour in the baseline profiles i.e., baseline concepts (or concepts emerging directly from an expert's micro-contributions) were better represented in the *Expert* profile and formed a decreasingly smaller group in the *Competent* and *Novice* profiles.

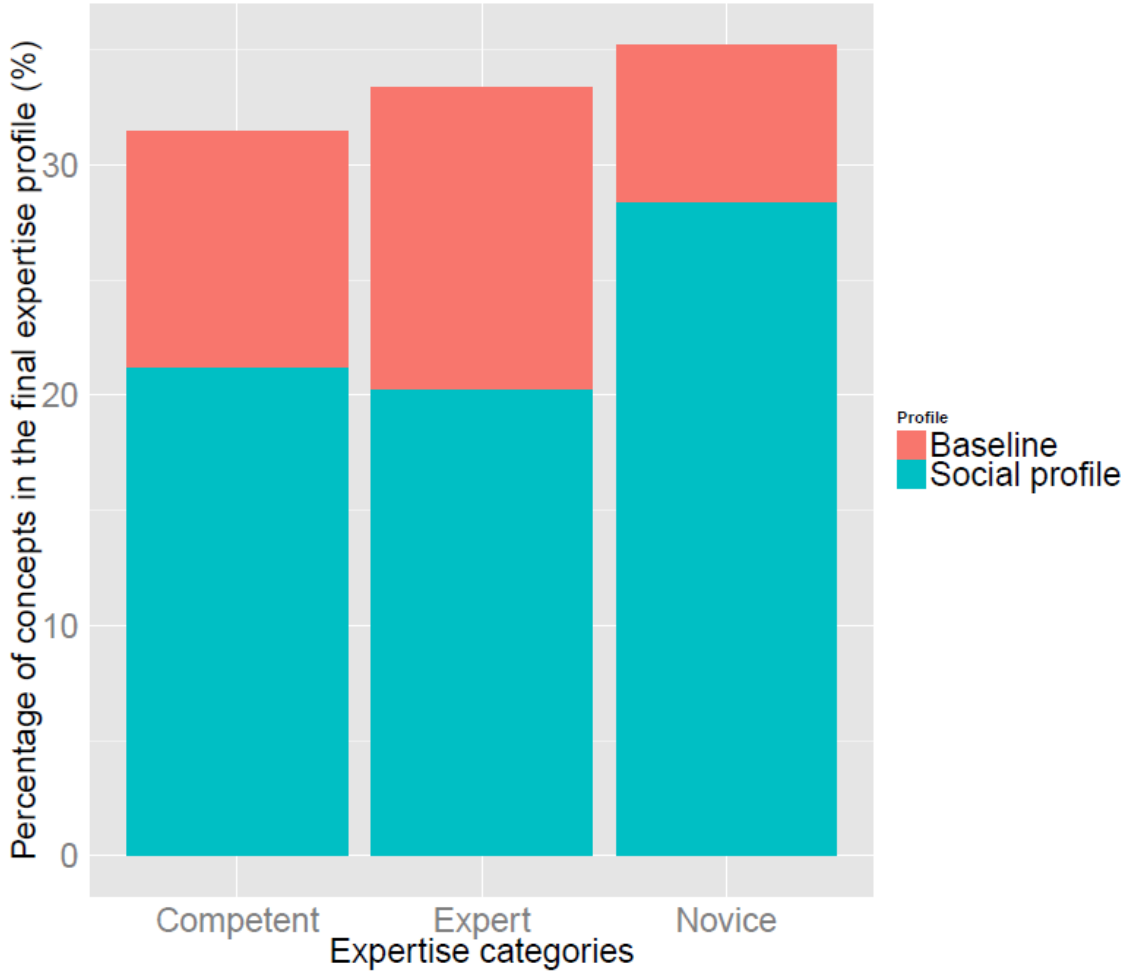

Figure 7-6: Contribution of the social component in building expertise profiles mapped to the three expertise categories. Each category is shown from an overall perspective – i.e., 33% of concepts were associated with Competent, 34% with Expert and the rest with Novice. Furthermore, each category is split into percentage of concepts contributed by using social factors and percentage of concepts emerged from the baseline profile.

7.6 Discussion

The experimental results presented above clearly demonstrate the utility of the expertise profiling method, as well as the added value contributed by incorporating *social factors* with the STEP methodology. Unlike content-based factors, such as the micro-contributions of an expert, social factors are embedded in a platform at diverse levels. The research presented in this chapter regards participation in a Q&A exchange as a social factor – i.e., it considers an implicit relationship between the experts asking and answering questions, based on the assumption that experts who contribute to the same topics, have similar or related expertise and interests. This study also considers explicit relations among experts, e.g., "*following"* or "*co-authorship*", in addition to positive or negative votes on the existing micro-contributions (Figure 7-1).

The study presented in this chapter identified the following limitations of the proposed model:

(i) The model does not consider all social factors embedded in the network. In order to perform an exhaustive study of the impact of contextual factors, all such factors should be identified

and used in refining expertise profiles. For example, this study could have also considered the implicit relationship between experts resulting from reciprocal citations.

(ii) The model uses micro-contribution contexts, as a contextual factor. The context of an expert's micro-contribution, i.e., a question raised or an answer provided to a question, comprises the question and all of its associated answers. This context is used to identify collaborators, the type and strength of their relationships, and the semantic relationships between domain concepts in their profiles and micro-contributions. This entire set of contextual factors, is used to refine the profiles of collaborators, resulting in an increase in the accuracy of expertise profiles. However, the combination of such factors, also led to the inclusion of unexpected expertise topics in profiles, i.e., experts evaluated themselves as novices in some of the expertise topics included in their profiles. These topics resulted from "noise" introduced by including the expertise of all collaborators in the profile refinement process.

(iii) The micro-contributions, i.e., questions and answers provided by the designated experts, were all discussed in the context of Q&A forums in the biomedical domain. In order to ensure that the proposed model is domain-agnostic, its applicability should be verified in the context of social expert networks and micro-contributions in various domains.

7.7 Conclusion and Future Work

This chapter demonstrated the integration of *contextual factors* embedded in social networks, with the STEP methodology and the way in which *micro-contribution contexts* and *intrinsic* and *extrinsic social factors* can be leveraged to enhance profile accuracy. The aim is to achieve one of the main objectives O5, described in Section 1.5 of Chapter 1; i.e., development of a *Profile Refinement Model*, by integrating contextual factors embedded in social expert networks, with the STEP methodology, in order to improve the accuracy of expertise profiles. Manual evaluation results computed with the help of nine ResearchGate experts show an encouraging 33.18% precision when considering the highest category of expertise judgement – i.e., the *Expert* level. Moreover, around 65% of the concepts listed in Expert-level profiles emerge from the *social factors*. These results clearly highlight the significance of incorporating social factors when building expertise profiles.

A number of recent studies have emerged which extend content-based, expert finding approaches with contextual factors. In particular, a study by Hoffman et al [14] identifies and combines contextual factors with content-based retrieval models. Experiments demonstrate that models combining content-based and contextual factors can significantly outperform purely content-based models. However, this study uses a large corpus of static documents associated with experts (e.g., publications), as the information sources. In addition, *SmallBlue* [124], a social-

context-aware expertise search system, mines an organisation's electronic communication to provide expert profiling and expertise retrieval. It uses both the textual content of messages and social network information (patterns of communication). Similarly, *K-net*, a social matching system, uses social networks to provide recommendations. It aims to improve sharing of tacit knowledge by increasing awareness of others' knowledge [183]. The system uses information on the social network combined with existing skills, and the required skills, both of which are provided explicitly by the users. All these approaches are similar to the work described in this chapter, in terms of the use and application of diverse social factors. The major difference is given by the underlying nature of expert contributions (i.e., micro-contributions in Q&A forums in this case), in addition to the associated processing mechanism.

Future work will focus on performing an exhaustive discovery of various social factors and their impact on expertise profiling. Furthermore, the identified social / contextual factors will be studied in the context of various social expert networks, e.g., Google Scholar, Biomed Experts or Academia.edu. This in turn provides the means to determine if the structure of the underlying networks influences the impact of social factors on profile refinement. Finally, the integration of the proposed Profile Refinement Model into ResearchGate and other social networks will be studied. For example, in addition to collaborators' suggesting / endorsing other experts in various expertise topics, the proposed Profile Refinement Model, could suggest a series of topics, based on contextual / social factors. Experts' response to these suggestions, i.e., whether an expert accepts or rejects expertise topics suggested by the profile refinement process, could be used as feedback on the performance of the proposed model, based on which the model could be improved. Finally, profiles refined by the proposed model will be integrated with the Profile Explorer visualisation tool, proposed by this thesis (Chapter 8), in order to facilitate visualisation and comparative analysis with profiles created by only using content-based factors. The results of this analysis could in turn be used to improve the Profile Refinement Model.

The next chapter, Chapter 8, presents a framework to support the visualisation, search and comparative analysis of expertise profiles created by the Semantic and Time-dependent Expertise Profiling methodology.

Chapter 8 Temporal Analysis and Visualisation of Expertise Profiles

8.1 Introduction

One of the main objectives of the research presented in this thesis is to develop a methodology for creating semantic and time-aware expertise profiles – by analysing micro-contributions to collaborative evolving knowledge platforms (Section 1.5 - Chapter 1). To this end, Chapter 4 presented the STEP methodology, while Chapter 5 and Chapter 6 investigated the application of this methodology to unstructured and structured micro-contributions, respectively. Chapter 7 demonstrated the value of integrating social and contextual factors with STEP. Regardless of the source of knowledge used in expertise profiling, (i.e., unstructured or structured micro-contributions) or whether or not contextual/social factors are combined with micro-contributions for profile refinement, STEP captures the *temporality* of expertise by differentiating between *short term* and *long term* expertise profiles.

The temporal aspect of micro-contributions, i.e., the *evolution* of knowledge in collaboration platforms, can be used to analyse and track the changes in individuals' expertise and interests over time. One of the main goals of this research is to facilitate the analysis and tracking of evolving expertise and interests over time (O6, Section 1.5 in Chapter 1). Towards this goal, this chapter presents *Profile Explorer*, the profile visualization paradigm for exploring and analysing *time-dependent* expertise and interests which *evolve* over time. Tracking the evolution of micro-contributions enables one to monitor the activity performed by individuals, which in turn, provides a way to show not only the change in personal interests over time, but also the maturation process of an expert's knowledge (similar to some extent to the maturation process of scientific hypotheses, from simple ideas to scientifically proven facts). *Profile Explorer* [35] facilitates comparative analysis of evolving expertise, independent of the domain or the methodology used when creating the profiles.

Visualizing the temporal aspect of expertise profiles captured by STEP facilitates the following:

(i) Tracking how individuals' expertise evolves over time; e.g., tracking experts' level of activity or bursts of activity in particular topics over time or the amount of time / time-windows that an expert has spent contributing to specific topics.

(ii) Identifying an individual's contributions to the evolution of knowledge, in the context of specific articles / host documents; e.g., identifying how recently an expert made

contributions to a specific topic in a particular article; the most/least active experts in particular topic/s; (e.g., the group of authors who have had the highest level of activity in a particular topic over a period of time in the context of an article or document within a collaboration platform).

(iii) Determine the domain which best describes micro-contributions to a particular evolving article or document in a knowledge-curation platform; i.e., viewing the evolution of an article or document from the point of view of different domains.

(iv) Conduct comparative analysis of expertise profiles; e.g., compare the expertise of authors contributing to multiple articles or comparison of expertise contributed to multiple documents in a collaboration platform.

Profile Explorer leverages virtual concepts created by STEP, in order to provide a consolidated view of the different textual representations of expertise topics that are semantically similar. The role of virtual concepts in Profile Explorer is described in Section 8.2, followed by a description of the technologies employed in implementing the Profile Explorer tool in Section 8.3. Section 8.4 illustrates the utility of Profile Explorer for *browsing*, *searching* and *tracking* expertise and interests over time, using the short term and long term profiles created for a use case from the Molecular and Cellular Biology (MCB) Wiki Project [38]. A usability study performed on the Profile Explorer identified a series of real world use cases, one of which has been implemented and described in Section 8.5. More concretely, a method is proposed for automatically *detecting* periods of peak activity in particular topics over time. Section 8.6 presents the usability study and outlines the benefits and limitations of the work described in this Chapter. Finally, Section 8.7 concludes the chapter by summarizing the outcomes and identifying areas requiring further research.

8.2 The Role of Virtual Concepts in Profile Explorer

Profile Explorer facilitates browsing, search, tracking and comparative analysis of individuals' evolving expertise and interests, by analysing short term and long term profiles created by STEP. As described in Chapter 4, a short term profile represents a collection of concepts extracted from micro-contributions over a specified period of time (time window). The goal of the long term profile, on the other hand, is to capture the collection of concepts occurring both *persistently* and *uniformly* across all short term profiles of an expert. The importance of *virtual concepts* in creating, visualising and analysing short term and long term profiles, was also highlighted. This section describes the significance of virtual concepts in detecting and analysing the trends and changes in experts' activities over time.

As outlined in Chapter 4, a *virtual concept* represents an *abstract* entity with multiple *manifestations*. For example, expertise topics identified in an expert's micro-contributions may be defined and annotated using concepts that include "*mRNA*", "*Messenger RNA*", and "*RNA Messenger*", all of which are manifestations of the abstract entity, "*mRNA*".

Virtual concepts are used by the Profile Explorer platform to identify *semantically similar* concepts that have different *textual groundings*, or lexical representations. For example, a search for the topic, "*mRNA*", in an expert's long term profile, will not only look for short term profiles in which this abstract entity is present, but also for short term profiles that contain any of the different manifestations of this abstract entity. Consequently, the Profile Explorer platform uses *virtual concepts* to perform *semantic* search and comparative analysis of the *time-aware* expertise profiles created by STEP.

8.3 Implementation

The aim of the Profile Explorer is to provide a user friendly and intuitive framework that facilitates the *visualization, search and comparative analysis* of an individual's *evolving expertise and interests over time.* It facilitates visualization of both short term and long term profiles, in addition to, *comparative analysis* by linking an expert's long term profile with his/her short term profiles and underlying contributions. Profile Explorer has been de-coupled from the methodology used in creating the expertise profiles (e.g., STEP), and the domain in which the profiles have been generated (e.g., the biomedical domain). The goal is to provide a visualization paradigm for analysing expertise that is *independent of methodology or domain*.

Profile Explorer has been built using TimelineJS [185]. TimelineJS is an open source tool for building Web-based user friendly and intuitive timelines built in JavaScript. It has built-in support for embedding media from a variety of sources such as Twitter, Google Maps, YouTube, Wikipedia and more. In the case of Profile Explorer, the short term and long term profiles created by the STEP methodology and captured and represented by the Fine-grained Provenance Ontology for Micro-contributions (described in Chapter 3) are uploaded to TimelineJS in JSON format. Profile Explorer is deployed as a Web application to the Apache Tomcat Web Server.

Profile Explorer also utilizes Data-driven Documents (D3) [186] to display the content of expertise profiles (i.e., concepts from domain-specific ontologies) as word clouds. Data-driven Documents (D3) is a JavaScript library for binding data to graphics using HTML, SVG and CSS, animation and interaction.

8.4 Functionality/User Interface

The following section illustrates the Profile Explorer tool using snapshots from the online system [35] and a use case (username: *JonMoulton*) from the MCB [38] Wiki project.

The STEP methodology has been applied to the micro-contributions made by this user to the MCB project, in order to create short term and long term profiles. In the context of the following examples, the goal is to find all time intervals in which this person has contributed to / exhibited expertise in the topic of "*mRNA*". As outlined in Section 8.2, *virtual concepts* play an important role in performing these searches, as periods of activity in the topic "mRNA" are identified by looking for short term profiles where the expert has made contributions to this topic, or *semantically similar* topics, such as "*Messenger RNA*" or "*RNA Messenger*".

Figure 8-1 depicts a portion of the profile timeline for this user. The system has been configured to create short term profiles expanding over two-week intervals. The timeline displays all short term profiles and the long term profile for the user; each short term profile is labelled with a date corresponding to the start of the two-week period that it represents; e.g., 21 May 2012 represents the expertise topics derived from contributions made by this author over the two-week interval starting from 21 May 2012. Please note that only a small section of the timeline is depicted for space reasons.

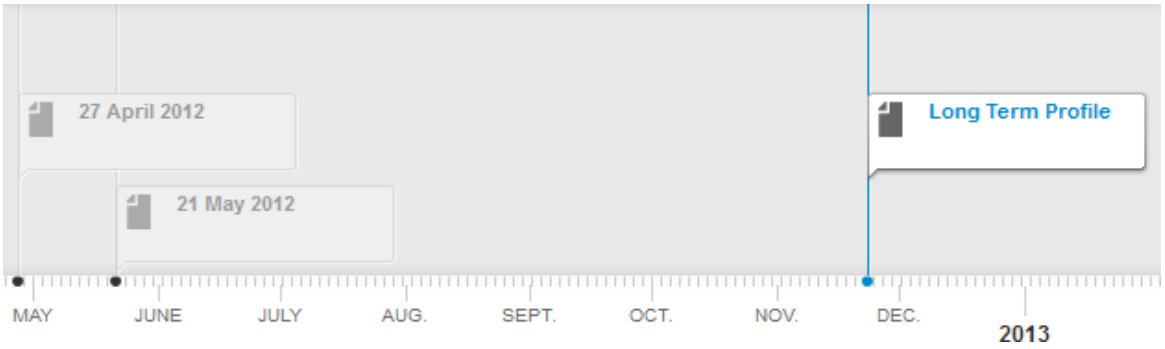

Figure 8-1: A portion of the profile timeline for user JonMoulton

Selecting the "Long Term Profile" label in the timeline will display the long term profile cloud for this expert, as depicted in Figure 8-2. Each label in the cloud represents a domain concept and its size is proportional to the weight of the concept in the expert's profile.

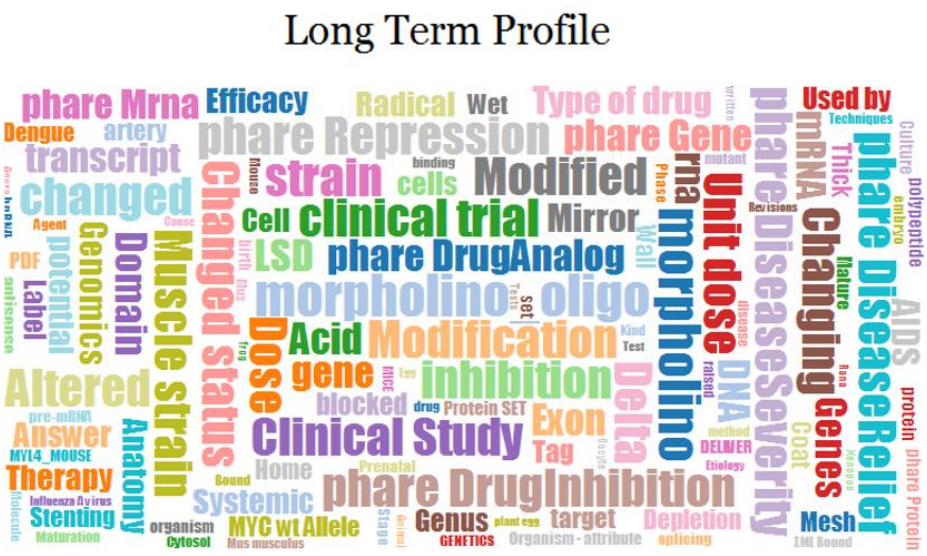

Figure 8-2: Long term profile for user JonMoulton

In addition to visualization, Profile Explorer provides profile search functionality. Search terms can be selected from the word cloud representing the long term profile of an expert. For example, in order to find short term profiles, which represent contributions/expertise on the topic "*mRNA*", this term is selected in the word cloud of the long term profile. Figure 8-3 depicts the screen that is displayed when the search term has been selected in the word cloud. Selecting a search term and invoking the search functionality will display the profile timeline, highlighting all short term profiles which contain concepts that represent the search term; *i.e.*, "*mRNA*" or other concepts which represent terms that are *semantically similar* to the selected term. This is achieved through *virtual concepts*, the building blocks of all profiles, as described in Section 8.2.

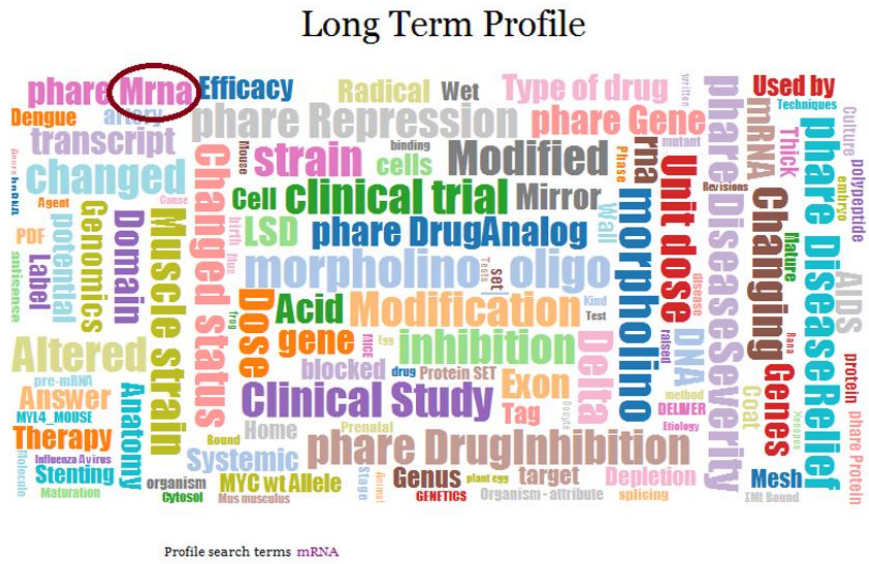

Figure 8-3: Selected search term in the long term profile

Figure 8-4 depicts a portion of the profile timeline for this expert, after the search functionality has been invoked. This provides a birds-eye view over the bursts of activities of the user in the expertise topic, "*mRNA*".

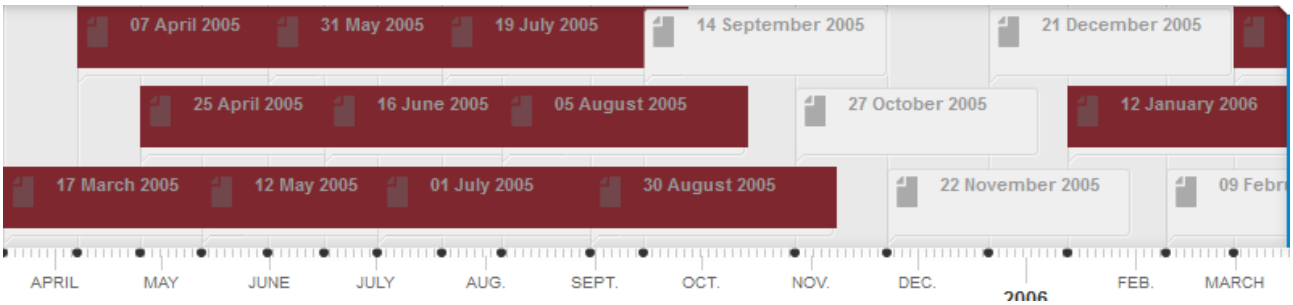

Figure 8-4: Profile timeline — search

Selecting a highlighted short term profile in the timeline displays its corresponding word cloud. Moreover, the label of ontological concepts that match the search term/s are highlighted with a dark border. Figure 8-5 depicts the word cloud for the short term profile, 07 April 2005. As mentioned above, concepts that belong to the *family of concepts* (virtual concept) representing "*mRNA*", have been highlighted. As depicted in Figure 8-5, three concepts have been highlighted; "*mRNA*" (the selected term), "*Messenger RNA*" and "*RNA Messenger*" representing other manifestations of the search term. Therefore, the search function selects profiles which contain concepts representing the *semantics* of search term/s, rather than only those that contain labels that precisely match the search term/s.

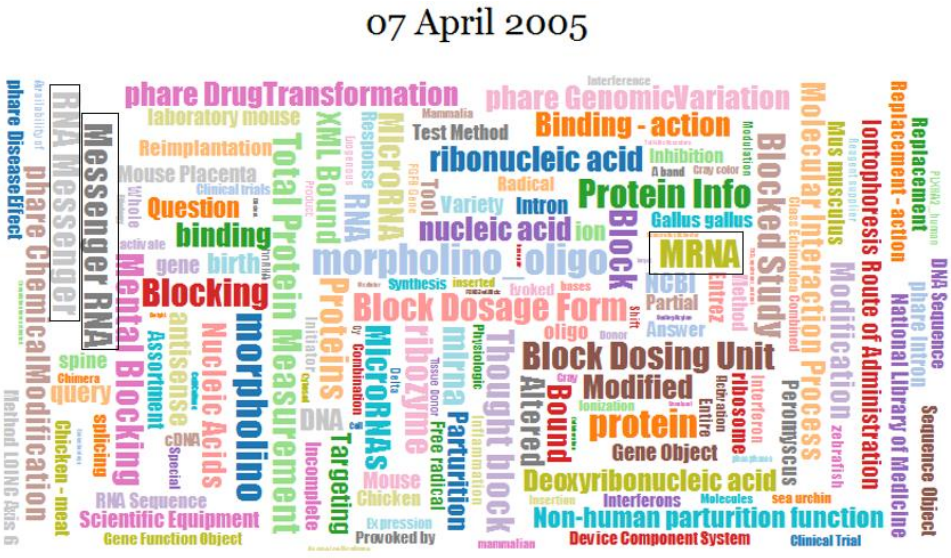

Figure 8-5: Short term profile cloud—search

Profile Explorer also facilitates search in the context of short term profiles. Search terms are selected from the word cloud of a short term profile for the expert. This will display a timeline of all the micro-contributions made by the author within the time period corresponding to the short term

profile. Micro-contributions containing the search term/s are highlighted in the timeline. As with the search performed in the context of the long term profile, this search is also based on the *semantics* of search terms; *i.e.*, micro-contributions are selected and highlighted based on the presence of concepts, which are semantically similar to the concepts selected for search. In this example, the term "*mRNA*" is selected in the word cloud of this short term profile. The timeline of micro-contributions for this period is displayed with each micro-contribution labelled with the topic that it represents. Micro-contributions containing the search term, "*mRNA*" or semantically similar terms, are highlighted (Figure 8-6).

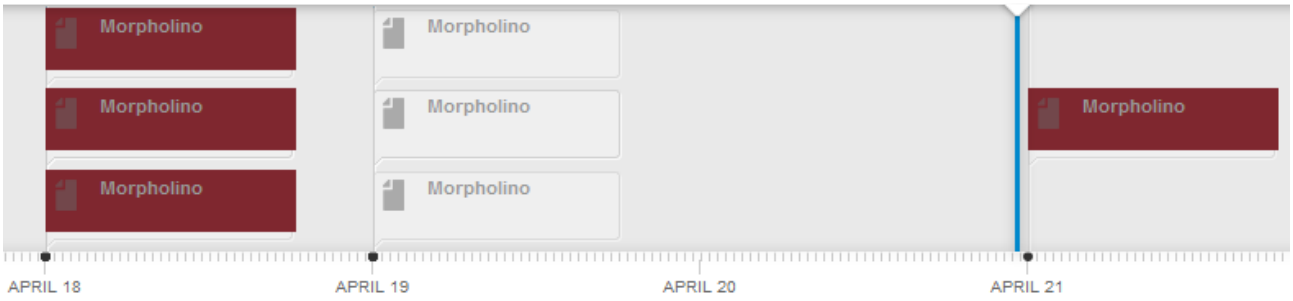

Figure 8-6: Micro-contribution timeline

Selecting/clicking on an individual micro-contribution displays its entire content, with terms matching the search terms highlighted and underlined (Figure 8-7).

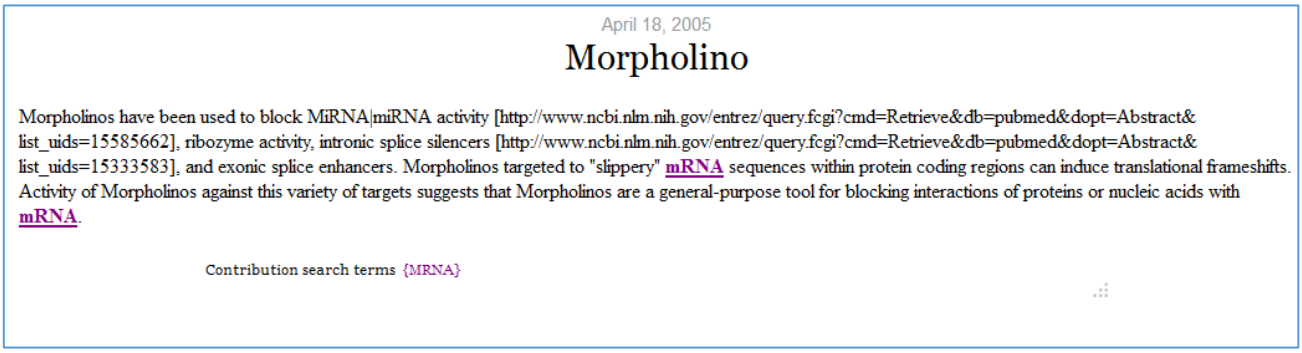

Figure 8-7: Micro-contribution content

8.5 Expertise Peak Detector

The previous section demonstrated Profile Explorer, using short term profiles, each of which represented topics of expertise inferred from micro-contributions made within a pre-configured time-window (e.g., two-week time windows). While this functionality provides significant value in revealing the changes and trends in an expert's activity, it does not automatically "*detect*" the highs and lows in an expert's activity in specific topics of expertise over an arbitrary time period. Based also on the feed-back received in the usability study of the Profile Explorer (see Section 8.6), here, a method is proposed for *detecting* the time-windows where an expert demonstrates "*peak activity*" in

particular topics of expertise. This method is integrated with Profile Explorer in order to facilitate the visualisation of these peak activity time-windows. There are a variety of real world application scenarios where such functionality is useful. For example, it may streamline the process of team-building by demonstrating and visualizing the level of experts' activities in particular topics throughout time. Similarly, it may enable a more efficient detection of experts who are more up-to-date in the given topic – based on the analysis of the more recent peak activity time windows.

Towards this goal, the minimum time window (in days) between an expert's contributions is determined and used as the interval for creating short term profiles. This ensures that changes in the expert's activities are captured within the smallest window of time, which will in turn provide the means for identifying the peaks and troughs of an expert's activity in particular topics, in addition to changes in interests and expertise over time.

For a given topic of expertise, represented by domain concept, C, the short term profiles in which C is among the highest ranked concepts, is identified. The window differences between the identified short term profiles are calculated, in addition to the mean of all window differences (Eq. 8-1).

$$M_{ST}(C) = \frac{1}{N_s} * \sum_{i=1}^{N_s}(ST_i - ST_{i-1}) \qquad \text{(Eq. 8-1)}$$

Where N_s represents the total number of short term profiles, $ST_i - ST_{i-1}$ represents the window difference between short term profiles in which C appears, and $M_{ST}(C)$ is the mean of all window differences.

In order to find timeframes in which the expert demonstrates peak activity in C, the method identifies intervals in which the window difference between consecutive short term profiles (containing C) is less than or equal to the mean of all consecutive window differences in which C appears (i.e., $M_{ST}(C)$). An interval where the window difference between any of the consecutive short term profiles containing C is greater than the mean, designates the end of the peak interval. In other words, an interval in which the expert exhibits activity in the topic at discrete points in time, marks the end of peak activity in the topic.

Figure 8-8 depicts the weight of the concept "*proteins*" in the short term profiles of an expert over a time window of one month. In this example, the mean of all consecutive short term profiles in which C has the highest ranking is $M_{ST}(C) = 2$. As depicted in Figure 8-8, the consecutive window difference between the short term profiles representing 8th June, 10th June and 12th June, in which the expert exhibits peak activity in C (i.e., C is among the highest ranked concepts in these short term profiles), is equal to the mean of all window differences. Therefore, the interval 8th June

– 12th June represents a time-window in which the expert exhibits peak activity in C. However, the window difference between short term profiles representing 12th June and 18th June, in which the expert also exhibits peak activity in C, (i.e., C is among the highest ranked concepts in these short term profiles), is greater than the mean of all window differences. Therefore, 12th June designates the end of a peak activity interval for C, i.e., 8th June – 12th June, while 18th June will mark the beginning of the next period of peak activity in C.

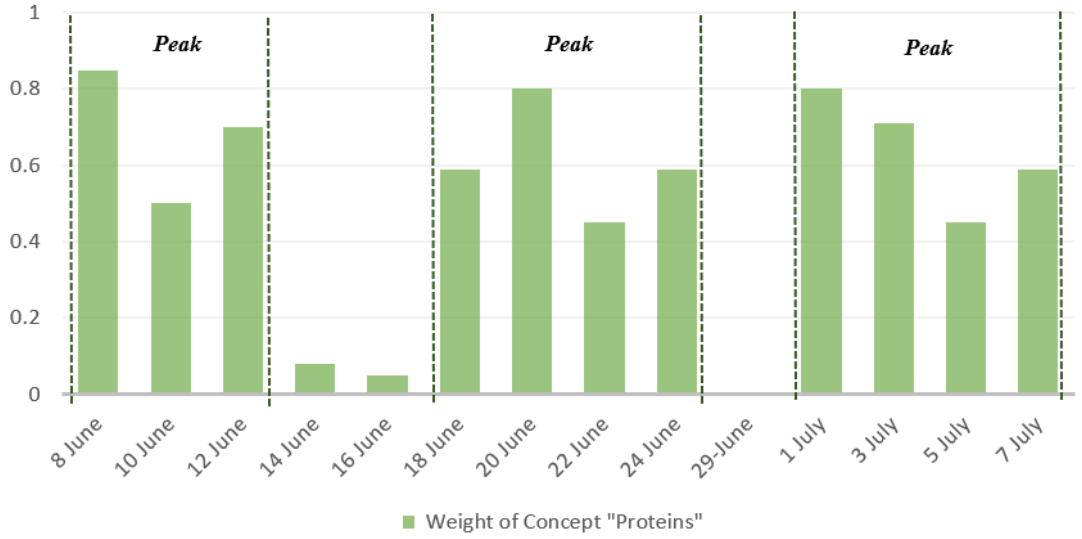

Figure 8-8: The weight of concept "Proteins" in all short term profiles of the expert

The method proposed in this section for detecting time-windows of an expert's peak activity in specific topics of expertise, has been integrated with Profile Explorer. Figure 8-9 depicts the same example using the profile timeline for this expert in Profile Explorer. Every rectangle represents a short term profile. The highlighted sections designate timeframes in which the expert demonstrates high activity in the topic *proteins*. The periods of peak activity in this topic are 8 June – 12 June, 18 June – 24 June and 1 July – 7 July.

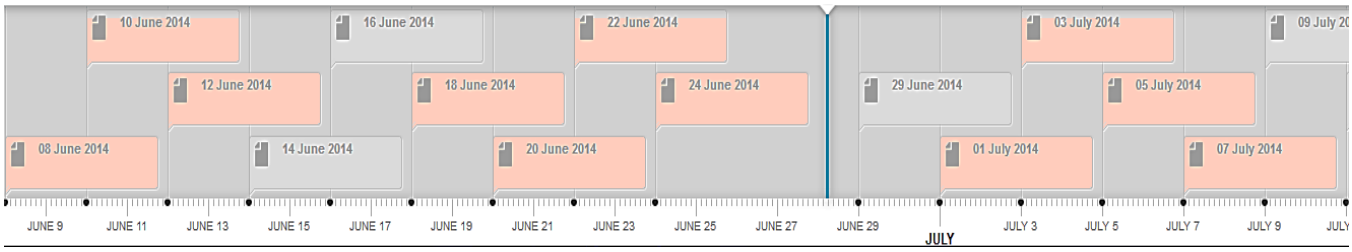

Figure 8-9: Example of peaks and troughs of an expert's activity in the topic "proteins" over time

As illustrated in Figure 8-9, timeframes identifying peak activity in a topic, may represent different lengths of time. Furthermore, peaks and troughs of activity in a particular topic are clearly distinguished. This is in contrast to the pre-configured intervals of equal length, which could

contain periods of inactivity for a highly ranked topic; e.g., if 30 day intervals were chosen for creating short term profiles, a single profile would have represented the time window 08 June – 07 July. While the expert exhibits high activity in the topic "*proteins*" in this interval, periods of inactivity are also present (14th June – 16th June and 25th June – 29th June).

8.6 Discussion/Evaluation

The Profile Explorer has undergone usability testing with the help of a group of 6 users. The resulting feedback indicated that the visualization interface provided a very useful tool for quickly and intuitively analysing, searching and exploring micro-contribution data. Furthermore, it has shown that the Explorer is useful in identifying interesting trends or patterns that require further investigation.

The usability study included nine tasks (Appendix 1), which had to be performed by the users. The tasks ranged from simple browsing to locating concepts in particular short term profiles, or identifying active contribution periods. Once each of the tasks were finalized, users were asked to score, using a 5-point Likert scale, the task difficulty (from 1=*Very difficult* to 5=*Very easy*) and their confidence in performing the task successfully (from *Not at all confident* to *Very confident*). The nine tasks were designed to evaluate three major aspects of the Profile Explorer: *browsing*, *search* and *analysis*.

Figure 8-10 depicts the results of the usability study. For each of the evaluated aspects, i.e., *browsing*, *search* and *analysis*, Figure 8-10 demonstrates the percentage of scores for each aspect on the Likert scale. For example, it shows that *browsing* is rated 3 (average difficulty) by 25% of users, 4 (easy) by 25% of users and 5 (very easy) by 50% of the users (none of the users have rated this aspect as 1 (very difficult) or 2 (difficult)). Similarly, 42% have rated *search* as 4 (easy) and 58% of the users have rated *search* as 5 (very easy), while *analysis* is rated 2 (difficult) by 8%, 3 (average difficulty) by 8%, 4 (easy) by 42% and 5 (very easy) by 42% of the users respectively.

It can be observed that more than 50% of the users have found all three categories very easy, with a very high confidence. As depicted in Figure 8-10, the browsing aspect of Profile Explorer was rated 3, by 25% of the users. This was partly due to: (i) the position of the long term profile, as the users had to scroll right to the end of the timeline in order to access the long term profile, and (ii) the small size and font of dates representing the start of time-intervals, made it difficult for some users to easily locate a particular short term profile.

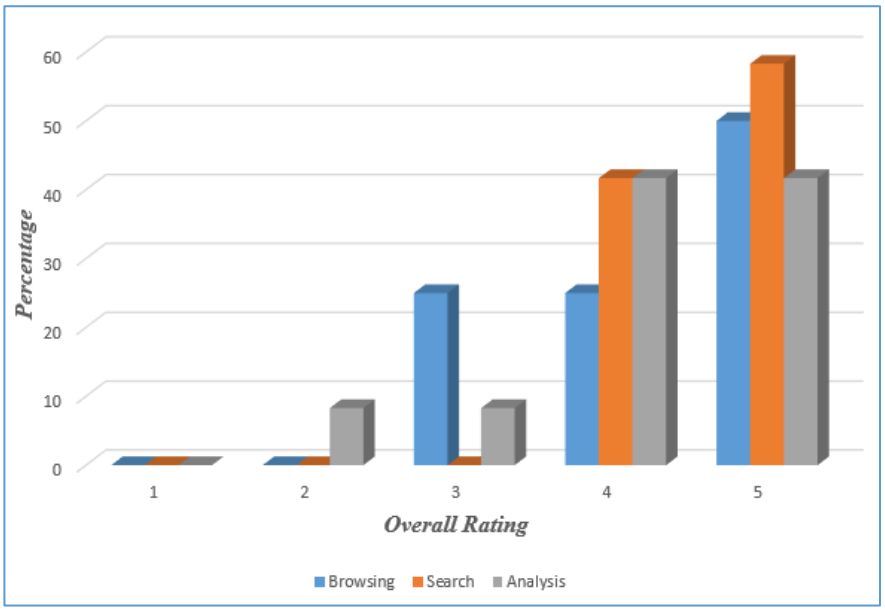

Figure 8-10: Results of the usability testing of Profile Explorer

The strengths as identified by users included: (i) visualisation of link between the long term profile of an expert, with his/her short term profiles and micro-contributions; (ii) inclusion of semantically similar terms in the search functionality; (iii) identifying the actual expertise topics in micro-contributions, representing domain concepts in the word cloud of long term and short term profiles and (iv) user friendliness.

However feedback from users also indicated that the Profile Explorer would be even more useful with the incorporation of a number of changes/improvements, as outlined below:

- Ability to display concept/term-frequency graphs as an alternative to word clouds and to support searches by selecting concepts from such concept/frequency graphs;
- Ability to overlay or display expertise profiles (both long term and short term) for multiple contributors simultaneously to enable comparisons between experts. This would help to determine time periods in which two experts were focused on a common set of expertise topics;
- Ability to display all micro-contributions (from multiple contributors) on a given concept or for a given time period;
- Ability to identify experts on a particular topic in a particular time period via visualizations;
- Ability to generate other types of graphics (e.g., pie charts) that illustrate for example, the percentage breakdown of contributors to a particular topic or the percentage breakdown of topics per contributor;
- Ability to attach annotations to profiles displayed via the Profile Explorer; e.g., textual annotations/notes, semantic tags, or links to related resources such as publications.
- Ability to highlight a search term in the word cloud of profiles, after the search is completed.

One of the major outcomes of this usability study was the development of the *Expertise Peak Detector*, which automatically identifies irregular time windows in which experts demonstrate peak activity in a particular concept. This represents the foundation of some of the temporal analysis improvements requested by the users. As discussed in Section 8.5, the current method focuses on a single concept. Future work will, however, include the visualisation of Expertise Peak Detector for a given number of concepts, and for multiple experts, in order to facilitate comparative analysis.

8.7 Conclusions and Future Work

This chapter presented *Profile Explorer*, a framework to enable *visualization*, *search* and *comparative analysis* of expertise profiles, independent of the methodology or domain. Profile Explorer uses the *temporality* of expertise captured by the STEP methodology, to track and analyse the evolution of individuals' expertise and interests over time (one of the main objectives of this research, O6 in Section 1.5 of Chapter 1).

The most important features that clearly distinguish Profile Explorer from other expertise visualization tools and networks are its ability to: (i) capture and visualize time-dependent aspects of expertise; (ii) conduct comparative analyses based on *semantics* represented by ontological concepts; and (iii) provide evidence of expertise by linking profiles to the actual underlying micro-contributions (Figure 8-7).

A number of recent studies have focused on temporal expertise profiling. One such study by Rybak et. al. [36], proposes the concept of a hierarchical expertise profile, where topical areas are organised in a taxonomy and snapshots of hierarchical profiles are then taken at regular time intervals. Tools such as SciVal Experts [187], BiomedExperts [37] and Expertise Browser [188], provide a visual interface to experts' profiles; however, they only provide an overall view of an individual's expertise and are therefore unable to track and analyse the *evolution* of expertise and interests over time. Profile Explorer analyses experts' *micro-contributions* and provides the flexibility of identifying expertise within specific time intervals, in addition to *detecting* time-windows in which an expert exhibits peak activity in specific topics of expertise. These time intervals *are not specified*, rather, they are *detected* by capturing and analysing the *patterns* and *changes* in an expert's activities over time. Furthermore, Profile Explorer facilitates tracking and analysis of evolving expertise and interests over time, by visualising time-aware expertise profiles.

ExperTime [189] is a web-based system for tracking expertise over time, which visualizes a person's expertise profile on a timeline, where changes in the focus or topics of expertise are detected and characterized. It also provides the flexibility to examine the underlying data (i.e., publications) as supporting evidence. However, Profile Explorer provides evidence for expertise

associated with individuals, by linking an expert's long term profile with his/her short term profiles and micro-contributions. In other words, it visualises the link between the comprehensive, overall view of an individual's expertise, with expertise exhibited in specific time-windows. In addition, Profile Explorer visualises the *individual attribution* captured by the STEP methodology, by linking an expert's profiles to his/her *micro-contributions*, as opposed to authored or co-authored publications.

In addition, VIVO [190], is an open-source, Semantic Web application used to manage an ontology and populated with linked open data representing scholarly activity. VIVO provides its users with faceted semantic search for expert and opportunity finding, rich semantic linking for research discovery, and profiles of people, organizations, grants, publications, courses, and much more. Furthermore, VIVO facilitates four visualisations, which are an integral part of its software (release 1.3 and higher). In particular, VIVO supports *Sparkline*, a line chart used to give a quick overview of a person's number of publications per year or the number of co-authors and publications; *Temporal Graph*, to identify and compare temporal trends of funding intake and publication output activity; *Map of Science*, shows the structure and interrelations of 554 sub-disciplines of science in a spatial format, each of which represents a specific set of journals. The science map is used as an underlying base map, allowing users to overlay the publication-based expertise profiles of people, departments, schools, institutions, and other nodes in the organizational hierarchy; and *Network Visualization*, representing collaboration networks extracted from papers (co-author networks) and grants (co-investigator networks).

While VIVO supports sophisticated visualisations, Profile Explorer specifically targets browsing and analysis of the *evolution of knowledge*, by linking the long term profile of an expert to his/her short term profiles, each of which represents expertise embedded in the expert's micro-contributions within a time interval. In addition, Profile Explorer facilitates the *detection* of time-windows in which an expert demonstrates higher levels of activity in a topic of expertise. Furthermore, the short term profiles of an expert are linked to the underlying micro-contributions (Figures 8-6 and 8-7) and therefore every topic of expertise included in an expert's profile, can be linked to the micro-contributions, in which evidence of the expertise topic was identified. Moreover, comparative analysis of expertise profiles is performed using the *semantics* of expertise topics (through the use of ontologies and virtual concepts), rather than lexical (text-based) comparisons.

Future work will focus on providing additional functionality in Profile Explorer, such as comparative analysis of expertise profiles; e.g., determining time periods when two experts were focused on a common set of expertise topics, clustering micro-contributions based on concepts and

clustering experts based on expertise. Furthermore, future research will aim towards resolving the following limitations of the Profile Explorer framework:

- The framework should be deployed and its applicability evaluated in other domains;
- Profile Explorer relies on virtual concepts to extend the search and comparative analysis of profiles to semantically similar concepts. Therefore, its functionality should be verified in the context of structured micro-contributions, where virtual concepts are not created. As mentioned in Chapter 6, structured micro-contributions target ontological concepts, thus, consolidating different textual groundings of semantically similar expertise topics through virtual concepts, which is performed in the context of unstructured micro-contributions, is not applicable to structured contributions. In this context, semantic similarity measures could be applied to facilitate the identification of semantically similar concepts for the search and comparisons performed by Profile Explorer.
- The framework should also be extended to include the contextual/social factors discussed in Chapter 7, in the visualisation of expertise profiles.
- Including the visualisation of peak activity periods across a number of concepts and multiple experts, in a comparative manner.
- Incorporation of tooltips and help menu, to assist with navigation and functionalities.
- Incorporating statistics; e.g., contribution counts/year, number of contributions in which a term occurs, total number of short term profiles, etc.
- Facilitating search through a search box, in addition to selecting terms from the Word Clouds of profiles.

The next chapter, Chapter 9, concludes the thesis by: summarizing the work presented in Chapters 1-8; assessing the extent to which the objectives outlined in Chapter 1 have been met; and suggesting promising directions for future research.

Chapter 9 Conclusion

9.1 Introduction

The research presented in this thesis was motivated by the emergence of Web 2.0, which has resulted in: an increasing trend in online participation and knowledge sharing; the growing importance of online profiles in generating reputations and raising one's visibility in particular communities; and the increasing use of such online profiles by head hunters, employment agencies and global organizations. Web 2.0 platforms, such as Wikis, blogs, folksonomies and social networks, provide individuals with the opportunity to share their knowledge and expertise through *micro-contributions* to community-generated, evolving knowledge bases. *Micro-contributions* or incremental refinements to collaborative knowledge platforms provide a dynamic environment where knowledge is subject to ongoing *evolution*. This growth in Web-based platforms in which users interact and collaborate with each other through social media, and create user-generated content, presents new opportunities for mining *expertise* from the tacit knowledge embedded in such platforms. This thesis presents an *Expertise Profiling Framework,* which advances the state of the art in expertise profiling by analysing *micro-contributions* to *living* documents (i.e., documents in which knowledge *evolves* over time) in order to capture the *temporality* of expertise.

Traditional approaches to expertise profiling are inadequate when applied to *micro-contributions* in the context of collaborative knowledge platforms, for the following reasons:

- Traditional approaches adopt a *document-centric* approach, which assumes that an individual is an expert in all topics that emerge from the documents which he/she has co-/authored. This *document-centric* approach is unable to match each contributing author to the expertise associated with his/her individual contributions.

- The *macro-perspective* of documents adopted by traditional approaches associate a document with expertise topics embedded in its *entire content;* thus, it cannot provide sufficient evidence for expertise topics associated with the contributors. Rather, a *fine-grained perspective* of the document is required that links authors with the content which they have contributed. The *fine-grained perspective* of documents can then be used as evidence for expertise associated with an expert.

- Analysis relies on *large corpora* of *static* documents, while micro-contributions to collaboration platforms consist of *short* and *sparse* contributions to *dynamic* documents.

- The *temporal* aspect of expertise cannot be captured, as analysis is performed on *static* content, such as publications and reports. The *Expertise Profiling Framework* presented in this thesis captures the *temporality* of expertise by capturing the *evolution* of knowledge in *micro-contributions*, in order to facilitate the tracking and analysis of *changes* in expertise and interests over time.

- Extensive use of *unstructured* data results in very limited inference capabilities. By employing ontologies, ontological relationships can be exploited to identify previously undetected expertise.

Section 9.2 describes the objectives identified in Chapter 1 for overcoming the above-mentioned challenges, and contributions made by this research towards meeting these objectives. Section 9.3 outlines the lessons learned from the application of the proposed Expertise Profiling Framework, to different types of collaborative knowledge platforms and their associated micro-contributions. Section 9.4 identifies remaining open challenges and potential areas for future investigation, before concluding the thesis in Section 9.5.

9.2 Objectives and Contributions

The following outlines and discusses contributions made by the research proposed in this thesis towards meeting the objectives identified in Chapter 1.

O1. Development of a *comprehensive and fine-grained Provenance Model* for capturing *structured* and *unstructured* micro-contributions, by combining coarse and fine-grained provenance, change management and concepts from domain-specific ontologies.

Chapter 3 introduced the *Fine-grained Provenance Model for Micro-contributions* and presented the *Fine-grained Provenance Ontology* for capturing the *fine-grained provenance* of micro-contributions in the context of platforms, where knowledge *evolves* over time. The model combines *coarse* and *fine-grained* provenance modelling to capture and represent *micro-contributions* and their *localisation* in the context of their *host living* documents. It also represents *revisions* resulting from such incremental refinements to the host documents at different levels of *granularity*, e.g., paragraph, sub-section, section, page and document. Three types of information are used by the proposed Expertise Profiling Framework, to create semantic and time-aware expertise profiles:

- Micro-contributions and their fine-grained provenance;
- Change management aspects of the platform such as *actions* (addition, updates, deletions) that lead to the creation of micro-contributions;
- Document *revisions*.

The resulting model makes the following significant contributions to the field of expertise profiling:

1. The model captures and represents the *evolution* of knowledge within micro-contributions by an individual, which in turn facilitates capturing and tracking the *changes* in individuals' expertise and interests over time.

2. Fine-grained provenance modelling facilitates the analysis of micro-contributions using the encapsulating content, thus providing adequate context to enable the *semantic* analysis of the *short* and *sparse* content of contributions.

3. The fine-grained provenance of micro-contributions can be used as *evidence* of expertise in topics represented by domain concepts in individuals' profiles.

4. The model facilitates *fine-grained attribution*, by providing a *contribution-oriented* view of the knowledge base. This contribution-oriented view enables the expertise of an individual to be profiled by analysing the content of his/her individual contributions. As outlined in Chapters 1 and 2, this is in contrast to traditional approaches, which profile expertise by associating individuals with expertise topics that emerge from the entire content of the authored or co-authored documents.

5. Instances of the model are not only useful for expertise profiling, but can also act as a personal repository of micro-contributions, which are captured in a standardized machine-processible, interoperable format that can be published, discovered, reused or integrated within multiple evolving, heterogeneous knowledge bases.

O2. Development of a *Semantic and Time-dependent Expertise Profiling methodology* by linking the textual representation of expertise topics in micro-contributions to weighted concepts from domain ontologies, whilst capturing the *temporality* of expertise.

Chapter 4 presented the *Semantic* and *Time-dependent Expertise Profiling* methodology, i.e., STEP, for creating *semantic* and *time-aware* expertise profiles from micro-contributions made to evolving knowledge platforms. Furthermore, STEP captures the *temporality* of expertise and serves as the foundation upon which the Expertise Profiling Framework proposed in this thesis is built. Moreover, the STEP methodology makes the following significant contributions to the field of expertise profiling:

1. The STEP methodology creates expertise profiles using concepts from domain ontologies, by tapping into the *semantics* conveyed by micro-contributions. As discussed in previous chapters, *semantic analysis* of micro-contributions is essential, because such contributions don't provide sufficient content for applying traditional approaches, which rely on the analysis of large corpora. As described in Chapter 4, the weighting associated with a virtual concept takes into account all of

its different manifestations. This is in contrast to traditional approaches, where a consolidated view of semantically similar terms cannot be created, because different manifestations of semantically similar terms are treated as separate entities.

2. The ontological concepts contained in the STEP expertise profiles facilitate the application of reasoning techniques developed by the Semantic Web community. Furthermore, semantic similarity techniques can be applied to these ontological concepts, in order to customise the granularity of expertise profiles and compare and evaluate profiles describing expertise at different levels of granularity.

3. STEP creates profiles that capture the *temporality* of expertise, by differentiating between *short term* and *long term* profiles. This in turn facilitates tracking and analysing changes in expertise and interests over time. Furthermore, the *long term* profile of an expert captures the collection of concepts that occur both *persistently* and *uniformly* across all *short term* profiles of the expert. Unlike other expertise profiling approaches, *uniformity* is considered as important as *persistency*; i.e., an individual is considered to be an expert in a topic if this topic is present persistently (over a long period of time) and its presence is distributed uniformly across all short term profiles for that expert. This provides the flexibility of computing expertise profiles that focus on uniformly behaving concepts or on concepts that are uniformly present throughout time.

O3. Application of the Semantic and Time-dependent Expertise Profiling methodology to different types of community-driven, dynamic knowledge platforms; i.e., both *unstructured* and *structured* micro-contributions in the context of a range of knowledge domains.

The STEP methodology was applied and evaluated in the context of multiple knowledge platforms and domains, each of which provided a different perspective on the methodology's applicability. This process also facilitated the design of an abstraction layer that ensures the final Expertise Profiling Framework is domain-agnostic. Chapter 5 demonstrated the application of the STEP methodology to *unstructured* micro-contributions in the context of the *Molecular and Cellular Biology* (*MCB*) [38] and the *Genetics* [39] Wiki projects. Similarly, Chapter 6 showcased the application of STEP to *structured* micro-contributions in the context of the collaborative authoring of the *International Classification of Diseases, revision 11, ontology (ICD-11*) [24].

Furthermore, Chapter 5 proposed and demonstrated the integration of two Statistical Language Modelling techniques with the STEP methodology, in order to reduce the effects of domain-specific concept extraction tools on the accuracy of resulting profiles. The pluggable architecture of STEP enabled the integration of the Concept Extraction phase – comprising *Lemmatization* as a pre-

processing step, followed by *topic modelling* and *n-gram modelling*. The research described in Chapter 5 made the following contributions to the field of expertise profiling:

1. Experiments and evaluation results (Section 5.6) confirmed a significant improvement in the accuracy of profiles generated by incorporating Language Models into STEP. More specifically, by setting an appropriate threshold, i.e., concept weight threshold of 1, the n-gram modelling approach delivered a significantly improved accuracy (F-score: 31.94%). These results illustrate that by incorporating *domain-independent* methods (Language Models), the accuracy of profiles can be enhanced and the reliance on domain-specific concept extraction tools can be minimized.

2. Evaluation results (Section 5.7, Table 5-1) confirmed that STEP creates profiles with higher accuracy (i.e., higher F-Score, considering both Precision and Recall) in comparison with two traditional IR methods (Saffron and EARS), both of which rely on the analysis of a large corpora of static documents.

O4. Development of a mechanism for *customising* the *granularity* of ontological concepts in expertise profiles in order: (i) to describe expertise with a level of specificity that accurately represents the knowledge embedded in micro-contributions, and; (ii) to facilitate the *comparison* and *evaluation* of profiles which describe expertise at different levels of abstraction.

The Expertise Profiling Framework proposed in this thesis represents expertise topics embedded in micro-contributions, using concepts from domain ontologies. The use of ontologies provides the flexibility to take into account more than just the specific domain concepts, by also considering ontological parents and children. Chapter 6 proposed an approach for creating expertise profiles at various levels of granularity, using expertise centroids - ontological concepts that act as representatives for an area of the ontology by aggregating highly similar concepts for all micro-contributions in close proximity to the centroid concept. These centroids provide a more accurate perspective of the actual expertise and facilitate comparison of profiles which describe expertise at different levels of abstraction. The research described in Chapter 6 made the following contributions to the field of expertise profiling:

1. Experimental results demonstrated that STEP could usefully be applied to *structured* micro-contributions, to generate high quality expertise profiles. In particular, semantic similarity measures were proposed for: (i) creating baseline profiles from experts' structured micro-contributions; experimental results demonstrated a 64.45% decrease in the number of concepts included in the baseline profiles, compared to the concepts to which experts had contributed, from an average of 33.5 concepts to 11.91 concepts per author; and (ii) evaluating the STEP profiles using the baseline

expertise profiles, demonstrated that even when using the highest concept weight threshold of 0.15 (at which only 8.24% of concepts are included in the STEP profiles), an almost 23% similarity is achieved. In addition, comparison of these results to the results achieved by applying STEP to unstructured micro-contributions (described in Chapter 5), demonstrated that semantic similarity methods and ontological relationships, result in more accurate comparisons than simply identifying exact matches between the content of profiles.

2. New methods for *customising* the *granularity* of ontological concepts in expertise profiles, using *semantic similarity measures* and *ontological relationships* were developed, evaluated and validated.

3. Facilitated *comparison* and *evaluation* of profiles that describe expertise at different levels of abstraction. In particular, fine-grained baseline profiles were created at a level of abstraction comparable with the STEP profiles, in order to study the alignment and coverage between STEP and baseline profiles. STEP profiles exhibited an almost constant behaviour in terms of coverage, independently of the imposed threshold. This confirmed that weights associated with concepts in a STEP profile, represent the true level of an author's expertise in the topics represented by those concepts.

4. Provided experts with the ability to complement their existing online profiles with fine-grained domain concepts that represent the implicit knowledge embedded in their micro-contributions to collaboration platforms.

O5. Development of a *Profile Refinement Model* by integrating contextual factors from social expert networks, with the Semantic and Time-dependent Expertise Profiling methodology, in order to improve the accuracy of expertise profiles.

Chapter 7 demonstrated the integration of *social factors* embedded in social expert networks, with the STEP methodology, in order to enhance profile accuracy. More specifically, it proposed the *Profile Refinement Model,* which uses a set of social factors to refine the expertise profiles created by using only content-based factors (i.e., micro-contributions). This study uses the social mechanisms provided by the *ResearchGate* network [27]; in particular, it uses social factors embedded in the ResearchGate Q&A forums. The study considers the implicit relationship between experts who participate in the same Q&A forums. This is based on the assumption that experts participating in the same Q&A forums have similar or related expertise and interests. Furthermore, it considers two explicit relationships, i.e., "*following*" and "*co-authorship*" between experts, in addition to positive and negative voting on micro-contributions, i.e.,

question and answers in the studied Q&A forums. The research described in Chapter 7 made the following contributions to the field of expertise profiling:

1. A Profile Refinement Model was developed which takes into consideration the contextual and social factors within a social network of contributors, to refine and improve the accuracy of expertise profiles.

2. The added value of incorporating contextual and social factors with the STEP methodology, was demonstrated. Contextual factors represent the *context* within which every micro-contribution is made (e.g., in the context of a Q&A forum, a question's context comprises the question and all the answers provided to the question; similarly, an answer's context comprises the question to which this is an answer and all other answers to the question). Social factors represent the *implicit* (the number of votes on questions and answers) or *explicit* ("*Following*"/"*Co-author*") relationships between experts who contribute to these contexts. Evaluation results (Section 7.5, Chapter 7) showed an encouraging 33.18% precision when considering the highest category of expertise judgement – i.e., the Expert level. Moreover, around 65% of the concepts comprised in Expert-level profiles emerge from social factors.

3. The value of incorporating *social relationships* formed during participation in discussions and Q&A forums for complementing profiles of collaborators; and *semantic relationships* among domain concepts in collaborators' micro-contributions and profiles; to refine the expertise of contributors was validated.

4. The proposed STEP methodology was validated in the context of a new type of collaborative knowledge platform – a scientific online community (ResearchGate). By applying and evaluating STEP in the context of a range of different types of *evolving* knowledge platforms, the domain-independent applicability of the framework is further validated.

O6. Development of a Profile Visualization service to facilitate analysis and tracking of evolving expertise and interests over time

Chapter 8 presented the *Profile Explorer*, a service that enables the *visualization*, *search* and *comparative analysis* of expertise profiles, independent of the methodology or domain. Profile Explorer uses the *temporality* of expertise captured by the STEP methodology, to track and analyse the evolution of individuals' expertise and interests over time. Chapter 8 also presented the *Peak Detector* service that enabled time windows associated with peak activities by individual experts to be automatically detected and then highlighted within the Profile Explorer.

The research described in Chapter 8 made the following significant and original contributions to the field of expertise profiling:

1. The first domain-independent, timeline-based visualization tool that enables both short term and long term expertise profiles for selected experts, to be displayed, browsed, searched and retrieved – to facilitate the quick and easy identification of experts in specific topics at given times.

2. Facilitates *semantic* search and comparative analysis of expertise profiles, using the comprehensive view of expertise topics generated by the STEP methodology, through virtual concepts.

3. Facilitates visualisation of time-windows in which an expert exhibits peak activity in particular topics of expertise, through the Expertise Peak Detector interface.

4. Visualises and clearly demonstrates the *evolution* of expertise *over time*, through linking the long term profile of an expert with his/her short term profiles and micro-contributions.

5. Visualises evidence of expertise by linking the time-aware profiles created by STEP to the underlying micro-contributions.

9.3 Insights

In addition to the original contributions to the field of expertise profiling (described above), the following insights have been gained from generating expertise profiles from micro-contributions.

9.3.1 Fine-grained provenance modelling of micro-contributions

The fine-grained provenance of micro-contributions proved to be one of the most important elements of the Expertise Profiling Framework proposed in this thesis. The *Fine-grained Provenance Model for Micro-contributions* proposed in Chapter 3, provided the means to adopt a micro-contribution-oriented approach to expertise profiling (rather than the document-centric view adopted by traditional IR approaches). The model also enabled the capture of both micro-contributions together within their surrounding context, which in turn enabled the analysis of short and sparse content.

Furthermore, the model captures the changes in experts' contributing activity, i.e., the evolving micro-contributions and revisions made to the host documents as a result of such incremental refinements. This provides the foundations for representing the evolution of knowledge over time, which in turn facilitates analysing and tracking the changes in individuals' expertise and interests over time.

The model provides provisions for representing micro-contributions using concepts from several ontologies, while capturing the exact placement and localisation of micro-contributions. This in turn provides evidence of expertise, as domain concepts representing topics of expertise in profiles, are linked to the underlying micro-contributions.

9.3.2 Representing Expertise Profiles as structured data

The STEP methodology represents the knowledge embedded in experts' micro-contributions using *weighted concepts* from domain *ontologies* (i.e., *structured content*). Representing the implicit knowledge embedded in micro-contributions using terms from machine-processible domain ontologies, provides the flexibility to integrate expertise profiles with the Linked Data Cloud [28] and apply reasoning techniques developed by the Semantic Web community.

From a technical perspective, building expertise profiles from concepts defined in widely adopted ontologies enables individuals to publish and integrate their profiles as *structured data* on the *Web*. This enables online "*expertise seekers*" and "*Web crawlers*" to discover and access published expertise profiles, consolidate profiles and seamlessly aggregate and compare profiles for communities of experts. Furthermore, the links between ontological concepts in expertise profiles and concepts in the Linked Data Cloud can be discovered and used to complement the published profiles, providing access to richer, more accurate and more up-to-date expertise profiles.

9.3.3 Semantic Analysis of Micro-contributions

As discussed in Chapter 4, the STEP methodology taps into the *semantics* conveyed by micro-contributions to create profiles representing expertise using concepts from domain ontologies. *Semantic* analysis of micro-contributions is essential as such contributions don't provide sufficient context for applying methods used by traditional approaches, which rely on the analysis of large corpora of static content.

In addition, *semantic* analysis of micro-contributions and the use of *ontologies* provides the means to identify the different lexical groundings of terms that are *semantically similar*. This results in more accurate and comprehensive expertise profiles because different manifestations of semantically similar expertise topics can be identified and accommodated (via *virtual concepts).* As discussed in Chapter 8, virtual concepts also play an important role in the *Profile Explorer* user interface, by facilitating search and comparative analysis of expertise profiles, using *semantics* rather than simplistic text-based comparisons.

9.3.4 Comparison of expertise profiles at different levels of granularity

The major lesson learned from the application of STEP to unstructured micro-contributions, presented in Chapter 5, was the need to create baseline profiles at a level of abstraction closest to the actual micro-contributions. Because an author of the MCB or Genetics projects typically describes his/her expertise using high-level concepts (such as *Genetics*, *Chemistry*, *Cell* and *Biology*) and the bottom-up profiles created by STEP represent expertise using low-level topics (such as *Metabolic pathways* and *Lipoprotein lipase),* direct comparison is particularly challenging.

This gives rise to another major objective of the Expertise Profiling Framework proposed in this thesis – the development of a mechanism for *customising* the *granularity* of ontological concepts in expertise profiles in order to: (i) describe expertise with a level of specificity that accurately represents the knowledge embedded in micro-contributions, and (ii) facilitate *comparison* and *evaluation* of profiles, which describe expertise at different levels of abstraction.

As discussed in Chapter 6, *semantic similarity* measures and the *structure of ontologies* (*subsumption* and *sameAs* relations) were used to customise the *granularity* of ontological concepts in expertise profiles, and facilitate *comparison* and *evaluation* of profiles which describe expertise at different levels of abstraction.

9.3.5 The impact of contextual factors in expertise profiling

Chapter 7 demonstrated the effects of *social factors* on expertise profiling, by integrating *contextual factors* embedded in *social networks*, within the STEP methodology. The proposed *Profile Refinement Model* integrates knowledge embedded in the relationship structure of collaborating experts in social networks for improving the accuracy of expertise profiles. It combines experts' micro-contributions (i.e., content-based factors), with contextual factors embedded in social expert networks. In the experiments presented in Chapter 7, social factors embedded in the ResearchGate Q&A forums were used to refine expertise profiles created by STEP. The *Context* of micro-contributions is represented by a *question*, and its associated answers, while *social factors* can be captured implicitly via the number of votes on questions and answers and relationships formed through participation in the same Q&A forums or explicitly via "*Following*"/"*Co-author*" relationships between experts. Experimental results demonstrate that on average, around 65% of the *Expert* and 75% of the *Competent* profiles, emerged from the social context, while in the case of the *Novice* category, the percentage increases to around 85%. These results show the value added by using the social context and relationships when creating expertise profiles.

9.4 Open Challenges and Future Research

Although this investigation into: "expertise profiling via the analysis of micro-contributions to evolving, collaborative knowledge platforms" solved a number of critical challenges, it also exposed a number of new problems and issues that require further research. The following subsections outline areas designated for future research and development.

9.4.1 Micro-contribution Quality

The Expertise Profiling Framework presented in this thesis, assumes that all micro-contributions of an expert are of equal quality. However, in practice, the quality of micro-

contributions varies across the micro-contributions from a single expert and varies from expert to expert. This variability in quality should ideally be taken into account when ranking expertise topics that emerge from an expert's micro-contributions.

The Fine-grained Provenance Model for Micro-contributions described in Chapter 3, aims at creating a comprehensive model for capturing and representing the fine-grained provenance of micro-contributions to evolving knowledge platforms. In particular, the SIOC-Actions module [152] is used to capture the actions that lead to the creation of micro-contributions; e.g., add, delete, update. Future work will focus on leveraging this information to determine the quality of micro-contributions and adjust the weight of concepts in expertise profiles, accordingly. For example, an expert could modify a document by making a series of micro-contributions. All or some of these micro-contributions may subsequently be rolled back or deleted by another expert. This would then result in a lower ranking of concepts emerging from the rolled-back or deleted contributions, in the expert's profile.

9.4.2 Concept Recognition

The focus of the Semantic and Time-dependent Expertise Profiling (STEP) methodology (introduced in Chapter 4) is on the *concept consolidation* and *profile creation* phases, involved in creating expertise profiles that capture the *temporality* of expertise. The *concept extraction* phase is performed using tools provided by the biomedical domain, i.e., the domain of relevance to the applications, content and experiments. The biomedical domain was chosen specifically because the associated ontologies and concept recognition tools (e.g., NCBO Annotator) are mature, robust, proven and widely adopted.

While the STEP methodology is domain-agnostic (i.e., none of the phases are restricted to the use of domain-specific tools or techniques), applying and evaluating STEP to other domains in which the concept extraction tools are less mature or reliable, will present challenges. Chapter 5 presented a solution for minimising the effects of domain-specific tools on the resulting expertise profiles, by integrating Language Models with the STEP methodology. Experimental results presented in Chapter 5, demonstrated a significant improvement in the accuracy of profiles generated by the *enhanced* STEP methodologies. More specifically, the best profile accuracy (identified by the F-score measure) was achieved at the concept weight threshold of 1, by the n-gram modelling approach (i.e., F-score = 31.94%), followed by the topic modelling approach (i.e., F-Score = 28.75%), followed by the original approach, i.e., the generic STEP methodology (F-Score = 27.81%). These results demonstrate that the effects of domain-specific concept extraction tools can be minimised by enhancing the STEP methodology with domain-agnostic concept recognition models, such as topic and n-gram models.

However, future work will focus on developing mechanisms to *de-couple* the concept extraction phase of the STEP methodology and the resulting profiles, from domain-specific tools and techniques; thus, providing a domain-agnostic solution for profiling expertise using micro-contributions to collaboration platforms, independent of concept extraction tool support in a domain. Towards this future initiative, the STEP methodology will be applied to collaborative knowledge platforms in other domains such as astronomy (e.g., Astronomy Wiki [191]), earth sciences (Earth Sciences Portal [192]) or chemistry (Chemistry Portal [193]), in order to verify that the mechanisms developed for *decoupling* STEP from domain-specific concept extraction tools, are effective and provide an Expertise Profiling Framework that is applicable to all domains.

9.4.3 Ontology Lenses

Chapter 6 described methods for customising the granularity of ontological concepts representing expertise topics in profiles. These methods were applied to *structured* contributions in the context of collaborative authoring of the ICD-11 ontology.

In order to verify the applicability of these methods to *unstructured* contributions, future research will focus on leveraging *ontological lenses*. An *ontology lens* provides a domain-specific view over the expertise of an individual by considering concepts that emerge from the annotation of the expert's contributions using a given ontology; e.g., all concepts from the SNOMED-CT ontology, that have emerged from annotating an expert's contributions, will constitute a SNOMED-CT lens; alternatively all concepts that emerge from annotating an expert's contributions using the Gene Ontology (GO) will constitute the GO lens. The ontology lens that best describes the expertise of the expert is then identified – i.e., the one that contains the highest number of concepts. The structure of the corresponding ontology will then be used to apply the semantic similarity methods proposed in chapter 6 for customising the granularity of expertise concepts in the profile.

9.4.4 An Alternative Measurement of Scientific Productivity

Assessment of the quality of scholarship products is a critical component of the research process. As the volume of academic literature explodes, scholars rely on filters to cherry-pick the most relevant and significant sources from large online corpuses. The evaluation of research has traditionally focused on scholarly journal articles and — particularly in the humanities and social sciences — books or book chapters. While the focus on these traditional outputs is critical in the assessment of scholarship, the significance of other emerging research outputs is increasingly recognized. Traditional metrics, such as peer-review, citation counts and impact factors, are primarily based on print processes and are increasingly failing to keep pace with changes in the form and usage of research outputs [194].

In growing numbers, scholars are transferring their everyday work practices to the Web. New forms of scholarly outputs, such as research data sets, scientific software, posters and presentations, blogs, Wikis, lectures, classes and other activities shared online, are not assessed by the traditional metrics. Nano-publications [29] (in which assertions, data, or discovery elements, are shared with minimal additional context) also represent an alternative form of scholarly output. These new forms of scholarly output also reflect and transmit scholarly impact. *Alternative metrics*, or *Altmetrics*, represent alternative measurements of scientific productivity [194]. *Altmetrics* provide an extended view of what impact looks like, but also of what's making the impact. This matters because expressions of scholarship are becoming more diverse [194].

The National Information Standards Organization [195], NISO, has recently undertaken the *Altmetrics initiative*, an important step in the development and adoption of new assessment metrics, which include usage-based metrics, social media references, and network behavioural analysis. One of the main areas of focus is to define relationships between different research outputs and to develop metrics for this aggregated model [195].

The research presented in this thesis, proposes an Expertise Profiling Framework, which creates semantic and time-aware expertise profiles for individuals who contribute to the evolution of knowledge in collaborative platforms. *Micro-contributions* to collaborative knowledge platforms also represent an alternative form of scholarly output. An area of potentially valuable future research is to develop and validate alternative assessment and evaluation metrics that assess the value/quality and impact of researcher's micro-contributions to collaborative knowledge platforms. Such an assessment tool could contribute towards assessing the expert's overall research and scholarly output.

9.4.5 A Foundation for Novel Trust and Reputation Metrics

Research into trust and reputation models has attracted significant interest in fields such as sociology, economics, psychology, and computer science [197]. Within the context of collaborative knowledge platforms, computational models of reputation mainly consider two sources of information: (i) direct interactions between individuals; and (ii) ratings/votes provided by other members of the platform. Other studies complement the reputation model by incorporating information obtained from analysis of social networks [196].

Chapter 1 described the increasing trend in the adoption of *nano-publications* [29] and *liquid publications* [30], where hypotheses or domain-related assertions are published in the form of *short statements* in online knowledge bases. In this new environment, mapping such micro-contributions to expertise will be essential in order to support the development of reputation metrics.

Furthermore, while platforms such as *WikiGenes* [2], a collaborative knowledge base for the life sciences, links every contribution to its author, the Expertise Profiling Framework presented in this thesis complements authorship recognition by attaching semantics to authored content and building profiles based on authored micro-contributions. Expertise profiles will therefore provide authors with due recognition for their contributions, which can in turn be used to complement existing trust and reputation metrics.

As mentioned in Chapter 1, the research presented in this thesis, focuses only on building expertise profiles from micro-contributions. However, a fruitful future research focus would be to use the resulting expertise profiles to provide a robust foundation upon which novel trust and reputation models can be developed and applied.

9.4.6 Enhancement of the Profile Explorer Visualisation Platform

The usability testing of Profile Explorer described in Chapter 8, highlighted a number of interesting directions for future research. Based on the outcomes of this study, future work will focus on: (i) improving the profile browsing and search functionalities (e.g., displaying all micro-contributions on a given concept or for a given time period, facilitating search for experts on a particular topic within a particular time period, displaying expertise profiles for multiple contributors simultaneously to enable comparisons between experts); (ii) facilitating comparative analysis of expertise profiles in order to identify the optimum set of experts for performing a particular task (i.e., team building) or determine experts with the most up-to-date knowledge in particular topic/s (temporal expert finding) and (iii) incorporating the visualisation of peak activity periods across a number of concepts and multiple experts, in a comparative manner.

9.4.7 Enhancement of the Profile Refinement Model

Experimental results presented in Chapter 7, clearly highlight the significance of incorporating social factors into the STEP methodology for building and refining expertise profiles. However, the social factors considered in this study, represent a subset of diverse social factors that exist within social expert networks. In addition, the results represent the effects of social factors in the context of one particular social expert network, i.e., ResearchGate [27].

Future work will focus on identifying, investigating and comparing the effects of various social factors in building expertise profiles. For example, reciprocal citations could be considered as an implicit relationship, based on which experts' profiles can be refined and its effects compared with other implicit relationships, such as the relationships formed through participating in forums and discussions. Furthermore, future work will investigate and compare the effects of various social factors in the context of different social networks, e.g., Google Scholar, Biomed Experts and

Academia.edu, in order to determine if the performance of these factors is influenced by the structure or processes embedded in the underlying networks. Moreover, the proposed Profile Refinement Model will be integrated with various social expert networks, for recommending expertise topics that result from the analysis of contextual / social factors. Experts' response to these recommendations will in turn be used as feedback for improving the proposed model. Finally, profiles refined by the proposed model, will be integrated with the Profile Explorer tool (described in Chapter 8), in order to facilitate visualisation and comparative analysis with profiles created using only micro-contributions (i.e., content-based factors). This analysis will also be used to improve the Profile Refinement Model.

9.5 Summary

This research proposed a possible solution for modelling expertise using *micro-contributions* to *community-driven knowledge-curation* platforms, where knowledge evolves over time. While significant open issues and enhancements remain to be explored or implemented, the research provides solid evidence that high quality expertise profiles, that capture the temporality of expertise, can be generated by analysing *micro-contributions* made to collaborative knowledge platforms. Moreover, the resulting short and long term profiles can be exploited via additional visualization and analysis tools to track the *evolution* of expertise and interests *over time*.

From a *conceptual perspective*, this thesis presented the *Fine-grained Provenance Model for Micro-contributions*, a comprehensive model for capturing and representing the fine-grained provenance of micro-contributions to evolving knowledge platforms. The model represents coarse and fine-grained provenance of micro-contributions through the adoption of a set of existing, established vocabularies from the Semantic Web for capturing micro-contributions and their localisation in the context of their dynamic host documents. More specifically, coarse and fine-grained provenance modelling, are combined using the *SIOC* ontology [151], with change management aspects captured by the *SIOC-Actions* module [152]. The *Annotation Ontology* [153] bridges the textual grounding and the ad-hoc domain knowledge, represented by concepts from domain-specific ontologies. The *Simple Knowledge Organization System* (*SKOS*) [154] ontology is used to define the links to, and the relationships that occur between, these concepts. Furthermore, ontology mappings are defined between the *Open Provenance Model Ontology* [155] and the fine-grained provenance model using the SKOS vocabulary.

In addition, the proposed model captures the textual grounding and domain concepts representing expertise topics embedded in micro-contributions. Finally, the model captures the *temporal* aspect of micro-contributions, providing the flexibility to *track* and *analyse changes* in expertise and interests *over time*. The main contribution of the model is that it facilitates *individual*

attribution, by providing a *contribution-oriented* view of expertise (as opposed to the traditional course, document-centric view that assumes all co-authors have the same expertise).

From an *implementation perspective*, this thesis presented the *Semantic and Time-dependent Expertise Profiling Methodology*, *STEP*, which analyses the fine-grained provenance of micro-contributions (captured by the Fine-grained Provenance Model) to represent the textual grounding of expertise topics, using weighted *concepts* from domain ontologies. Furthermore, the STEP methodology uses the change management aspects of the platform, captured by the Fine-grained Provenance Model, to create *time-aware* expertise profiles represented by *short term* and *long term* profiles.

From an *application perspective*, this thesis verified the applicability of the proposed STEP methodology to a variety of collaborative knowledge platforms that comprised both *unstructured* and *structured* micro-contributions. Application use cases included: the MCB project, Genetics Wiki and the ICD-11 Ontology. Furthermore, the proposed *Profile Refinement Model* (which integrates *contextual and social factors* from social expert networks, with the STEP methodology) was implemented, applied and validated in the context of the ResearchGate social expert network.

The hypothesis that underpins the research described in this thesis is that a *comprehensive*, *fine-grained provenance model*, that is able to capture and consolidate *structured* and *unstructured micro-contributions* made within the context of multiple host documents, will improve expertise profiling in *evolving*, *dynamic* knowledge bases. Evaluations of the proposed Expertise Profiling Framework, reported in this thesis, provide evidence to support this hypothesis.

Bibliography

1. “Collaboration and Expertise Networks,” HP Autonomy, [Online]. Available: http://www.ndm.net/archiving/HP-Autonomy/collaboration-and-expertise-networks. [Accessed 01 September 2014].
2. Hoffmann, R. (2008). A wiki for the life sciences where authorship matters. Nature genetics, 40(9), pp. 1047-1051.
3. M. Sampson, "Expertise Profiles - How Links to Contributions Changed the Dynamics at IBM," [Online]. Available: http://currents.michaelsampson.net/2011/07/expertise-profiles.html. [Accessed 02 September 2014].
4. Balog, K. (2008). People Search in the Enterprise (Doctoral dissertation).
5. Zhang, J., Tang, J., & Li, J. (2007). Expert finding in a social network. In Advances in Databases: Concepts, Systems and Applications, pp. 1066-1069. Springer Berlin Heidelberg.
6. "Saffron," National University of Ireland Galway. [Online]. Available: http://saffron.deri.ie/ [Accessed 17 October 2014].
7. Petkova, D., & Croft, W. B. (2008). Hierarchical language models for expert finding in enterprise corpora. International Journal on Artificial Intelligence Tools, 17(01), pp. 5-18.
8. Fang, H., & Zhai, C. (2007). Probabilistic models for expert finding. In Proceedings of the European Conference on IR Research, (ECIR’07), Berlin, Heidelberg, pp. 418–430. ISBN 978-3-540-71494-1.
9. Ng, A.Y., Jordan, M.I. (2001). On discriminative versus generative classifiers: a comparison of logistic regression and naive Bayes. In Proceedings of the Advances in Neural Information Processing Systems, (NIPS ’01), MIT Press, pp. 841–848. ISBN 0-262-02550-7.
10. "Text REtrieval Conference (TREC)," TREC. [Online]. Available: http://trec.nist.gov/ [Accessed 30 September 2014].
11. Balog, K., Azzopardi, L., & De Rijke, M. (2006). Formal models for expert finding in enterprise corpora. In Proceedings of the 29th annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (SIGIR ’06), New York, USA, pp. 43–50, 2006. ISBN 1-59593-369-7.
12. Balog, K., Azzopardi, L., & de Rijke, M. (2009). A language modeling framework for expert finding. Information Processing & Management, 45(1), pp. 1-19.
13. Balog, K., & De Rijke, M. (2007). Determining Expert Profiles (With an Application to Expert Finding). In Proceedings of the 20th International Joint Conference on Artificial Intelligence (Vol. 7, pp. 2657-2662).

14. Hofmann, K., Balog, K., Bogers, T., & De Rijke, M. (2010). Contextual factors for finding similar experts. Journal of the American Society for Information Science and Technology, 61(5), pp. 994-1014.
15. Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. Scientific American, 284(5), pp. 28-37.
16. Fazel-Zarandi, M., & Fox, M. S. (2013). Inferring and validating skills and competencies over time. Applied Ontology, 8(3), pp. 131-177.
17. Paquette, G. (2007). An ontology and a software framework for competency modelling and management. Educational Technology & Society, 10(3), pp. 1-21.
18. O'Reilly, T., & Musser, J. (2006). Web 2.0 principles and best practices. Retrieved March, 20, 2008.
19. Clark, T., & Kinoshita, J. (2007). Alzforum and SWAN: the present and future of scientific web communities. Briefings in bioinformatics, 8(3), pp. 163-171.
20. "Gene Wiki," Wikipedia, [Online]. Available: http://en.wikipedia.org/wiki/Gene_Wiki. [Accessed 04 September 2014].
21. Zankl, A., Groza, T., Li, Y. F., Ziaimatin, H., Paul, R., & Hunter, J. (2011). The SKELETOME Project: Towards a community-driven knowledge curation platform for Skeletal Dysplasias. In 10th Biennal Meeting of the International Skeletal Dysplasia Society.
22. "WikiProject Medicine," Wikipedia, [Online]. Available: http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine. [Accessed 02 September 2014].
23. "Stack Overflow," Stack Overflow, [Online]. Available: http://stackoverflow.com/. [Accessed 02 September 2014].
24. "The International Classification of Diseases 11th Revision," World Health Organization, [Online]. Available: http://www.who.int/classifications/icd/ICDRevision/. [Accessed 04 September 2014].
25. "World Health Organisation," [Online]. Available: http://www.who.int/about/en/ [Accessed 09 October 2014].
26. Walk, S., Singer, P., Strohmaier, M., Tudorache, T., Musen, M. A., & Noy, N. F. (2014). Discovering Beaten Paths in Collaborative Ontology-Engineering Projects using Markov Chains. Journal of biomedical informatics.
27. "ResearchGate," ResearchGate, [Online]. Available: https://www.researchgate.net. [Accessed 02 September 2014].
28. Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. International journal on semantic web and information systems, 5(3), pp. 1-22.

29. Mons, B., van Haagen, H., Chichester, C., den Dunnen, J. T., van Ommen, G., van Mulligen, E., Singh, B., Roos, M., Hammond, J., Kiesel, B., Giardine, B., Velterop, J., Groth, P., Schultes, E. "The value of data," NATURE GENETICS, 29 March 2011. [Online]. Available: http://www.nature.com/ng/journal/v43/n4/full/ng0411-281.html [Accessed 30 October 2014].
30. Casati, F., Giunchiglia, F., & Marchese, M. (2007). Liquid publications: Scientific publications meet the web.
31. "Language model," Wikipedia, [Online]. Available: http://en.wikipedia.org/wiki/Language_model . [Accessed 04 September 2014].
32. Blei, D. "Topic modeling," Princeton University, [Online]. Available: http://www.cs.princeton.edu/~blei/topicmodeling.html. [Accessed 04 September 2014].
33. De Kok, D., Brouwer, H. Natural Language Processing for the Working Programmer. Available online: http://nlpwp.org/book/index.xhtml (accessed on 04 September 2014).
34. "BioPortal," National Center for Biomedical Ontology, [Online]. Available: http://bioportal.bioontology.org/. [Accessed 04 September 2014].
35. Ziaimatin, H. Profile Explorer. [Online]. Available (tested on Firefox): http://skeletome.metadata.net/dpro/handler/profile/explorer [Accessed 04 September 2014]
36. Rybak, J., Balog, K., & Nørvåg, K. (2014). Temporal expertise profiling. In Advances in Information Retrieval (pp. 540-546). Springer International Publishing.
37. "Biomed Experts", [Online]. Available: http://www.biomedexperts.com/ [Accessed 17 September 2014].
38. "Wikipedia:WikiProject Molecular and Cellular Biology," Wikipedia, [Online]. Available: http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Molecular_and_Cellular_Biology [Accessed 05 September 2014].
39. "Wikipedia:WikiProject Genetics," Wikipedia, [Online]. Available: http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Genetics. [Accessed 05 September 2014].
40. Devedzic, V., & Gašević, D. (Eds.). (2009). Web 2.0 & Semantic Web. Springer.
41. Gruber, T. (2008). Collective knowledge systems: Where the social web meets the semantic web. Web semantics: science, services and agents on the World Wide Web, 6(1), pp. 4-13.
42. "Social web," Wikipedia. [Online]. Available: http://en.wikipedia.org/wiki/Social_web [Accessed 26 September 2014].
43. "Wikipedia:WikiProject RNA," Wikipedia. [Online]. Available: http://en.wikipedia.org/wiki/Wikipedia:WikiProject_RNA [Accessed 26 September 2014].
44. "LabTalk," AstraZeneca. [Online]. Available: http://www.labtalk.astrazeneca.com/ [Accessed 26 September 2014].

45. "Social network," Wikipedia. [Online]. Available: http://en.wikipedia.org/wiki/Social_network [Accessed 29 September 2014].
46. "ResearchGate," CrunchBase. [Online]. Available: http://www.crunchbase.com/product/researchgate [Accessed 10 October 2014].
47. "my experiment," myExperiment. [Online]. Available: http://www.myexperiment.org/ [Accessed 30 September 2014].
48. "Quora The best answer to any question," Quora. [Online]. Available: https://www.quora.com/ [Accessed 30 September 2014].
49. Tudorache, T., Nyulas, C. I., Noy, N. F., & Musen, M. A. (2013). Using Semantic Web in ICD-11: Three Years Down the Road. In The Semantic Web–ISWC 2013 (pp. 195-211). Springer Berlin Heidelberg.
50. Tudorache T, Nyulas C, Noy NF, Musen MA (2013) WebProtégé: A collaborative ontology editor and knowledge acquisition tool for the Web. Semantic Web Journal 4: pp. 89-99.
51. Gruber, T. R. (1993). A translation approach to portable ontology specifications. Knowledge acquisition, 5(2), pp. 199-220.
52. "Ontology (information science)," Wikipedia. [Online]. Available: http://en.wikipedia.org/wiki/Ontology_(information_science) [Accessed 02 October 2014].
53. "Web Ontology Language," Wikipedia. [Online]. Available: http://en.wikipedia.org/wiki/Web_Ontology_Language [Accessed 02 October 2014].
54. "Resource Description Framework," Wikipedia. [Online]. Available: http://en.wikipedia.org/wiki/Resource_Description_Framework [Accessed 02 October 2014].
55. "W3C," World Wide Web Consortium (W3C). [Online]. Available: http://www.w3.org/ [Accessed 02 October 2014].
56. Draganidis, F., & Mentzas, G. (2006). Competency based management: a review of systems and approaches. Information Management & Computer Security, 14(1), pp. 51-64.
57. Houtzagers, G. (1999). Empowerment, using skills and competence management. Participation and Empowerment: An International Journal, 7(2), pp. 27-32.
58. Zhu, J., Gonçalves, A. L., Uren, V. S., Motta, E., & Pacheco, R. (2005). Mining web data for competency management. In Proc. of Web Intelligence (WI 2005), IEEE Computer Society pp. 94–100
59. Buitelaar, P., & Eigner, T. (2008). Topic extraction from scientific literature for competency management. In Personal Identification and Collaborations: Knowledge Mediation and Extraction (PICKME2008)

60. Sure, Y., Maedche, A., & Staab, S. (2000). Leveraging Corporate Skill Knowledge-From ProPer to OntoProPer. In Proceedings of the third international conference on practical aspects of knowledge management, pp. 30–31.
61. Paquette, G. (2007). An ontology and a software framework for competency modelling and management. Educational Technology & Society, 10(3), pp. 1-21.
62. Heath, T., & Motta, E. (2008). The Hoonoh ontology for describing trust relationships in information seeking. Personal Identification and Collaborations: Knowledge Mediation and Extraction (PICKME2008).
63. Hunter, L., Lu, Z., Firby, J., Baumgartner, W. A., Johnson, H. L., Ogren, P. V., & Cohen, K. B. (2008). OpenDMAP: an open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression. BMC bioinformatics, 9(1), p. 78.
64. Müller, H. M., Kenny, E. E., & Sternberg, P. W. (2004). Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS biology, 2(11), e309.
65. Van Landeghem, S., Hakala, K., Rönnqvist, S., Salakoski, T., Van de Peer, Y., & Ginter, F. (2012). Exploring biomolecular literature with EVEX: Connecting genes through events, homology, and indirect associations. Advances in bioinformatics, 2012.
66. "PubMed," PubMed. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed [Accessed 03 October 2014].
67. "BioTxtM 2014," LREC 2014. 5 June 2014. [Online]. Available: http://www.nactem.ac.uk/biotxtm2014/ [Accessed 03 October 2014].
68. Lindberg, D. A., Humphreys, B. L., & McCray, A. T. (1993). The Unified Medical Language System. Methods of information in medicine, 32(4), pp. 281-291.
69. Bank, Hazardous Substances Data (1998). National Library of Medicine. Bethesda, Maryland (TOMES CPS# CD-ROM)
70. "The open biological and biomedical ontologies," Foundry, O. B. O. [Online]. Available: http://www.obofoundry.org/ [Accessed 03 October 2014].
71. "The National Center for Biomedical Ontology," National Center for Biomedical Ontology. [Online]. Available: http://www.bioontology.org/ [Accessed 03 October 2014].
72. "ICD Information Sheet," World Health Organisation. [Online]. Available: http://www.who.int/classifications/icd/factsheet/en/ [Accessed 07 October 2014].
73. SáNchez, D., & Batet, M. (2011). Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. Journal of biomedical informatics, 44(5), pp. 749-759.

74. Pesquita, C., Faria, D., Falcao, A. O., Lord, P., & Couto, F. M. (2009). Semantic similarity in biomedical ontologies. PLoS computational biology, 5(7), e1000443.
75. Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In Proceedings of The 1995 International Joint Conference on Artificial Intelligence, pp. 448–453, Montreal, Canada.
76. Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. IEEE Transaction on Systems, Man, and Cybernetics, 19(1), pp. 17–30.
77. Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. WordNet: An electronic lexical database, 49(2), pp. 265-283.
78. Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. In Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pp. 133-138. Association for Computational Linguistics.
79. "Pattern recognition," Wikipedia. [Online]. Available: http://en.wikipedia.org/wiki/Pattern_recognition [Accessed 09 October 2014].
80. "Text mining," Wikipedia. [Online]. Available: http://en.wikipedia.org/wiki/Text_mining [Accessed 09 October 2014].
81. Hur, J., Schuyler, A. D., States, D. J., & Feldman, E. L. (2009). SciMiner: web-based literature mining tool for target identification and functional enrichment analysis. Bioinformatics.
82. Tudor, C. O., Arighi, C. N., Wang, Q., Wu, C. H., & Vijay-Shanker, K. (2012). The eFIP system for text mining of protein interaction networks of phosphorylated proteins. Database: The Journal of Biological Databases and Curation, 2012.
83. Campos, D., Matos, S., & Oliveira, J. L. (2013). A modular framework for biomedical concept recognition. BMC bioinformatics, 14(1), p. 281.
84. Settles, B. (2004). Biomedical named entity recognition using conditional random fields and rich feature sets. In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, pp. 104-107. Association for Computational Linguistics.
85. Campos, D., Matos, S., & Oliveira, J. (2012). Current Methodologies for Biomedical Named Entity Recognition. In Biological Knowledge Discovery Handbook: Preprocessing, Mining And Postprocessing Of Biological Data, pp. 839-868.
86. Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. Journal of the American Medical Informatics Association, 18(5), pp. 544-551.
87. "GATE Information Extraction," The University of Sheffield. [Online]. Available: http://www.gate.ac.uk/ie [Accessed 09 October 2014].

88. "Apache Unstructured Information Management Architecture," The Apache Software Foundation. [Online]. Available: http://uima.apache.org ([Accessed 09 October 2014].
89. Shah, N. H., Bhatia, N., Jonquet, C., Rubin, D., Chiang, A. P., & Musen, M. A. (2009). Comparison of concept recognizers for building the Open Biomedical Annotator. BMC Bioinformatics, 10(Suppl 9), S14.
90. Whetzel, P. L., Noy, N. F., Shah, N. H., Alexander, P. R., Nyulas, C., Tudorache, T., & Musen, M. A. (2011). BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic acids research, 39(suppl 2), W541-W545.
91. Musen, M. A., Noy, N. F., Shah, N. H., Whetzel, P. L., Chute, C. G., Story, M. A., & Smith, B. (2012). The national center for biomedical ontology. Journal of the American Medical Informatics Association, 19(2), pp. 190-195
92. Jonquet, C.; Shah, N.; Musen, M. (2009). The Open Biomedical Annotator. In Proceedings of the Summit of Translational Bioinformatics, San Francisco, CA, USA, pp. 56–60.
93. "Experimental Factor Ontology," BioPortal. Online]. Available: http://bioportal.bioontology.org/ontologies/EFO [Accessed 08 October 2014].
94. Jonquet, C., Musen, M. A., & Shah, N. H. (2010). Building a biomedical ontology recommender web service. Journal of Biomedical Semantics, 1(S-1), S1.
95. Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), pp. 77-84.
96. Balog, K., Azzopardi, L., & De Rijke, M. (2006). Formal models for expert finding in enterprise corpora. In Proceedings of the 29th annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (SIGIR '06), New York, USA, pp. 43–50, 2006. ISBN 1-59593-369-7.
97. Petkova, D., & Croft, W. B. (2008). Hierarchical language models for expert finding in enterprise corpora. International Journal on Artificial Intelligence Tools, 17(01), pp. 5-18.
98. Balog, K., Fang, Y., de Rijke, M., Serdyukov, P., & Si, L. (2012). Expertise Retrieval. Foundations and Trends in Information Retrieval, 6(2-3), pp. 127-256.
99. Serdyukov, P., & Hiemstra, D. (2008). Modelling documents as mixtures of persons for expert finding. In Proceedings of the European Conference on IR Research, (ECIR '08), Berlin, Heidelberg, pp. 309–320. ISBN 978-3-540-78645-0.
100. Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In Machine learning: ECML-98, pp. 4-15. Springer Berlin Heidelberg.

101. Sahlgren, M., & Cöster, R. (2004). Using bag-of-concepts to improve the performance of support vector machines in text categorization. In Proceedings of the 20th international conference on Computational Linguistics, p. 487. Association for Computational Linguistics.
102. Zhu, J., Song, D., & Rüger, S. (2009). Integrating multiple windows and document features for expert finding. Journal of the American Society for Information Science and Technology, 60(4), pp. 694-715.
103. Yang, L., & Zhang, W. (2010). A study of the dependencies in expert finding. In Knowledge Discovery and Data Mining, 2010. WKDD'10. Third International Conference on, pp. 355-358. IEEE.
104. Thiagarajan, R., Manjunath, G., & Stumptner, M. (2008). Finding experts by semantic matching of user profiles, Technical Report HPL-2008-172, HP Laboratories.
105. Pflugrad, A., Jurkat-Rott, K., Lehmann-Horn, F., & Bernauer, J. (2013). Towards the Automated Generation of Expert Profiles for Rare Diseases through Bibliometric Analysis. Studies in health technology and informatics, 198, pp. 47-54.
106. Tang, J., Zhang, J., Zhang, D., Yao, L., Zhu, C., & Li, J. Z. (2007). ArnetMiner: An Expertise Oriented Search System for Web Community. In Proceedings of Semantic Web Challenge'2007.
107. Crowder, R., Hughes, G., & Hall, W. (2002). An agent based approach to finding expertise. In Practical Aspects of Knowledge Management, pp. 179-188. Springer Berlin Heidelberg.
108. Liu, P., Ye, Y., & Liu, K. (2008). Building a Semantic Repository of Academic Experts. In Wireless Communications, Networking and Mobile Computing, 2008. WiCOM'08, pp. 1-6. IEEE.
109. Stankovic, M., Wagner, C., Jovanovic, J., & Laublet, P. (2010). Looking for experts? What can linked data do for you? In: Bizer, C., Heath, T., Berners-Lee, T., Hausenblas, M. (eds.) Linked Data on the Web (LDOW 2010). CEUR Workshop Proceedings.
110. "W3C," World Wide Web Consortium (W3C). [Online]. Available: http://www.w3.org/ [Accessed 02 October 2014].
111. Balog, K., Bogers, T., Azzopardi, L., De Rijke, M., & Van Den Bosch, A. (2007). Broad expertise retrieval in sparse data environments. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 551-558. ACM.
112. Demartini, G. (2007). Finding Experts Using Wikipedia. In Proc. of the ExpertFinder Workshop, co-located with ISWC 2007, Busan, Korea, 290, pp. 33-41.

113. "SemEval-2007," SemEval, 2007. [Online]. Available: http://nlp.cs.swarthmore.edu/semeval/ [Accessed 30 September 2014].
114. Fuhr, N., Gövert, N., Kazai, G., & Lalmas, M. (2002). INEX: INitiative for the Evaluation of XML retrieval. In Proceedings of the SIGIR 2002 Workshop on XML and Information Retrieval, Vol. 2006, pp. 1-9.
115. Price, S., Flach, P. A., Spiegler, S., Bailey, C., & Rogers, N. (2010) "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.
116. Richardson, L., & Ruby, S. (2008). RESTful web services. "O'Reilly Media, Inc."
117. J.L.Neto, A.D.Santos, C.A.A. Kaestner, and A.A.Freitas. (2000) Document Clustering and Text Summarization. 4th International Conference on Practical Applications of Knowledge Discovery and Data Mining, London, 2000.
118. Mitchell, J., & Lapata, M. (2008). Vector-based models of semantic composition. In Proceedings of ACL-08: HLT, pp. 236–244, Columbus, Ohio. Association for Computational Linguistics.
119. Aleman-Meza, B., Bojars, U., Boley, H., Breslin, J. G., Mochol, M., Nixon, L. J., Polleres, A. & Zhdanova, A. V. (2007). "Combining RDF vocabularies for expert finding", In Proc. of the 4th European Semantic Web Conference, Innsbruck, Austria, 2007, pp. 235-250.
120. Michelson, M., & Macskassy, S. A. (2010). "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, pp. 73-80.
121. Abel, F., Gao, Q., Houben, G. J., & Tao, K. (2011). Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web. In Proc. of the 8th Extended Semantic Web Conference, pp. 375-389. Springer Berlin Heidelberg.
122. Balog, K, "Entity and Association Retrieval System". [Online]. Available: http://code.google.com/p/ears/ [Accessed 30 September 2014].
123. Stankovic, M., Jovanovic, J., & Laublet, P. (2011). Linked data metrics for flexible expert search on the open web. In The Semantic Web: Research and Applications, pp. 108-123. Springer Berlin Heidelberg.
124. Ehrlich, K., Lin, C. Y., & Griffiths-Fisher, V. (2007). Searching for experts in the enterprise: combining text and social network analysis. In Proceedings of the 2007 International ACM SIGGROUP Conference on Supporting Group Work, (GROUP '07), New York, USA, pp. 117–126. ISBN 978-1-59593-845-9.

125. Carrington, P. J., Scott, J., & Wasserman, S. (Eds.). (2005). Models and methods in social network analysis (Vol. 28), Cambridge University Press, Cambridge.
126. Campbell, C. S., Maglio, P. P., Cozzi, A., & Dom, B. (2003). Expertise identification using email communications. In Proceedings of the twelfth international conference on Information and knowledge management, pp. 528-531. ACM.
127. Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford InfoLab.
128. Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. In Proceedings of the Nineth Annual ACM-SIAM Symposium on Discrete Algorithms, 46(5), pp. 604-632.
129. "SciVal Experts," Elsevier. [Online]. Available: http://www.elsevier.com/reviewers/reviewers-update/archive/issue-9/scival-experts [Accessed 01 October 2014].
130. "Elsevier Fingerprint Engine," Elsevier. [Online]. Available: http://www.elsevier.com/online-tools/research-intelligence/products-and-services/elsevier-fingerprint-engine [Accessed 01 October 2014].
131. "Scopus," Elsevier. [Online]. Available: http://www.elsevier.com/online-tools/scopus [Accessed 01 October 2014].
132. "Professional Networking and Expertise Mining for Research Collaboration," The Harvard Clinical and Translational Science Center. [Online]. Available: http://profiles.catalyst.harvard.edu/?pg=home. [Accessed 30 September 2014].
133. Yan, E., Ding, Y., Milojević, S., & Sugimoto, C. R. (2012). Topics in dynamic research communities: An exploratory study for the field of information retrieval. Journal of Informetrics, 6(1), pp. 140-153.
134. Felzenszwalb, P. F., & Huttenlocher, D. P. (2006). Efficient belief propagation for early vision. In IEEE Conference on Computer Vision and Pattern Recognition, 70(1), pp. 41-54.
135. Lin, L., Xu, Z., Ding, Y., & Liu, X. (2013). Finding topic-level experts in scholarly networks. Scientometrics 97 (3), pp. 797-819.
136. Bordea, G. (2013). Domain Adaptive Extraction of Topical Hierarchies for Expertise Mining. Ph.D. thesis, National University of Ireland, Galway.
137. Edwards, P., Pignotti, E., Eckhardt, A., Ponnamperuma, K., Mellish, C., & Bouttaz, T. (2012). ourSpaces–design and deployment of a semantic virtual research environment. In The Semantic Web–ISWC 2012, pp. 50-65. Springer Berlin Heidelberg.
138. "eagle-i," National Institutes of Health. [Online]. Available: https://www.eagle-i.net/ [Accessed 17 October 2014].

139. "An interdisciplinary network," VIVO. [Online]. Available: http://vivoweb.org/ [Accessed 17 October 2014].
140. "Clinical and Translational Science Awards," National Center for Advancing Translational Sciences. [Online]. Available: http://www.ncats.nih.gov/research/cts/ctsa/ctsa.html [Accessed 17 October 2014].
141. "CTSAconnect," National Center for Advancing Translational Sciences. [Online]. Available: http://www.ctsaconnect.org/ [Accessed 17 October 2014].
142. Simperl, E., & Luczak-Rösch, M. (2014). Collaborative ontology engineering: a survey. The Knowledge Engineering Review, 29(01), pp. 101-131.
143. "Bone Dysplasia Ontology," BioPortal. [Online]. Available: http://bioportal.bioontology.org/ontologies/BDO [Accessed 08 October 2014].
144. "Human Phenotype Ontology," BioPortal. Online]. Available: http://purl.bioontology.org/ontology/HP [Accessed 08 October 2014].
145. Ziaimatin, H., Groza, T., Bordea, G., Buitelaar, P., & Hunter, J. (2012). Expertise profiling in evolving knowledge-curation platforms. Global Science and Technology Forum Journal on Computing, 2(3), pp. 118-127.
146. Groza, T., Tudorache, T., & Dumontier, M. (2013). Commentary: State of the art and open challenges in community-driven knowledge curation. Journal of Biomedical Informatics, 46(1), pp. 1-4.
147. Groza, T., Handschuh, S., Breslin, J. G., & Decker, S. (2009). An abstract framework for modelling argumentation in virtual communities. International Journal of Virtual Communities and Social Networking (IJVCSN), 1(3), pp. 37-49.
148. Rector, A. L. (2003). Modularisation of domain ontologies implemented in description logics and related formalisms including OWL. In Proceedings of the 2nd international conference on Knowledge capture, pp. 121-128. ACM.
149. "Portal:Gene Wiki," WikiPedia. [Online]. Available: http://en.wikipedia.org/wiki/Portal:Gene_Wiki [Accessed 22 October 2014].
150. "Online Mendelian Inheritance in Man," Johns Hopkins University. [Online]. Available: http://omim.org/ [Accessed 09 October 2014].
151. Breslin, J. G., Decker, S., Harth, A., & Bojars, U. (2006). SIOC: an approach to connect web-based communities. The International Journal of Web-based Communities, 2(2), pp. 133-142.
152. Champin, P. A., & Passant, A. (2010). SIOC in action representing the dynamics of online communities. In Proceedings of the 6th International Conference on Semantic Systems, pp. 1-7. ACM.

153. Ciccarese, P., Ocana, M., Garcia-Castro, L. J., Das, S., & Clark, T. (2011). An open annotation ontology for science on web 3.0. Journal of Biomedical Semantics, 2(S-2), S4.
154. "SKOS Simple Knowledge Organization System," W3C. [Online]. Available: http://www.w3.org/TR/skos-reference [Accessed 22 October 2014].
155. Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P. and Myers, J. (2011). The open provenance model core specification (v1. 1). Future Generation Computer Systems, 27(6), pp. 743-756
156. "W3C Provenance Incubator Group Wiki," W3C. [Online]. Available: http://www.w3.org/2005/Incubator/prov/ [Accessed 22 October 2014].
157. Ogden, C., Richards, I.A. (1923). The Meaning of Meaning: A study of the influence of language upon thought and of the science of symbolism. Magdalene College, University of Cambridge.
158. Niwa, Y., & Nitta, Y. (1994). Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In Proceedings of the 15th International Conference on Computational Linguistics, COLING'94, pp. 304-309. Association for Computational Linguistics.
159. "Systematized Nomenclature of Medicine - Clinical Terms," National Center for Biomedical Ontology BioPortal. [Online]. Available: http://bioportal.bioontology.org/ontologies/SNOMEDCT [Accessed 27 October 2014].
160. "MedlinePlus Health Topics," National Center for Biomedical Ontology BioPortal. [Online]. Available: http://bioportal.bioontology.org/ontologies/MEDLINEPLUS [Accessed 27 October 2014].
161. "Radiology Lexicon," National Center for Biomedical Ontology BioPortal. [Online]. Available: http://bioportal.bioontology.org/ontologies/RADLEX [Accessed 27 October 2014].
162. "Read Codes, Clinical Terms Version 3 (CTV3)," National Center for Biomedical Ontology BioPortal. [Online]. Available: http://bioportal.bioontology.org/ontologies/RCD [Accessed 27 October 2014].
163. Ziaimatin, H., Groza, T., & Hunter, J. (2013). Semantic and Time-Dependent Expertise Profiling Models in Community-Driven Knowledge Curation Platforms. Future Internet, 5(4), pp. 490-514.
164. Jonquet, C., Shah, N., Youn, C., Callendar, C., Storey, M. A., & Musen, M. (2009). NCBO annotator: semantic annotation of biomedical data. In International Semantic Web Conference, Poster and Demo session, Washington, D.C., WA, USA.

165. Dai, M., Shah, N. H., Xuan, W., Musen, M. A., Watson, S. J., Athey, B. D., & Meng, F. (2008). An efficient solution for mapping free text to ontology terms. AMIA Summit on Translational Bioinformatics, San Francisco.
166. Xuan, W., Dai, M., Mirel, B., Athey, B., Watson, S. J., & Meng, F. (2007). Interactive Medline Search Engine Utilizing Biomedical Concepts and Data Integration. BioLINK SIG: Linking Literature, Information and Knowledge for Biology; Vienna, Austria. pp. 55–58.
167. "National Center for Integrative Biomedical Informatics (NCIBI)," The University of Michigan. [Online]. Available: http://portal.ncibi.org/gateway/ [Accessed 05 November 2014].
168. Gudgin, M., Hadley, M., Mendelsohn, N., Moreau, J. J., Nielsen, H. F., Karmarkar, A., & Lafon, Y. (2003). Simple object access protocol (SOAP) 1.2. World Wide Web Consortium.
169. "Stemming and Lemmatization," [Online]. Available: http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html [Accessed 29 October 2014].
170. "Lemmatisation," Wikipedia. [Online]. Available: http://en.wikipedia.org/wiki/Lemmatisation [Accessed 29 October 2014].
171. Liu, H., Christiansen, T., Baumgartner Jr, W. A., & Verspoor, K. (2012). BioLemmatizer: A lemmatization tool for morphological processing of biomedical text. Journal of Biomedical Semantics, 3, 3:1–3:29.
172. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. The Journal of Machine Learning Research, 3, pp. 993-1022.
173. Groza, T., Zankl, A., Li, Y. F., & Hunter, J. (2011). Using Semantic Web Technologies to Build a Community-Driven Knowledge Curation Platform for the Skeletal Dysplasia Domain. In Proceedings of the 10th International Semantic Web Conference, Bonn, Germany, 23–27 October 2011; pp. 81–96. Springer Berlin Heidelberg.
174. "MAchine Learning for LanguagE Toolkit," University of Massachusetts Amherst. [Online]. Available: http://mallet.cs.umass.edu/topics.php [Accessed 04 November 2014].
175. Monaghan, F., Bordea, G., Samp, K., & Buitelaar, P. (2010). Exploring Your Research: Sprinkling some Saffron on Semantic Web Dog Food. In Proceedings of the Semantic Web Challenge at the International Semantic Web Conference, Shanghai, China, 7–11 November 2010.
176. "Markov chain," Wikipedia. [Online]. Available: http://en.wikipedia.org/wiki/Markov_chain [Accessed 24 November 2014].
177. "International Classification of Diseases," World Health Organization. [Online]. Available: http://www.who.int/classifications/icd/en/ [Accessed 05 November 2014].

178. Ziaimatin, H., Groza, T., Tudorache, T. & Hunter, J. (2014). Modelling expertise at different levels of granularity using semantic similarity measures in the context of collaborative knowledge-curation platforms. Manuscript submitted for publication.
179. Noy, N. F., Chugh, A., Liu, W., & Musen, M. A. (2006). A framework for ontology evolution in collaborative environments. In Proceedings of the 5th International Semantic Web Conference, pp. 544-558, Springer Berlin Heidelberg.
180. Seco, N., Veale, T., & Hayes, J. (2004). An intrinsic information content metric for semantic similarity in WordNet. In Proceedings of the 16th European Conference on Artificial Intelligence, pp. 1089–1090.
181. Zhou, Z., Wang, Y., & Gu, J. (2008). A new model of information content for semantic similarity in WordNet. In Proceedings of the Second International Conference on Future Generation Communication and Networking Symposia, pp. 85-89. IEEE.
182. Ziaimatin, H., Groza, T. & Hunter, J. (2014). Building expertise profiles from micro-contributions and social collaboration factors. Manuscript submitted for publication.
183. Shami, N. S., Yuan, Y. C., Cosley, D., Xia, L., & Gay, G. (2007). That's what friends are for: facilitating 'who knows what' across group boundaries. In proceedings of the 2007 International ACM Conference on Supporting Group Work, (GROUP '07), New York, NY, USA, pp. 379–382, 2007. ISBN 978-1-59593-845-9.
184. Bhandari, M., Einhorn, T. A., Swiontkowski, M. F., & Heckman, J. D. (2003). Who did what? (Mis) perceptions about authors' contributions to scientific articles based on order of authorship. The Journal of Bone & Joint Surgery, 85(8), pp. 1605-1609.
185. "Timeline JS," Northwestern University. [Online]. Available: http://timeline.verite.co [Accessed 10 November 2014].
186. "Data-Driven Documents". [Online]. Available: http://d3js.org [Accessed 10 November 2014].
187. Vardell, E., Feddern-Bekcan, T., & Moore, M. (2011). SciVal experts: A collaborative tool. Medical reference services quarterly, 30(3), pp. 283-294.
188. Mockus, A., & Herbsleb, J. D. (2002). Expertise browser: a quantitative approach to identifying expertise. In Proceedings of the 24th international conference on software engineering, pp. 503-512. ACM.
189. Rybak, J., Balog, K., & Nørvåg, K. ExperTime: Tracking Expertise over Time. In Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '14, pp. 1273-1274

190. Börner, K., Conlon, M., Corson-Rikert, J., & Ding, Y. (2012). VIVO: A semantic approach to scholarly networking and discovery. Synthesis Lectures on The Semantic Web: Theory and Technology, 7(1), pp. 1-178.
191. “Portal:Astronomy,” Wikipedia, [Online]. Available: http://en.wikipedia.org/wiki/Portal:Astronomy [Accessed 21 November 2014].
192. “Portal:Earth sciences,” Wikipedia, [Online]. Available: http://en.wikipedia.org/wiki/Portal:Earth_sciences [Accessed 21 November 2014].
193. “Portal:Chemistry,” Wikipedia, [Online]. Available: http://en.wikipedia.org/wiki/Portal:Chemistry [Accessed 21 November 2014].
194. "altmetrics: a manifesto," altmetrics. [Online]. Available: http://altmetrics.org/manifesto/ [Accessed 16 November 2014].
195. "NISO Alternative Assessment Metrics (Altmetrics) Initiative," National Information Standards Organization. [Online]. Available: http://www.niso.org/topics/tl/altmetrics_initiative/ [Accessed 16 November 2014].
196. Sabater, J., & Sierra, C. (2002). Reputation and social network analysis in multi-agent systems. In Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1, pp. 475-482. ACM.
197. Golbeck, J. (2008). Computing with social trust. Springer.
198. "Natural Language Toolkit (NLTK)," Natural Language Toolkit. [Online]. Available: http://www.nltk.org/ [Accessed 11 February 2015].
199. Serdyukov, P., Rode, H., & Hiemstra, D. (2008). Modelling multi-step relevance propagation for expert finding. In Proceedings of the 17th ACM conference on Information and knowledge management, pp. 1133-1142. ACM.
200. "The Initiative for the Evaluation of XML Retrieval (INEX)," Saarland University. [Online]. Available: https://inex.mmci.uni-saarland.de/ [Accessed 27 February 2015].

Appendix 1: Tasks Evaluated in the Profile Explorer Usability Study

Description: The Profile Explorer underwent a usability study involving 6 users who were each asked to perform a set of 9 tasks and then to rank the difficulty of performing each task on a 5-point Likert scale (1=Very difficult; 2=Difficult; 3=Average difficulty; 4=Easy to 5=Very easy). The users also had to rank their confidence in performing the task successfully (from *Not at all confident* to *Very confident*). The nine tasks were designed to evaluate three major aspects of the Profile Explorer: *browsing, search* and *analysis*.

List of Tasks that users were asked to perform and score:

1. Open project participant AaronM's timeline (*Browsing*)

2. Browse to the week starting '22 September 2006' (*Browsing*)

3. Cilium is a prominent term for this week, find (and write down) a term that is less prominent than Cilium (*Search*)
Term: _____

4. For the week starting '22 September 2006', search for AaronM's contribution involving 'Cilium' (*Search*)

5. Find (and write down) on which date AaronM made a contribution about Cilium (*Analysis*)
Date: _____

Go back to the main timeline

6. Browse to 'Long term profile' (*Browsing*)

7. Using the long term profile, search for weeks in which 'eukaryote' is mentioned (*Search*)

8. Find (and write down) the first week in which 'eukaryote' is mentioned (*Analysis*)

9. Identify in which year AaronM most actively contributed about the topic ‘eukaryote’ (*Analysis*)