UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

Electronics and Computer Science

Exploiting Linked Open Data (LoD) and Crowdsourcing-based semantic annotation & tagging in web repositories to improve and sustain relevance in search results

by:

Arshad Ali Khan

ORCID ID: 0000-0002-6062-0098

Thesis for the degree of Doctor of Philosophy

November 2018

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES
Electronics and Computer Science

Doctor of Philosophy

EXPLOITING LINKED OPEN DATA (LOD) AND CROWDSOURCING-BASED SEMANTIC ANNOTATION & TAGGING IN WEB REPOSITORIES TO IMPROVE AND SUSTAIN RELEVANCE IN SEARCH RESULTS

by: Arshad Ali Khan

Online searching of multi-disciplinary web repositories is a topic of increasing importance as the number of repositories increases and the diversity of skills and backgrounds of their users widens. Earlier term-frequency based approaches have been improved by ontology-based semantic annotation, but such approaches are predominantly driven by "domain ontologies engineering first" and lack dynamicity, whereas the information is dynamic; the meaning of things changes with time; and new concepts are constantly being introduced. Further, there is no sustainable framework or method, discovered so far, which could automatically enrich the content of heterogeneous online resources for information retrieval over time. Furthermore, the methods and techniques being applied are fast becoming inadequate due to increasing data volume, concept obsolescence, and complexity and heterogeneity of content types in web repositories. In the face of such complexities, term matching alone between a query and the indexed documents will no longer fulfil complex user needs. The ever growing gap between syntax and semantics needs to be continually bridged in order to address the above issues; and ensure accurate search results retrieval, against natural language queries, despite such challenges. This thesis investigates that by domain-specific expert crowd-annotation of content, on top of the automatic semantic annotation (using Linked Open Data sources), the contemporary value of content in scientific repositories, can be continually enriched and sustained. A purpose-built annotation, indexing and searching environment has been developed and deployed to a web repository, which hosts more than 3,400 heterogeneous web documents. Based on expert crowd annotations, automatic LoD-based named entity extraction and search results evaluations, this research finds that search results retrieval, having the crowd-sourced element, performs better than those having no crowd-sourced element. This thesis also shows that a consensus can be reached between the expert and non-expert crowd-sourced annotators on annotating and tagging the content of web repositories, using the controlled vocabulary (typology) and free-text terms and keywords.

Contents

List of Figures

List of Tables

List of Algorithms

Listings

List of Publications

1. Khan. A., Tiropanis. T and Martin. D.,*Paper published*-Crowd-annotation and LoD-based semantic indexing of content in multi-disciplinary web repositories to improve search results, *Being reviewed for the conference proceedings of the AWC 2017:The Australasian Web Conference 2017-January 31-February 3,2016*

2. Khan. A. and Tiropanis T. and Martin. D., *Paper published*-Exploiting Semantic Annotation of Content with Linked Open Data (LoD) to Improve Searching Performance in Web Repositories of Multi-disciplinary Research Data, *In proceedings of the 9th Russian Summer School in Information Retrieval (RuSSIR 2015)*, Springer International Publishing

3. Khan. A., Tiropanis. T and Martin. D.,*Using Semantic Indexing to Improve Searching Performance in Web Archives*,2013, *In proceedings of the First International Conference on Building and Exploring Web Based Environments (WEB2013), 27 Jan - 01 Feb, 2013*, **Awarded Best Paper Award**

4. D. Byatt and A. Khan and I. Stark and W. White, *Embedded librarians in the National Centre for Research Methods (NCRM) and in Chemistry*, organised by the University of Southampton in July 2015, *In proceedings of the IT as a Utility Network+ community conference, 6th-7th July 2015*, Available at http://eprints.soton.ac.uk/377553/

Declaration of Authorship

I, Arshad Ali Khan , declare that the thesis entitled *Exploiting Linked Open Data (LoD) and Crowdsourcing-based semantic annotation & tagging in web repositories to improve and sustain relevance in search results* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published as: Arshad Ali Khan

Signed:...

Date:..

Acknowledgements

First and foremost, I am very grateful to my supervisors Dr.Thanassis Tiropanis and Prof. Hugh Davis for their support and guidance through the course of this PhD. I am sure to say that without their support and supervision, it won't have been possible to make it quite this far. I am also immensely grateful to Prof. David Martin for his critical and profoundly effective review of my PhD thesis in terms of proof-reading and general discussions. David has been really helpful from the outset and provided very timely suggestions and advice throughout my PhD studies.

I am very thankful to my examiners Prof. Carole Goble and Prof. Les Carr for taking time to understand and critically evaluate my thesis.

To the ESRC National Centre for Research Methods (NCRM) and ReStore project teams, many thanks for your outstanding financial and moral support during the course of my PhD study. To Samantha Cockings, many thanks for your lunch-time discussions over my PhD progress and for your constant encouragement.

I am deeply appreciative of the study participants who participated in various experiments at various stages of this research. I sincerely appreciate and acknowledge your cooperation. I am thankful to all the people, I have met in various conferences and workshops, for their insightful suggestions, criticism and encouragements.

I also acknowledge the support of *Annotateit.org* team, *IBM Watson Alchemy API team*, *Library services professionals* in Southampton and UK Data Service for cooperating in understanding various typologies, vocabularies in the context of crowd-annotation in web repositories.

To my wife Shagufta Arshad, no words to sum up your cooperation in balancing *work-PhD-family* troica but I am deeply indebted to you for your unstinting support and encouragement throughout my PhD. I love you so much. To my daughter Hira, thank you very much for bearing with me until now, as I know you want me to set up a new rabbit hutch and to decorate your room, after my thesis submission.

To my parents, I love you very much and very many thanks for your prayers and wishes throughout this PhD journey. And last, but by no means least, I am so grateful to my family and friends for their immense emotional support to keep me motivated and focused on my PhD alongside full-time work.

Abbreviations

RQs	**R**esearch Questions
NCRM	**N**ational Centre for Research Methods
ESRC	**E**conomic and Social Research Council
SRR	**S**earch Results Retrieval
KOS	**K**nowledge Organization Systems
IR	**I**nformation Retrieval
IE	**I**nformation Extraction
KB	**K**nowledge Base
RDBMS	**R**elationship Database Management System
ES	**E**lasticsearch
JSON	**J**avaScript Object Notation
QAC	**Q**uery Auto-Completion
TF	**T**erm Frequency
IDF	**I**nverse Document Frequency
LoD	**L**inked Open Data
AnnoTagger	**A**nnotation & Tagging tool
SW	**S**emantic Web
AP	**A**verage Precision
MAP	**M**ean Average Precision
MAR	**M**ean Average Ranking
AR	**A**verage Ranking
OWL	**W**eb Ontology Language

SPARQL	**S**PARQL Protocol and RDF Query Language
OWLIM	**S**emantic repository for OWL
VSM	**V**ector Space Model
CSM	**C**osine Similarity Metrics
HVSM	**H**yper Vector Space Model
RDF	**R**esource Description Framework
RESTFUL	**R**epresentational State Transfer (Stateless client-server)
KIM	**K**nowledge Information Management
RM	**R**elevance Maximization

To. . . My lovely parents

Chapter 1

Online searching in multi-disciplinary web repositories

In the present age it is practically impossible to find anything on the web without employing a search engine to assist us and they are the primary gatekeepers of the Web (Levene, 2011). Searching for relevant information in multi-disciplinary web repositories[1] is emerging a topic of immense interest among the computer science and social research communities. Scholars from various disciplines use the Web every day to search, read and collaborate to satisfy their specific or general information needs. To date, methods and techniques to extract useful and relevant information from the online repositories of research data, have largely been based on static full text indexing which entails a 'produce once and use forever' strategy. That strategy is fast becoming inadequate due to increasing data volume, concept obsolescence, and complexity and heterogeneity of content types in web repositories.

Current searching techniques are predominantly based on keywords instances which are matched against the content in web resources repositories while paying little attention to analysing the semantics, types of content, context and relationship of keywords and phrases. In addition, poor or almost no linkages amongst web pages and other web documents restricts the users to merely rely on the presence of the keywords and phrases in web pages. This is now becoming a challenge for users due to the *information and data* deluge phenomenon of the current digital age. This issue is further complicated when we look at it across time and disciplines, where

[1] A web repository stores and provides long term online access to a collection of web sites or web resources (containing static and dynamic web pages), research papers, presentations, experimental code scripts, reports etc. funded by UK research councils. Examples include http://www.data-archive.ac.uk/, http://www.restore.ac.uk/, http://www.ncrm.ac.uk/, http://www.timescapes.leeds.ac.uk/, http://www.icpsr.umich.edu/icpsrweb/ICPSR/

changes in concepts, language and terminological obsolescence, change the meanings of content in today's web resources thereby compromising the performance of retrieval systems in the future.

Web searching in a specific domain, is increasingly emerging a topic of substantial interest and there have been impressive developments taking place in that area, so far focusing mainly on semantic searching and retrieval. The complex information needs of users, expressed predominantly in natural language, lay bare the semantic gap between the data stored in search indices and the search results retrieved against users' queries. Contemporary research users struggle to filter out irrelevant information, especially in a scientific discipline where relevance and precision are of great importance to support ongoing research studies. Time is another factor, on top of social, technological, scientific and socio-economic changes, which alters the meanings of various concepts, terminologies and things over time. This results in making it difficult for the search engines, which use the same Boolean search model, to satisfy the information needs of online users in a particular research domain.

Most of the research initiatives have been focusing upon methods and techniques to extract meaningful and relevant information from large repositories of text, structuring them in a searchable platform and offering them in a web container on which users can perform online search. The fact that the vast majority of search engines still use the lexical or literal matching of terms in the user's query with content in the documents, necessitates the transformation of these techniques. The primary factors that drive such transformation include, but are not limited to, phenomenal data volume, fast concept obsolescence in a particular domain of interest, and complexity and heterogeneity of content types in a repository of multi-disciplinary online resources. Such content is prone to continual changes in terms of textual content(full-text) as well as associated internal and external content rendered on the local and remote platforms.

For example a page having content on *data collection* should also be a relevant page when query terms include *survey and questionnaire design* as well as *mail surveys, email surveys*, and *web-based questionnaire.* Similarly, when *online data collection* is searched, pages having content on *Big data*, *mobile digital data*, *online communities*, *e-social science* and *online forums* should also be retrieved. *Construct validity* is another synonym which is associated with *evaluation of research* and is a sub concept of *Quality in qualitative research.* Another example would be *multilevel models* which is a contemporary synonym of a broader concept *small area estimation* whereas *Writing research blogs* is a sub concept of *Writing skills* derived from a broader concept *Research skills, communication and dissemination.* These terms are not discoverable by lexical or literal matching strategies. According to Fernández et al. (2011), search engines have experienced impressive enhancements in the last decade but information searching still relies on the keywords-based searching which falls short of meeting users' needs due to the insufficient description of content. Only a decade ago, many of the leading web search engines, were still mostly lexical i.e. they looked mostly for literal matches of the query words. Modern search engines go more and more in the direction of accepting a broader variety of queries, understanding them and providing the most appropriate answer in the most appropriate form (Bast et al., 2016). However, in order to understand users' complex queries, the search engines and search results retrieval systems have to be constantly evolving over time and amenable to changes taking place in various aspects of life. They include, but are not limited to, advancements in technology and its impact on everyday life, changes in terminologies and concepts, economic and cultural shifts due to global, political and geographical events etc. Part of the problem is the objective nature of knowledge retrieval from the web due to extreme variability of information(Levene, 2011) while the information needs are expressed subjectively. Similarly Wu et al. (2006a) terms the basic Web search as inadequate when it comes to finding contextually relevant information in web archives or collection of web sites like the ReStore repository[2]. Contextual relationship between content must be an essential component to search results retrieval in such repositories but it is often missing due to the full-text keywords-based searching. To address these issues, this thesis focuses on 3 main questions to see if the performance of search applications in web repositories can be improved, which are widely used by users, especially researchers in various scientific domains and disciplines e.g. Social & Human Sciences, Life Sciences, Web & Internet Sciences etc. These questions are:

1. Whether obsolescence in terms and concepts in online repositories of Social & Human Sciences could be addressed by periodic semantic augmentation of keywords for better searching?

2. Can a shift from domain-specific ontology-based annotation to distributed and wider data spaces like linked open-data(LoD)–based semantic annotation address the issues of entity, concepts and relations disambiguation, thus leading to better search results retrieval?

3. Can crowd-tagging and annotation (assigning semantic tags to relevant search results) be employed to address the issues of content heterogeneity, concepts

[2] ReStore is an online repository of web resources developed as part of Economic & Social Research Council (ESRC) funding. Average users per month accessing the repository ranges between 17,000 to 20,000 - Available at http://www.restore.ac.uk

obsolescence and semantic augmentation of content for sustainable relevance in search results?.

LoD-based semantic annotation refers to the exploitation of LoD by Named Entity Extraction (NEE) systems for annotating web documents. NEE is the process of identifying entities in web documents (web pages, PDF, Word files etc.) and linking them to relevant semantic resources in a single or multiple Knowledge Bases (KB). There are several approaches and tools that offer LoD-based NEE service such as OpenCalais[3], DBPedia Spotlight[4] and Alchemy API[5] or it's new successor service called IBM Watson Natural Language Understanding APIs[6] etc. These and other tools have been further analysed in terms of performance and suitability, in this thesis, in Chapter 3 and Chapter 4. The term *"crowd"* here and in the entire thesis refers to the online users community of multi-disciplinary web repositories, who have genuine interest in searching, browsing and consuming information in these repositories, to meet specific information needs. *Crowdsourcing*-based annotation or *crowd-annotation* therefore, are defined in this thesis, as annotation and tagging activities in web repositories, aimed at adding extra semantics to the content for better search results retrieval.

1.1 Economic & Social Research Council (ESRC) research initiatives and repositories

In 2004, ESRC set up the National Centre for Research Methods (NCRM) initiative at the University of Southampton, UK. NCRM aims to increase the quality and range of methodological approaches and techniques used by the UK social scientists, across all sectors and career stages, through training, capacity building and methodological innovation research. The centre also offers rich online resources on social sciences research methods, which can be accessed online through the centre's main repository website as well as through the sub investments projects that NCRM funds during the tenure of its funding award. The dissemination and engagement of the centre's research is carried out through the biennial research methods festival, the centre website, the ReStore repository, digital media, conferences, dissemination events, summer schools and publications. Like NCRM, the Administrative Data Research Network (ADRN) is a UK-wide partnership

[3] http://www.opencalais.com

[4] http://demo.dbpedia-spotlight.org/

[5] This API has been retired in March 2018 and the equivalent of Alchemy API is now IBM Watson API called *Natural Language Understanding* which can be accessed at https://www.ibm.com/watson/services/natural-language-understanding-3/

[6] https://www.ibm.com/watson/services/natural-language-understanding-3/

[6] http://www.restore.ac.uk

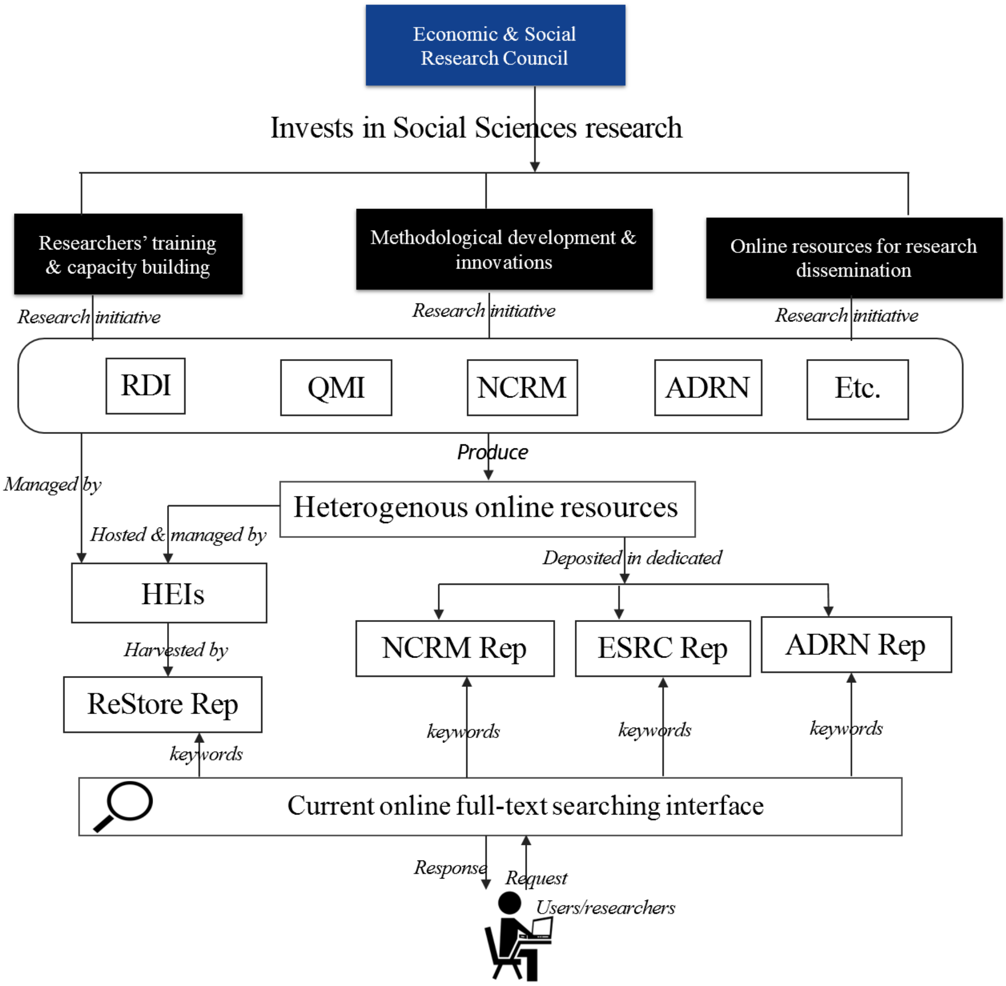

Economic & Social Research Council
Invests in Social Sciences research
Researchers' training & capacity building
Methodological development & innovations
Online resources for research dissemination
Research initiative
Research initiative
Research initiative
RDI
QMI
NCRM
ADRN
Etc.
Produce
Managed by
Heterogenous online resources
Hosted & managed by
HEIs
Deposited in dedicated
NCRM Rep
ESRC Rep
ADRN Rep
Harvested by
ReStore Rep
keywords
keywords
keywords
keywords
Current online full-text searching interface
Response
Request
Users/researchers

Figure 1.1: An overview of various ESRC research methods initiatives leading to the creation of various online repositories (Rep) and interaction of online users with them through the full-text search

between universities, government departments and agencies, national statistics authorities, the third sector, funders and researchers. ADRN is funded by ESRC until September 2018 and it coordinates four Administrative Data Research Centres, all at different locations across the UK. ADRN has a dedicated repository for online resources created by the various sub research projects funded by ADRN as illustrated in Figure 1.1. Research Methods Programme (RMP), Researcher Development Initiative (RDI) and Quantitative Methods Initiative (QMI) are others research initiatives funded by the ESRC, which don't have dedicated repository service and their online resources have been migrated to the ReStore repository service. The funding awards for all these programmes have ended and their research outputs are available in the form of various online resources in the ReStore repository. The principle aim of RMP was to develop qualitative and quantitative methods within the context of substantive research. Specific objectives included:

the development of new methodological tools; the encouragement of new initiatives in methodological training; and the improvement of methodological practice. Similarly RDI initiative supported the training and development of researchers in the social sciences at all stages of their career. RDI's sub projects included events, training and development of new web resources for researchers' training. Some of the RDI web resources are archived in the ReStore repository, which is now hosted and maintained by the NCRM IT team. QMI, on the other hand, aims to create a national training infrastructure, which builds quantitative skills development and create a flexible framework to meet the particular skills requirements of social science researchers. The research outputs from QMI are disseminated and deposited in a dedicated web resource[7], which is managed by NCRM. Further details on these ESRC-funded research projects have been given in Table A.1 in Appendix A.

1.1.1 Key stakeholders

ESRC supports independent, high quality research which has an impact on business, the public sector and civil society. The key audiences and stakeholders of the ESRC-funded projects include national and international academic and non-academic audiences and stake holders. The academic audiences further include researchers in higher education institutions (HEIs), including PhD students, research centres and think tanks etc. The non-academic audiences and stake holders include central and local government, publicly funded non-governmental bodies, third sector organisations, and private sector organisations.

1.1.2 Social Science Research repositories

In order to investigate inadequacy of keywords-based search results retrieval in a domain-specific perspective, it is important to explore the need for having multi-disciplinary web repositories and their usage in a particular domain of interest e.g. Social & Human Sciences, Web & Internet Sciences and Life Sciences etc. ESRC invests heavily in Social Science research methods projects which create, as part of their activities, online training and resource materials, often with considerable interactive or reference-value content for the UK social scientists across all sectors and career stages. Typically, the development of an on-line resource is time-consuming and expensive and the full value of the resource only comes into play close to the point at which funding ends. To ensure that the resource remains available online for a considerable period of time, ESRC has been funding various

[7] http://www.quantitativemethods.ac.uk/

online repositories initiatives which aim to maintain and host such online resources for maximum return on their initial investment. Such valued repositories of online resources collect and harvest online resources under a specific web address but do not provide state of the art structured or semantic search platform which could be used by the research community and other stake holders to derive maximum values from the existence of such repositories.

This thesis proposes semantic structuring (incorporating automatic and crowdsourcing-based methods) of content for better search results retrieval, as a sustainable process in the ReStore repository, which is one of the ESRC-funded repositories of online resources[8]. Figure 1.1 illustrates various ESRC's research initiatives and the resultant repositories, available for online full-text searching including the ReStore repository. All repositories typically contains web resources in the field of research methods e.g. web pages, workshop materials (PDF, Doc, Excel files), tutorial materials, audio/video files, methodology papers and presentations, sample data sets, etc.

1.1.2.1 Searching in web repositories

The primary goal of any searching or retrieval system is to structure information so that it is useful for people in finding desired and relevant information effectively and efficiently. Current searching techniques in discipline-specific or multi-disciplinary repositories[9] predominantly use keyword instances in web documents where they rely on the presence of keywords and phrases in those documents. Furthermore, issues arise when *users or researchers* struggle to filter out irrelevant information to support the ongoing research projects. The time factor (with the passage of time) further changes the meanings of various concepts, terminologies thus making it difficult for search engines to serve online users by producing research-specific search results. In the current era of information abundance, "keywords context" has become very important to a *researcher user*, searching for specific content on a particular research topic. Researchers in Social & Human Sciences, Arts & Humanities, Biological and Environmental Sciences, produce data or research outputs at the end of research projects, which are published as a standalone repository of content; only findable through Google search or internal full-text search application. As shown in Fig. 1.1, these web resources are created and published to the Web by researchers as part of funded research projects in various

[8] This award was a continuation of the *Sustaining Online Resources in Research Methods* award, which created the prototype ReStore repository for online research methods resources- Available at https://www.researchcatalogue.esrc.ac.uk/grants/RES-576-25-0023/read

[9] Multidisciplinary repositories contain data or research outputs that are produced by researchers within disciplinary domains e.g. Social & Human Sciences, Web & Internet Sciences etc.

disciplines in order to increase the value of their research outputs. Figure 1.1 also illustrates a typical web repositories creation process, which may involve research funding bodies, multi-disciplinary team of researchers, HEIs and publication of research outputs in a dedicated online space either provided by the hosting institution or website hosting company. The users of such repositories (*according to our website survey in 2011 and 2013*) are predominantly research students and fellows, academics, industry professionals and even funding bodies. Figure 1.1 elicits the process flow starting from a funding research initiatives or projects in HEIs; particular research groups work on the project (typically for 3-6 years) and publish research outputs (an online resource in the form of a website) usually on the hosting institution's website. However, following the cessation of funding award, the web resource ends up later on in a Web repository, which hosts it for long term online access. Figure 1.1 also indicates that the pool of web resources created out of specific funding initiatives focus on specific research disciplines. Hence, while working on any similar or related disciplinary project in the future, a researcher would ideally want to search over the existing online resources using natural language keywords and expecting contextually and semantically relevant search results. For example, a user enters *"Randomised control trials"* in the *Google search box* to find relevant materials on this topic and related topics like *Experimental design*, *Laboratory studies* or *Experimental research*. Almost all of the top 10 search results (a) contain the term *randomised control trials* with no presence (at least in the highlighted text) of the above-mentioned related terms and (b) all of the results are somehow related to the *medicine* discipline. The related topics have been specifically mentioned above because they are part of a classification system called NCRM typology[10] , which is widely used for labelling the research outputs in all of the above-mentioned research initiatives. These research initiatives and projects use the typology to categorise their research activities, training events, journal articles, working papers and presentation etc. Such formal structuring of content not only caters to the continually evolving programmes of research methods investments by ESRC but also act as the major source of online searching and knowledge discovery. It is therefore vital to consider the typology terms in search results retrieval for maximum user satisfaction. A searcher in a scientific domain would ideally expect the top 10 search results to have a contextual linkage among them based on such classification and structure in order to inform their specific research projects. Based on the above scenario, when similar terms are used in the search box of the NCRM website [11], results are retrieved having titles

[10] National Centre for Research Methods typology is extensively used in the UK HEIs for the classification of social science research outputs. I as a member of the NCRM team, participated in the up-gradation of the typology completed in 2015. Available at http://eprints.ncrm.ac.uk/3721/

[11] NCRM website hosts various Social Science-related content created out of various specific research projects funded by various research councils in the UK. Available at https://www.ncrm.ac.uk

like *Experimental and quasi-experimental design, Impact Evaluation Methods:use of randomised control trials in social research, the use of RCTs to test interventions and measure* etc. To a Social Scientist or someone having similar interests, these titles would appear more appealing than those based on mere definitions of *randomised control trials*. The basis for such scenarios will be further elaborated in the *Methodology* (4) and *Evaluation* (8) chapters of this thesis.

1.1.3 Analysis of web content

Along with the issues, highlighted so far in this chapter, another problem is the analysis of web content for accurate and contextual retrieval, which demands sufficiently accurate representation of the content in web documents inside a web repository. A web page for example may have three types of content: (a) local and static textual content (b) dynamic content fetched from a remote database server and (c) content in an *iframe* container embedded in a web page by a third party component. A holistic approach is therefore needed to address the issue of scalability at the time of indexing and storage of the content (bits). The author understands that such an approach would also facilitate universal access to all content and improve the capacity of retrieval systems, leading to more efficient ranking of relevant search results at the time of searching. Alongside a robust annotation system, a search retrieval system is also needed which can inform about the existence and location of a document, which contains the desired information. A perfect retrieval system would retrieve the relevant documents and not the irrelevant documents, however such a perfect system does not presently exist. This is in part because search statements are necessarily incomplete and relevance depends on the subjective opinion of the user (Hiemstra, 2009). In order to know the subjective opinion of users, they have to be involved both at the text analysis and retrieval stages in order to sustain the contextual linkage amongst various web documents over time. This thesis further discusses annotation and retrieval models along with the author's own implementation strategy in Chapters 4 and 5.

1.2 The methodology

This thesis investigates the incorporation of automatic semantic annotation of content in web repositories (using Linked Open Data or LoD sources as knowledge resource) without creating domain-specific ontologies, to assess the performance of online search systems in terms of retrieving highly relevant search results. This thesis also examines that by expert crowd-annotation of content on top of automatic semantic annotation, the semantic index can be enriched over time to augment the contextual value of content in web repositories; thus ensuring the content remain *findable* despite changes in language, terminology and scientific concepts. Figure 1.2 presents an overview of the proposed annotation, indexing and searching framework, which aims to address the gap between syntax and context at the time of searching. The figure also elicits the annotation of web documents in a specific domain of interests along with query formulation and searching processes. The figure further explains that the *search-tier* starts from *information need* and the *semantic indexing* is a parallel process which has been further subdivided into two sub processes *automatic annotation* and *manual annotation*. The *retrieved documents* component in the figure highlights the integration of *annotation* and *searching* with a view to assess the need for further alignment between the two processes. *User evaluation* finally determine whether further alignment is needed to optimise the retrieval of relevant search results. Furthermore, Figure 1.2 shows

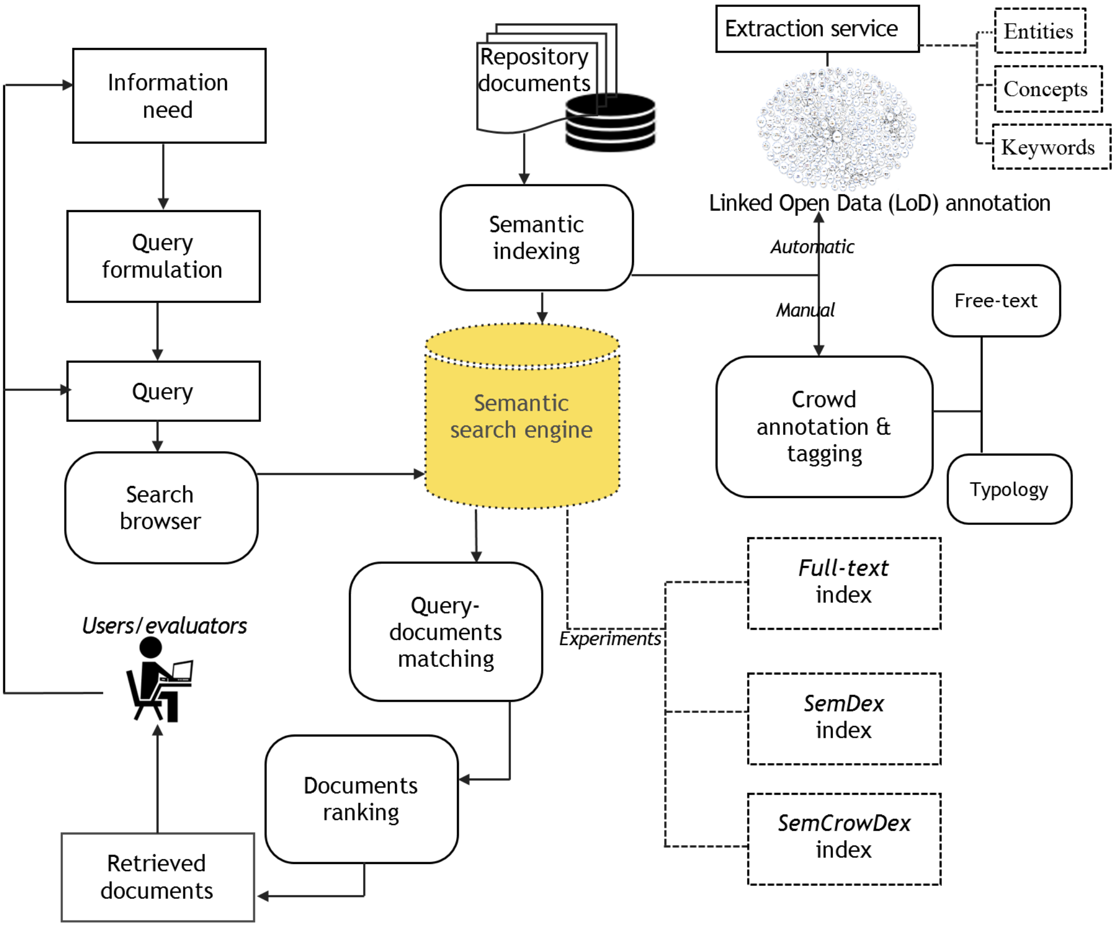


Figure 1.2: An overview of the proposed methodology showing various components: starting with information needs, query formulation, semantic annotation, indexing and search results retrieval

3 experiments to be run against 3 different indices, all packaged into a semantic search engine, in order to assess the system's performance in terms of relevant search results retrieval. Crowd annotation & tagging in Figure 1.2 refers to expert

and non-expert crowd-annotators, which are two categories of annotators, who participated in various experiments as part of this research. However, *experts* refer to those annotators who have a certain degree of command on the subject in question and have proper understating of a particular scientific field of interest e.g. a final year PhD student, post-doc researcher, lecturer, academic, etc. Non-expert annotators, on the other hand, refer to any online user including non-researcher students, industrial users, etc., having a particular interest in a collection of online resources in a repository or group of repositories. Both participant groups use free-text and typology classification terms to annotate and tag content as part of various experiments, which have been detailed in Chapter 6

The methodology adopted here broadly involves (a) implementation of an annotation, indexing and searching framework for contemporary searching in repositories of scientific data (b) automatic and manual annotation of content in the ReStore repository (c) the deployment of a purpose built search engine called Elasticsearch [12] for evaluating search results broadly in three situations (1) Keywords-based search results evaluation (using Full-text index) (2) Keywords, Concepts, Entities-based search evaluation (using SemDex index) and (3) Social Science-specific typology-based search results evaluation (using SemCrowDex index). The technical deployment for various annotations and indexing approaches is also discussed in the context of ReStore repository; while considering appropriate evaluation benchmarks, at the time of searching and evaluation.

1.3 Research Questions (RQs)

Scholarly research in any domain extends knowledge and reconceptualises the researchers' understanding of the world. Such contemporary evolution and formalisation of knowledge in a particular research domain needs to be aligned with the community or institutional process of knowledge acquisition in order to maximise its impact on future research activities in terms of accurate knowledge classification and retrieval. Evolving the consensus between formal classification and community-based knowledge formalisation takes time before an agreement is gradually evolved. Such consensus leads to the adoption of a dynamic classification system for knowledge discovery and online searching. It is, however, important that before the research activities are classified, and their scholarly impact is agreed and codified, multidisciplinary scholarly archives and repositories are needed to capture and make accessible the outcomes of various research initiatives and investments. As outlined in the previous sections, such multi-disciplinary repositories in the

[12] Elasticsearch is a flexible and powerful open source, distributed, real-time search and analytics engine. available at http://www.elasticsearch.org

Social Science domain exist and have been in use for showcasing a plethora of research outputs: online resources, training materials and sample data sets etc. as part of several research initiatives. However, such repositories are characterised by:

- Heterogeneous contents, covering wide variety of subject domains and purposes
- Highly technical contents, with a significant depth of knowledge in specific areas
- Complex contents that consist of interwoven reports, presentations, data sets, evaluations
- The constituent parts of a repository used for discovery - its interlinked contents, its classification schemes and the mark-up between the schemes and contents are subject to change and to re-interpretation over time.

This thesis addresses the challenges, inherent in making the contents of multi-disciplinary repositories findable, and focuses on the following research questions.
RQ1: Can accuracy and relevance in search results be improved by employing automatic semantic annotations?
RQ2: Can the relevance of search results be further improved by periodically adding contemporary semantic annotations?
RQ3: Can the relevance of search results be further improved by expert and non-expert crowd-sourced semantic annotations & tagging?

1.3.1 Contribution of this thesis

Based on the knowledge gained through working in this research area, no research initiative has been found which integrates all the 3 components: annotation, indexing processes with search results evaluation in real time aimed at sustainable relevance in search results. The contributions of this thesis to LoD and crowd-annotation based information retrieval in web repositories include:

(a) The design and development of a full-fledged semantic annotation platform capable of importing content from multiple sources e.g. crawlers, RDBMS and automatically linking keywords, concepts, entities in the content to LoD sources (*Named Entity and Concepts extraction*).

(b) Development of a system to ascertain improvement and sustainable accuracy and relevance in search results in multi-disciplinary web repositories by employing automatic LoD-based semantic annotation methods as well as crowdsourcing-based tagging techniques

(c) An adaptable semantic search index enriched by expert crowd-sourced annotators through annotation and tagging of content in web repositories.

(d) Development of a complete search-application and KB built on the Elasticsearch platform (*in a client-server environment*) aimed at evaluating search results retrieved against users' natural text queries

To materialise the above, the custom-built semantic annotation framework has been extensively used to mass-annotate all types of content in the ReStore web repository. The actual content and annotation metadata are indexed and stored in the dedicated Elasticsearch distributed search. The deployment of online searching application, SRR (Search Results Retrieval) system, enables us to evaluate the search results in two different situations i.e.

- *Hypothesis 1:* Search results relevance and ranking improve when user queries are searched against 1. Full text index 2. Semantic index (*SemDex*)
- *Hypothesis 2:* Search results relevance and ranking further improve when user queries are searched against 1. SemDex (LoD-based semantic annotation) 2. Crowd-sourced semantic index (*SemCrowDex*)

As shown in Figure 1.2, three distinct experiments have been conducted i.e. searching performed by the expert evaluators using (1) full-text or inverted index and *SemDex* index (2) expert crowd annotation of content in the ReStore repository to build a crowd-sourced semantic annotation layer on top of *SemDex* and evolve it into *SemCrowDex* and (3) Search results evaluation of results produced by our search engine from two indices i.e. *SemDex* and *SemCrowdex*. All the three experiments are thoroughly discussed namely Experiment A (Exp.A), Experiment B (Exp.B) and Experiment C (Exp.C) in Chapters 4 and 6. By conducting these experiments, the author has:

- Compared the ranking of search results retrieved against two sets of benchmark queries i.e. 20 and 33 in Chapter 7 and Chapter 8 respectively. Both sets of queries have been given in Table A.3 and Table A.7.
- Evaluated the relevance of top 10 search results against each query from both sample queries set.
- Assessed the level of agreement between the expert annotators and crowd-sourced annotators & taggers based on the amount of annotation and tagging; they performed in similar and dissimilar documents in the ReStore repository.

These contributions have led to the publication of the following peer reviewed papers:

1. Khan. A., Tiropanis. T and Martin. D.,*Paper published*-Crowd-annotation and LoD-based semantic indexing of content in multi-disciplinary web repositories to improve search results, *In proceedings of the AWC 2017:The Australasian Web Conference 2017-January 31-February 3,2016*
2. Khan. A. and Tiropanis T. and Martin. D., *Paper published*-Exploiting Semantic Annotation of Content with Linked Open Data (LoD) to Improve Searching Performance in Web Repositories of Multi-disciplinary Research Data, *In proceedings of the 9th Russian Summer School in Information Retrieval (RuSSIR 2015)*, Springer International Publishing
3. Khan. A., Tiropanis. T and Martin. D.,*Using Semantic Indexing to Improve Searching Performance in Web Archives*,2013, *In proceedings of the First International Conference on Building and Exploring Web Based Environments (WEB2013), 27 Jan - 01 Feb, 2013*, **(Awarded Best Paper Award)**

A custom-built crowd and expert annotation, indexing and searching environment is displayed in a web repository website and has been used by expert and crowd annotators to annotate web pages using free text and vocabulary terms. The terms vocabulary and NCRM Typology have been interchangeably used to refer to the classification system, that was originally produced by the NCRM researchers aiming to classify research outputs from various ESRC funded projects in the UK.

1.3.2 Typology-based annotation and tagging

During the course of this PhD, the author worked alongside academic social scientists and library sciences professionals to upgrade the typology, which has been extensively used, in the classification of social science research outputs in the UK. The typology upgrade was completed after six months' review in January 2015[13]. However, the upgrade was not performed as part of this PhD research. The new version of the typology has been thoroughly deployed in the annotation and retrieval framework to assess the effectiveness of typology-based annotation and tagging vis-a-vis free-text tagging. The outcome of that latest research has been compiled into a paper in preparation at the time of thesis submission.

[13] The classification hierarchy and paper is available at http://eprints.ncrm.ac.uk/3721/

1.3.3 Thesis structure

The rest of this thesis is structured as follow: In **Chapter 2**, the background of the problem is discussed, which motivated this research in the first place. That includes but is not limited to (a) exploring domain specific classification systems e.g. disciplinary taxonomies, typologies and their potential role in building sustainable information retrieval systems (b) obsolescence in terms and concepts affecting relevance in search results and (c) ascertainment of the role of a discipline-specific research community (as crowd-annotators and taggers) for annotating and tagging web content.

Chapter 3 provides a review of relevant literature, addressing the issues relating to the ontology-based semantic annotation, crowdsourcing-based annotation and crowdsourcing frameworks. The chapter also sheds light on the social media platforms used for annotation and classification of content aimed at better information retrieval.

Chapter 4 elaborates the process of content annotation in repositories of heterogeneous content and methodologies to cater to the issues of content heterogeneity and concepts obsolescence, affecting the retrieval of relevant search results. Furthermore, this chapter broadly discusses the overall *modus operandi* for conducting various experiments and evaluation exercises and the justification for developing particular technical infrastructure as part of this research.

Chapter 5 describes the building blocks of a complete annotation, indexing and retrieval framework aimed at sustaining the relevance in search results through automatic and crowdsourcing-based annotation and tagging. The chapter further explains, how to implement a complete information retrieval system based on LoD-based semantic annotations and users' tagging.

Chapter 6 elaborates the experimentation process, the infrastructure that has been built for different types of experiments along with different benchmarks. There are 3 experiments which are described in this chapter: (a) Search results retrieval against Lexical vs. Semantic index (Exp.A) (b) Expert Crowd-sourced annotation and tagging of heterogeneous content in the ReStore repository (Exp.B) and (c) Search results retrieval against *SemDex* vs. *SemCroDex* and evaluation of results with discussion and lessons learnt (Exp.C). Furthermore, the recruitment of participants for all the 3 experiments has also been detailed in this chapter.

Chapter 7 evaluates and discusses the performance of SRR system, based on users' participation in Experiment A (6.4). This chapter also elaborates the impact of automatic semantic annotation in web repositories in terms of most relevant search results retrieval and optimum users satisfaction.

Chapter 8 presents analysis of results from various experiments detailed in Chapter 6. The performance of SRR system has been figuratively detailed in this chapter in terms of relevant and highly ranked search results specifically taking into account the crowd element in Exp.B (6.5). This chapter also presents the analysis of participants' feedback in terms of crowd-sourced semantic annotation vis-a-vis automatic semantic annotation to ascertain the efficacy of annotation and tagging in terms of sustainable, highly relevant search results.

Finally **Chapter 9** concludes this thesis by summarising the amount of work, the author has done to answer the research questions. This chapter also describes the prospects of web-scale LoD-based annotations and crowd-sourced annotation, required to sustain the performance of search engines over time. Furthermore, the chapter also highlights the possibility of *RDFising* the storage of automatic annotation as well as crowd-sourced annotation for interoperable search results retrieval in a distributed cross-platform searching environment.

Chapter 2

Knowledge representation, organization and retrieval in web repositories: *Background & Motivation*

This chapter reviews the methods and techniques used for web content annotation, Knowledge Organisation Systems (KOS) used to represent knowledge in web documents and various Information Retrieval approaches used for search results retrieval. It draws on the analysis of various KOSs and their usage specifically in web repositories for annotation and retrieval purposes with a view to highlighting their limitations and strengths. The use of typologies, taxonomies and ontologies in modelling the searchable knowledge bases (KB) is investigated for sustainable search results retrieval. A discipline-specific typology i.e. NCRM typology is also presented as a case study in the context of a classification system and as a source of crowdsourcing-based annotation in web repositories. Moreover, discussions are held to ascertain the role of contemporary disciplinary research communities in augmenting the automatically generated KBs in the face of fast changing informational and digital landscapes. The augmentation entails experts' and crowd annotation of content in web repositories on top of the automatic semantic annotation of heterogeneous web content using one or more KOSs.

2.1 The motivation

The motivation for this research is driven by the potential of automatic annotation techniques as well as crowd-sourcing-based annotation in the multi-disciplinary web repositories for sustainable search results retrieval due to changes in language, technology, cultures, environment, geography, politics and natural phenomena. In addition, it is important to look at these areas in the context of inter-disciplinary and multi-disciplinary research where a research community uses online searching as a primary tool for searching. They engage in online searching using the mainstream search engines as well as enterprise searching platforms depending upon their information needs.

The nature of contemporary Internet-based research is cross disciplinary in that researchers come from different academic backgrounds and approach puzzles with a multitude of assumptions (Karpf, 2012). Karpf (2012) elaborates that underlying concepts such as "The Blogosphere" are not fixed in place and they change even while we study them. Shirkey (2013) wrote that "at some point, *weblog technology* will be seen as a platform for so many forms of publishing, filtering, aggregation and syndication that blogging will stop referring to any particularly coherent activity". Commenting on the above, Karpf (2012) concludes that Shirkey (2013) still referring to blogs as "weblog technology" in 2003 i.e. blogging was still new enough that writers had to explain that blog is short for "weblog". Similarly another example would be that of *"Applied psychometric"* which is associated with diagnosis of psychiatric depression and learning difficulty. But at the same time, it could also be interpreted as *"psychometric test"* aimed at staff recruitment and development. Similarly in terms of different scientific disciplines, different academics in these disciplines attach a wide range of meanings and interpretations to the terminology of research (Grix, 2002).

This chapter focuses more on the background of the above in the context of this research. Moreover, the recent related work done in the areas of semantic annotation, indexing, crowd-sourcing-based annotation and search results retrieval with evaluation methods are discussed in the next chapter (Chapter 3).

2.2 Information Retrieval (IR) systems

Before going further, it is important to explain the key steps in a typical information retrieval process, which is one of the main components of this research. The terms, Information Retrieval (IR), Semantic Web (SW) and Ontology are used differently but they are interconnected with each other. IR technology and web based Indexing contribute to the existence of SW which in turn provides a platform for understanding vast amount of web documents and retrieving knowledge out of it. Sanderson et al. (2010) describes Information Retrieval (IR) as, "*.."finding material (usually documents) of an unstructured nature (usually text)* *that satisfies an information need from within large collections (usually stored on computers).."*.

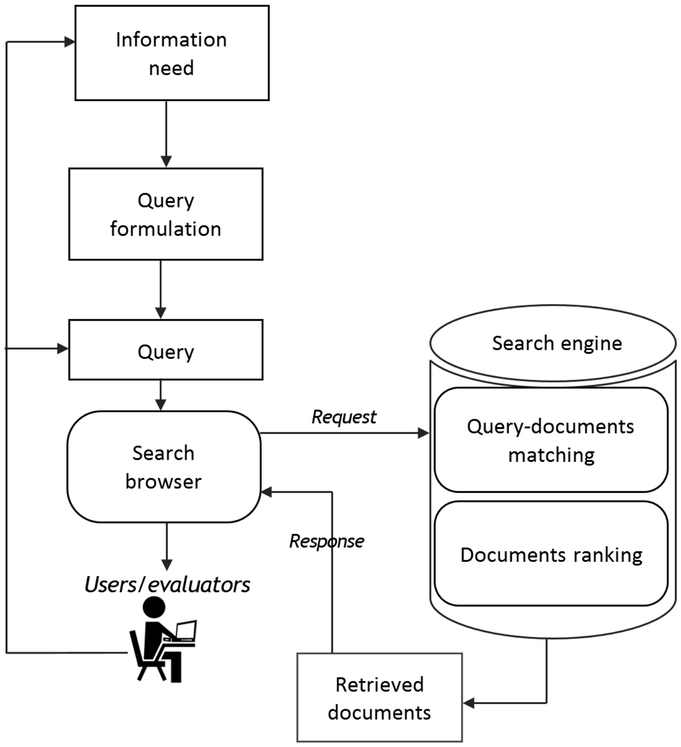


Figure 2.1: An overview of IR architecture and components: Retrieval of documents and assessment of retrieved results/documents based on query-document similarity and fulfilment of user's information needs

In simple terms, Weiss et al. (2010) elaborate the task of Information Retrieval (IR) as the retrieval of relevant documents in response to a user query. Figure 2.1 has been sliced off from Figure 1.2 and it sums up the explanation, given by Weiss et al. (2010), highlighting the objectives of query-based IR; (a) a general description is given in the form of a query (b) the document collection is searched, and (c) subsets of relevant documents are returned (in this case top 10 relevant documents). The basic building blocks of the information retrieval process has been illustrated in Figure 2.1, where the need is expressed in terms of query and the output is a number of results. Most of the search results retrieval models adhere to the standard information access model, which entails an interaction cycle comprising of query formulation, query submission and search results evaluation. The last step is either to stop further searching or reformulate/refine the query. This cycle continues repeatedly until the information need is satisfied and the desired result is retrieved. The standard information seeking model, however, can differ in some way from the standard model; as the specification of query formulation and searching in some search applications may adapt to the changing requirements in a certain domain. The retrieval search engine (whether semantic or full-text search engine), therefore doesn't influence the basic IR process, as illustrated in Figure 1.2 and Figure 2.1. Furthermore, Weiss et al. (2010) observe that the fundamental technique of information retrieval is measuring similarity between query and indexed documents. Terms in *query* and *document* are transformed into vectors of values and the values are compared at the time of ranking (query vector against all indexed document collection) for presenting most relevant search results to the user.

It is worthwhile and pertinent to mention here that the two terms i.e. IR and Search Results Retrieval (SRR) have been used interchangeably in this thesis to refer to the retrieval system in domain-specific web repositories and not in the web-scale searching. However, SRR particularly refers to the system the author has evolved, and which has been incorporated in the experimental and evaluation analysis in this thesis.

2.2.1 Information Retrieval (IR) architecture

Semantic searching-based IR requires a search engine to return documents against user query from a KB rather than an invested index based on similarity of keywords and concepts in the query to those of KBs. IR facilitates the retrieval of documents based on their tokens which are associated with either instances of the domain ontology or data sources in the LoD cloud. The *terms-document* relevance model is widely adopted due to the straightforwardness of terms frequency computation. In such a model, the words that occur in the title, author list, section headings are given more importance than those occurring later in the document or set of documents. Very common words such as "a", "an", "the", "it" etc. which are collectively called *stop words* are eliminated during the indexing of documents before the content and metadata are stored for information retrieval. Proximity is another known feature in such predominantly Boolean retrieval models which entails that if keywords in query occur close together in the document, the document has a higher importance than if they occur far apart.

The final element is the documents retrieval, matching the terms in the query, in decreasing order of relevance score at the time of ranking of relevant documents by the search engine. The selection of a particular retrieval model is driven by the specific information needs, domain of interest, structuring as well as indexing and storage process of the actual content for information retrieval purposes. Figure 2.1 clearly shows various components, joined together to construct a simple information retrieval system. The "information need" and "retrieved documents" components are the input and output respectively enabling the users where to start and end the search process. The *query formulation* helps the user to formulate queries representing their information needs using natural language keywords and phrases. The ranked list is then produced by the *search* process employing specific internal (such as *Lucene's*[1] search algorithms) or external query-based algorithms (e.g. *field boosting and Boolean filters etc.*). The ranked search results then either leads to the users clicking on a particular search result (depending upon the subjective information needs of the user) or restarting the whole process from the start i.e. formulating another query to maximise the satisfaction level and to meet the information needs.

2.2.2 Dynamicity of the IR architecture: *Berry-picking style*

In real life online searching, however, users may begin with just one feature of a broader topic, or just one relevant reference, and move through a variety of sources. Each new piece of information they encounter gives them new ideas and directions to follow and consequently, a new conception of the query. At each stage they may feel the need to submit the evolved query before reaching the end of the cycle in Figure 2.1. At each stage they are not just modifying the search terms used in order to get a better match for a single query, rather the query itself (as well as the search terms used) is continually shifting, in part or whole. This type of search is here called an evolving search (Bates, 1989). In other words, Bates (1989) explains, the query is satisfied not by a single final retrieved set, but by a series of selections of individual references and bits of information at each stage of the ever-modifying search. A *bit-at-a-time retrieval* of this sort is here called *berry-picking*. The *berry-picking* model suggests that interesting information is scattered like berries among bushes. The actual path of satisfying information needs starting from query formulation to getting the desired results transform the sketch in Figure 2.1 to more or less Figure 2.2. The *zigzag* path in Figure 2.2 is highly likely to be true in many retrieval models. The searchers' information needs are not satisfied by a single, final retrieved set of documents, but rather by a series of selections and bits of information found along the way (as shown in Figure 2.2 as documents icons). In other words, the continuous exploration of new information corresponding to the shifting query gradually satisfies the actual information needs. This in essence stands in contrast to the notion that the search process has to do with the retrieval of documents perfectly matching the original information needs. This element of *haphazard satisfaction of information needs* is addressed in experiments to establish proximity level between actual retrieved document and suggested topics/tags, discussed in Chapters 6 and 8.

[1] Apache Lucene is a high-performance, full-featured text search engine library written entirely in Java. It is a technology suitable for nearly any application that requires full-text search, especially cross-platform. Available at https://lucene.apache.org/core/

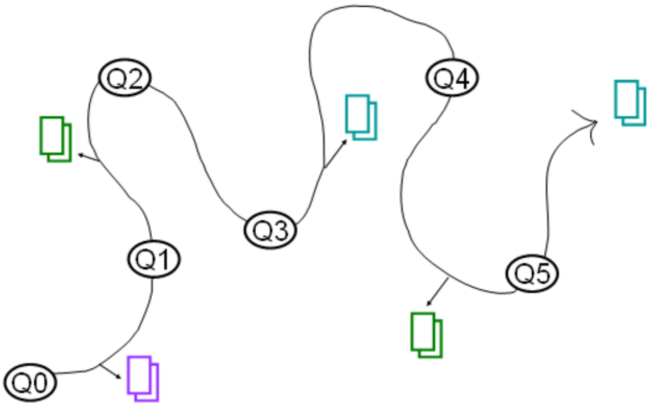


Figure 2.2: A IR path of an information seeker engaged in "berry-picking" style information seeking process, in which the query shifts as relevant information and documents are found along the way

2.2.3 Retrieval evaluation models

Various evaluation models are used by the research community to get the subset of relevant documents following the submission of a user query. The need for having a retrieval model in a generic or specific search environment is a must in order to assess the efficacy and soundness of a search results retrieval system. Hiemstra (2009) for example describes two reasons for having models for IR. The first, he elaborates, is that models guide research and provide the means for academic discussion and secondly models can serve as a blueprint to implement an actual retrieval system. Moreover, a model of IR predicts and explains what a user will find relevant given the user query. This component has been illustrated in Figure 2.1 as "*User/evaluators*". Without this component, in the IR process, it would not be possible to establish the performance of the retrieval system nor would one be able to measure users' satisfaction following his/her searching experience based on defined information needs.

2.2.4 Key processes

There are three basic processes an IR system has to support (Hiemstra, 2009) and which can be inferred from Figure 2.1:

- The representation or interpretation of the documents in a given document collection (e.g. a web repository or web archive etc.)

- The representation of the user's information needs, as illustrated in Figure 2.1.

- The comparison of the two representations i.e. *document* and *query* e.g. *Query-document matching* component in Figure 2.1.

There are various IR models which could be employed to assess the efficiency, accuracy and scalability of an IR system. Hiemstra (2009) have particularly mentioned *Exact match model*, *Boolean model*, *Vector Space Model (VSM)*, and *Probabilistic retrieval model*. However, the focus here is on the VSM (Vector Space Model), which is a quite effective retrieval model when it comes to web searching and when the evaluation is based on top 10 search results retrieved against a given set of benchmark queries. Moreover, the modification and conversion of this model to *Hybrid* VSM for achieving our specific research-driven objectives are explained in Chapters 5 and 8.

Chapter 3 addresses document representation or semantic annotation, various approaches being used for ontology-based and crowdsourcing-based annotations, semantic search and KB, a couple of case studies and usage of social media platforms for semantic tagging aimed at retrieving relevant search results.

2.2.5 Similarity based retrieval

Similarity-based retrieval of documents, given a query of terms, is another model widely used in the development of domain specific or web scale information systems. Similarity may be defined on the basis of common words e.g. find k terms in a collection of documents with highest *TF (d,t) / n(t)* and use these terms to find relevance of other documents. Document-to-documents and documents-to-query similarity is further discussed in the following sub sections; implementation and results evaluation are examined in Chapters 5 and 8 respectively.

2.2.6 Relevance feedback

Another similarity-based retrieval model is based on *relevance feedback* from online users. According to this, similarity in documents can be achieved based on users feedback in terms of initial selection of relevant documents from the result set. After having selected a few relevant documents from those retrieved by the first keyword query, the system then finds other documents similar to these. This retrieval model is closed in proximity to the crowdsourcing-based feedback but the former more or less relies on the post-query submission scenario after the results have been retrieved. The crowdsourcing based annotation or experts feedback leads to the classification (using controlled or uncontrolled vocabulary or even natural language free text) of content so that the existing full-text content are augmented with the extra semantic classificatory annotation before the query submission.

2.2.7 VSM-based retrieval: Document representation

The most popular approach for relevance ranking in IR is the so-called Vector Space Model (VSM). The underlying assumption is that documents are considered as objects which are characterized by the tokens appearing in them (Davies et al., 2008). The tokens represent a set of features of all individual documents and are indexed for search engines to match against user's queries before retrieving ranked search results. It is important that the degree of association is considered based on characteristics, so that some tokens are considered more characteristic for a document than the others. Davies et al. (2008) further note that the abstraction used to formally model these weight-based characterisation is a geometrical one.

All this means that documents are represented in a space, where each dimension is associated with a token based on various characteristics. Thus if there are 10 million different tokens appearing in the indexed documents, the space used as a model for relevance ranking will have 10 million dimensions. Each document is modelled as a vector, where the number of occurrences of a specific token in it is taken as a coordinate value for the corresponding dimension. The search engine or retrieval system represents the query as a vector in the same space and then takes the cosine of the angle between the query and the document as a relevance measure. In practice, IR engines implement quite a number of normalisation on top of the basic VSM in order to handle issues like very popular terms, documents of varying size, fields of special importance (e.g., title), etc.

The author further elaborates the utilization of VSM in the results evaluation and relevance calculating criteria based on the multiple characterisation of documents, as a result of crowdsourcing-based annotation in Chapters 5 and 8.

2.2.7.1 VSM: Similarity-based retrieval

This retrieval model defines an *n-dimensional space*, where n is the number of words in the document set and similarity is determined based on two vectors i.e. document vector and query vector. Vector for document d goes from origin to a point whose i^{th} coordinate is $TF(d,t) \,/\, n(t)$. The cosine of the angle between the document and query (*in document-query similarity*) or document A and document B (*in document-document similarity*) is used as a measure of their similarity.

2.2.8 Term Frequency-Inverse Document Frequency-*tf-idf*

Basic search is essentially what is often called full-text retrieval, which entails that all the words in each document (in tokenized form searchable via index) are potential keywords, retrievable against users' queries. Query expressions, amenable to the retrieval system, include not only the actual terms or words but also the connectives like *and*, *or* and *not* where *and* are usually implicit even if not explicitly specified. Like other retrieval systems, full text retrieval demands ranking of documents on the basis of estimated relevance to a query which is based on factors such as *Term Frequency (tf)*, *Inverse Document Frequency (idf)* or simply *tf-idf.*

f-idf, is a well known method to evaluate how important is a word in a document and an interesting way to convert the textual representation of information into a VSM. *tf* deals with the frequency of occurrence of a query keyword in a document *where* as the *idf* is concerned with the number of documents in which the *query keyword* occurs. *tf-idf* implies that the more frequent a "rare" term appears in a document, the greater chance that the document is more relevant about the topic. Rare here means that a term has low frequency in overall documents. Low *idf* frequency of a keyword means fewer instances of the keyword in the entire set of documents retrieved against the users' query hence more importance is given by the ranking process to the document in the result set and vice-versa. This technique will be further discussed in relation with the implementation methodology in Chapter 5 (Section 5.2.3) as part of implementation of the search results ranking methodology. The relevance ranking of terms is therefore given by:

$$TF(d,t) = log\left(1 + \frac{n(d,t)}{n(d)}\right) \tag{2.1}$$

where *n(d)* is the number of terms in the document *d*, *n(d,t)* is the number of occurrences of term t in the document d and *TF (d,t)* refers to the relevance of a document d to a term t. The log factor is to trade off the excessive weight assigned to frequent terms in the document or set of documents. Likewise, the relevance of a document to *query* Q is measured as:

$$r(d,Q) = \sum_{t \in Q} \frac{TF(d,t)}{n(t)} \tag{2.2}$$

where *n(t)* is the number of terms t in *query* Q.

2.2.9 *tf-idf* with cosine similarity

tf-idf provides a representation for a given term in a document or a set of document. Cosine similarity, on the other hand, gives a score for two different documents that share the same representation. Tf-idf is a transformation, applied to texts to get two real-valued vectors. One can then obtain the cosine similarity of any pair of vectors by taking their *dot product* and dividing that by the product of their norms. That yields the cosine of the angle between the vectors. In this thesis, *Cosine Similarity* and *tf-idf* have been extensively used to calculate the similarity between different documents, based on queries submitted by online users as part of various experiments.

2.2.10 Cosine similarity between query and documents

Let us assume one has to calculate similarity of two documents in a VSM, and to do that they need to convert each document to vectors, which can then be visualized in a vector space. Quantification of similarity between two document vectors and query vectors in a given vector space would be an arduous challenge due to the magnitude of the vector differences as two documents with very similar content may have significant vector difference simply because *documentA* is longer than *documentB*. In other words, the relative distribution of terms in two documents may be the same but the absolute term frequencies of *documentA* may be larger than *documentB*. The length of documents is determined by the amount of content it contains as well as semantic annotation as a structured metadata to add meanings to the content. Along with the core information in a document, there are stop words, acronyms, jargons and contemporary terms which the users in a particular domain may use to refer to various things in their domain of their interests. Stop words need to be defined to restrict them from being indexed and searched. Such arrangements address the document length issue from the outset which helps at the time of search results retrieval.

2.2.11 Cosine similarity vs. Dot Product (DP)

It is understood that cosine similarity only cares about angle difference, while *dot product* cares about angle and magnitude between two documents or query (q) and document (d). Sometimes it is desirable to ignore the magnitude or reduce its weightage, hence cosine similarity is preferred, but if magnitude plays a role, dot product remains a better option as a document-query similarity measure. DP for two vectors $\overrightarrow{a} = (a1, a2, a3, ...) and \overrightarrow{b} = (b1, b2, b3, ...)$ can be calculated as a simple multiplication of each component from both vectors added together. a_n and b_n in the vectors are the components of the vector (TF-IDF values for each word of the document) and the n value is the dimension of the vectors:

$$\overrightarrow{a} . \overrightarrow{b} = a_1 b_1 + a_2 b_2 + ... + a_n b_n$$

. The calculated value of a DP for two vectors is a singular value (not a vector), which is also referred to as a scalar.

2.2.12 Document-Document & Document-Query Cosine Similarity (CS) computation

Cos(q,d) or the *dot-product* measures the cosine of the angle between q and d but the problem with *dot-product* is that it is longer if document vector is longer in |V| dimensional vector space. In the hybrid semantic vector space, which is the core of this evaluation analysis, document-query similarity does not have to be sensitive to word frequency in the |V| dimension vector space.

2.2.12.1 Vector normalization

Normalizing refers to the process of making something "standard" or "normal." In the case of vectors, a standard vector has a length of 1. To normalize a vector, is to take a vector of any length and, keeping it pointing in the same direction, change its length to 1, or what is called a unit vector. Hence the normalized cosine similarity between *documentA* and *documentB* is given as:

$$Cos(\overrightarrow{docA}, \overrightarrow{docB}) = \frac{\sum_{i=1}^{|V|} docA_i docB_i}{\sqrt{\sum_{i=1}^{|V|} docA^2}\sqrt{\sum_{i=1}^{|V|} docB^2}} \tag{2.3}$$

The above equation 2.3 underscores the similarity based ranking on factors other than only the Boolean model which results in retrieving relevant documents with far higher accuracy. The similarity angle obtained between query and documents leads to document ranking i.e. the smaller the angle the more relevant the document. That way the relative distribution of terms is offset in a set of documents and the proximity and relevance to the query is computed accordingly. In other words, long and short documents' vectors now have comparable weights after new annotations were added automatically or by the crowd. Like the *document-document*

CS, the CS between query and document is computed using:

$$Cos(\overrightarrow{q}, \overrightarrow{d}) = \frac{\overrightarrow{q}, \overrightarrow{d}}{|\overrightarrow{q}||\overrightarrow{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}} \tag{2.4}$$

where q_i is the TF-IDF score of term i in the query vector and d_i is the the TF.IDF score of term i in the document vector. $|\vec{q}|$ and $|\vec{d}|$ here are the lengths of $\vec{q}$ and $\vec{d}$ respectively. So the normalized vector in the semantic vector space model would be equivalent to the dot product only if $|\vec{q}|$ and $|\vec{d}|$ are length normalized i.e.

$$Cos(\vec{q}, \vec{d}) = \sum_{i=1}^{|V|} q_i d_i \tag{2.5}$$

2.3 Web searching: basic vs. structured

Current search engines are increasingly struggling to help the user in the tasks that go under the umbrella of exploratory search. Here, the user needs not only to perform a look up operation but also to discover, understand and learn novel contents on complex topics while searching Mirizzia et al. (2010). According to Fernández et al. (2011) search engines have experienced impressive enhancements in the last decade but information searching still relies on keywords-based searching which falls short of meeting users' needs and defining content meaning. Similarly Wu et al. (2006a) term the basic web search as inadequate when it comes to finding contextually relevant information in web archives. Despite achieving efficient results for common queries, keywords-based search has exhibited limitations particularly in dealing with more complex queries (Lashkari et al., 2017). Web search can be defined as an information-seeking process, conducted through an interactive interface that involves searching over heterogeneous Web resources, either directly or via an information portal (Gal et al., 2003) depending on the requirements of an online community. In addition, Wu et al. (2006a) term the basic web search as inadequate when it comes to finding contextually relevant information in web archives or repository of web sites like ReStore [2] and NCRM (National Centre for Research Methods)[3] etc. In an analysis of the query log from an academic searching engine *SemanticScholar.org*, Xiong et al. (2017) note the inability of the

[2] ReStore is an online repository of web resources developed as part of Economic & Social Research (ESRC) council

[3] https://www.ncrm.ac.uk

term-frequency-based search engine to understand the meaning of research concepts in queries e.g. whether a query is about a particular concept in Computer Science or a research topic.

Relationships between content should be an essential component to search results retrieval in scientific repositories, but it is often missing due to the full-text keywords-based searching. In order to structure the web content through semantic annotation, methods and processes are needed, which could be applied on the existing web without too many complexities to enable users retrieve relevant search results. Some problems are being discussed in the following, which are still commonplace when it comes to semantic web searching research. As illustrated in Figure 1.2 and Figure 2.1, a typical searching process starts with query formulation, followed by its submission to a search engine, which retrieves a list of ranked search results. Online users enter keywords in a search box with the expectations of finding relevant documents or results presented in a ranked order ready to be explored for meeting their information needs. The keyword or keywords themselves is a type of document or *information need* which keeps changing in the form of new queries searched against the document index in order to retrieve relevant or similar documents. A measurement process takes places inside the search engine which computes the relative similarity between the new documents (*query*) and all the *documents* in the document collection. However, the objective of the search engine is to rank the documents and not to assign label or annotations to the content in the document collection.

Structured search, on the other hand, takes into account another layer of semantic description of the first layer i.e. full-text bit or actual content. The semantic index or *SemDex*, enables the search algorithm in the search engine to perform retrieval of results based on the keyword mentions as well as their contextual meanings. The search in that case, not only preforms the retrieval process based on the explicit presence of keywords in documents, but also their semantic and contextual meanings along with individual relevance score.

Different criteria and corresponding approaches have been adopted over the years for different semantic search classification systems. Fernández et al. (2011) have presented these in Table 2.1. Analysis of all of the approaches in Table 2.1 in comparison with the aim and objectives of this research is beyond the scope of this thesis. However, such classification of semantic search classification will help us to understand and analyse the related classification systems under the criteria set out for various approaches in Table 2.1. This thesis evaluates and discusses semantic searching in the light of these criteria focusing on *web searching in domain repositories, natural language queries, information retrieval, keywords-based ranking and semantic-based ranking* approaches.

Table 2.1: Semantic search systems classification [(Fernández et al., 2011)]

Criterion	Approaches
Semantic knowledge representation	Statistical Linguistic conceptualization Ontology-based
Scope	Web search Limited domain repositories Desktop search
Query	Keyword query Natural language query Controlled natural language query Structured query based on ontology query languages
Content retrieved	Data retrieval Information retrieval
Content ranking	No ranking Keyword-based ranking Semantic-based ranking

2.4 Users' searching criteria

Information retrieval, in the context of web searching, is often an iterative process where the user refines his/her query to focus on highly relevant documents. But this process implies that the end-user has a precise understanding of the results proposed by the search engine and that interaction techniques allow him/her to reformulate the query, select interesting documents and give some hints to the system about his/her application needs. Visualisation techniques may also be considered as key components of this process since they play a mediating role in this understanding (Ranwez et al., 2013). This phenomenon has been illustrated in Figure 2.2 (*Berry-picking style searching*). However, query processing by a semantic search system, may have its own limitations. For example, in some ontology-based semantic search systems, users are expected to use formal query languages to express their information needs; in order to enhance the conceptual representation of user queries beyond plain keywords [Fernández et al. (2011); Fatima et al. (2014)]. Such interpretation of user queries, based on a limited domain-specific ontology, is likely to affect the *query-document* relevance, especially, when the ontology is not populated with new concepts and their descriptors at the time of searching.

Moreover, Figure 2.3 shows the *congruity gap* emanating from the *creation* and *consumption* of information by the *authors* and *users* respectively. The lesser the gap, the better would be the quality of search results retrieval and optimal user satisfaction. Refining the search results retrieval process and search results retrieval performance is, therefore, subject to optimum congruity between *information seeker* and *authors or knowledge producer*, as demonstrated in Figure 2.3.

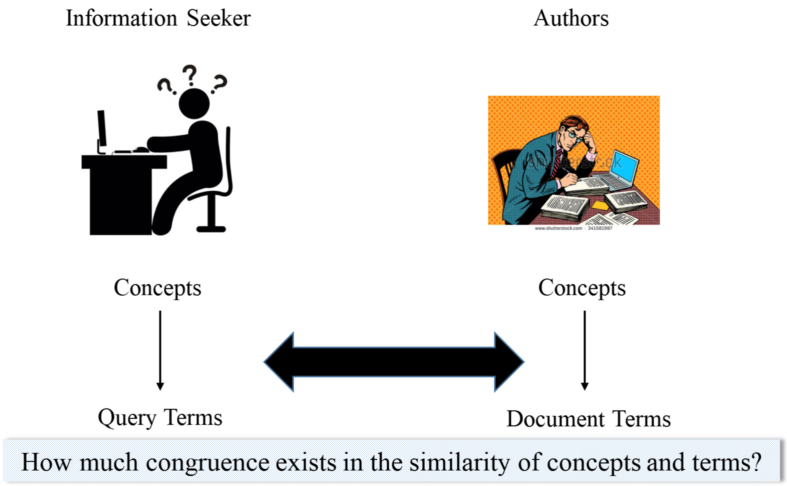

Information Seeker
Authors
Concepts
Concepts
Query Terms
Document Terms
How much congruence exists in the similarity of concepts and terms?

Figure 2.3: An overview of the abstract relationship between information authors and information seekers using online search

2.4.1 Precision-Recall measures

The performance of two search results retrieval systems is predominantly determined by the precision/recall measures. *Recall* is the ratio of relevant results returned divided by all relevant results and *precision* is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant results retrieved. The measure determines the extent to which the overhead of a user or online searcher is minimised by locating the needed information against a single or multiple queries.

One can think of precision as focusing on the garbage (the not relevant) documents that a retrieval system returns, while recall emphasizes the missed opportunities i.e. low precision means lots of garbage in the retrieved set and low recall means lots of missed opportunities i.e. many relevant documents were not found. Recall is how well the system performs in finding relevant documents. The best situation is to have high precision and high recall, which is very hard but not impossible to attain. However the inverse relationship which exists between precision and recall dictates that high precision implies low recall which means there are only a few documents for which the system can be very certain that they are correct. On the other hand high recall implies low precision, which means if you require to retrieve most of the relevant documents you have to tolerate lots of noise or garbage.

Precision at rank r is denoted by P_r and the Recall at rank r by R_r. These are defined by:

$$P(r) = \frac{retrieved_r \cap relevant}{retrieved_r} \in \{0, 1\} \qquad (2.6)$$

and

$$R(r) = \frac{|retrieved_r \cap relevant|}{relevant} \in \{0, 1\} \tag{2.7}$$

Where $retrieved_r$ denotes a set that consists of the r top-ranked documents and *relevant* is the set of all documents that are relevant to a query in the benchmark queries set. *P(r)* measures the fraction of top-10 highly ranked documents that are relevant to the query (accuracy) and R(r) measures the fraction of all relevant documents that show up in the top-ranked r documents (completeness).

2.4.2 Degree of satisfaction

The level of satisfaction, achieved by online searchers, may vary, subject to the fulfilment of their information needs. The satisfaction may be partial or complete after having analysed the search results to answer a search query individually or collectively. Jiang et al. (2015) stress that estimating satisfaction with search engines is an important aspect of evaluating retrieval performance. They argue however that several studies indicate that there are imperfect correlations between evaluation metrics and searchers' actual ratings of their search experience. They define *satisfaction of searcher* as the searcher's subjective preference on the utility of search systems and their search results. The higher the system utility is for the searcher, the more satisfied the searcher would be. Some studies measure satisfaction in terms of *frustration, satisfaction, success* while other classify user satisfaction after following the retrieval of search results in the form of *satisfied* or *dissatisfied.* Since utility is usually considered as the value of search results compared to the effort spent on searching (Jiang et al., 2015), the factors associated with effort level need to be considered in the measurement of satisfaction. Given an information need, Verma et al. (2016) hypothesise that the effort needed to satisfy the information need is affected by three primary factors:

- *Findability*: Effort needed to find the relevant information in a document e.g. extraction of useful and relevant information.
- *Readability*: Effort required to read a document e.g. terse expression of information, easy vocabulary or classification, etc.
- *Understandability*: Effort required to understand a document to satisfy the information need e.g. coherence, fluid description of information, etc.

Addressing the above factors will provide criteria for determining representative factors of users' efforts while assessing the retrieval performance of a search system. A consistent search results satisfaction and ranking model has been adopted during all search results evaluation cases which will be discussed in Chapters 7 and 8. The core reason for considering the *satisfaction level* in parallel with *search results evaluation*, the gap between search results evaluation metrics and users' levels of satisfaction have been thoroughly discussed and analysed.

2.4.3 Query auto-completion

One of the tools for assessing the quality of search results retrieval system is to use an auto-complete feature, embedded in the input of a search application. Having an auto-complete feature assists users to formulate their queries which leads to a more satisfactory search experience. Query auto-completion or QAC has attracted an increasing level of interest from the research community with proposals that take into account personalisation, context, time-awareness and user behaviour in QAC systems (Vargas et al., 2016). The algorithms used for implementing QAC are either based on popularity-based terms or context as well as users behaviour.

The features employed in this research included popularity-based tags (free-text and vocabulary) which were provided to expert crowd-annotators during the time of annotation. Further observations from the author's experiments focusing on query-auto-completion feature are detailed in Chapters 6 and 7.

2.5 Online searching scenarios

Although words with multiple meanings give English a linguistic richness, they can also create ambiguity: putting money in the bank could mean depositing it in a financial institution or burying it by the riverside; drawing a gun could mean pulling out a firearm or illustrating a weapon (Clark, 2013). The inability to designate unambiguously the rapidly growing number of new concepts generated by the growth of knowledge and research in Social Sciences (Riggs, 1981) is a lingering issue causing traditional search engines to fail. Furthermore (Riggs, 1981) argues that the justification of new concepts in the social sciences is much more problematical than it is in the natural sciences and technology. The research further suggests that innovative social scientists typically resort to the practice of using a familiar word for any new concept that they want to put forward. The disadvantage of this practice, however is that anyone who fails to recall a new meaning that has been stipulated for a word is likely to assume that it signifies one of its earlier meanings and therefore to misunderstand the author's message. Searching for *"transactional citizenship"* in a web archive like UK Web Archive[4], retrieves a list of results showing the instances of "citizenships" with no sign of showing (at least) both words together e.g. "transactional citizenships". Besides this, there is no *"see also"* link, which might have helped explore other relevant concepts like "Multiculturalism", "immigration", "naturalisation" etc. Similarly searching for *"preschool education"*, retrieves a list of results showing "preschool" and "preschool education" in a variety of files e.g. `html, PDF, PowerPoint presentation` etc. However, no instance of related concept is found relevant e.g. *childcare, kindergarten and head start etc.* although they were present in close proximity to the words being searched. If this semantic gap between keywords and meaningfulness is stretched over a century (*temporal factor*), another scenario can be presented. If *"Globalisation"* were to be searched in the library of ancient manuscripts before the 19th century, nothing would have been discovered but if searched in contemporary Social Science disciplines, it could be associated with a variety of developments in a broader set of fields: e.g. *communication, trade (with the emergence of multinational corporation), media and public opinion, the globalising political system, the globalisation of political system, the globalisation of culture and the spread of human rights as a global standard of behaviour*. In other words, today's and tomorrow's Internet searches have to cater for the ever changing advances in the transportation and telecommunication infrastructure, the rise of the Internet, which are the major factors of globalisation.

In an another instance the author searched for *"Deprivation index"* in UK Data archive[5] and there were five results in total although quite a lot of studies have been conducted in the UK over the years investigating *poverty and deprivation index* as part of various research projects. This is not to criticise but to make a point that when a phrase *"underprivileged area score"* is searched; nothing comes up which shows that the search system does not know the meaning of *deprivation index* or anything relating to measuring poverty and/or deprivation. According to Wikipedia, *"Underprivileged area score"* is an index to measure socio-economic variation across small geographical areas. Likewise, searching for *"Ontology"* in a web archive, retrieves results relating to "Ontologies" in a particular discipline of Computer Science, which shows biasness of its search engine towards relating the term to a particular discipline. Ontology in Social Science refers to *"What's out there"* which is the first building block in Social Science research (Grix, 2002).

Likewise, a user in an educational discipline of Social Sciences uses three queries to search for relevant content in the online search facility of a web repository. The

[4] http://www.webarchive.org.uk/ukwa/
[5] http://data-archive.ac.uk/

first query is *critical thinking* and the second query is *discourse analysis of text* and the third query is *presentation and communication skills.* The first query is pretty simple but a semantic search has to offer more in terms of contextual meaning to *critical thinking* e.g. web resources having content on *discussion in a non-threatening climate*, *careful analysis* and *evaluation of reasons and evidence* etc. Similarly, the second query is slightly longer than the first query but specifically looks for results having *discourse analysis of text* within certain disciplinary research area. *Critical discourse analysis or CDA* is a branch of linguistics that seek to understand how and why certain texts affects readers and hearers. The second query, therefore in a sense should encompass the first query too, which is about *critical thinking* as the second query does require the results to show *critical discourse analysis of text.* The third query *writing and presentation skills* would ideally expect the search engine to retrieve results having content on how best someone could critically and eloquently analyse and present a certain discourse and how communication and linguistics skills could play a role in enhancing such skills.

A Binary information retrieval model would certainly struggle to list the most relevant search results in all these three cases due to the two weaknesses 1. Such models do not rank documents based on relevance and 2. A document's score is calculated, based on keyword terms, in a user's query with no attention given to the semantics of content. Besides, technical writing on information retrieval is, understandably , heavily engaged with natural language processing, especially named entity extraction, parsing to identify adjective-noun phrases and all manner of frequency counts and statistical association (Buckland, 2012). Such approaches, however, fail to make much sense of the same topics referenced with different domain vocabularies without using (mentioning) the same terms. The issue arises at the time of search by the contemporary cross-lingual online searchers.

The purpose of using these scenario is to highlight the importance of moving away from "mention" to "meaning and relationship" with respect to the "passage of time" and gradual "terminological obsolescence". The purpose of describing these scenarios from the outset is to stress the classification of the inherent relationship in the categories of terms and concepts (in a meaningful but less complicated way) with a focus on sustainable relevance and accuracy in search results later. For example, in Statistics, "General Linear Modelling" is equivalent to "Latent variable modelling" which is a super class of "Statistical models". *"Statistical models"* is a further super class of *"Correlation Model"* and *"Multiple Regression Model"*. Relying only on the explicit presence of these concepts in web documents (whether part of a document collection or a disciplinary web repository of online resources) compromises the essence of meaningful search based on the semantic and contextual relevance in the existing and future online resources .

2.5.1 *Enterprise Searching* approach in inter-disciplinary research

Users in discipline-specific search increasingly view searching as *enterprise searching* which allows users in an enterprise to retrieve desired information through a simple search interface. That disciplinary search environment is thought of as the primary means of information retrieval, which is mainly driven by two factors: (1) larger volumes and more varieties of information within an enterprise or domain which are difficult to organise hierarchically, and (2) users are getting used to retrieving any information they want through *search box-based* search which has become familiar from Internet search

A research student, for example, engaged in a particular research is interested in "what is known about the problem" to prevent unwitting duplication of work that has already been done (Repko, 2008). In order to understand the background of the problem by tracing its development over a period of time, the student has to: (a) narrow down the amount of literature on that problem (b) contextualise the problem i.e. its relationship with other similar problems and (c) build a path of linkage between the past and present multi-disciplinary perspectives on that and similar problems. In today's world of the interconnected Web, the student would expect the results to have been produced by some sort of algorithms or techniques taking care of multi-disciplinary terminologies, concepts and contemporary discipline-specific classification. However, this may not be the case in reality as language evolves within each community of discourse and evokes that community. Since each community has at least slightly different linguistic practices, no one index will be ideal for everyone and, perhaps, not for everyone to search against using natural language keywords or phrases (Buckland, 2012).

2.5.2 Contemporary terms searching

To continue with the discussion in the earlier section, a search for "Global Warming" in the UK Data archive website[6] only retrieved one result. However, more results were found when "*climate change*" was searched but less of them having *climate change* instances appearing in them together. Interestingly enough, one of the results was a link to a page having a mention of *"Economic Climate"* which

[6] UK Data Archive is Economic and Social Research Council (ESRC)'s initiative, which curates and provides access to the UK's largest collection of social and economic data. Find out more at http://data-archive.ac.uk/

is far from relevant (in top 10 results) compared to the query terms. The query terms here were formulated to emphasise on the weather related global warming or climate change phenomenon.

The reason for using *"Climate Change"* rather than *Global Warming"* was that in some areas temperature change causes little change in weather pattern e.g. some individual places may actually get cooler and other will just experience changes in the entire weather pattern. Such changes in concepts and phrases occurring either in a particular discourse or perhaps in a multi-disciplinary domain necessitates the micro level classification of natural language text (on sustainable basis) with an agreed upon extensible, broader vocabulary (*derived from conceptual typology and empirical taxonomy*) by the domain experts from time to time. The availability, usability, compatibility and accessibility to such vocabulary must be assessed in relation to the domain-specific annotation and retrieval framework before the actual implementation. The approach, to address this and related issues has been detailed in Chapters 4 and 5.

2.6 Semantic annotation of documents: *Knowledge Extraction (KE) from text*

One of the most intuitive methods to transform a web into a semantic or structured web is through semantic annotation (Wu et al., 2006a). Semantic annotation is about adding formal description to web content and making it more efficient for information management (Kiryakov et al., 2004). The process of joining semantic concepts to natural language is referred as semantic annotation (Oliveira and Rocha, 2013) and is thus termed the source of KE. Documents are annotated with concepts instances from the KB by creating instances of the entities, concepts in a KB (may comprise of ontology, multiple ontologies (Müller et al., 2016), LoD sources etc.). Figure 1.2 clearly shows the *semantic indexing* component in the proposed system, where annotation is sub-divided into automatic and manual annotation of content with different sources of annotation. The instances and annotated documents are stored in a KB for web searching and/or data retrieval purposes. The emphasis here will remain on the matching of annotated documents with users' keywords for search results retrieval. It is also important to mention that the resultant annotated data in a traditional semantic annotation process is *machine readable* in order to facilitate other semantic crawlers and agents to make sense of the data for augmenting linked data Knowledge Bases (KBs). In this thesis, however, the focus will remain on the periodic semantic annotation, knowledge representation and crowdsourcing-based annotation of web content for sustainable information retrieval in a specific domain of interests. The *machine* *readability* aspect has been discussed in Chapter 9 as one of the feasible future works using the research infrastructure built in this thesis.

2.6.1 Types of web documents

Semanticizing website content (for adding meaningful annotation metadata) involves not only static HTML pages but the rest of content such as dynamically generated web pages (PHP (Hypertext Preprocessor), ASP (Active Server Pages), JSP (Java Server Pages) etc.), software script and code files, PDF and Word files etc. In other words, seamless content availability regardless of the type of content in an online repository is a must to ensure holistic indexing. Navas-Delgado et al. (2004) have elaborated two different annotation methods for static and dynamic web page annotation using ontology instances. Navas-Delgado et al. (2004) however assume that the originating static page having links to the dynamic page should be annotated first and the results from running the parameterized queries against the dynamic page should also be annotated, which would be cumbersome and not scalable when it comes to annotating a large number of dynamic pages written in different scripting languages. Query annotation only(instead of the entire dynamic web page), which retrieves dynamic web page as described by Benjamins et al. (2002) would certainly limit the effectiveness of annotation as it will not annotate the entire content for Keywords, Concepts and Entity extraction. Keywords, Concepts and Entities extraction are discussed in details in Chapter 4. A common issue in dealing with dynamic Web content is that page presentation changes over time: pages can be restyled, with respect to layout and mark-up, and content can be reorganised and moved to different pages. Also, the content within a web page has commonly a certain degree of granularity and the very same content can often be repeated in several pages (Grassi et al., 2013). Grassi et al. (2013) moreover note that along with the actual content (e.g. a document, a digitised manuscript, a picture), Web pages contain accessory content like navigation menus, advertising banners, and page headers. These accessory elements of web pages have been considered in the semantic annotation framework presented here and will be discussed in detail in Chapter 5. Thinking about all possible elements of web documents in terms of semantic annotation would be a challenge but ensuring deep and thorough access to such resources is a key to long-lasting relevant search results retrieval.

The emphasis in this research has been on (a) access to all the elements of the content which need to be analysed for concepts, entities and topical keywords extraction (b) presenting the web content in most usable *annotable* form to crowd-annotators for best possible expert annotation; to ascertain the impact of change on the relevance of documents over time. Recently applied methods and approaches adopted by the semantic search and LoD research communities are discussed in Chapter 3.

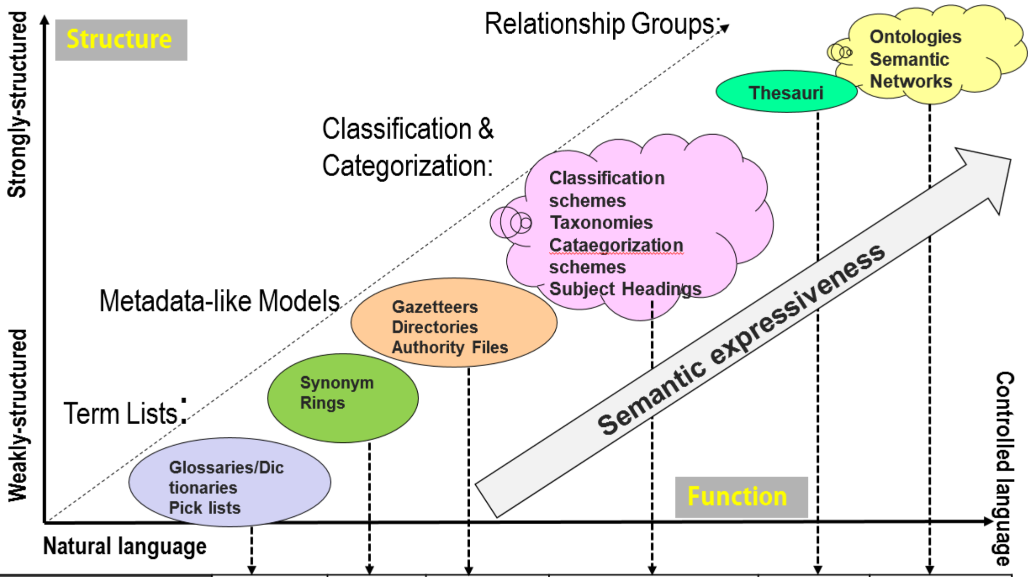
2.6.2 Formal vs. informal annotation

Annotation is the process that creates a function from a document to a formal or informal representation. Creating such a function involves three sub-processes: choice of a document or a part of document to be annotated (source); choice of the element of representation that is the result of the function (target) and finally definition of properties of the function itself. Consequently, automatic annotation means that the three annotation sub-processes are performed automatically by a software agent; manual annotation means that they are performed by a human agent, even if he/she uses software tools for that. Likewise, semi-automatic annotation implies that the human agent is helped by the software tools to perform at least one of the three annotation sub-processes (Azouaou et al., 2004).

In both types of annotation, it is paramount to establish evaluation methods in order to compare system and human responses with the benchmark output. The end product is to ascertain whether a combination of both could help in sustaining the performance of a search results retrieval system. Further related work in relation to both types of annotation will be discussed in Chapter 3.

2.7 Sources of knowledge organisation

A container of information includes any physical artefact, book, box, CD, web page, record (data or document form) or person. These have been collectively termed web content or web documents or objects. Classification and/or categorization of this information can produce metadata which can be searched and used to boost search ranking, in addition to the textual content held within the containers (Cleverley and Burnett, 2015). In order to interpret knowledge from the increasingly heterogeneous and predominantly unstructured content in web repositories for *searchability* purposes, Knowledge Organisation Systems (KOS) are exploited by domain experts and knowledge engineers. KOS, include classification systems, gazetteers, lexical databases, ontologies, taxonomies, typologies and thesaurus as can be seen in Figure 2.4 and these will be discussed in the remainder of this section to highlight their usage and limitations in the context of searchable knowledge bases. Classification systems like taxonomies, typologies and ontologies are discussed in this chapter to ascertain their efficacy in terms of knowledge representation in documents, contained in web repositories for search results retrieval purposes.



Eliminating ambiguity	XXX		XXX	XX	XXXX	XX
Controlling synonyms		XXXX	XXX	XX	XXXX	XX
Establishing hierarchical relationships			X	XXXX	XXX	XXX
Establishing associative relationships					XXXX	XXXXX
Presenting properties						XXXXX

Figure 2.4: Overview of various types of Knowledge Organisation Systems (KOS) showing semantic expressivity of content based on structuredness [(Lei Zeng, 2008)]

KOSs vary enormously in format and display, but they share the general characteristic of aiding knowledge elicitation and organisation, aiming at promoting the *retrievability* of information(Souza et al., 2012). KOS are not new and they have been in use over centuries for catalogues, bibliographic classification systems and taxonomies. Souza et al. (2012) further note that KOS have received special attention nowadays in context like the Semantic Web given the need for vocabulary disambiguation and the highly formalised structures needed to allow machine process "semantics" and "understanding".

2.7.1 Classification schemes

The term classification is used to refer to both the system or process of organising objects of interest and the organisation of the objects according to a system (Nickerson et al., 2013). Bailey (1994a) uses the term classification as the process of 'ordering entities into groups or classes on the basis of similarity'. He further observes that classification can be uni-dimensional or multidimensional, and that it can be done conceptually or empirically.

Classification systems are a very effective way of representing knowledge about the domain of discourse (McCloskey and Bulechek, 1994). Classifications are powerful technologies which must have the capacity to demonstrate that categories are tied to things that people do; to the worlds to which they belong. Several sets of classification schemes can be quoted here e.g. classifications of diseases, viruses, tuberculosis, race and of nursing work or things on the Web in a particular domain or the entire Web for that matter. A classification system demands to have an organisation (government, institution, individual) or a body responsible for maintaining it at present with a view to managing knowledge of a particular domain in the future. Kamnardsiri et al. (2013) employ the two KO schemes to represent tourist and tourism information in their recommender system for forecasting users' preferences for the desired destinations. Kamnardsiri et al. (2013) have amalgamated the classical scheme of typology and taxonomies with the more structured scheme of Ontology for achieving clarity in using tourism-related terms in the recommender system. The relationship between different classification groups can be achieved by having specific goals in mind. Figure 2.5 shows a an overview of classification systems split into 3 sources of classification i.e. *typology, taxonomy and ontology*. *Typology* is a conceptual classification using general labels or names and *Taxonomy* is a theoretical study of classification, comprising of principles, procedures and rules arranged in a a hierarchy. Ontology on the other hand, is a specification of conceptualisation in a knowledge domain which is an abstract, easy view of the world that is represented for a certain purpose. In other words, *Ontologies* are similar to faceted *taxonomies* but they have richer semantic relationships among terms and attributes and contain rules to specify terms and relationships (Bailey, 1994a). The flexibility and capability, offered by the three sources of classification, in knowledge interpretation, largely depends on domain-specific requirements. The frequency, with which heterogeneous content are annotated and

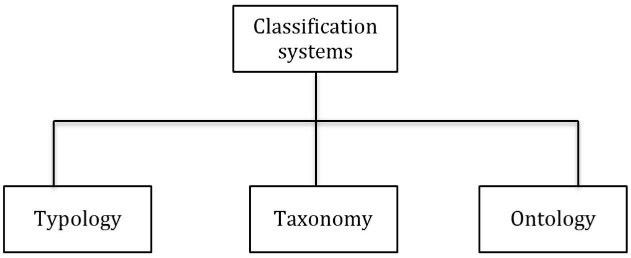


Figure 2.5: KOS Classification schemes [(Kamnardsiri et al., 2013)]

structured for searching; the types of content, which need to be structured and indexed, ultimately determine the adoption of a suitable classification system aimed at knowledge annotation and discovery.

2.7.2 Classification sources: Typologies, taxonomies, ontologies ...?

Two characteristics distinguish typologies from generic classifications i.e. a typology is generally *multidimensional* and *conceptual*. Typologies generally are characterised by labels or names in their cells (Bailey, 1994b). A hypothetical example by Bailey (1994b) explains it further using a two-dimensional classification system. The two dimensions are intelligence (dichotomised as intelligent/unintelligent) and motivation (dichotomised as motivated/unmotivated). Combining these two dimensions creates a fourfold typology, as shown in Table 2.2. These four categories can be defined as the types or type concepts. In other words, a motivated and intelligent person can be labelled as successful; an intelligent but unmotivated person is likely to be an underachiever; while a motivated but unintelligent person is an overachiever; and one who lacks both intelligence and motivation is likely doomed to failure. As shown in Table 2.2 the number of categories and dimensions are what determine the completeness of a typology. The problem arises when the number of dimensions gets larger and along with it the number of categories in each dimension increases as well. That would then lead to a typology containing a great many cells or types.

	Motivated	Unmotivated
Intelligent	Success 1	Underachiever 2
Unintelligent	Overachiever 3	Failure 4

Table 2.2: A hypothetical fourfold typology

For example, even if all dimensions are dichotomies, the formula for determining the number of cells is 2^M, where M is the number of dimensions. Thus, for five dichotomous dimensions the typology will contain only 2^5 or 32 cells, but for 12 dichotomous dimensions the number of cells is 2^{12} or 4,096. If the dimensions are polytomous rather than dichotomous, as is often the case, the number of cells expands much more rapidly. Because the number of types can be so large, researchers have often found it helpful to use partial or shorthand typologies. These can be formed either by constructing the full typology and then selecting only certain types for use in the analysis or by merging some types together.

For example, depending on requirements, it may be that only a few chief types are found to be really important so that these can become the focus and neglect the remainder. In such a situation, it is common to utilise a shorthand typology by first constructing only key criteria types and then locating all other types in reference to these criteria. This methodology has been explained in the following Section 2.9 where NCRM typology has been presented as an example.

The above-mentioned situation arises in the case of domain-specific ontologies when they are used as a *source of knowledge* for semantic annotation. Unlike typologies, the categories in ontologies are classes and sub classes and the dimensions are the multitudes of triples expressing the relationships between different classes and sub classes through *subsumption* and *inferencing.* However, unlike short hand typology, the interrelationships between classes, instances properties and relations will not be possible in the form of a shorthand or sub ontology given that each new category (class, sub class) and dimension should conform to the previously established subsumption and inheritance rules of the ontology at the time of semantic annotation of unstructured text in different documents.

2.7.2.1 Ontology composition: implicit/explicit expressivity challenges

The expression of an implicit fact in an ontology during semantic annotation still remains a challenge given that such expressivity depends on the inference and reasoning of explicit formal semantics. For every given inference or reasoning-based instantiation of a new concept in a domain specific ontology, the instance statement must be true for all the possible instatiations of the domain. For example consider a formal statement expressed in a domain-specific ontology, <*Regression methods* is a subclass of *Quantitative Data Handling & Data Analysis*> and <*Logistic-linear regression* is a subclass of *Regression methods*>. To instantiate the text *ANOVA-based regression analysis* with a subclass of *Regression methods*, it must be clear to reasoning and inference agents that *ANOVA-based regression analysis* is a type of *Quantitative Data Handling & Data Analysis*>. However, the problem arises when further concepts are introduced or the existing ones evolve over time. For example, *Non-Parametric Approaches* is a related term to *ANOVA, ANCOVA, Linear Regression, Logistic regression.* If it is known that the latter terms are individual instances of *Regression methods* class, then *Non-Parametric Approaches* should be related to *Regression methods* which is a subclass of *Quantitative Data Handling & Data Analysis.*

The example given is true in many other types of ontology-model-based inferencing and reasoning at the time of semantic annotation of unstructured text. These types include, but are not limited to:

- Classification: This type of inference statement is made if the reasoner thinks that a particular individual in a set of documents is an instance of a particular class. In the example given above, micromanaging such relationships between the existing and new concepts will remain a challenge during semantic annotation aimed at relevant search results retrieval.
- Subsumption: This type of inference deduces all the subclass relationships or sub subclass relationships between the existing classes in the ontology and the new concept class.
- Equivalence of classes: A relationship of this kind in inferred if the two classes are the same in terms of explicit/implicit mutual subsumption.
- Consistency of concepts: This type of inference-based relationship deduction entails whether a concept is consistent i.e. there is no contradictory axiomatic statements about a concept; that is to say that an inconsistent concept can have no instances or annotations in the text. This kind of inferencing particularly poses a significant challenge when envisaged in terms of fast-paced changes in terminologies and discipline-specific concepts.

Evolving ontologies to absorb this level of intricacies arising out of the evolving concepts (general vs. specific (Müller et al., 2016)) or the introduction of new concepts will remain a challenge for both the ontology engineers and multi-disciplinary research community. Moreover continuously monitoring the emergence of new concepts and terminologies in a particular domain specific ontology like Social Science Research Methods is in itself unsustainable; as an agreement over a group of concepts and the context in which they are used will be needed between the subject experts and ontology experts.

2.7.2.2 Typology vs. Taxonomy

This section is not intended to delve into the discussion of *typology/taxonomy* but to emphasise the usage of hierarchical and conceptual classification in the form of typology for collaborative or manual semantic enrichment and search results retrieval.

A fundamental problem in many disciplines is the classification of objects in a domain of interest into a taxonomy. Developing a taxonomy, however, is a complex process that has not been adequately addressed in the information systems (IS) literature (Nickerson et al., 2013). The term taxonomy is perhaps the most confused. As with classification, taxonomy is sometimes used for the system or process and sometimes used for the result of applying the system (Bailey, 1994a). Doty and Glick (1994) equate taxonomy with classification scheme, although they note that classification scheme, taxonomy, and typology are often used interchangeably. Gregor (2006) echoes this thought, stating that 'the term typology is used more or less synonymously for taxonomy and classifications'. Bailey (1994a) distinguishes the taxonomies (*classification systems derived empirically*) from typologies (*classification systems derived conceptually*). However, Bailey (1994a) also presents a methodology for developing taxonomies/typologies that is a combination of conceptual and empirical approaches.

(Nickerson et al., 2013) explains that taxonomy development in the social sciences has also been well studied. Nickerson et al. (2013) elaborates (citing (Bailey, 1994a) the distinction between a typology and a taxonomy, saying that the former is derived conceptually or deductively and the latter is derived empirically or inductively. In the conceptual typology approach, the researcher proposes a typology of categories or types based on a theoretical ideal or model. In the process, the researcher could define an ideal type, which Bailey (1994a) (citing (Weber, 1949)) explains is the 'extreme' of types. The ideal type is used to examine empirical cases in terms of how much they deviate from the ideal (Nickerson et al., 2013). The researcher may conceive a single type and then add dimensions until a satisfactorily complete typology is reached, a process called *substruction* ((Bailey, 1994a), p. 24). Alternatively, the researcher could conceptualise an extensive typology and then eliminate certain dimension in a process called reduction ((Bailey, 1994a), p. 24) until a sufficiently parsimonious typology is reached.

2.7.3 Why typology-based classification for manual annotation?

The typology-based classification aims to classify different themes and outputs of research and the prioritisation of various research methods in a domain like Social & Human Sciences in a certain period of time. Such classification provides a useful framework for the identification of gaps in current research, the identification of needs for further research and the current focus of certain research methods, the evaluation of recent developments, the classification of research projects and the establishment of further research funding (Beissel-Durrant, 2004). The typology classification could also be used to monitor the evolution of a certain research area or the development of new scientific research methods. For example, to identify the needs for the deliverance of training in the future to researchers in a particular area, the typology would help first to understand the literary gaps in the current methods and requirements of training. Du Toit and Mouton (2013) suggests that a typology needs to be a reflection on methodologies used within a discipline and draws on Teddlie and Tashakkori (2006) five benefits of a typology in the area of *mixed methods research design.* Luff et al. (2015a) term the first three as one of the most relevant to the NCRM typology which are summarised as:

- Help researchers decide how to proceed with their respective research
- Establish a common language for the respective research field
- Provide the field with organisational structure

Nickerson et al. (2013) further explains that, in terms of defining organisational structure, both taxonomies and typologies can be combined to avoid the more constructed nature of types in taxonomies and "ideal types" in traditional typologies. That way, the researchers in a particular domain can make the best use of a classification system. Using a classification system in their own restricted silos with less tendency to extensibility restrict the true purpose of classifying things, which may lead to less satisfactory results at the time of searching.

2.7.3.1 *Practical attributes of typology*

The attributes and characteristics, Nickerson et al. (2013) have listed as the building blocks of a good taxonomy, typology or a classification system that has the capacity of describing the objects in a particular domain include:

- *Flexibility* which entails that it should be amenable to alternative approaches borrowed from other systems that is appropriate for the domain of interest.
- *Conceptual and empirical* which implies that a taxonomy or the mixture of taxonomy and typology should have dimensions and characteristics based on conceptual and/or empirical grounds and not on the basis of *ad hoc* or arbitrary dimensions
- *Efficiency in completion* which states that it must be revised/completed in a reasonable period of time
- *Ease of use* which refers to its straightforwardness in terms of application. That is quite important as a typology or taxonomy is developed by researchers/expert having different levels of understanding of the typology/taxonomy development literature. The application of categories (during annotation of content or documents) particularly in a scientific discipline has to be done by researchers and non-researchers and they must be easily understood without reference to the literature.
- *Usefulness* it must lead to a useful system which is widely applicable and accessible.

2.7.4 *Example*: Enterprise searching based on KOS

Enterprise search is the subset of web search that refers to IR technology which automatically indexes enterprise content (including web pages, document and people expertise profiles) providing a single place for staff to search without necessarily knowing where content is located (Wilson, 2000). The theme of our research has more in common with the enterprise search where all the elements of web searching are taken into account in a specific domain of interest to facilitate a certain community retrieve relevant information more efficiently and accurately. In enterprise search most queries are single word (*lookup* and often portrayed as not working well compared to Internet search engines (Andersen, 2012). The crowd using *enterprise search* is very small compared to the Internet, hampering the effectiveness of using statistical crowdsourced usage data for all but the most common search queries (Cleverley and Burnett, 2015). Cleverley and Burnett (2015) further note *lookup* search is likely to account for between 80-90% by volume of all enterprise search tasks including accurately locating definitive documents. An example of such queries is *NCRM/NCeSS Collaborative Fund Report* where a prticular report is being looked up having association with an *organization* called *NCRM or NCeSS*. On the other hand, *Exploratory search* queries are open ended e.g. *What do instrumental variable models deliver with discrete dependent variables*. Different KOS methods may be required to meet the browsing and searching needs for these two types of search goals i.e. *lookup* and *exploratory* searches yet the KOS literature rarely differentiates between these two search goals (Cleverley and Burnett, 2015).

In essence, the author has pitched various annotation/retrieval experiments around the above-mentioned goals by exploiting the structured and hierarchical KOS in order to ascertain whether sustainable relevance in search results could be continually achieved.

2.8 Challenges of semantic searching

In this section some of the challenges are highlighted which need to be addressed in order to achieve, implement and streamline an effective semantic searching strategy in domain-specific web repositories. A semantic search engine, which is one of the components of IR system, is responsible for retrieving the results based on user queries. The two factors that impact a semantic search engine, include *managing huge amount of data* and *providing very precise results for queries* quickly (Lashkari et al., 2017). Table 2.3 sums up various such challenges based on requirement criteria.

2.8.1 Limitation of semantic search approaches

The large number of dimensions involved in semantic indexing and semantic searching pose myriad challenges to those who are involved in the knowledge representation and knowledge creation tasks. Fernández et al. (2011) present 6 points criteria for assessing the effectiveness of a particular semantic search approach, given in Table 2.3.

Criterion	Limitation	IR	Semantic
Semantic knowledge representation	No exploitation of the full potential of an ontological language, beyond those that could be reduced to conventional classification schemes	X	(partially)
Scope	No scalability to large and heterogeneous repositories of documents		X
Goal	Boolean retrieval models where the information retrieval problem is reduced to a data retrieval task		X
Query	Limited usability		X
Content retrieved	Focus on textual content: no management of different formats (multimedia)	(partially)	(partially)
Content ranking	Lack of semantic ranking criterion. The ranking (if provided) relies on keyword-based approaches	X	X
Additional limitations			
Coverage	Knowledge incompleteness	(partially)	X
Evaluation	Lack of standard evaluation frameworks		X

Table 2.3: Limitation of semantic search approaches (Fernández et al., 2011)

Besides the limitations mentioned in Table 2.3, the natural, cultural or global phenomena *per se.* are also likely expected to require specialised approaches for better knowledge representation, classification and retrieval. Such specialised approaches limit the scalability and scope of semantic search whether it be search results evaluation, document ranking , usability interfaces or query formulation etc.

2.8.2 Subject naming & cultural changes

In today's digital world, the pace of change has gone faster. It is increasingly becoming a challenge to keep track of the connotation and context of the past discourse and natural language literature in a particular discipline or domain. Buckland (2012) observe that there are cognitive developments: new ideas and new inventions need new names. The shift from traditional library-based naming to more online Internet user-based annotation of things is taking over the primary role of libraries and archives. The need for delegating the naming and classification role from libraries and archives to the online crowd has never been greater. Subject classification and naming by the libraries are increasingly coming under pressure when the naming and classification of artefacts does not address (Buckland, 2012) the elements of denotation and connotation. In other words, Buckland (2012) explains that some linguistic expressions are socially unacceptable or what is deemed acceptable to one cultural group may not be acceptable to another cultural group. Such disparity if left un-addressed during the classification of terms and concepts in a particular domain, will affect the retrieval when terms are searched (using multiple linguistic expressions) in the future. For example, the phrase "yellow peril" was widely used to denote what was seen as an excessive immigration from East Asia, but it is now considered too offensive to use even though there is no convenient and acceptable replacement name and the phrase remains needed in historical discussion. Such phrases need to be linked to the related and synonymous terms and phrases of today's linguistic expressions both in terms of connotation and context in order to enable the future researchers make sense of the past, present and the future.

2.8.2.1 Interpretation of new subjects: *expert annotation perspective*

The exploitation of the current digitally connected knowledge community could be a brilliant opportunity for subject experts, librarians and technical experts to interpret the semantic ambiguity in concepts, terms and subjects. The interpretation based on cultural connotation and contextual annotation by the online crowd (from multiple domains and disciplines) having subject expertise and knowledge could provide a headway for more relevant and accurate information retrieval. A contemporary vocabulary or typology or a mix of typology and taxonomy, fed into the landscape of current digitally connected crowd of experts and users, could provide fodder to address this and related issues to a satisfactory level leading to evolving a sustainable base.

Furthermore, the classification systems such as typology or taxonomy can help in categorising the present knowledge with a focus on the future retrieval but fall short of addressing multiple contemporary discourses in a discipline at the time of categorization. It is not simply that content in a new document has to be positioned in relation to both past discourse and future needs, but additional complexity arises due to multiple discourses (Buckland, 2012). Language evolves with each community of discourse and produces and evokes that community, leading to the establishment of myriad more or less specialised, stylised practices of language. Addressing these kind of elements by automatic semantic annotators of natural language text is almost impossible.

2.8.2.2 Interpretation of new subjects: *Annotator-user collaboration*

Collaboration between annotators and users (now and in the future) is therefore inevitable to make sense of the evolving terminology at every stage of discourse development to avoid multiplicity and instability of meanings. In terms of retrieval, searching against one document index comprising of content produced by multiple communities, having different linguistic practices, would seriously undermine the fulfilment of the information needs of future communities. So, (Buckland, 2012) sums up that multiple, dynamic indexes, one per community, would be ideal to ensure context and perspective-based satisfaction of information needs expressed by multi-disciplinary communities. They exemplify this by saying that a rabbit can be discussed as a pet, as a pest or as a food in present and future community discourses. Similarly, in medicine, specialists in anaesthesiology, geriatrics, and surgery might all ask for recent literature on, say, cardiac arrest, but because they are interested in different aspects they will not, in practice, want the same documents Buckland (2012) citing (Petras, 2006) and (Buckland et al., 2001).

2.8.3 Search engine challenges

The unambiguous naming of new concepts in a multi-disciplinary domain is posing a huge challenge to the performance of search engines due to knowledge growth in the current digital age. Various scenarios and examples have been discussed in Section 2.5 to highlight this challenge from the outset. These challenges have partly been considered by the keywords based searching and addressed to an extent where plain keyword queries are converted into equivalent semantic queries followed by syntactic normalization, word sense disambiguation (Snow et al., 2008) and noise reduction. To do that, the use of dictionaries (e.g. Wordnet), thesauri and other library classification systems have been exploited in collaboration with domain specific ontology to express keywords in more structured language. Semantic keywords are then matched with ontology terms and various semantic agents are applied to disambiguate terms and words before retrieving the results (Royo et al., 2005). Similarly, Mestrovic and Calì (2017) talks about automatically extracting a domain taxonomy from ontologies and then populating them with base

concepts of the domain for information retrieval. However, mapping multiple ontologies back into the domain taxonomy is highly prone to creating ambiguity in terms and concepts thus affecting the ranking and accuracy of the search results (IR).
Moreover, as described above, like other information domains, in scientific research disciplines terms change over time due to cultural, social, technological, scientific and socio-economic etc. factors which compromise relevance and accuracy in search results. All this suggests that semantic expressions and matching terms with ontologies classes/properties (linguistic) and instance data (semantic information) will get further complicated and would need frequent and regular expert human intervention both at annotation and retrieval times. Further ontology-based annotation and retrieval is discussed in Chapter 3.

2.8.3.1 Terminological interpretation of topics

It may sound trivial, but given the fact that many students and even seasoned academics, have difficulty in differentiating between crucial terms such as *ontology* (e.g. what is out there to know about) and *epistemology* (e.g. what and how can one knows about it), their subsequent research is bound to suffer, as knowledge of these terms and their place in research is essential to understanding the research process as a whole Grix (2002). Buckland (2012) explains that the challenge of creating description is to enable those to be served to identify and select the best documentary means in the future. Will some future searcher consider this document "on topic" and better yet "relevant", is not simply a matter of what the document is about, but of how it might be viewed in an imagined future. Familiarity with the community and its purposes, ways of thinking, and terminology are important building blocks for those classifying today's knowledge for the future retrieval. Topical description is a matter of naming what a document is about but in practice, descriptions summarise.

Stating that a subject heading represents a topic or a concept is valid but unhelpful as this merely points to another name, but does not add to explanation. For example, saying that the subject heading "Dowsing" is 133.32 in the Dewey Decimal Classification provides an alternative name but does not explain what dowsing is. An explanation of what a subject heading (and, therefore, a document) is "about" must be derived from the discourse with which the name is associated (Fairthorne, 1971). The emphasis on discourse, discussion, dialogue rather than topics is what drives the futuristic classification of things. Buckland (2012) further points out that meanings are established by usage, and so always draw on the past with a focus on the future. This might seem difficult but reality is made much worse by time, by technology, by the nature of language and by social change.

Buckland (2012) observes that the biggest problem is finding and using an "agreed upon" vocabulary that could be used by those concerned with the classification, annotation and tagging of all types of content in all disciplines and domains. As described by the panellists in a panel talk (Berners-Lee et al., 2012) "*integrating different scribbles subject to an agreed upon vocabulary is beyond imagination which might be the reason for many organisations that they put together lots of different stuff together*". Research has repeatedly revealed that different indexers will commonly assign different subject index terms to the same document, as will a single indexer at different times (Olson and Wolfram, 2006). The multiplicity and fluidity of natural language vocabulary can be very unpredictable in terms of classifying natural language topics or words (Buckland, 2012). Searching for *violin* or *fiddle* or both may be one of the many situations, the users may find themselves engaged in while looking up relevant information. Buckland (2012) describes that the multiplicity of natural language terminology can be mitigated by adopting a restricted vocabulary either a "controlled vocabulary" of natural language terms (e.g. "Fiddles see Violins") or an artificial notation for the descriptive names (e.g. "787.1" in the Dewey Decimal Classification[7]).

2.9 The NCRM Typology: A classification system in Social Sciences

The NCRM research methods typology is presented here as a case study in the context of semantic annotation of online resources in web repositories containing multi-disciplinary content in Social Sciences.

The typology was initially developed to classify, tag and label various research outputs developed by (Luff et al., 2015a). It provides a hierarchical classification of research methods used in the Social Sciences and has been used by the NCRM to categorise training events, research activities and other outputs. The typology has become one of the most frequently downloaded items from the NCRM web repository[8].

[7] Dewey Decimal Classification (DDC) is a proprietary library classification system first published in the U.S. by Melvil Dewey in 1876. Available at https://en.wikipedia.org/wiki/Dewey_Decimal_Classification

[8] https://www.ncrm.ac.uk

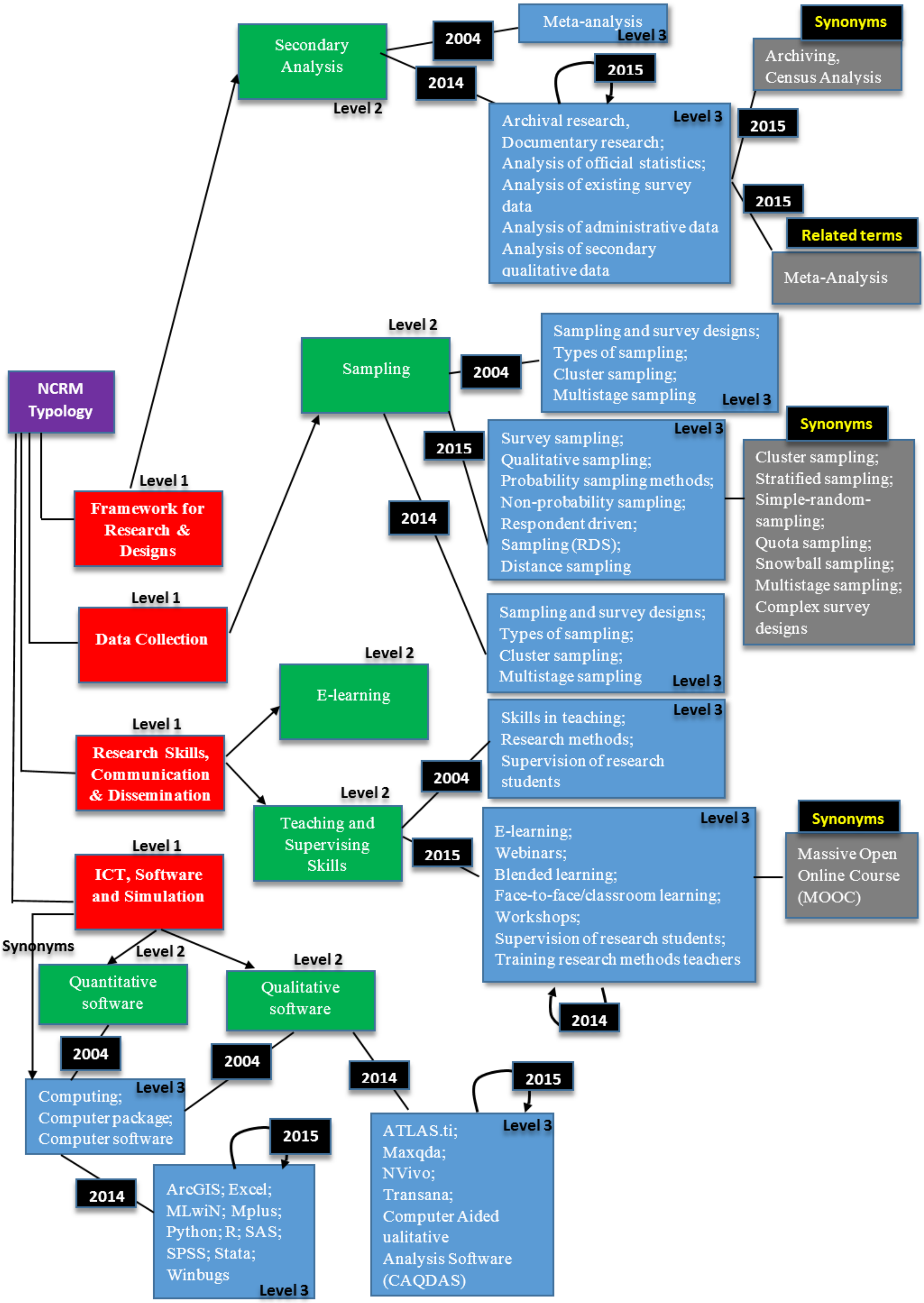

Secondary Analysis
Level 2
2004
Meta-analysis
Level 3
2014
2015
Synonyms
Archiving, Census Analysis
Archival research, Documentary research; Analysis of official statistics; Analysis of existing survey data Analysis of administrative data Analysis of secondary qualitative data
Level 3
2015
2015
Related terms
Meta-Analysis
Level 2
Sampling
2004
Sampling and survey designs; Types of sampling; Cluster sampling; Multistage sampling
Level 3
NCRM Typology
2015
Survey sampling; Qualitative sampling; Probability sampling methods; Non-probability sampling; Respondent driven; Sampling (RDS); Distance sampling
Level 3
Synonyms
Cluster sampling; Stratified sampling; Simple-random-sampling; Quota sampling; Snowball sampling; Multistage sampling; Complex survey designs
2014
Level 1
Framework for Research & Designs
Level 1
Data Collection
Sampling and survey designs; Types of sampling; Cluster sampling; Multistage sampling
Level 3
Level 2
E-learning
Level 1
Research Skills, Communication & Dissemination
Skills in teaching; Research methods; Supervision of research students
Level 3
2004
Level 2
Teaching and Supervising Skills
2015
E-learning; Webinars; Blended learning; Face-to-face/classroom learning; Workshops; Supervision of research students; Training research methods teachers
Level 3
Synonyms
Massive Open Online Course (MOOC)
Level 1
ICT, Software and Simulation
Synonyms
Level 2
Quantitative software
Level 2
Qualitative software
2014
2004
2004
2014
2015
Computing; Computer package; Computer software
Level 3
2015
ATLAS.ti; Maxqda; NVivo; Transana; Computer Aided ualitative Analysis Software (CAQDAS)
Level 3
2014
ArcGIS; Excel; MLwiN; Mplus; Python; R; SAS; SPSS; Stata; Winbugs
Level 3

Figure 2.6: Evolution of NCRM typology main and sub categories along with synonyms and related terms over the years.

The typology contains elements of the classification developed by the Data Documentation Initiative (DDI[9]) and the classification used by the ESDS (Economic and Social Data Service(ESDS[10]))/UK Data Archive. The NCRM typology is currently being used to label all items in the NCRM web repository including research publications, training courses and materials, funding and conference calls and news articles informing the larger research community of the developments taking place in Social Science methods research. The use of the typology not only supports the NCRM itself, but supports hosting content relating to the continually evolving programme of research methods investments by the ESRC, the UK's largest social science research funder. The typology therefore is an inevitable source of classification providing important elements of Social Science research infrastructure. The classification mechanism covers a wide range of research outputs which are widely used in the repository for online searching and information retrieval purposes. The typology has been adopted for experimentation as part of this PhD research because of the evolution of terms and concepts taking place regularly in the development of this typology.
In Figure 2.6, it can be clearly seen that there are various hierarchical levels between main and sub categories of terms along with synonyms. However, there needs to be a balance in the structure of the typology i.e. categories in the typology should be more conceptual in nature with only the lowest level (*Descriptor*) having an attribute of specificity for ease of use and search. For example, *Data Collection* is the main category or *Level 1* (in Red) in the typology structure in Figure 2.6 connected to the sub category or *Level 2* (in Green). The Blue box or sub sub category of *Data Collection* or *Level 3 or descriptor* (in Blue) contains the categories from different versions of the NCRM typology i.e. 2004, 2014, and 2015.

A major revision was completed in 2015 after only a few categories were changed in 2014. It is often only at *Level 2* that more specific research methods types are specified e.g. *"Regression Analysis"* or *"Secondary Analysis"*. The category value at *Level 2* is less subject to contemporary trends and fashions and short-lasting approaches than at level 3. Annotation and retrieval on the basis of level 2 offers a variety of online resources to users but with less specificity. Annotation or categorisation at level 2 means that a user could label any type of regression analysis, including *linear*, *ordinary* or *logistic* as *"Regression Analysis"*. However, the authors in (Luff et al., 2015a) suggests that based on feedback from researchers they do not think on that broader level while describing their work. They, rather

[9] The Data Documentation Initiative (DDI) is an international standard for describing the data produced by surveys and other observational methods in the social, behavioural, economic, and health sciences. Available at https://www.ddialliance.org/

[10] ESDS has moved to the UK Data Service accessible at https://sp.ukdataservice.ac.uk/introduction.asp

describe the key elements of their work in more detail like someone teaching *logistic regression* will not necessarily label various aspects of the course using just *regression analysis*. They may even skip using the term *logistic regression* or even the term *logistic* in the title of the course. This level of specificity would therefore ideally and necessarily sit in the descriptor and connected term as referred to by (Doty and Glick, 1994) and *connected term (synonyms)* and *related term* in the up-gradated version of the typology (Luff et al., 2015a).

The *level 2* descriptors in blue boxes in Figure 2.6 have changed over time both in terms of new categories and the corresponding synonyms. This behaviour actually defines the granularity of the main category and needs to be addressed at the time of embedding the typology in a web repository. The aggregation of terms at the time of faceted search would then clearly visualise the obsolete and contemporary terms under one main category with the relationship shown from older *level 3* or synonym item to the newer or contemporary *level 3* item along with new synonyms or related terms.

2.9.1 Evolution of the NCRM Typology

The need to upgrade a typology for better representation of concepts in documents necessitates regular review and update of those systems for consistency and contemporariness. The upgrade was initiated by the researchers and academics, associated with the NCRM, based on the genuine need to refine the existing categories and sub categories of the typology along with adding new concepts, terminologies, synonyms, related terms etc. The initial proposal was made by the researchers and librarians, having interests in the NCRM Typology as a classification system, arguing that there has been a constant growth in the usage of new terms and concepts in the literature and various discourses of Social Sciences. In order for the typology to accommodate those new concepts and terminologies in the existing classification hierarchy, and for the researchers to take advantage of it at the time of tagging and classifying their research outputs, the upgrade was inevitable. (Luff et al., 2015a) observe that there are four reasons for developing and upgrading the NCRM typology:

- Advances within the research methods and research methods technologies in order to make the typology more relevant to today's digital landscape
- The application of the typology has expanded over time from one set of documents classification relating to training materials to a wide variety of

subject classification in areas like research project funding, research publications, myriad number of online and offline courses and news articles in social research methods.

- The inconsistency in usability of the old version of the typology prompted researchers, experts and librarians to upgrade the typology so that all categories, sub categories, related terms, synonyms and narrower/broader terms reflect the current digital and academic landscape. For example, to label a particular training event before depositing it in a repository e.g. training or workshop for researchers, the *depositor* had to select all broader and main categories terms in order to get to the desired term. In other words, the granularity of the terms and accessing them efficiently and precisely was the third issue prompting the update.

- The lack of depth was another reason which implies that users are unable to locate the term they want to assign to their item. For example, *questioning* and not *questionnaire* was sometime sought by the user to classify their item with the typology term.

The above points highlight the needs for having a consistent, compliant and easily usable typology that could be deployed in a web repository environment for both annotation and retrieval. Luff et al. (2015a) further observe citing (Gregor, 2006) that the evaluation of the success of a typology is based on the category labels being meaningful, the logic of the dimensions being clear and the ability to completely and exhaustively classify being demonstrable. The correct assignment of typology category or multiple categories to a particular research methods, theme or even a research project highly depends upon the consensus of the wider social sciences community. The introduction of new and the upgrading of existing categories must also conform to established and contemporary theories and practices in that particular discipline or domain. Lack of the above mentioned characteristics in the earlier typology essentially led to the upgrading of the old NCRM typology (Beissel-Durrant, 2004) to the new version (Luff et al., 2015a).

2.9.2 Discussion

The *typology evolution* has to continue over time in order to sustain the relevance of all the categories and their relationships with the actual content items stored in a database, document collection or a searchable web repository. Highlighting the stark differences between the older and newer version of the NCRM typology, the authors further argue (citing (Given, 2008)) that typologies shouldn't be hierarchical, but rather categories should be related to one another, rather than some being subsidiary to others. This characteristics is quite important when one looks at different terms and concepts across time and disciplines as assigning a subsidiary terms to a piece of text would limit the relationship element of the web page for example having that text in the rest of the related content on other pages or web documents per se. The theoretical ideal characteristics in a purely hierarchical typology (Beissel-Durrant, 2004) conflict with the practical consideration especially at the time of annotating content with those typology terms or retrieving relevant information through faceted search in web repositories.

The updated NCRM typology needs to meet the broader aims, discussed in Section 2.7.3.1, relating to supporting the wider research community by providing a coherent frame of reference. In addition, the existence of an up to date typology leads to the development of broader applications e.g. classification of literature in the Social Sciences, classification of computer software or data analysis tools, the identification of relationships between research methods and understanding and clarification of these relationships (Beissel-Durrant, 2004). Such contribution could then be packaged into a full-fledged annotation or web repository classification system to enable the wider research community take part in the organisation of knowledge accumulated over the years in a particular multi-disciplinary research environment.

2.10 Scalability in indexing and searching

Semantic indexing and exploring the resulting knowledge base in scalable manner still remains an issue. Scalability within the context of a search engine is the ability of a system to handle a rapidly growing amount of data(Fatima et al., 2014). Given the volume of data being published on the web and the processes (automatic as well as manual) aimed at structuring that data for better search results retrieval, have to be scalable enough for trust building, sustainable efficiency and effectiveness. The users' involvement in real time semantic metadata generation and relevant search results retrieval necessitates the fault-free interaction with the content at all times. Along with the crowd and expert semantic annotation, the system should continually process absorbing new annotations, establishing relationships based on the new annotation and subsequent modification of weights for each token and keywords in the multi-dimensional space of searchable KBs.

2.11 Summary

This chapter has provided a detailed overview of various classification systems, taxonomies, typologies, complex information needs of users and challenges of semantic searching. It is proposed that the knowledge of the designated research community in particular and online multi-disciplinary users in general need to be exploited; to tackle the issues, arisen out of recurring changes in disciplinary concepts and terminologies. Neither a purely ontology-driven retrieval model suffices the representation of all content in a particular domain nor a fully automatic annotation system can entirely structure the huge amount of information, with total precision/accuracy, in web repositories. The crowdsourcing-based semantic annotation would, therefore, be best placed to play a pivotal role in evolving the terms and concepts in a particular domain or discipline. The *"web-embeddable* vocabulary, which is *"upgradable"* and *"web-embeddable"* , when used in conjunction with the automatic semantic annotators, would enable the experts, in a particular domain, to actively engage in annotating and tagging the content for better search results retrieval. It is also observed that the sustainability element, both in the upgrading of the typology (classification system) and the act of annotation and tagging of content (crowd-annotation & tagging) must be central to the preservation of relevance and accuracy in search results.

Citation	Description	Section
(Karpf, 2012), (Shirkey, 2013), (Fernández et al., 2011), (Gal et al., 2003)	Multi-disciplinary Internet-based research, Search engine impressive enhancement in the last decade,Searching contextually relevant information in web archives/repositories	2.3
(Wu et al., 2006a), (Kiryakov et al., 2004), (Navas-Delgado et al., 2004),(Benjamins et al., 2002)	Transformation of web into semantic or structured web through semantic annotation, static and dynamic web page annotation using ontology instances	2.6
(Cleverley and Burnett, 2015), (Azouaou et al., 2004)	Classification and organization of semantic metadata, automatic/manual annotation	2.6.1, 2.6.2
(Souza et al., 2012), (Nickerson et al., 2013), (Bailey, 1994b), (McCloskey and Bulechek, 1994)	Knowledge elicitation and organization aimed at retrievability, representation of domain-specific knowledge	2.7, 2.7.1
(Kamnardsiri et al., 2013),(Lei Zeng, 2008)	KO schemes and recommender system using taxonomies and typologies, KOS classifications	2.7.1
(Nickerson et al., 2013), (Doty and Glick, 1994), (Gregor, 2006), (Müller et al., 2016)	Multidimensional and conceptual typologies for knowledge representation, empirical taxonomy development as a classification system in social sciences vs. conceptual typology classification	2.7.2, 2.7.2.1, 2.7.2.2
(Beissel-Durrant, 2004), (Du Toit and Mouton, 2013), (Teddlie and Tashakkori, 2006), (Luff et al., 2015b)	Framework for the identification of research gaps, focus of certain research methods, evaluation of certain research methods, use of typology in social science research themes	2.7.3
(Buckland, 2012), (Marshall, 1977), (Lashkari et al., 2017), (Xiong et al., 2017)	Efficient indexing for semantic search, Challenges of semantic searching, Subject naming, cultural changes affecting search engine performance	2.3,2.8.1, 2.8.2
(Clark, 2013), (Riggs, 1981), (Snow et al., 2008), (Royo et al., 2005), (Mestrovic and Calì, 2017)	Ambiguous interpretation of terms, search engine challenges, justification of new concepts in a research discpline, syntactic normalization, word sense disambiguation using ontologies, natural language challenges	2.8.3
(Luff et al., 2015b), (Doty and Glick, 1994), (Gregor, 2006), (Given, 2008)	NCRM Typology (vocabulary) label/tags classification, Broader/narrower categorization of terms, specificity and generality of terms assignment using connected/descriptors terms in annotation, meaningfulness of typology category label	2.9
(Wilson, 2000), (Andersen, 2012), (Cleverley and Burnett, 2015), (Fernández et al., 2011)	Web search vs. enterprise search, exploratory vs. lookup searches based on KOS, Limitation of semantic searching	2.5.1, 2.7.4, 2.8.1

Table 2.4: Summary of citations along with corresponding topics cited in this chapter

Chapter 3

Review of related & relevant literature

As outlined in the previous chapter, traditional web search has led users towards obtaining search results based on lexical string matches in web documents. The semantic analysis of documents and the resultant representative metadata is what actually augments the keywords-based search by enabling the retrieval system to get and rank precise results against a user query. (Ranwez et al., 2013) explains that keyword-based retrieval relies on an exact match, an approximate match or a string distance between words within documents and query indexing. When a query is submitted, the lexical retrieval system will retrieve documents indexed by exact query keywords or some of their lexical variations e.g. *tumorous* instead of *tumour*; thereby missing documents having keyword synonyms in their indexing e.g. *carcinoma* instead of *tumour*. In addition to the synonymy problem, keywords-based retrieval also fails to consider various kinds of semantic relationship between words i.e. *hyponyms, hypernyms.*

The dimensions of Knowledge Organization Systems (KOS) and the role of typologies, taxonomies and ontologies have been explored in the context of knowledge representation and mass semantic annotation of web content. Various challenges were highlighted affecting the performance of search engines due to fast-changing terminological and conceptual evolution in particular domain of interests. The architecture of the NCRM typology was presented as a case study, showing temporal terms evolution over time. It was further argued that it can be employed (through the support of designated research community) for semantic annotation of content and information retrieval in a particular domain of interest.

This chapter, reviews research carried out in 3 main areas i.e. (1) automatic semantic annotation and retrieval e.g. LoD or ontology-based (2) crowd-sourced or manual semantic annotation and search results retrieval and (3) analysis of Knowledge Extraction techniques aimed at enhanced retrieval of relevant search results. Social and non-social annotation platforms are also considered as part of the discussion aimed at sustainable automatic and manual semantic annotation.

3.1 Web searching: ontology perspective

It is recognised that some substantial work has already been done where the emphasis has been on collecting; storing and maintaining web resources in multidisciplinary web repositories. Ontologies have been valuable in knowledge extraction technologies, especially in the aggregation of knowledge from unstructured documents (Pech et al., 2017). However, searching across research repositories of different disciplines remains an open challenge. Fernández et al. (2011) highlights the limitations of keywords-based models and proposes Ontology-based information retrieval by capitalising on SW. Looking into the context of repositories and web archives or any other website hosting a set of online resources or a number of archived websites, changes the semantic searching perspectives. For example, web archives contain complex collections of materials on various subjects that can serve distinct communities, including social scientists and/or historians (Wu et al., 2007). The contextual and classification information of these collections, which are essential, as Wu et al. (2007) argue, are not made apparent or taken into account at the time of ranking and retrieving the search results. The reason for this is that much of the information is buried deep within the archives and unlikely to be discovered. The context in which users use a web repository is different to using a typical search engine in that users are largely interested in the contextual relationship of various collections or web pages corresponding to a time line in order to understand the past and inform their current or future research practices. Such is one of the main purposes of using web archives which would be challenging to achieve through the exploitation of keywords-based or ontology-tuples-based Boolean search model.

Earlier research has been focusing on user's query expansion-based searching and ontology-based information retrieval model proposed by [Sicilia et al. (2007), Fernández et al. (2011), (Chauhan et al., 2013)], ontology-extension model based on adding further classes to the root ontology (Georgiev et al., 2013), ontology classes/properties matching between LoD cloud datasets (DBPedia[1], Freebase[2], Factforge[3] etc.) and domain independent ontology like PROTON[4] (Damova et al., 2010). However, Mukherjee et al. (2014) note that the automatic query expansion system allows too little opportunity for a user to participate in the query development process. In addition the level of complexity and time it takes to refine takes to refine the classes and their relationship with external sources of data (keeping in view concepts disambiguation, over linking, word sense and terms stemming), it proves to be of less benefit to the online users, searching the scientific web resource repositories. Focused crawling of tagged web resources, using domain ontology is another approach, proposed by Bedi et al. (2013). However, the query term expansion at search time, relies on the richness of domain-specific concept ontology. All such approaches have a tendency towards distorting actual user queries (Shabanzadeh et al., 2010) thus turning the terms in queries ambiguous leading to the retrieval of less relevant search results. On the other hand, De Virgilio (2011) proposes key phrase extraction based on semantic blocks which entails pre-selecting blocks of information that have higher coherence in terms of extracting the most meaningful key terms from a web page. That step is then followed by discovering the concepts in semantic knowledge base (e.g. DBPedia) based on the identified terms in a web page. Such an approach further complicates the domain-ontology-based entity and concepts extraction (discussed above); by adding another assumption that a more coherent space in a web page or web documents will be pre-selected before annotating the content inside.

3.1.1 Ontology-based information retrieval

Ranwez et al. (2013) describe an ontology-based retrieval system as a hybrid when it manages document indexes of different granularities (ontology based and keyword based) during indexing and matching processes or during the results presentation stage. This methodology is usually true for many semantic search retrieval systems when they incorporate some sort of knowledge representation structure based on a domain ontology formalism. A domain ontology (or domain-specific

[1] DBpedia (RDFized version of Wikipedia) is a project aiming to extract structured content from the information created in the Wikipedia project. This structured information is made available on the World Wide Web- See https://wiki.dbpedia.org/

[2] All freebase.com URLs, as of now, redirects to the Google developer page, where data dump of Freebase data is available. Google acquired Freebase.com in 2014- See https://arxiv.org/pdf/1805.03885.pdf

[3] http://factforge.net/

[4] PROTo ONtology (PROTON) is designed (by Ontotext) as a lightweight upper-level ontology for use in Knowledge Management and Semantic Web applications. Available at http://ontotext.com/documents/proton/Proton-Ver3.0B.pdf

ontology) represents a set of concepts which are specific to a domain and the relationships among these concepts (Ou et al., 2008). Moreover, in terms of user interface, a considerable amount of semantic research has been focusing on natural language interfaces querying ontologies, which is different from querying aimed at retrieving semantic annotations and associated documents (Bontcheva et al., 2014). The latter entails retrieving the actual instances of information matching the ontology classes from the KB with the user's query. Another related issue is that, in many cases, semantic search systems suffer from the lack of optimal semantic annotation of content in web documents; due to using a small set of predefined domain ontologies and data sets (Fernández et al., 2011). Using DBPedia spotlight discussed in Mendes et al. (2011), for instance, assumes that users should be able to opt for preferred or alternative labels while searching for things in the DBPedia spotlight web application. The author understands, that such assumptions compromise the soundness of semantic data-based search application as the majority of users still prefer to use free-text keywords without pre-specifying advanced search options Sicilia et al. (2007). As illustrated in Figure 3.1, the classical keywords-based retrieval (e.g. Figure 2.1) has been evolved into an ontology-based semantic KB retrieval system (Castells et al., 2007). The system takes *RDQL query* as as in input which is generated from the users' natural language keywords via *Query UI* interface. The query is executed against the KB which returns a list of tuples that satisfy the query. Castells et al. (2007) describe the retrieval as purely Boolean-based in that the returned instances must strictly hold all the conditions in the formal query. The documents that are finally returned are annotated with the instances retrieved in the previous step. The documents are then presented to the user ranked according to their individual relevance score. The obvious problem

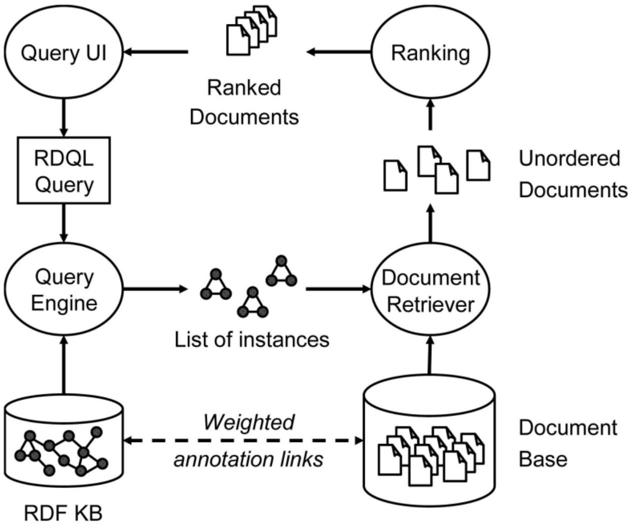


Figure 3.1: A typical overview of the ontology-based IR (Castells et al., 2007) model

with such an approach is the low *recall* (not enough results are returned) due to to strict formal condition in the query for the sake of high *precision* (subject to instance availability in the KB). The dynamicity and evolution of ontologies are other issues, as noted by Alba et al. (2017), which pose major challenges to all applications that rely on SW technologies. Alba et al. (2017) has amalgamated LoD bootstrapping and human-in-the-loop to optimise multi-lingual concept annotation and extraction from unstructured text. These and related studies, further highlight the the need for examining other dimensions like *annotatable* document types in a corpus, ontology vs. document annotation/population for retrieval, in order to realise the true potential of ontology-based annotation and retrieval.

3.1.2 Contemporary search engines & Knowledge Bases (KB)

Over the past few years, major web search engines have introduced *knowledge bases* to offer popular facts about people, places and things in the search results (Bi et al., 2015). Gonzalo et al. (2014) claim that significant progress has been made in research on what is termed *semantic matching* in web search, question answering, online advertisement, cross-language information retrieval and other tasks. They note that the web search, supported by advancement in Machine Learning (ML) and other techniques still heavily rely on the term-based approach where relevance is calculated on the basis of the degree of matching between query terms and document terms. This issue, they think, becomes further complicated due to query-document mismatches when *searcher* and *author* use different terms caused by the changing nature of human languages.

Bi et al. (2015) further elaborate that the major search engines recommend relating entities based on their similarities to the main entity which the user searches for. However, (Bi et al., 2015) use the *user feedback* approach for determining the relatedness of recommended entities which further draws value from users' click logs. The user click is interpreted as relevant (*positive observation* following a click on the link in search results, while non-clicked results are not considered in determining the recommended relevant result selection. Such approaches in most cases rely upon the clicking behaviour of users and do not necessarily address the natural language text in a typical query aimed at *domain-specific* web searching.

This phenomenon has been witnessed in Google Analytic (GA) *Bounce rate* measurement which is one of the behaviour attributes of GA for monitoring the usage of a particular website. The bounce rate for a particular web page, for example, increases when a user lands on that page following a click (on one of top 10 search results) but leaves it immediately after a quick scan if the discovery does not match with the requirements. The gradual decrease in bounce rate represents users' interest in that page as they start staying on that page for a longer period of time. A large number of keywords-based query searches in Google search sometimes leads to the retrieval of a particular web page having content related to the users' keywords. But a proportionally higher bounce rate in percentage terms, for that page, represents the worthlessness of the page's content (in terms of relevance with users' queries) which is manifested in the form of immediate exist of users form that page. Interpretation of this phenomenon in the work of Bi et al. (2015) would be a challenge especially when the query being searched does not contain related entities in the query text. In addition, the change in users' interests in the future or the typical *cold-start* issue emanating from dealing with new users are likely to impact the relevance criteria and documents ranking proposed by (Bi et al., 2015).

Cameron et al. (2014) note, while describing their hybrid approach to semantic searching, that semantic search has gained credibility due to structured ontology-based KBs but there is often misalignment between the information needs of users and the searchable knowledge contained in those KBs. Domain ontologies surely provide means for interpreting some elements of complex information needs, but not all aspects of such needs (Fernández et al., 2011) are considered, especially when it is important to sustain an infrastructure aimed at continual semantic annotation and enhanced search results retrieval. Castells et al. (2007) views semantic search as a tool that gets formal ontology-based queries in the form of RDQL[5], SPARQL[6] from a client (human and/or machine), executes them against a KB and returns tuples of ontology values that satisfy the query. The authors observe that these kinds of techniques typically use the Boolean search models based on an ideal view of the information space as consisting of non-ambiguous, non-redundant, formal pieces of ontological knowledge which reduces the IR process to a data retrieval task. This assumption is quite central to many web-based semantic search results retrieval systems as such models assume that the content is somehow fully represented by the domain-specific ontology or common sense upper level ontology for retrieval purposes. An upper ontology describes common concepts that are generally applicable across a wide range of domains. Several standardized upper ontologies are available for public use, such as WordNet, OpenCyc etc. (Ou et al., 2008). Butt et al. (2015), however, distribute retrieved result types into three categories : *document-centric, entity-centric and relation-centric*, to facilitate users in data exploration. The complexity of heterogeneous unstructured text demands a more sophisticated, flexible, sustainable and 'easy to implement' solution that would meet the information needs of contemporary users via the web-based search results retrieval systems.

[5] RDQL—A Query Language for RDF," W3C member submission, http://www.w3.org/Submission/RDQL

[6] "SPARQL Query Language for RDF," W3C working draft, http://www.w3.org/TR/rdf-sparql-query

3.2 Ontology-based annotation and retrieval: *KIM platform development*

Ontology-based searching has been discussed in the preceding section with a view to information searching in web repositories. In this section, an existing annotation and indexing platform called KIM (Knowledge & Information Management)[7] is elaborated to further highlight the contribution of the SW community towards semantic searching. KIM platform is also discussed here to further highlight the differences between ontology-based and LoD-based knowledge representation approaches. KIM is used for ontology-based annotation and retrieval (Rujiang and Xiaoyue, 2010). However, integrating the built in KIM ontology (PROTON) with domain specific ontology followed by *gazetteer-based annotation* requires huge efforts both on the part of developers and ontology designers. Popov et al. (2004) and Popov et al. (2003) have used the KIM platform to model various lexical resources in the ontology such as `currency, dates, abbreviations` which were used for document annotation and subsequent entity searching. They further use *pattern-matching grammar* based on GATE[8], which recognises relations in text, by gleaning entity associations from predicates in the ontology schema. Adopting the KIM-based model, for semantic annotation and subsequent search results retrieval, may however lead to more "un-instantiable" and ambiguous concepts and entities. Those new entities could only be incorporated in the semantic annotation once the common sense ontology called PROTON is extended to include the new terms and concepts. KIM initially relies on PROTON ontology which is extensible to include the necessary domain knowledge depending on the required conceptualisation granularity (Georgiev et al., 2013).

The issues with such approach (in terms of semantic annotation of web content aimed at web searching) (Georgiev et al., 2013), based on the author's earlier experimental work (Khan et al., 2013) include but are not limited to:

- Evolving an infrastructure which includes a full-fledged repository called OWLIM[9] repository as a KB

[7] KIM provides a semantic service platform architecture and applications on this framework, including: Web content semi-automatic semantic annotation, ontology deployment, content-based semantic indexing, retrieval and knowledge navigation and knowledge

[8] General Architecture for Text Engineering. Available at https://gate.ac.uk/

[9] OWLIM is a semantic repository or semantic publishing platform (now known as GraphDB) for storing and manipulating huge quantities of RDF data and is made up of three components 1. RDF database, 2. Inference engine and 3. Query engine. Available at https://www.w3.org/2001/sw/wiki/GraphDB and https://goo.gl/ZKnY8y

- Mapping of LoD ontologies like DBPedia, Geonames, Freebase etc. to KIM's common sense PROTON ontology for high quality reasoning and consistency of the modelled knowledge in the repository.
- Internal mapping arrangements between DBPedia, Freebase, Geonames[10] to resolve duplicated information in the three datasets
- Correction of the ambiguous mappings by the curating specialists, as a result of automatic PROTON ontology based annotation.

The task of simultaneous mapping of concepts in the PROTON ontology (last bullet point above) to LoD-based datasets, is too costly to sustain. The intervention of a specialist annotator in the semantic annotation of online resources, especially in an academic environment, would be unrealistic to sustain. In such an environment, funding for maintaining any particular online resources tends to cease after a stipulated time period as indicated in Figure 1.1. The time-critical nature of financial resources in academic disciplines necessitates a low-cost sustainable model aimed at ensuring long term access to online resources in a repository with a view to benefiting the research community in general. In that situation, maintaining such a framework for online knowledge delivery will pose an enormous challenge. Using the KIM platform, Popov et al. (2004) evaluate the precision and recall of annotation types rather than actual results of semantic search. They have used the document corpus that contains articles on news and media which implies that in order to tune in the KIM's semantic publishing architecture to incorporate multi-disciplinary web content in a repository website, another round of mapping and classes/properties disambiguation might be needed to refine the semantic instantiation (semantic enrichment) of newer or future content in a KB.

3.2.1 KIM vs. Client-Server platforms

The KIM's generalised semantic publishing architecture has been given in Figure A.1 in Appendix A. It is noted that some features of this approach adopted by the KIM are similar to the framework proposed here, which is designed to:

- Implement the annotation, indexing and retrieval in a local repository in a client-server environment instead of using OWLIM
- Use the LoD-based semantic annotation via industrial strength Keywords, Concepts, Entities APIs for seamless and sustainable semantic annotation

[10] The GeoNames Ontology makes it possible to add geospatial semantic information to content in a website on the World Wide Web. Last accessed on 20/09/2018 at http://www.geonames.org/ontology/documentation.html

which stand in contrast with the KIM's *common sense* PROTON ontology which demands mapping of classes to the publicly available ontologies like Freebase, GeoNames, DBPedia etc.

- Achieve optimal retrieval of relevant search results using the crowd-supported typology-based semantic annotation of web documents in web repositories on sustainable basis while exploring those repositories via their browsers. The professional curator (KIM) vs. Crowd (proposed here) element improves cost effectiveness although it is recognised that crowd-annotating large sections of a typical web repository would be time consuming and still incur expenses, however small they may be.

However, the author notes that, KIM platform-based search application is still far from implementation in a typical client server architecture, as reported in (Khan et al., 2013). Nonetheless, it is recognised that the KIM's domain-specific annotation and search results retrieval share the same goal i.e. to dynamically annotate the online resources in multi-disciplinary-repositories to sustain the performance of search results retrieval systems over time.

3.3 LoD-based annotations vs. domain-specific ontology

This section presents LoD-based and ontology-based annotation approaches with a focus mainly on the semantic representation tools and approaches for knowledge extraction from text. Most automatic semantic annotation tools exploit ontology-based Information Extraction (IE) to extract information by matching the extracted information with domain ontology (De Virgilio, 2011). Pech et al. (2017) present a novel semantic annotation approach based on ontologies for the improvement of information search in unstructured documents, using the similarity of entities of an ontology. The expansion of the web and the increasing engagement of web users throughout the world have given rise to a need for mapping the archival description metadata from bibliographic (author, creator, title, etc.) and/or physical (size, shape, material, etc.) descriptions to more meaningful ontology-based semantic metadata (Zervanou et al., 2011). However the acquisition of semantic metadata through the development of automatic annotation tools is still a major challenge for the semantic web community whose emphasis until recently has been on ontology-based annotations (Gagnon et al., 2013). They further note that recent trends have emerged, with the development of semantic annotators, that are based on the Linked Open Data (LOD) cloud, which extract/link entities and concepts in a given set of documents to established LOD datasets like

DBPedia, Freebase and Yago[11] etc. For example, the Google's Knowledge Graph is powered (partly) by the Freebase linked dataset, which is now being used by Google (as an LoD enrichment dataset) to enhance its search results (Subhashree et al., 2018). Ristoski and Paulheim (2016) note that domain ontologies have been impressive in representing knowledge in a specific domain and knowledge discovery. However, Ristoski and Paulheim (2016) put a stronger focus on the usage of Linked Open Data (LOD) in the process of knowledge discovery, which represents a publicly available interlinked collection of datasets from various topical domains (unlike single domain). Furthermore, Bizer et al. (2009b) argue that the DBpedia KB can be used in the context of classic Web search engines, to relate search terms to entities and to improve search results based on DBpedia's conceptual structure. DBpedia (accessible via LoD) provides approximately 4.7 billion pieces of information and covers multiple domains such as geographic information, people, companies, films, music, genes, drugs, books, and scientific publications (Bizer et al., 2009b). A very interesting work has recently been done by Alobaidi et al. (2018), which has extensively exploited the LoD framework for automatic bio-medical ontology generation. Alobaidi et al. (2018) have cited *complexity* and *availability of ontology experts*, as the primary reasons for using LoD-based sources instead of building a domain-specific ontology, for concepts and relationship extraction.

Another fact worth highlighting is that most domain ontology construction methods do not hold lexical information (in entirety) from which its concepts are taken. The formalism used to represent an ontology, such as OWL, focuses on intrinsic description of concepts, property classes and the logical constraints on them. The domain-specific ontology has to be supplemented with lexical resources so as to be able to identify document passages that are related to domain ontology concepts (Ranwez et al., 2013). Given the limitations identified by Fernández et al. (2011), a more flexible approach is proposed here to semantic annotation of web pages containing heterogeneous content. Supervised approaches for semantic annotations do not address the issue of scalability and precision as they rely on a specific set of extraction rules learnt from a pre-defined knowledge base, based on a specific domain ontology (Sánchez et al., 2011). Ristoski and Paulheim (2016) present a very interesting study on the use of ontology and LoD data sources (as background knowledge) for data mining and interpretation of entities, semantic relations and patterns in a large amount of text documents. Authors like Kamnardsiri et al. (2013) explain a use case involving a *recommender system* which aims to change the way people filter out large amounts of information due to the

[11] YAGO is a huge knowledge base, derived from Wikipedia, WordNet and GeoNames, developed under supervision of Max Planck Institue in Saarbr Acken(Boiński and Ambrożewicz, 2017). Available at https://goo.gl/dyAqcy

exceptional growth in web-based online resources. They further elaborate that the introduction of LoD-based recommender systems are an emerging research area that extensively uses LoD as background knowledge for extracting useful data mining features that could improve recommendation results. It has been argued in the study((Kamnardsiri et al., 2013)) that the LoD can improve recommender systems towards a better understanding and representation of user preferences, item features and the contextual signs with which they deal. Ristoski and Paulheim (2016) present an interesting summary of other research studies on the use of LoD and ontologies in the context of data mining in Table 3.1. The ontology-LoD comparison drawn in Table 3.1 clearly shows that with the advent and growth of LoD, information from the SW can be used beneficially in data mining, LoD-based semantic indexing and search results retrieval. The *reasoning* feature has been a core selling point of the SW for years but Ristoski and Paulheim (2016) note that they are rarely combined with data mining and knowledge discovery processes. They also claim that the types of domain and problems studied so far provide further support to the hypothesis in this thesis which aims to extensively use LoD in document annotation and semantic searching for sustainable search results retrieval. Given that the LoD knowledge representation sources have been recommended by several studies in Table 3.1 where the complexity of ontology still remains very high, it would be an ideal solution for semantic indexing and sustainable searching.

Approach	Domain problem	Ontology complexity	Reasoning	LoD Links	Semantics
[(Paulheim, 2012a)]	Sociology, Economy	High	No	Yes	Yes
[(Paulheim, 2012b), (Ristoski and Paulheim, 2013), (Ristoski and Paulheim, 2015)]	Statistics	High	No	Yes	Yes
[(d'Aquin and Jay, 2013),(Jay and d'Aquin, 2013)]	Students, Medicine	High	No	Yes	Yes
[(Tiddi, 2013), (Tiddi et al., 2013)]	Books, Publications	High	No	Yes	No
[(Tiddi et al., 2014a), (Tiddi et al., 2014b)]	Education, Publication	High	No	Yes	No
[(Vavpetič et al., 2013)]	Finance	High	No	Yes	No

Table 3.1: Summary of approaches used in the data mining step using LoD data sources and ontologies (Ristoski and Paulheim, 2016) (Table 5)

3.4 Knowledge extraction tools analysis

In a landscape analysis of Knowledge Extraction Tools (KET), Gangemi (2013a) observes that Knowledge Extraction (KE) from text has become a key semantic technology and a key to the SW. The term knowledge extraction is used to refer to the semantic annotation of documents to extract meaningful entities, concepts and topics etc. from text in documents. Furthermore, Gangemi et al. (2017) highlight the significance of transforming natural language text into formal structured knowledge and present a comparison of knowledge extraction tools including Alchemy API[12]. Gangemi (2013a) claims that the limitation of KE stems from the fact that it was initially limited to the SW community, which preferred to concentrate on manual design of ontologies as a seal of quality. However, he claims, citing (Bizer et al., 2009b), that things started changing after Linked Data bootstrapping, provided by DBPedia and the consequent need for a substantial population of knowledge bases, schema induction from data, natural language access to structured data and in general all applications that make joint exploitation of structured and unstructured content.

The selected tools, for comparison ((Gangemi et al., 2017)), include (but are not limited to) AIDA[13], Alchemy[14], DB Spotlight[15], FRED[16], NERD[17], Open Calais[18], Semiosearch[19], Wikimeta[20], and Zemanta[21]. The author is also interested to analyse the techniques exploited during this analysis, which entails mapping of NLP basic tasks to SW ones and to assess the soundness of the Alchemy API functionalities in accomplishing those tasks. Table 3.2 clearly shows that Alchemy API has done well in terms of *topics(topic extraction), NE (Named Entity),NER (Named Entity Recognition), NE-RS (NE & Resolution) TE(Terminology Extraction), TE-RS(TE & Resolution) Senses* and *Relationships.* There are other attributes, which can be used as metrics, to assess the performance. For example, the Alchemy service links entities to the LoD dataset while Open Calais, on the other hand, links expressions in text to Open Calais ontology: thus limiting the scope of entity and/or concepts when it comes to searching. It is understood that unlike the majority of tools, which provide machine interpretable RDF encoding, Alchemy

[12] This API will be retired in March 2018 and the equivalent of Alchemy API is IBM Watson API called Natural Language Understanding which can be accessed at https://www.ibm.com/watson/services/natural-language-understanding-3/

[13] http://www.mpi-inf.mpg.de/yago-naga/aida

[14] https://www.ibm.com/watson/services/natural-language-understanding/

[15] http://demo.dbpedia-spotlight.org/

[16] http://wit.istc.cnr.it/stlab-tools/fred

[17] http://nerd.eurecom.fr/

[18] http://www.opencalais.com/

[19] http://wit.istc.cnr.it/stlab-tools/wikifier/

[20] https://www.w3.org/2001/sw/wiki/Wikimeta

[21] http://www.zemanta.com/blog/

Tool	Topics	NER	NE-RS	TE	TE-RS	Senses	Rel
AIDA	-	+	+	-	-	+	-
Alchemy	+	+	-	+	-	+	+
DB Spotlight	-	+	+	-	-	+	-
FRED	-	+	+	+	+	+	+
NERD	-	+	+	-	-	+	-
Open Calais	+	+	-	-	-	+	-
Semiosearch	-	-	+	-	+	-	-
Wikimeta	-	+	-	+	+	+	-
Zemanta	-	+	-	-	-	-	-

Table 3.2: Task-based analysis of contemporary knowledge extraction or semantic annotation tools with a focus on Alchemy "API vs. the rest" (Gangemi, 2013a), (Gangemi et al., 2017)

API provides REST service to populate the knowledge base. However, the extracted knowledge can be turned into RDF triples as a separate activity which is briefly described in Chapter 9 of this thesis. Another related work is [Rizzo et al. (2012a), Tab.1], where most of the above mentioned tools have been analysed on the basis of factual information extraction tasks, with the exception of Alchemy API. Another interesting annotation tool survey [Oliveira and Rocha (2013), Tab. 1] has extensively listed the most commonly referenced tools found in a literature review of the automatic semantic annotation. Rizzo et al. (2014) have presented a thorough study of Named Entity Recognition and disambiguation (NERD) for populating knowledge bases, analysing various NER extractors including Alchemy, DBPedia and OpenCalais. The richness of tool sets represent the interest of the research community in this area which shows that the potential for using such tools for knowledge extraction is greater than ever before.

3.5 Social platforms as annotation tools

In this section, the role of human annotation is highlighted, aimed at refining the automatic semantic annotation of unstructured content in web repositories. An account is given of social annotation platforms, followed by social semantics and social tagging as potential sources of automatic semantic annotation aimed at improved search results retrieval. By analysing the social annotation platforms, a comparison is drawn with the purpose-built expert and crowd annotation environment (*AnnoTagger*) which has been used in a web repository by participants in various experiments as part of this research.

3.5.1 *Social annotations*

Social annotations are emergent useful information that have been used as part of web search in terms of folksonomy[22], visualization and semantic web (Bao et al., 2007) but to the author's knowledge annotations and vocabulary-based tags have not been used as part of a full-fledged semantic indexing and searching environment. Several studies have been conducted to explore social annotation as one of the enabler platforms for implementing semantic annotation and information retrieval. However, the fact that tags are chosen by the user without conforming to a *priori* dictionary, vocabulary ontology or taxonomy (Wu et al., 2006b), means that this approach has not been widely adopted for in-house multidisciplinary search applications. The obvious problem with social annotations is that they are made by a large number of ordinary web users without reference to a pre-defined ontology or classification system such as the case of Delicious (Wu et al., 2006b). Social bookmark services no doubt provide a pragmatic user interface for users to annotate content, but the challenge remains that without clear semantics, social annotations would be of less use for web agents and applications on the Semantic Web (Wu et al., 2006b). All this implies that the web searching platforms, built on top of the social annotation-based platforms, are unlikely to yield relevant search results in the face of an ever-increasing and ephemeral web of information.

3.5.2 Social semantics

Social semantics is another area of research, defined by the interaction and socialisation of users with user-generated content, which in most cases does not conform to a classification system. The tacit agreement on their usage and understanding (due to ease of use), however, make social semantics an important element of web search but they stand in contrast to the more logical semantic web (Halpin, 2013). Furthermore Halpin (2013) elaborates that there are concrete benefits to the tagging approach compared to the Semantic Web's traditional focus on formal ontologies. He further observes that the flexibility of tagging systems is thought to be an asset, which is a categorization process such as expert-generated taxonomies. It is also a fact that to sustain taxonomic or ontological classification, a number of experts are required to review the axiomatic expression of new terms and concepts and then to populate the document corpora with the new instances leading to the creation of implicit relationships between different entities and concepts in the KB.

[22] Folksonomies are increasingly adopted in web systems. These "social taxonomies", which emerge from collaborative tagging, contrast with the formalism and the systematic creation process applied to ontologies (Alves and Santanchè, 2013).

Another issue, worth highlighting here, is that social bookmark service providers enable tagging on web pages, but analysing tags (e.g. free from stop words, redundancy and ambiguity) and mapping them on to the relevant record in a semantic index still remains an issue. The level of noise in the resulting tag clouds does not usually produce a meaningful or semantically related tag cloud that could lead to efficient web-based search results retrieval.

A case in point: DBPedia-based approach: The DBPedia-based approach to tagging and information retrieval is a promising work done by Mirizzia et al. (2010) but the over-reliance on Wikipedia and the fact that all tag suggestions have to come from Wikipedia's labels, categories and abstracts makes their approach somewhat restrictive. The fact that every tag suggestion has to come via a RESTful endpoint from DBPedia only, and not from a domain-specific tagging environment, makes it potentially ineffective in a specific domain. In a domain specific environment, the source of tagging and retrieval may be a local or domain-specific classification system.

3.5.3 Social tagging

It is important to consider the social tagging aspects of annotation to ascertain whether tagging could augment the semantic value of content in web repositories. The popularity of tags especially grew with the advent of social media and networking websites and brought an innovative element to what can generally be referred to as document description in which users describe their own or someone else's documents within the World Wide Web.

Today tags are a dominant force that makes the long and difficult task of searching for information, especially "personal information" within the Web easier (Gerolimos, 2013). In other words, today's online search users have become part of the subject description process where their role has shifted from merely searchers or browsers to contributors. Another aspect of social tagging and one of the resulting products created out of it is folksonomies, which better represents specific target groups' understanding of the world. By enabling users to tag content and the tags later on being combined into the semantic index for searching could address the issue of the changing needs of users, which could give rise to serendipitous discovery of content in online repositories.

Social tagging is reminiscent of Web 2.0 as an enabler of tagging whereby users are able to assign a non-hierarchical term, keyword or phrase to a piece of text in a web page or the entire web page. The use of tagging has given way to user-generated classification of online resources generally called *folksonomy* by Derntl et al. (2011) but also referred to as collaborative tagging or social indexing. The problem with tagging however, arises when mass tagging produces noise or semantic noise (Suchanek et al., 2008), especially when these are subjective or personal tags (Lawson, 2009). Other studies suggest that the skills and approaches of taggers and annotators vary, whereby expert annotators/taggers are more consistent, compared to novice or non-expert taggers or annotators.

Furthermore and most importantly in the context of this research, Lu and Kipp (2014) found that tagging-based noise effectively reduced precision in information retrieval, especially in single word searches. User-generated content on top of the lexical content as a semantic overlay leads to better search results retrieval, but evolving this into a sustainable model comes at a cost, especially when considering non-commercial academic domains and disciplines. However it is proposed that the use of social tagging could potentially be exploited to form contemporary consensual classification system which can then be utilised by the contemporary online resources users to augment the semantics of heterogeneous resources for better search results retrieval.

3.6 Manual semantic annotation & tagging: *Non-social platforms*

In general terms, semantic annotation is conceived as the process of discovering the recognised entities in the text and assigning links to their semantic descriptions, which are usually defined in a KB or knowledge resource (Nebot and Berlanga, 2014). However this section presents the human involvement in the annotation tasks or manual tagging using KO or classification systems aimed at further disambiguating the automatically generated entities/concepts for retrieval purposes. Expert and non-expert approaches on top of the automatic semantic annotation of content are also discussed.

3.6.1 Expert vs. non-expert tagging

The biggest difference between labelling web content with a fixed typology classification and tagging is that tagging on a social media platform does not offer a vocabulary or typology for users to chose from. As a results, the user generated data absorb lots of noise and bias making it too crude to be used for search results ranking and retrieval. Luff et al. (2015a) explain that while tagging-related noise can be reduced as more meaningful tags are added, significant inconsistencies would likely remain from discipline to discipline. Furthermore they highlight the disparity of skills between expert and non-expert taggers. The experts are believed to be more consistent and better than the novice tagger identifying key points in the text being tagged and ignoring unrelated or low relevance materials. The divergence in background and knowledge level of those involved in tagging is highly likely to produce folksonomies of tags full of biases, noise and ambiguity which pose a significant challenge to relevant search results retrieval at the time of searching in web repositories in particular and other domains in general.

In short, it may be worth offering a tagging service to enhance the classification structure of typology for better annotation and retrieval. However, the refinement of categories, sub categories vis-a-vis their related terms, broader and narrower terms and synonyms would still need the intervention of expert annotators to ensure relevance and retrieval are least affected over time.

3.6.2 Ontology-based human annotation

A human-annotated gold standard, or ground truth may be used for training, testing and evaluation of information extraction, however such a process can be expensive and time consuming due to the cost associated with expert annotators and Ontology engineers (Dumitrache et al., 2015). Ontology-supported crowd-sourcing, for example, has been used before in a variety of tasks including for collecting semantic annotation, medical entity extraction (Zhai et al., 2013), (Finin et al., 2010); clustering and disambiguation (Lee et al., 2013), relation extraction (Kondreddi et al., 2014), and ontology evaluation (Noy et al., 2013). However, some issues are still presenting hard challenges to researchers relating to ontology-based semantic annotation, particularly ontology engineering and evaluation of large-scale information retrieval. The lack of experts and willingness to refine the ontology classes, on regular basis, scalability of the crowdsourcing system and large scale disagreement amongst Mechanical Turk crowd-sourced force, are additional issues, which necessitate an alternative to the collaborative annotation and tagging of web content.

3.6.2.1 Risk of crowd-spamming

In domains like Social Sciences or Medical Sciences, filtering out spam users, spam and identical responses and formulating qualification questions would always necessitate the intervention of experts who might have produced the Gold standard questions. The cost of crowd-sourced tasking is another issue which varies from place to place and between crowd-sourced workers. The unrealistic expectations of crowd-sourced workers such as *Mechanical Turker*, as outlined by Kondreddi et al. (2014), where there is no relationship between entities and text but the worker is expected to suggest and phrase the relationships. Such an attitude by the crowd leads to producing potentially ambiguous *term-entity* relevance relationships manifested in the form of retrieved search results based on the annotation metadata.

3.6.3 Sustainability: Ontology development and structured input

Some of the problems with the utilisation of a sustainable ontology-based semantic annotation and subsequent information retrieval, are that (a) SW specific ontology engineering expertise is always needed in every domain of interests especially if it requires a comparative approach, as described by Fernandez-Lopez et al. (2013). (b) the application of the full or short-hand ontology (assuming a short-hand version exists) by the expert or non-expert crowd-sourced taggers/annotators will pose a huge challenge at the time of searching. The later point demands and assumes that the human annotators are aware of the basic *mereology*[23] *axioms* such as *reflexivity, anti-symmetry, transitivity, Disjointness* etc (Fernandez-Lopez et al., 2013). It is imperative, therefore for any mass-human annotation of existing knowledge, that the *"structuredness"* be reduced to *"unstructuredness"* from a usability perspective in order to tap into the collective intelligence of information consumers. The annotation feedback can then be structured by intelligent agents for both human and machine consumption, especially in web searching and ranked information retrieval.

3.7 Crowd-sourced annotation and tagging

In this section, various contemporary web-based annotation tools are discussed which aim to augment automatically created semantic annotation techniques. New knowledge can be continually created by enabling the designated research community to semantically annotate and tag web content in a specific domain particularly, and on the Web generally. A case study is also presented to elicit the potential of *crowd-tagging* in BBC World Service archive aimed at informing today's events with historical perspectives.

In order to fully understand the potential of web-based crowd-sourced annotation, it is necessary to understand the concept of *"crowdsourcing"*. Estellés-Arolas and González-Ladrón-de Guevara (2012) elaborate while citing (Howe, 2008), that the

[23] According to the Stanford Encyclopedia of Philosophy, the abstract study of the relations between parts and wholes is called mereology- Available at https://stanford.library.sydney.edu.au/archives/fall2008/entries/mereology/

word "crowdsourcing" is used for a wide group of activities that take on different forms. The adaptability of crowdsourcing allows it to be an effective and powerful practice, but makes it difficult to define and categorise. Howe (2008) further explains that the term *"Crowdsourcing"* is formed from two words : *crowd*, making reference to the people who participate in the initiatives; and *sourcing*, which refers to a number of procurement practices aimed at finding, evaluating and engaging suppliers of goods and services. In other words, the author affirms that *crowdsourcing* is a business practice that means literally to outsource an activity to the crowd.

Crowd-sourcing techniques are considered as enablers of web-based manual annotation aimed at sustaining relevance in semantic searching. During the course of this research, it has been investigated whether a particular annotation tool could be enabled into a web repository (built on *client-server platform*) and offered to experts and/or crowd to annotate content which could then be used to improve search results retrieval on sustainable basis. The *Annotator*[24] tool has been used in various experiments which is not based on semantic technologies and not supporting classical semantic annotation (RDF based machine readable triples etc.) but could be used as the back-end to harvest expert annotations. These annotations can then be used to modify the scoring criteria for ranking and retrieving search results. Annotation-based scoring and ranking are discussed in Chapters 5 and 8.

3.7.1 Crowdsourcing: enabler of semantic annotation augmentation

Crowdsourcing or crowd-annotation, as it is termed here, involves outsourcing a number of tasks to a distributed group of people online, typically in the form of micro tasks (Chi and Bernstein, 2012). Cataloguing, and now annotating or tagging content, aim to augment the meaning of content in online archives and to support reuse (Raimond et al., 2013). One means of such reuse could be the utilisation of crowd-sourced annotations in retrieving more relevant information in burgeoning repositories of multi-disciplinary research.

Furthermore, Grassi et al. (2013) term it a new knowledge when crowd-sourced annotators create links from the Web of documents (made up of natural language text and digital media) to existing Web of Data (made up of structured resources datasets like DBPedia and Freebase). The aim here is not to delve into the ascertainment of evolving such knowledge bases but rather to investigate the impact of

[24] The Annotator is an open-source JavaScript library and tool that can be added to any web page to make it annotatable. Annotations can have comments, tags, users and more. Available at http://okfnlabs.org/projects/annotator/

crowdsourcing-based semantic annotation on relevant search results retrieval. This research focuses on the means of annotation and tagging in order to evolve a sustainable search results retrieval model in online resources repositories containing content produced by multi-disciplinary research communities.

3.7.1.1 Mechanical Turk vs. custom-built annotators

In further relevant research, aimed at search engine optimisation in ever growing information repositories, Grady and Lease (2010) have used Amazon Mechanical Turk [25] to evaluate search engine accuracy. While Mechanical Turk or MTurk has become popular as a means of obtaining data annotations quickly and inexpensively through Human Intelligence Task (HITs) (Grady and Lease, 2010), the interface design and efficacy of HIT for all sorts of annotation still impacts the quality and quantity of annotations. Another issue in using third party tools such as MTurk for annotation purposes is that annotation of web pages in a repository of research data could be better performed by users having a genuine interest in the content, rather than by the general public. In the former case, using a tool like MTurk and accessing relevant HITs for annotation would therefore be less attractive as compared to built-in annotation tools in a website which offers annotation as functionally integrated with the searching activity. For example, a user genuinely intending to search for a specific query in a repository search is more likely to annotate one of the relevant search results (following the satisfaction of information needs) compared to a *MTueker* who has no genuine information searching needs but only annotate as part of an HIT.

3.7.1.2 BBC World service archive: *A case study*

More recently, BBC World Service Archive [26] has setup an online prototype aimed at tagging BBC legacy programmes spanning over 45 years to help journalists connect today's events with historical programmes to give a historical perspective to viewers and listeners. In addition, this initiative aimed to improve upon the machine-generated ranking of search results generated by BBC hosted search applications. The tagging service has been launched to refine the automatic algorithms (Raimond et al., 2014). However, the algorithms rely heavily on users' interpretation of content and a one-word description of the entire page without conforming to a domain-specific classification, which makes it a relatively ambiguity-prone

[25] http://aws.amazon.com/mturk

[26] http://www.bbc.co.uk/blogs/researchanddevelopment/2012/11/developing-the-world-service-a-1.shtml and http://www.bbc.co.uk/blogs/researchanddevelopment/2012/11/the-world-service-archive-prot.shtml

choice for semantic annotation. A significant part of this archive has been manually catalogued by professional archivists but the coverage of such metadata is not uniform across the BBC's archive (Raimond et al., 2013). Raimond et al. (2013) further claim that little reuse is made of such parts of the BBC archives as there is little or no metadata to help locate content within them. They propose a system comprising of semantic web technologies, automated interlinking, user feedback and data visualisation to ensure the past is connected with the present. However the vocabulary they use for *"upvoting"* and *"downvoting"* (at the footing of Facebook's Like/Unlike) particular items available online, is based on DBPedia (Wikipedia) which limits the efficacy of tagging when it comes to information retrieval in a scientific archive e.g. the ReStore repository. The "*wikification*" process, as Milne and Witten (2008) call it, entails linking topics in unstructured text to a Wikipedia article for better explanation and information enrichment. The downside of this approach is its dependence on Wikipedia-based disambiguation of terms at the time of harvesting annotations. However, such attempts are perceived to be heading in the right direction i.e. involving more interested online users to annotate content in a simple and sustainable manners aimed at better search results retrieval.

3.7.2 Analysis of collaborative(manual) semantic annotation tools

While the role of crowdsourcing has been quite encouraging in semantic search and Information Retrieval studies, many questions still remain as to how crowdsourcing methods can be most effectively and efficiently employed in practice (Lease and Yilmaz, 2012). Some innovative annotation tools have been listed in Table 3.3 that are currently in use to understand and create semantic connection among objects that could be used in search results retrieval processes. Such tools are however different from the likes of *clipboard*[27], *Pinterest*[28] and *Bundler*[29] etc. which aim to help users organise and share online resources.

The *Annotator* tool has been used during experimentation as part of this research and compared with relevant contemporary annotation tools (Grassi et al., 2013) in

[27] https://clipboard.com

[28] http://pinterest.com/

[29] http://bundlr.com/

terms of performance and features, as shown in Table 3.3. The selected comparator tools include Pundit[30], ECMAP[31]. *One Click annotation* [32], CWRC Writer[33], LORE[34] and the *Annotator*.

As Table 3.3 shows various tools including the *Annotator* tool are being used in different capacities for collaborative or manual annotation. The *Annotator* tool is highlighted in grey column in Table 3.3 as the tool has been used extensively during various experiments; to store annotation metadata, obtained through *in-page content annotation* and tagging. A common issue suffered by most of the collaborative annotation tools is the ambiguity of natural language found in the annotations/comments and tags which further affects the accuracy and efficiency in search results retrieval at search time. The main problems using *Pundit* for example, are (a) slightly disproportionate expectation from users to comment, tag different resources using vocabularies and ontologies and (b) using third party tools to visualise the annotations and tags form various users and make sense of the new knowledge created as a semantic overlay on top of the actual content.

[30] http://thepund.it/annotator-web-annotation/

[31] Europeana Connect Media Annotation Prototype (ECMAP) is an online media annotation suite that allows users to augment textual comment linking DBpedia resources (Haslhofer et al., 2010)

[32] One Click Annotator is a WYSIWYG Web editor for enriching content with RDFa annotations (Heese et al., 2010)

[33] The Canadian Writing Research Collaboratory (CWRC) is developing an in-browser text markup editor (CWRCWriter) for use by collaborative scholarly editing projects. Available at http://www.cwrc.ca/projects/infrastructure-projects/technical-projects/cwrc-writer/

[34] LORE (Literature Object Re-use and Exchange) is one of Mozilla Firefox's add ons that enables literary scholars to author, publish and search annotations and compound objects. Available at https://addons.mozilla.org/en-US/firefox/addon/lore/

Parametrs	**Tools**					
	Pundit	**ECMAP**	**One click annotation**	**CWRC Writer**	**LORE**	**Annotator**
Annotation purpose	Generic Web content annotation (special attention to DL)	Generic Web content annotation	Content editing	Content editing	Generic Web content annotation	Generic Web content
Representation of annotation metadata (context)	RDF/OWL OA data model	RDF	RDF	RDF	RDF	OA data model Different data stores-No RDF
Representation of knowledge expressed by the annotation	RDF/OWL Ontologies (Named Graph)	Plain Text, Semantic Tags, geo tagging	RDF	RDF	RDF	Different data stores No RDF
Annotatable resource types	Text-fragments, Images, Image fragments. Prototype for video and video fragment	Text-fragments, Images, Image fragments, video (temporal fragment), maps	Plain Text (copy-/paste from diff. sources	Plain text, TEI xml	Text fragment, images	Text-fragments, Images
Annotation storage	Dedicated annotation server based on Sesame Triplestore (inference)	Dedicated annotation server based on Mongo DB.	Dedicated annotation storage based on Triplestore	Local files	Dedicated annotation server	Server with endpoints (index, create, read, update, delete) No triple store
Annotation search	External application to explore Notebooks and Annotations (Ask the Pund)	Textual search	Faceted and textual search	No	Not clear	Yes (RESTFul search API
Installation in Web pages	Add the Javascript to the page or lauch the Bookmarklet. Configurable using a JSON file.	No	No	No	No	Add the Javascript to the page or launch the Bookmarklet.

Table 3.3: Features-based manual semantic tools qualitative comparison:*Annotator* vs. others (Grassi et al., 2013)

3.7.2.1 Manual annotation tools: *with a focus on searching*

As per the author's understanding based on various focus groups conducted as part of experimentation (detailed in Chapter 6) and other online experiments in this research, the noise level in data becomes higher when one web page is annotated by more than 5 people. Moderating large scale annotation and tags, assigned by non-expert in a specific scholarly discipline, for efficient search results retrieval, would be an enormous activity to engage in. Moreover, without the middle semantic layer i.e. automatically generated LoD-based semantic annotation (which is one of the components of the annotation framework in this thesis) before the collaborative annotation/tagging, the ranking of results will heavily rely on lexical mention and *pundit*-based annotation. The RDF triple model, applied on harvested annotation data, is impressive as it facilitates machine readability and subsequent information retrieval via RESTful API and SPARQL endpoint. However, the author hasn't come across any search results evaluation performed on a set of documents or websites exploiting any of these tools. Therefore, no comment can be made on the retrieval performance of a system which employs *pundit* or any of the above tools as a semantic annotation tool.

Hypothesis is another effort to implement an easy to use web-based annotation tool for collaborative annotation and tagging. The tool is the new face of *Annotator* tool but tends to be more robust and could be used in many different forms. The open-source project[35] aims to provide a conversation layer over the entire web that works everywhere, without needing implementation by any underlying site. Hypothesis offers users a lightweight way to annotate texts line by line and start shareable conversations in the margin. With a free account, a user can turn on annotations on any web page or online PDF using browser *bookmarklet* (Davis, 2016). It is another innovative and practical tool available to online users who may want to organise and maintain the scholarly online resources with a semantic overlay on top of it in the form of comments, tags and collaborative discussion. However, configuration of the tool would be a challenge to dynamically add discipline specific vocabulary and fine tune GUI(Graphical User Interface) of the tool from time to time to meet specific experts annotators requirements. Like *Pundit*, *Hypothesis* would also require the technical team responsible for a specific repository to review the annotation and tags collaboratively exchanged across the website, in order to augment the semantic meaning of existing content aimed at the retrieval of relevant search results in a repository search.

[35] https://web.hypothes.is/about/ (accessed 25/01/2018)

3.8 Summary

Various situations have been considered in which semantic annotation, indexing and searching is performed on web content for better search results retrieval. Ontology-based semantic annotation and retrieval have been playing a pivotal role in making the semantic search a reality despite the rapid growth in information publication, volume and scale.

However, the structure of the heterogeneous web content currently hosted on different platforms especially in web repositories, is too complex to be resolved by a domain-specific ontology-based semantic layer alone or social-annotation oriented collaborative annotation for better searching. Instead, adaptation of the current semantic annotation and searching model, is advocated on sustainable basis, to include more diversified LoD-based semantic annotation of web content in web repositories. That would address the issue of "reliance on one domain ontology" and increase the chances of enriching a KB with new concepts as and when they are introduced.

To cope with the large-scale structuring and classification of content in web repositories, adaptation of the classic vector-space model (VSM) is proposed by adding the crowd-annotation element to the model to influence the search results ranking at the time of searching. The challenges associated with the fast-paced information publications especially in terms of scale, volume and heterogeneity can be addressed by methods focusing on crowdsourcing-based annotation and tagging process on top of automatic processes. This annotation and retrieval framework is further discussed, in terms of methodology and implementation, in Chapters 4 and 5.

Citation	Description	Section
(Wu et al., 2007), (Fernández et al., 2011), (Sicilia et al., 2007), (Chauhan et al., 2013), (Shabanzadeh et al., 2010), (De Virgilio, 2011), (Ranwez et al., 2013), (Bontcheva et al., 2014),	Keywords search limitation, searching in web archives, ontology-based IR, ambiguous queries, Ontology-based phrase extraction, Ontology-based retrieval	3.1
(Castells et al., 2007), (Bi et al., 2015), (Gonzalo et al., 2014),(Cameron et al., 2014)	KB retrieval systems, semantic matching in web search, hybrid semantic search, structured ontology-based KB	3.1.1, 3.1.2
(Hinze et al., 2012), (Wang et al., 2006), (Wang et al., 2009), (Fensel et al., 2008), (Packer, 2011), (Rujiang and Xiaoyue, 2010), (Popov et al., 2004), (Georgiev et al., 2013), (Subhashree et al., 2018)	Ontology development, abundance of ontologies, reasoning/inferencing in ontologies, KIM platform case study, OWLIM, PROTON, LOD enrichment	3.2, 3.2, 3.3
(Bao et al., 2007), (Wu et al., 2006b), (Halpin, 2013), (Mirizzia et al., 2010), (Gerolimos, 2013), (Derntl et al., 2011), (Suchanek et al., 2008), (Lawson, 2009), (Lu and Kipp, 2014)	Social annotation & web search, social semantic annotation/tagging using vocabulary, ontologies, page-ranking based annotation/-tagging, tag folksonomy vs. ontologies, DBpedia-based approach, Noise/bias by social tagging, non-expert/ tagging	3.5
(Nebot and Berlanga, 2014),(Luff et al., 2015b), (Dumitrache et al., 2015), (Zhai et al., 2013), (Lee et al., 2013), (Fernandez-Lopez et al., 2013), (Kondreddi et al., 2014)	Experts vs. non-experts tagging, cost of crowdsourcing-based annotation, Ontology-based manual annotation, crowdsourced-annotation	3.6
(De Virgilio, 2011), (Zervanou et al., 2011), (Gagnon et al., 2013) , (Ranwez et al., 2013), (Sánchez et al., 2011), (Ristoski and Paulheim, 2016), (Kamnardsiri et al., 2013)	LOD vs. domain-speicifc ontology semantic annotation, LoD-based recommnder system, LoD vs ontology in research studies	3.3
(Gangemi, 2013a), (Rizzo et al., 2012a), (Oliveira and Rocha, 2013)	KE from text, KE tools analysis , Alchemy API vs. others analysis	3.4
(Chi and Bernstein, 2012), (Raimond et al., 2013), (Lease and Yilmaz, 2012), (Howe, 2008), (Grassi et al., 2013), (Grady and Lease, 2010), (Davis, 2016)	Crowd-sourced annotation/tagging, crowd-annotation , a source of semantic annotation, CA in optimized search results, Mechanical Turk HIT vs. web-page based tasks, analysis of collaborative annotation tools vs. Annotator, BBC World archive case study	3.7
(Sanderson et al., 2010), (Weiss et al., 2010), (Hiemstra, 2009) , (Fernández et al., 2011)	IR, query-documents retrieval, accuracy & scalability of IR, IR models, VSM	2.2, 2.2.1
(Ranwez et al., 2013), (Jiang et al., 2015), (Verma et al., 2016), (Vargas et al., 2016)	user searching criteria, degree of satisfaction, query auto-completion	2.4

Table 3.4: Summary of references in this chapter

Chapter 4

Designing a sustainable knowledge representation and retrieval framework

In this thesis, it is proposed that by incorporating dynamic LoD-based semantic indexing, enhanced by crowd-sourced annotations, it is possible to address the issues of content heterogeneity, volume of data and terminological obsolescence in multi-disciplinary web repositories. In this approach, in addition to automatic semantic annotation using Alchemy APIs, it is proposed that the knowledge of the designated research community should be exploited to tackle the changes in terminologies and scientific concepts. The crowd-annotation & tagging would therefore play a pivotal role in evolving the terms and concepts in a specific scientific domain e.g. Social Sciences over time, thereby, sustaining the effective *searchability* of research data in long term. In order to enable the research community to augment the contextual and literal meanings of terms and concepts expressed implicitly or explicitly inside web pages, free-text and vocabulary-based terms and concepts are used for labelling and tagging. Figure 4.1 is the detailed version of Figure 1.2, which collectively illustrates heterogeneous content acquisition, annotation, indexing and retrieval.

A vocabulary or typology in this case, in any scientific field is a collection of terms, concepts and terminologies, which contemporary researchers may use to refer to various things in that field. A vocabulary-based term or keyword is part of a particular classification system such as the NCRM Typology, used to classify research and training items for better searching and browsing. The data or research outputs, contained in web repositories, are produced by researchers within disciplinary domains; and they are partially tagged at the time of deposition into a web repository for classification purposes.

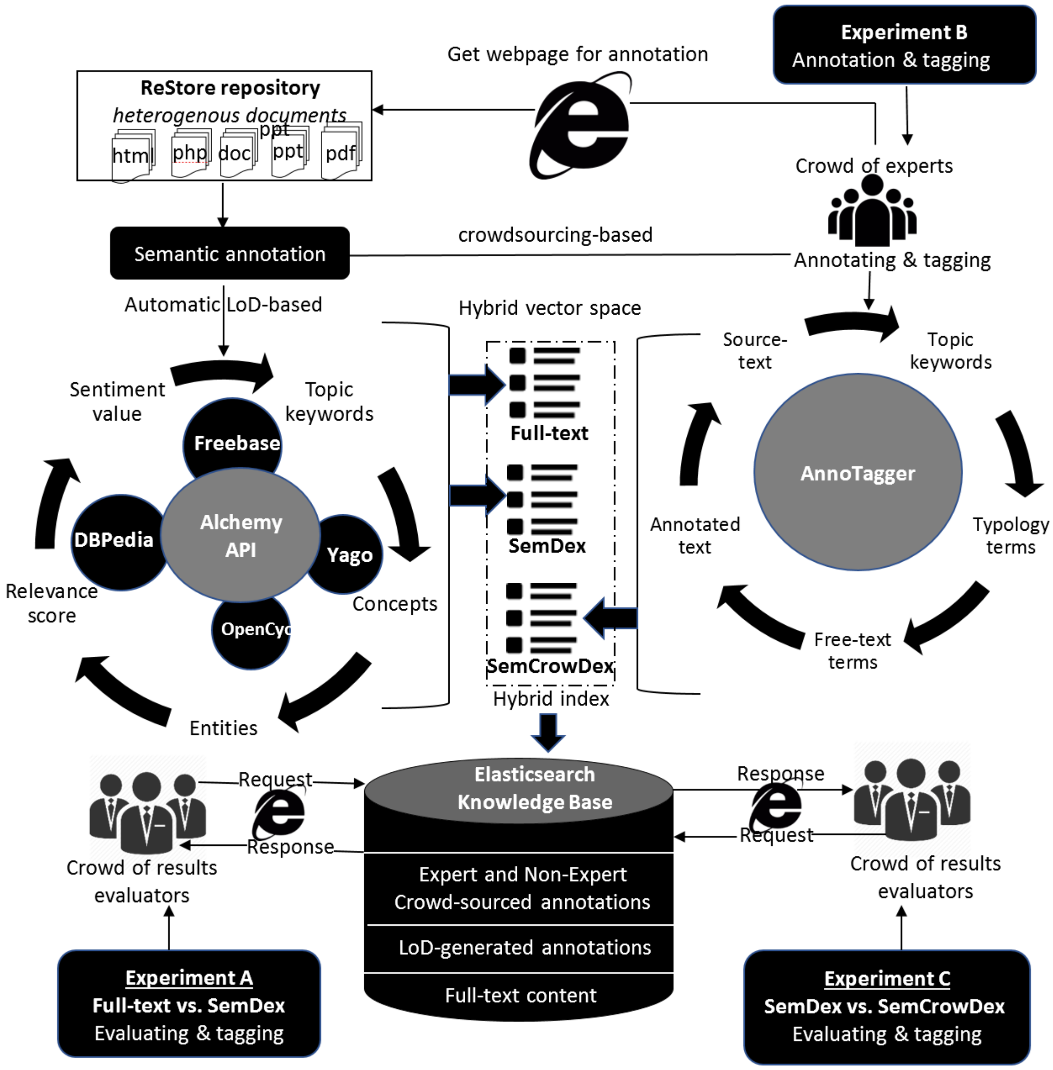


Figure 4.1: Overview of methodology: components of our knowledge representation and searching framework

In this chapter, the focus remains on augmenting the existing content metadata utilising automatic semantic annotation and indexing followed by crowd-annotation (free text and vocabulary annotation techniques). It is also investigated that continually evolving semantic KB in a particular knowledge domain, the relevance score of a set of search results can be optimised against natural text queries using TF.IDF algorithms and Vector Similarity metrics. Figure 4.1 illustrates the overall process flow of a document's journey from the repository to Elasticseach KB via Alchemy API and then the augmentation of each document with manual webpage-based annotation and tagging (using *AnnoTagger*).

This chapter is presented into two parts 1. justification of the tools and technologies, adopted for designing the experimental framework and 2. description and demonstration of the methodological framework, designed to investigate the RQs (Section 1.3) by incorporating participants-based experiments and search results retrieval systems.

4.1 *Alchemy APIs*: A source of LoD-based annotation

Alchemy API has been extensively used as the automatic Information Extraction (IE) tool, which combines machine learning algorithms in order to analyze content, and to extract semantic meta-data and entities e.g. people, organisations, places and more. Alchemy API incorporates LoD data sources namely DBPedia, Freebase, YAGO[1], OpenCyc[2] to extract Keywords, Concepts and Entities from text. The API uses proprietary ontologies whose instances are linked to the above mentioned KBs, through the *owl:sameAs* relationship (Pech et al., 2017). By using the Alchemy API service, the author obtains a Knowledge Base of structured documents which conforms to the data retrieval model elaborated by Fernández et al. (2011). Keywords, entities and concepts are extracted by Alchemy API using three different API services i.e. Keywords, Entity and Concepts. Keyword extraction focuses on topics in hypertext of web pages, Concepts API recognises and extracts high-level concepts that are related to the text in web documents, which may not be explicitly mentioned in the document. Table 4.1 summarises, keywords, concepts and entity extraction, from three documents Doc 1, Doc 2 and Doc3, using Alchemy APIs. Elasticsearch (ES) Custom analyzer has also been implemented for mapping keywords in text to synonyms, descriptors at the time of semantic annotation and indexing. Further details on custom analysers and synonyms-based retrieval are given in Chapter 5.

Manual crowd-sourced annotation is then performed by the expert and non-expert participants as part of annotation experiments which will be detailed later in this section. This approach has been adopted throughout the automatic annotation and indexing experiments and the technical implementation of these APIs as a service will be discussed in Chapter 5. The LoD KB accessible via Alchemy API contains over 200 datasets which span numerous domains such as media, geography, publications and life sciences etc. incorporating several cross-domain datasets Rusu et al. (2011). It is an open source of structured data, which so far has been employed for building Linked Data (LD) browsers, LD search engines and LD

[1] YAGO is a huge linked dataset constructed through automatic extraction from sources such as Wikipedia and WordNet(Subhashree et al., 2018)

[2] OpenCyc contains hundreds of thousands of general knowledge terms organized in a carefully designed ontology (Bollacker et al., 2008). Available at http://opencyc.org/doc/opencycapi

	Doc 1	Doc 2	Doc 3
Sources of text (Full-text)	ESRC website introduces an approach to social science data analysis for both qualitative and quantitative research, using computer software to assist in comparative analysis	Microdata methods in quantitative data analysis	geo-refer in social sciences
Keywords	social science data, quantitative data, comparative analysis	microdata methods, quantitative data analysis	social sciences, geo-refer
Concepts	Qualitative research, Scientific method, Quantitative research	scientific method, quantitative research qualitative research	science, sociology
Entities	ESRC	None	None
Custom-analyzed	**qualitative research**(*map to*) {qualitative research methodology}, **comparative analysis** (*map to*) {qualitative data analysis}	**microdata methods** (*map to*) {data linkage, micro-econometrics}	**geo-refer** (*map to*) {geographical referencing, geo referencing}

Table 4.1: Automatic information extraction using Alchemy API and ES custom analyzer

domain-specific applications such as semantic tagging (Bizer et al., 2009a). A number of web services have recently been developed to extract structured information from text (incorporating LoD) such as Alchemy API[3] , DBPedia Spotlight[4], OpenCalais[5] and Zemanta[6]. The Alchemy framework is used here due to its holistic approach towards text analysis and broad-based training set (250 times larger than Wikipedia) used to model a multi-disciplinary domain. It uses machine learning and natural language parsing algorithms for analyzing web or text-based content for named entity extraction, sense tagging and relationship identification (Gangemi, 2013b). Alchemy API was also one of the best in the performance evaluation review by Rizzo et al. (2012b) where the Alchemy API remained the primary option for NE recognition and over-all precision and recall of NEs and types inferences and URI disambiguation.

4.1.1 Why Alchemy API web service?

One of the most important building blocks of the semantic web is representation of knowledge through annotation of data (identifying entities of interest and linking them together based on inherent features and linked data web resources) in

[3] https://goo.gl/U7mTfc
[4] http://demo.dbpedia-spotlight.org/
[5] http://www.opencalais.com
[6] http://www.zemanta.com

a corpus of documents (HTML, PDF, DOC, etc). Knowledge retrieval, based on knowledge representation, then, determines the accuracy, precision and authenticity of search results retrieval systems. In order to address the research question (RQ2) i.e. obsolescence in scientific terminologies and concepts, it is necessary to employ natural language processing tools, which could take care of the historical as well as contemporary scientific terminologies and concepts in social science research data repositories.
The Alchemy service crawls billions of pages every month thus expanding its knowledge base through which entities and concepts are identified in web documents and linked to various linked data sources of data. In terms of named entity recognition, topic, relation extraction and accurate terminology extraction, the Alchemy API stood out in the knowledge extraction tools analysis carried out by Gangemi (2013a).

4.1.2 Knowledge extraction through Alchemy API

The Alchemy API suite analyses textual content by extracting semantic meta-data: information about people, places, companies, topics, languages, and more. It is a named-entity extraction tool for performing semantic tagging on URLs, HTML files, or text content. The objective is specifically the extraction of scientific concepts from text in web pages, which are essentially abstract ideas being discussed in the text. The text analysis includes keywords (explicit terms), concepts (implicit terms) and entities (explicit) extraction along with relationships (implicit and explicit association) from heterogeneous documents, currently contained in the ReStore repository.

The service exposes the semantic richness hidden in any content using named entity extraction, phrase/term extraction, document categorization, language detection, structured content scraping, and more. The author has used HTML API, and Text API against sets of URLs in ReStore repository, and text from proprietary documents e.g. MS Word, MS Powerpoint and MS Excel etc. The API calls extract content from the requested URL, extracts text from text and performs Entity, Concepts, topical Keywords extraction operations. Algorithm 3 details the implementation of the service in Appendix C. One of the best features, that the Alchemy API (and IBM NLU (Natural Language Understanding[7])) offer is the "*neat*" feature, that attempts to remove advertisement from a retrieved HTML page (Dale, 2018). Such features address the content structure issues in webpage-based scholarly documents, which is not usually the case in PDF-based frameworks

[7] NLU is an upgraded version or replacement of Alchemy API, part of IBM Bluemix cloud services - Available at https://www.ibm.com/watson/services/natural-language-understanding/

like the one proposed by Wu et al. (2015). To ensure access to all content types in the repository for document analysis, crawlers have been used for harvesting static pages, and custom-built scripts for accessing dynamic web pages on two different web servers. Keywords, Concepts and Entities are harvested depending upon the positive sentiment values along with individual relevance score for scoring purposes at the time of searching. The sentiment value filter has been used to ensure the extraction of keywords and entities, which have positive and neutral connotation or feelings associated to them by the public. Sentiment analysis[8] is the process of using Natural language Processing (NLP), statistics, or machine learning methods to extract, identify, or otherwise characterise the sentiment content of a text unit. Three possible sentiment values exist for characterising the sentiment content of a text unit, document or words i.e. *positive, negative and neutral.* The reason for filtering out entities and keywords having *negative* sentiments was to remain focused on only positive and neutral content at document level as well as text level annotation.

Figure 4.1 also shows the two semantic indices i.e. *SemDex* and *SemCrowDex*, which continue to be populated with more and more content as and when it is deposited into the repository web server. *SemDex*, in the hybrid vector space, is populated with automatically generated Keywords, Concepts and Entities and the *SemCrowDex* index contains *SemDex* and the crowd-annotations hence the name *SemCrowDex.* The Hybrid semantic index on the ES cluster now hosts two instances of the data ,one with the crowd annotation index and another with only automatically generated semantic index. The making of a hybrid semantic index can be formalised as follows:

$$SemanticIndex_{hybrid} = \sum_{i=1}^{n} \overrightarrow{doc}_k + \sum_{j=1}^{n} \overrightarrow{doc}_{k,s} + \sum_{k=1}^{n} \overrightarrow{doc}_{k,s,crowd} \tag{4.1}$$

$where\ crowd \in Annotation_{comments,ftTags,vocTags}$

$docs_{i,j,k} \in D$

and $ftTags \leftarrow freetextTags, vocTags \leftarrow vocabularyTags, comments \leftarrow annotationComments$

k, s, on the other hand represent keywords(k) and automatically generated semantic entities (s) extracted from document objects. It is evident in Figure 4.1 that the annotated document is finally indexed in the hybrid ES KB.

[8] https://www.ibm.com/blogs/insights-on-business/government/ibm-bluemix-sentiment-analysis/

4.2 Crowdsourcing-based annotation: *a source of semantic augmentation*

Having a general classification vocabulary of research themes (on which there is a consensus of the research community of the present time) and topics in a particular domain provides a platform for the manual semantic annotation of content in web repositories using knowledge of the concerned research communities.

However, as in Hinze et al. (2012) words, the success of semantic web depends on reaching out to a critical mass of users, creating and consuming semantic content, by employing simple-to-use tools without any complexity. The ease of use of such tools will play a decisive role in motivating a large number of web users to create contemporary metadata of content currently archived in repositories. For example, using the DBPedia resource, linguistics information can be populated in ontologies from time to time e.g. populating the earthquake with new instances such as 2007 Peru earthquake Fernandez et al. (2009) and/or 2005 earthquake in Pakistan etc.

Similarly, if the instance data against the new instances (sub classes and properties) continue to be populated by exploiting the designated research communities (through crowdsourcing), the harmony between linguistic and semantic information will remain stable. That will also contribute to sustaining the performance of search systems in terms of relevant search results retrieval. This research investigates whether utilising other reference sources such as NCRM typology through crowdsourcing techniques, semantic meanings of content can be continually evolved (using contemporary terms) in web repositories of research data. A particular focus here is on the NCRM typology which has been evolved over the years into a widely used classification system in Social Science research methods. The typology was one of the top 10 most downloaded items on the NCRM website for the last several years. Data available from the NCRM website[9] usage stats shows that in 2013 it was retrieved from a wide range of locations within and beyond the UK suggesting a wider use of the classification system. The recently launched search application on the NCRM website[10] and its consistent usage by online users every month, is testimony to the fact that typology-based categorisation of the content in a repository is becoming a popular aid to retrieving relevant search results.

4.2.1 Alchemy API and Crowdsourcing-based extraction: *The nexus*

As stated earlier, Alchemy API draws on several linked data resources in calls to Entity, Concept, and Keywords APIs. By expert crowd-annotation of content on

[9] https://www.ncrm.ac.uk

[10] https://www.ncrm.ac.uk/search

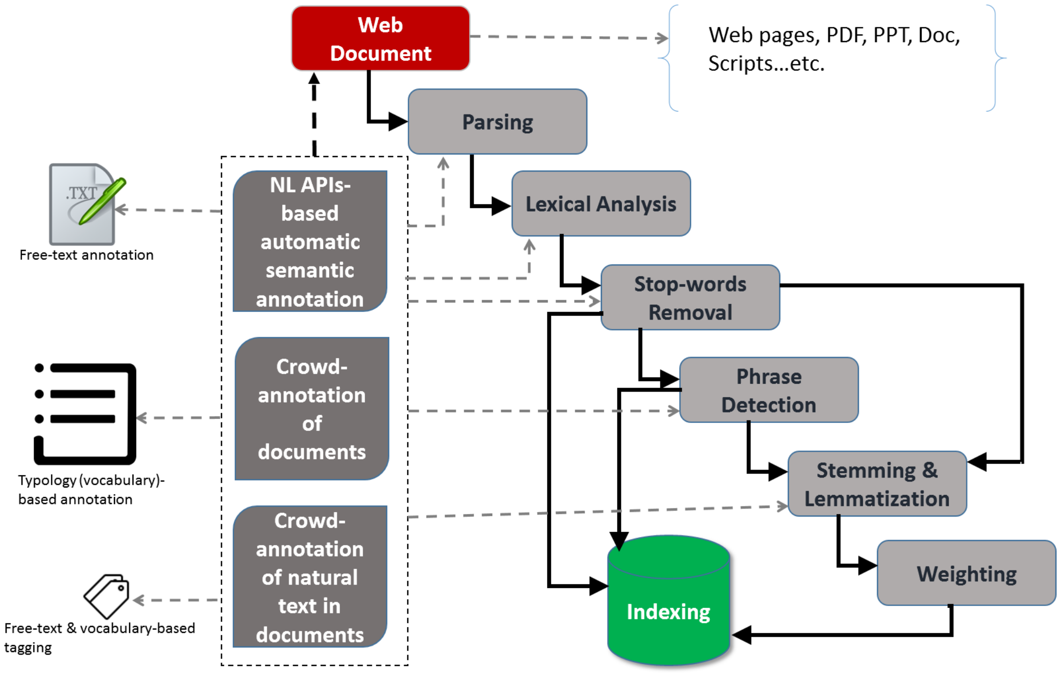
top of automatic semantic annotation, the semantic index can further be enriched over time to augment the contextual value of content in web repositories; so that they remain findable, despite changes in language, terminology and scientific concepts. A custom-built annotation, indexing and searching environment has been developed in a web repository website, and used by expert annotators to annotate web pages using free text and vocabulary terms.

The gap has been bridged between known and unknown relations and between sourced and annotated entities. For example, terms from vocabulary (NCRM Typology) are offered in a free text field expecting the annotators to choose the right classification against a particular search result. Similarly, vocabulary terms have been embedded inside the popular terms on textual level annotation to enable users chose from the established terms and phrases before resorting to formulating their own. That way, the noise generated from uncontrolled terms by different users remains at a minimum level and accuracy in terms of relevance documents retrieval at searching results stage increases.

4.3 Crowd-annotation and tagging methodology

Halpin (2013) elaborates that there are concrete benefits to the tagging approach compared to the Semantic Web's traditional focus on formal ontologies. The flexibility of tagging systems is thought to be an asset, which is a categorisation process as compared to pre-optimized classification process such as expert-generated taxonomies. In a tagging environment, however, users are enabled to order and share data more efficiently than using classification schemes, as associating free text with content in a web page is cognitively simpler than decisions about finding and matching existing categories (Halpin, 2013).

However, in this annotation and tagging framework (*AnnoTagger*), tags as well as prominent vocabulary tags have been supplied in the form of an auto-complete feature attached to the search box which maps users' cognitive thinking at the time of assignment of annotations. A number of annotators have been observed through the course of experimentation (in Exp.C (6.6) while annotating content and almost all made use of the auto-complete feature while they were typing the first few words of their queries. The suggested list of keywords (matching with the keywords typed in the search box) usually kick-started the thinking process of picking/assigning relevant keyword without awareness of whether the keyword was free-text or vocabulary-based (i.e. borrowed from the NCRM Typology). Figure 4.2 outlines the anatomy of semantic indexing of web documents, which entails systematic processes being applied on the content (*textual bits*) of a typical web document. Figure 4.5 presents the ReStore repository-specific view of the



Web Document
Web pages, PDF, PPT, Doc, Scripts...etc.
Parsing
Lexical Analysis
Stop-words Removal
Phrase Detection
Stemming & Lemmatization
Weighting
Indexing
NL APIs-based automatic semantic annotation
Crowd-annotation of documents
Crowd-annotation of natural text in documents
.TXT
Free-text annotation
Typology (vocabulary)-based annotation
Free-text & vocabulary-based tagging

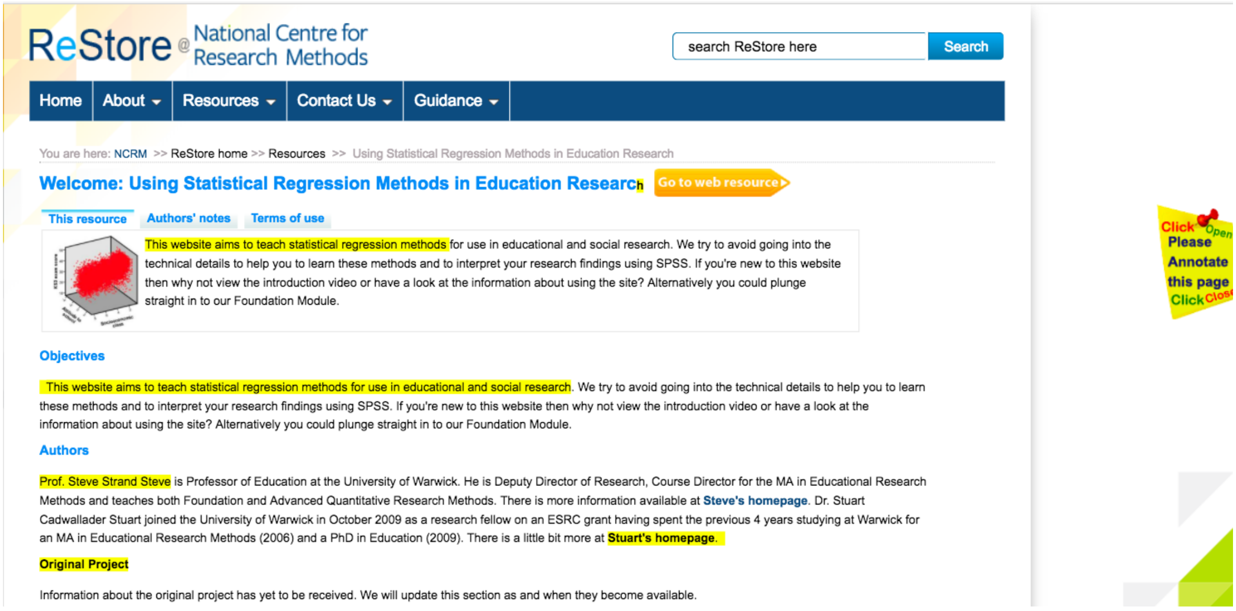
Figure 4.2: The text analysis and indexing stage in the Search Results Retrieval (SRR) system comprising of two layers i.e cascading layer of processes and sources of semantic annotation of documents (stack of 3 boxes)

processes, illustrated in Figure 4.2. The cascading layer of processes is applied on both the intrinsic full-text content of the documents as well as expert-generated annotations/tags created by expert annotators via *AnnoTagger*. The stack of 3 boxes shows the 3 sources of semantic annotation which are used to represent knowledge in each document object for search results retrieval.

4.3.1 Expert-crowd annotation:*Requirements consideration*

The aim is to enable expert annotator to annotate textual content in the ReStore repository by using the purpose-built annotation tool i.e. *AnnoTagger*. This has been considered as a requirement in designing and implementing the user interface of the *AnnoTagger* because (a) a non-expert user (*crowd*) is able to click a *star-rating* button, or tag the search result, but they are not subject experts (e.g Post-doc researcher (expert) vs. Master or Bachelor student (non-expert)) (b) expert users annotate content based on their research background and command on the subject including *taxonomic classification* and tagging categories. They are able to annotate the text inside a web page and tag the entire web page using free-text and vocabulary tags populated in the *AnnoTagger*. In order to streamline the usage of *AnnoTagger* in terms of *ease of use*, and *ease of understanding*, various pilot experiments were run before the actual ones for optimal feedback of the participants in the actual experiments. These requirements have further been analysed in the light of the following requirements compiled by Heese et al. (2010).

1. *Intuitive user interface.* The *AnnoTagger* is a very simple in-page plugin attached to the *Annotator* tool in the back-end for annotation storage and embedded in the web page front-end for providing an intuitive user interface to the user. The two dimensions of textual and page-level annotations adopt the "Two Click" approach: first clicking on the *piece of text* inside the page then annotate and secondly clicking on the page-level annotation handle in order to complete the task. Both clicks enable experts/non-experts to access controlled/uncontrolled terms for semantic annotation using auto-complete feature.



You are here: NCRM >> ReStore home >> Resources >> Using Statistical Regression Methods in Education Research

Welcome: Using Statistical Regression Methods in Education Research

This website aims to teach statistical regression methods for use in educational and social research. We try to avoid going into the technical details to help you to learn these methods and to interpret your research findings using SPSS. If you're new to this website then why not view the introduction video or have a look at the information about using the site? Alternatively you could plunge straight in to our Foundation Module.

Objectives

This website aims to teach statistical regression methods for use in educational and social research. We try to avoid going into the technical details to help you to learn these methods and to interpret your research findings using SPSS. If you're new to this website then why not view the introduction video or have a look at the information about using the site? Alternatively you could plunge straight in to our Foundation Module.

Authors

Prof. Steve Strand Steve is Professor of Education at the University of Warwick. He is Deputy Director of Research, Course Director for the MA in Educational Research Methods and teaches both Foundation and Advanced Quantitative Research Methods. There is more information available at Steve's homepage. Dr. Stuart Cadwallader Stuart joined the University of Warwick in October 2009 as a research fellow on an ESRC grant having spent the previous 4 years studying at Warwick for an MA in Educational Research Methods (2006) and a PhD in Education (2009). There is a little bit more at Stuart's homepage.

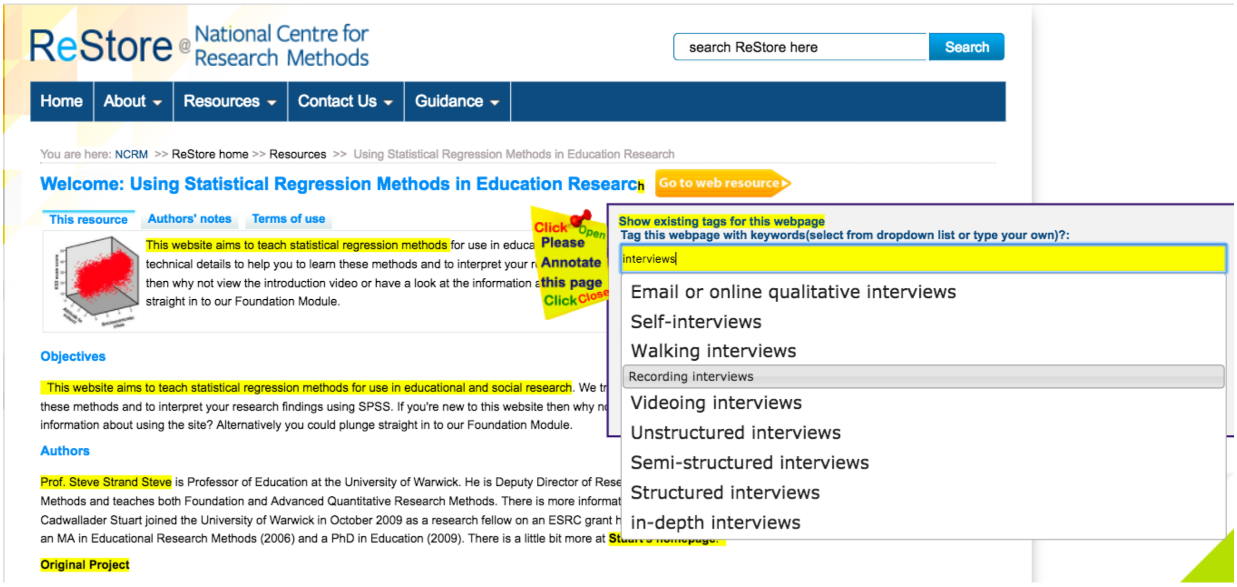
Original Project

Information about the original project has yet to be received. We will update this section as and when they become available.

Figure 4.3: A screen shot of the *AnnoTagger* in a web page showing highlighted textual annotation and a handle (right hand side) for annotating the entire web page.

2. *No technicalities in tags/annotation assignment* Popular tags and annotations become available to participants with chain-based drop-down menus containing auto-complete-based vocabulary annotation/tagging. In the case of no vocabulary item being found against the typed phrase, the free-text popular tag is suggested by the *AnnoTagger* with the option of creating new semantic tags. All of this is offered in a simple slide-in/slide-out panel embedded inside a web page. A demonstration page has been setup to show these components in action at

   http://www.restore.ac.uk/resources/statisticalRegression.html.


ReStore @ National Centre for Research Methods
search ReStore here
Search
Home
About
Resources
Contact Us
Guidance
You are here: NCRM >> ReStore home >> Resources >> Using Statistical Regression Methods in Education Research
Go to web resource
This resource
Authors' notes
Terms of use
Click Open
Please Annotate this page
Click Close
Show existing tags for this webpage
Email or online qualitative interviews
Self-interviews
Walking interviews
Recording interviews
Videoing interviews
Unstructured interviews
Semi-structured interviews
Structured interviews
in-depth interviews

Welcome: Using Statistical Regression Methods in Education Research

This website aims to teach statistical regression methods for use in educa
technical details to help you to learn these methods and to interpret your r
then why not view the introduction video or have a look at the information a
straight in to our Foundation Module.

Tag this webpage with keywords(select from dropdown list or type your own)?: interviews

Objectives

This website aims to teach statistical regression methods for use in educational and social research. We tr
these methods and to interpret your research findings using SPSS. If you're new to this website then why no
information about using the site? Alternatively you could plunge straight in to our Foundation Module.

Authors

Prof. Steve Strand Steve is Professor of Education at the University of Warwick. He is Deputy Director of Rese
Methods and teaches both Foundation and Advanced Quantitative Research Methods. There is more informat
Cadwallader Stuart joined the University of Warwick in October 2009 as a research fellow on an ESRC grant h
an MA in Educational Research Methods (2006) and a PhD in Education (2009). There is a little bit more at St

Original Project

Figure 4.4: A screen shot of the *AnnoTagger* in a web page a handle (right hand side) for tagging the entire web page.

3. *Focus on the user's task.* Annotation/tagging with *AnnoTagger* usually involves a user wanting to perform the task of (1) writing some text after selecting a portion of text inside the web page (2) assign vocabulary and free-text tags to the entire web page based on the content theme. Both of these activities are seamlessly embedded into a *pop out* text editor inside a web page followed by a benign handle clinging on to the web page for holistic web page annotation.

4. *Flexibility.* All annotation and tags assigned via the *AnnoTagger* are editable should the annotators change their mind or want to add further vocabulary terms or annotation to existing web pages or the text inside them.

4.4 Extensible *Annotation, Indexing, Searching* platform

The scale, at which the annotation, indexing and search results evaluation environment has been set up, is extensible (for example beyond ReStore content) and offers greater degree of freedom in terms of utilising annotation APIs e.g. Concepts, Entities and Keywords. Sentiment analysis could for example be used in order to determine the degree of relevance of concepts, entities and keywords in the documents collection across different LoD data sets. That in essence offers various search performance levels which could be evaluated at the time of searching to attain the required level of performance. The fact that the ES KB in this thesis is built on a non-relational(non-RDBMS) and schema-free platform (documents-based) for annotation, indexing, storage, retrieval, makes it flexible and elastic to accommodate future mergers and integration at structure level as well as index level. In addition, the KB responds to queries formulated in *noSQL or anti-SQL* language[11] but that feature could be extended by enabling SPARQL endpoint on the KB using GraphDB technology[12]. This kind of potential extension is discussed in Chapter 9. However, this aspect is not covered in the experimentation or evaluation sections in this thesis.

4.4.1 Why Elasticsearch(ES)?

Traditional relational database management systems (RDBMS) store data in tabular form i.e. rows and columns. The author understands that in order to implement the strategy of developing a sustainable, dynamic and "upgradable" indexing and retrieval system, it is important to adopt the document-oriented storage (unlike classical row-column relational records management system) and retrieval system. In other words, a system needs to be built where a change in the type and format of data does not require the entire database schema, rows and column types and constraints to be altered. In a document-oriented storage such as Elasticsearch, data objects are stored as documents, each document stores data and enable update/delete/insert processes on an as-and-when-needed basis. The biggest hurdle in using the traditional RDBMS-based storage for indexing and retrieval is that one must define a schema before adding/updating new or existing records to a database. When developing an application with a relational database, all the data elements are first mapped to an abstract entity-relationship model, which defines the data, relationships and structure that will then subsequently be used to create relational table and column definitions. This means that if (and when) the application data model changes, each of the corresponding table and column definitions need to be adjusted accordingly. Typically, these types of changes require the application developer to request the database administrator to aid in updating the schema.

Table 4.2: Comparison of RDBMS and Documents-based storage and retrieval structure e.g. MySQL vs. Elasticsearch

RDBMS	Database	Table	Row	Column
Elasticsearch	Index	Type	Document	field

[11] A NoSQL (originally referring to "non SQL" or "non relational") database provides a mechanism for storage and retrieval of data that is modelled in means other than the tabular relations used in relational databases.

[12] Graph database is a next generation database that uses graph structures for semantic queries with nodes, edges and properties to represent and store data.

In contrast, a document-oriented database contains documents as individual units (unlike components of a record spread over columns and rows in RDBMS), which are records that describe the data in the document, as well as the actual data. Documents can be as complex as you choose which necessitates nested data objects to provide additional sub-categories of information about documents object. That is one of the biggest reasons Elasticsearch has been adopted as an annotation storage and documents retrieval system as part of the experimental infrastructure in this thesis. The author has built and deployed the ES-based indexing, storage and retrieval platform based on the following points:

- Availability of information in the form of documents with multiple options to retrieve and filter out certain documents based on particular fields selection
- Single or several fields objects can be added/deleted/updated without the risk of affecting the retrievability or structure of other documents
- In case of crowdsourcing-based annotation, any number of sub-categories (futuristically speaking) can be added to a number of document objects with immediate retrievability based on *query-document* ranking score. A screen shot showing such nested elements of an ES document object can be seen in Appendix A in Figure A.6 and Figure A.7
- Horizontal scaling ensures optimum performance at the time of parallel indexing and retrieval
- Extensible clusters-based platform, which can cater to the needs and requirements as and when they arise. The platform can either be extended by acquiring more clusters within the current server or plugged into other networked server, discoverable via unique credentials
- Last but not least is the capability of ES to serve RDF-based SPARQL queries by using the GraphDB module[13] which makes our entire framework interoperable with the outside repositories having SPARQL speaking endpoints. This point has been detailed in the Conclusion Chapter 9.

4.4.2 The *inner* working of ES cluster

The cluster that runs the Elasticsearch node can be mounted on a dedicated server in order to serve any authenticated web server request using one of the many available client libraries having full community support. The front-end search application therefore does not necessarily have to run on a similar networked environment;

[13] A graph database is a database that uses graph structures for semantic queries with nodes, edges and properties to represent and store data.

rather, it can ping the Elasticsearch server as a remote server to serve online users enabling them to annotate and search using the legacy search applications.

Figure A.2 in Appendix A demonstrates that one Elasticsearch cluster can be combined into multiple clusters thus making it a domain-independent, multi-disciplinary annotation and search platform accessed by the users via a universal user interface.

4.5 Methodology: *Components of methodological framework*

The methodological framework explains methodology components, experimental setup and search results evaluation, using the tools and approaches, elaborated in the preceding section. The proposed framework comprises of four elements:i.e.

- Semantic structuring of the ReStore repository's content (Schematization)
- Mass annotation of documents addressing all types of content incorporating automatic semantic annotation as well as expert/crowd annotation/tagging.
- Storing and indexing actual full text content and semantic annotations (automatic and manual)
- Web-based search application using Elasticsearch to evaluate search results.

Figure 4.1 shows components of the proposed framework (first outlined in Figure 1.2), where the hybrid vector space takes centre stage in terms of the sources of feedback components i.e. *Annotations*, *Expert crowd*, *Crowd of results evaluators* and search results evaluation. The sources of full-text content are heterogeneous documents in ReStore repository. The number of repositories may be more than two depending upon the requirements. The *request-response* component facilitates *Querying* and *results* throughput against the hybrid vector space which is supposed to be continually evolving with the addition of new annotations and tagging.

4.5.1 Categorisation of processes

In order to evaluate the search results, various processes are undertaken to examine the four basic elements i.e. annotation, indexing, retrieval and crowdsourcing with different perspectives:

1. Keyword and semantic annotation of web documents using Alchemy API which analyzes each document by using built in NLP (Natural Language

processing) and Machine Learning (ML) and other complex linguistic, statistical, and neural network algorithms.

2. Assignment of the instances of NCRM Typology items to individual web documents (after the API-based automatic indexing) through expert crowdsourcing experiment to enable ourselves assess the degree of change in relevant search results retrieval. The sub-processes around *AnnoTagger* in Figure 4.1 illustrates this process.

3. The issue of query expansion and the resulting query distortion (before the query is processed by search engines) are addressed by empowering the users through the web-based front-ends to form sensible queries before they are submitted to the search engine. This is termed here *pre-search auto-complete* functionality, borrowing terms from available keywords, typology terms, concepts and entities based on user's research interests.

4. Finally, in order to address the research question RQ2 (1.3) relating to obsolescence in concepts and terminologies and to continually augment the content in ReStore repository, experiments have been conducted on semantic crowdsourcing aimed at tagging content (*in-side-page text annotation*) as well as web documents (*page-level annotation*). Moreover, such annotation and tagging experiments also enabled the author to investigate the issue of *enhanced recall* in terms of relevant documents retrieval especially when new documents are added in the semantic index. Further details on all the experiments are provided in Chapter 6.

4.5.2 Hybrid searchable index

As illustrated in Figure 4.1, the Hybrid Index, contains 3 sub indices i.e. *Full-text, SemDex and SemCrowDex*. For instance, if a user is interested in a concept "*Field Research*" and wants to retrieve all documents which have instances annotated with *Field Research*, all documents having the highest relevance score for *Field Research* in a document set will be presented in descending order i.e. highest score on top followed by documents having decreasing score. If document A has the instance of *Field Research* with cumulative relevance score of 0.8, another document B has the instance of *Field Research* but with a low cumulative score of 0.4 and document X has an instance of *Field Research* with higher cumulative score 0.75, then the order of the documents will be sorted as per cumulative relevance score before they are presented to users in the list of search results. The instance-based annotation in the hybrid index, has been implemented at both

Alchemy-based automatic annotation (using synonyms mapping) as well as crowd-annotation, aimed at improved search results retrieval. Various scenarios have been presented in Chapter 5 to further elaborate these criteria.

4.6 Required experiments

Figure 4.1 demonstrates the three experiments, which include Exp.A, Exp.B and Exp.C. The figure also clearly shows the flow of various processes performed as part of these experiments. The overall objective of conducting all these experiments is to prove the claim that (a) automatic LoD-based semantic annotation layer on top of full-text index brings improvement in online searching systems and (b) expert and non-expert crowd-annotators prove to be an effective annotation *cross-validation* force which need to be integrated for (c) preserving the performance and efficiency of online searching systems. Details and findings from the various experiments are presented in Chapters 6, 7 and 8 respectively in the following order:

4.6.1 *Experiment A (Exp.A (6.4))*: Semantic annotation and evaluation

This experiment focuses on the Hypothesis 1, as outlined in Section 1.3.1, which claims improvement in relevant search results retrieval when full-text index is augmented with LoD-based semantic annotations. As shown in Figure 4.1, content harvesting from the ReStore repository is performed by the *Semantic annotation* process before the analysis of Alchemy API annotators. The resultant content and annotations are indexed and stored in the ES KB for search results retrieval. The indexing and annotation of 3400 documents in the ReStore repository produces *SemDex*, which contains both full-text content fields as well as the associative semantic annotation fields. The next step is to perform search results evaluation using 20 baseline natural language queries. The criteria for preparing queries in this experiment, has been detailed in Section 6.2, Chapter 6. The evaluation involves the submission of randomly grouped queries (originally selected from Google search log) to the search engine and evaluation of the top 10 search results against each query in the set. Evaluation of search results, against each query, involves search results ranking (using the custom-built star-rating tool) and assigning further relevant complementary tags to the search result in question. The participants has no knowledge of whether the query is searched against the full-text index or semantic index. All the participants' feedback is stored in a separate database for Precision/Recall analysis as part of search results evaluation. Precision/Recall analysis proves improvement in system's performance in terms of relevance search results retrieval. The evaluation of complementary tags association in this experiment will be performed after getting similar feedback from participants in Exp.C.

4.6.2 *Experiment B (Exp.B(6.5))*:Crowdsourcing-based annotation & tagging

Crowd-sourced annotation of different online resources in the ReStore repository, involves *in-page text annotation* by expert academics, researchers, librarians and other professionals using free-text comment, typology tags and popular tags. Each expert participant accesses the desired *annotatable* web page after logging on to the *AnnoTagger*. Using the in-page annotation and whole-page annotation methods, a page is annotated using free-text and typology terms. The entire process finally augments the existing *SemDex* index, and the modified index is then called *SemCrowDex*. Annotation and tagging feedback from experts, in this experiment, is evaluated in two stages i.e. (a) Precision/Recall analysis based on feedback from search results evaluators in Exp.C and Exp.A and (b) Matching expert and non-expert tags to assess the agreement between expert and non-expert crowds.
All the "annotatable" web pages were pre-selected against participants' research interests and representative keywords proposed by the participant experts. Furthermore, the experiment required the participants to classify the entire web page using the NCRM Typology (controlled), free-text keywords (uncontrolled) and popular tags. The *AnnoTagger* was only used by the expert annotators as non-expert annotators were not invited to participate in this experiment in order to create a noise-free and authentic gold standard annotation and tags baseline. Moreover, a sample email has been demonstrated in Figure B.10 in Appendix B showing a list of URLs, sent out to a an expert participant, for annotation and tagging. Annotation and tagging in Exp.B enables the author to setup another search results evaluation experiment i.e. Exp.C to prove the second hypothesis, as described in Section 1.3.1.

4.6.3 *Experiment C (Exp.C (6.6))*: Search results performance evaluation

This experiment addresses Hyothesis 2, which claims that search results relevance and ranking further improve when user queries are searched against a more enriched semantic index i.e. *SemCrowDex*. The evaluation, in this experiment, is performed based on how relevant are the top 10 search results when a query is submitted to the ES KB, which has full-text content, *SemDex* and *SemCrowDex* indices. The non-expert participants were asked to star-rate each result in the set of 10 results with 5 star being the most relevant and 1 star indicating the non-relevant result. The star rating also indicated how satisfied the participant was in terms of how many stars were assigned to a particular result. The last task of each participant was to assign a minimum of 3 relevant tags to each result. The purpose of tag assignment was to find out the level of agreement between the expert annotators and non-expert results evaluators. It is important to highlight that the search results evaluation and *complementary tagging* of results was performed by non-expert. The set of benchmark queries for the evaluation of results was formed based on the query usage in Exp.A and Exp.C. A total of 33 queries were evaluated as part of this experiment. The criteria for bench-marking queries has been detailed in Section 6.2, Chapter 6.

4.6.4 Verifiable search results evaluation

In order to address users' searching needs, evolving a dependable and verifiable searching model is an important step of evaluation. The focus will remain on relevant content retrieval based on a set of benchmark queries in the concerned experiments. The users' ranking for a set of particular search results will also be evaluated in order to understand the linkage between the keywords entered and page-specific semantic tags (free text and vocabulary tags). The last thing to be analysed through the above experimentation is the '*degree of agreement*' between expert crowd and non-expert crowd annotators who participated in the experiments labelling the same pages with free-text and vocabulary tags depending upon their subject knowledge and understanding.

4.6.5 Prior knowledge and experience

In order to undertake this sort of experimentation and evaluation, having the following strengths, will help to build the proposed experimental and evaluation infrastructure:

- Access to a collection of online web resources (used as benchmark document corpus) including a live website i.e. www.restore.ac.uk, actively used by a high number of users every month.
- Free license from Alchemy API (now called IBM Watson Natural Language Understanding[14]) to use various annotation RESTful API services with 30, 000 documents annotation per day.

[14] NLU is an upgraded version or replacement of Alchemy API, part of IBM Bluemix cloud services - Available at https://www.ibm.com/watson/services/natural-language-understanding/

- Purpose-built client-side annotation tools for *automatically* annotating content in static and dynamic web pages
- The *AnnoTagger* tool, capable of *manually* annotating web page content using free-text and controlled vocabulary tags/annotations.
- Deployment of purpose-built search engine and client-side searching application, capable of searching over multiple indices, with complete control over documents versioning, adaptation and integration with another networked clusters.
- A vibrant Social Science research community having diverse backgrounds, from human geography to population and census and sociology to political Science and statistics, from which expert annotators can be recruited.

4.7 Semantic analysis of heterogeneous content

The current ReStore repository, like most other social science repositories (as outlined in Figure 1.1) with built-in search engines, relies on keywords-based searching. Essentially, the search tool crawls over the content of the site's pages and identifies terms and phrases (avoiding stop words such as "the", "in", "and" etc.). From this, an index table is built which ranks all the pages in terms of the frequency with which these terms appear on a page relative to their prevalence across the entire site. When searching, the pages with the highest density of the requested terms or rare terms in a set of searchable web pages are ranked highest in the search results.

For a repository site such as ReStore, which contains multiple heterogeneous resources, created by different authors, and for which a user may not be fully aware of the resources available about any particular research methods topic, it is particularly important that searching be made as effective as possible. The ReStore repository contains numerous resources over which there has been no single design influence or editorial control.

Searching can only really be enhanced if it is possible to provide more "intelligent" information about the contents of pages to the search tool. The key means of achieving this objective is to add the most appropriate annotation terms to ReStore site content, to provide a richer basis for searching the pages than is possible by analysing their textual content alone. For example, a page with a relatively low score on the term "*software simulation*" may in fact be all about this topic, but mention the exact term only once or twice. Annotating this page with "software simulation" identifies it as being highly relevant to the topic, such that it would achieve a higher score (and position in search results) than its retrieval based only on the actual content elements. Some examples concerning the typical search experiences in a multi-disciplinary web repository have already been presented in Chapters 1 and 2. As can be seen in Figure 4.5, a specific environment has been

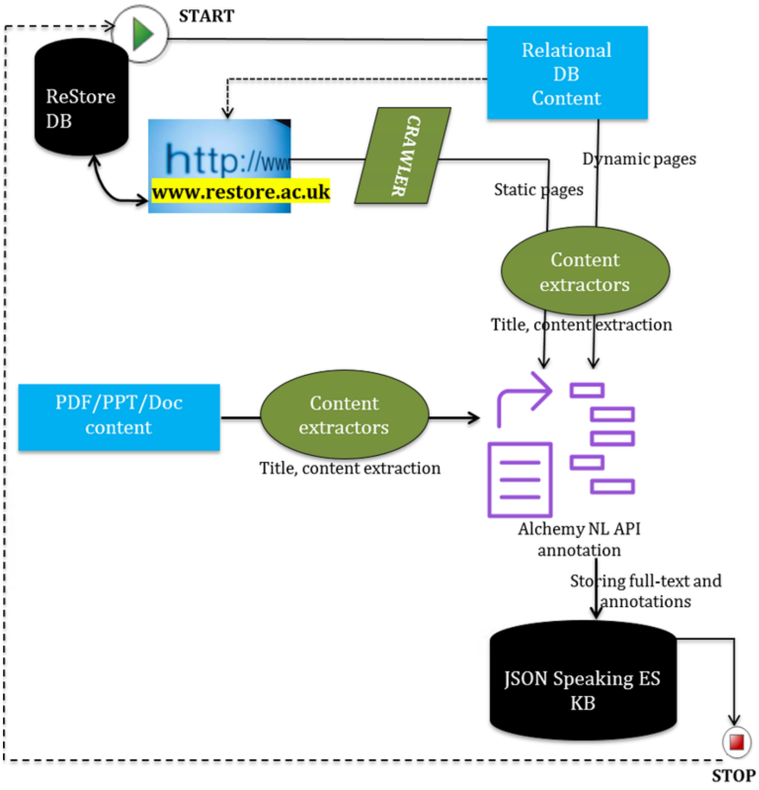


Figure 4.5: Information extraction architecture showing indexing and storage process flow of heterogeneous documents from the ReStore repository

set up where content from the ReStore website is extracted by using two distinct methods i.e. crawling static web pages from the website and extracting dynamic content and non-web page documents from relational databases e.g. MySQL, MS SQL Server etc. Semantic annotation of web documents is performed by the Alchemy Natural Language (NL) APIs suite, using the input received from the two *content extractors*. Figure 4.5 further shows that content is extracted from 3 different sources i.e. crawlable URLs, core documents e.g. MS Word, MS Excel etc. and content from RDBM databases. Algorithm 3 in Appendix C, further details the process in a step-by-step manner. Similarly topical keywords, concepts and entities have been added to 3400+ documents and the relevant TF.IDF score stored along with Alchemy APIs score against each of the individual items in a single record, to enable manipulation of precision and recall during various experiments and evaluation exercises in the results evaluation stage of this thesis. Data has been indexed in multiple indices based on the type of data, size of documents and the possibilities in which the indexed data could be searched and browsed. This has been achieved by designing unique schema maps for each index. The indexed item, at this stage, should be modifiable by the expert and non-expert crowd-sourced annotators which has been achieved in one of the three experiments. Finally, a searching application has been developed, which sits on top of the above to facilitate users, to search for their topic of interest as part of the evaluation exercises.

4.7.1 Entity extraction

Natural Language Processing (NLP) and information extractors aim to identify features of text called Entities and link them to other web resources by means of typed inferences (Rizzo et al., 2012a). This method brings structure to web content in order to enhance the meanings of text in web pages, leading to improvement in content searching based on mutual relationships between different sections of documents, keywords and entities. Based on this interpretation, it is proposed that (a) the entity or context extraction, in this thesis, will be applied on a number of documents i.e. web pages in order to pinpoint keywords, entities and concepts to a more meaningful contemporary LoD source of data (b) a Knowledge Base of data will be built, which comprises of actual documents and semantic metadata along with inter-documents relationships. To retrieve information from that KB, a web-based application (SRR system) has been implemented, which will be used by social science researchers (unlike ontology engineers) to help out in evaluating the accuracy and precision of information retrieval from continuously evolving KB. Further details on the methods, undertaken to build and use the KB as part of search results evaluation, are given in Chapters 5 and 7.

4.7.2 Syntactic markup vs. learning based NLP techniques

In order to mark up keywords in terms of entities, classes and 'things of interest', the syntactic marking feature offered by Alchemy API framework has been used. With such an approach, it is ensured that all the documents in the corpus get marked up with multiple linked data sources as well as domain-specific social science typology through the crowdsourcing element. This approach is in contrast with the ontology-based approach aimed at training semantic extractors to identify ontology instances and linking them to linked data sources. The crawler/spider-based data-rich web page marking up proposed by Jellouli and El Mohajir (2009) is based on the assumption that web pages are generated based on a pre-selected template which is filled with content from a remote database. In the ReStore repository case, web pages and other documents have been archived from multiple sources and thus, a tailor-made solution is needed to holistically markup data in these heterogeneous web pages in order to address the research questions (RQ1 and RQ2); which addresses content heterogeneity in multi-disciplinary repository of research data.

4.8 Search results evaluation

The performance of the search systems in terms of very relevant top 10 results is evaluated by using TREC[15] 11-point Recall-Interpolated precision, traditional Average Precision measure, MAP (Mean Average Precision), two-tailed-t tests and Mean Average Ranking (MAR). In addition, it is shown through search results evaluation that the crowd in various other scientific disciplines can perform just as well as expert crowd annotators by matching the expert annotations with the non-expert evaluation tags. The front-end application, aimed at search results evaluation, is one of the core components of the experimental and evaluation platform. For example, relevant concepts (initially tagged by the system) have been evaluated by presenting most relevant tags to evaluators, clustered around each search results link, as detailed in Section 6.6 . Unlike evaluating Ontology classes of a specific domain or assessing the appropriateness of a sub class in a domain Ontology before annotating content with it, the author is interested in annotating topic of interests in a range of topically diverse heterogeneous documents having different file formats. To assess the appropriateness of annotation, search results are presented to evaluators based on topic (term, concept, keywords) popularity and weight of relevance calculated at the time of annotation and indexing.
It is also investigated whether the skills and knowledge of online research communities can be exploited in their respective research domains by recruiting them in for various experiments, as illustrated in Figure 4.1. It is examined, whether the research community intervention as the annotators of online content in multi-disciplinary web repositories, can prove vital when it comes to sustainable search results retrieval by future search engines.

4.8.1 Findings

Findings are presented based on the annotation and tagging data on top of LoD-based annotations in Chapters 7 and 8. It has also been analysed and demonstrated that by adding expert annotations to the automatically generated semantic index, the relationship between *query* and *documents* can be improved. In order to analyse and recognise the research community contribution in terms of searching and

[15] Text Retrieval Conference is actually a set of several different relevance benchmarks organised by the U.S. National Institute of Standards and Technology (NIST) - Available at http://trec.nist.gov/overview.html - Last accessed: 28/08/2018

information retrieval, experiments have been conducted using star-based rating and click-based tagging to measure the system's performance, based on automatic and crowd-based classification of online resources. The requirements for a number of experiments and the outcomes from each of them are presented in Chapter 6 (See Figure 6.1) and Chapter 8. Furthermore, it is also demonstrated through experiments that there exists a substantial level of agreement, between the expert and non-expert annotators, on content tagging in a repository.

4.9 Summary

The requirements for setting up an automatic and semi-automatic knowledge representation and retrieval framework have been described at length. The proposed framework for knowledge annotation and searching aims to adapt to the changing needs of online users community in a multi-disciplinary environment by exploiting LoD-based semantic annotation and crowdsourcing-based annotation and tagging. The need for experimentation to assess the *feasibility, soundness and sustainability* of our semantic annotation and searching framework has been outlined. The content heterogeneity element and accessibility have been addressed by including content from several sources i.e. documents, RDBMS databases and directly from websites. The time factor has been identified as crucial when it comes to the enrichment of semantic index with expert crowd-annotations and non-expert crowd-tagging. The reason being is the management of two diverse group of users whose research interests might be the same but whose approach to annotation and tagging is different. The methodology of amalgamating two sources of knowledge acquisition in web repositories has been explained i.e. *automatic* and *manual*; and the assessment of relevance in a hybrid index has also been elaborated. The above points and the technical implementation of the building blocks of the knowledge acquisition and retrieval framework are elaborated in Chapter 5.

Chapter 5

Development of the knowledge annotation, searching and evaluation framework

This chapter describes components of the knowledge annotation and retrieval system along with the entire flow of annotation, indexing and searching across the ReStore repository. It presents the process of relevance score computation in different scenarios following the submission of a user query. The author explains 3 major annotation scenarios i.e. (a) Automatic LoD-based semantic indexing (b) NCRM Typology-based annotation in web pages and (c) Users' annotation in web pages using free-text and controlled vocabulary. A search results retrieval system:SRR is also described, which incorporates the VSM-based algorithms, to retrieve the most relevant search results in all of the above cases.

5.1 Implementation structure

The detailed system architecture for annotation, tagging, indexing and retrieval modules is illustrated in Fig 5.1. The figure details various components of the proposed system which includes a web repository, semantic annotation of documents (Alchemy Annotator), crowd-sourced annotation and tagging (AnnoTagger), experts annotation and tagging, indexing and storage of content with annotations in Elasticsearch server, and ranked search results retrieval. In addition, algorithm 3 in Appendix C shows the step by step process of keywords, entities and concepts extraction along with their storage in the Elasticsearch. As demonstrated in Figure 5.1 , the annotation process flow starts from the repository content, hosted on the ReStore website. The Alchemy Annotator in the diagram deals with the automatic annotation of content comprising of extracting keywords, entities and concepts using Alchemy APIs. The inner- workings of index creation including the definition of *indexAnalyzer or customAnalyser* and *SearchAnalyzer* are further explained in Section 5.2.2 and 5.3.

5.1.1 Experts' annotation and tagging

There are two phases of this study: (a) annotation and tagging of content by the research community in online repositories of multi-disciplinary research data and (b) exploiting annotation metadata obtained from (a) in order to improve searching in those repositories . It is assumed that web pages are semantically related if they are tagged by a number of users having similar research interests. The author also infers, based on observing many participants annotating and tagging web resources, that related web resources are usually tagged more than once by semantically related tags such as these groups of tags {*team management*, *group dynamics*, *team leader*, *project management*, *research team*, *leadership*, *research team leader*}, {*people skills*, *research data management*, *data sharing*, *data dissemination*} etc. A generic annotation page for demonstration purposes was set up at the time of pilot experimentation for greater understanding of the whole purpose of annotating and tagging. This page can be accessed at http://goo.gl/MEJIze. The initial participation confirmation page was also setup at the ReStore repository site, aimed at publicising the study, which is available at http://www.restore.ac.uk/focusgroup.

Figure 5.1 shows expert annotators contributing to the annotation process as part of *AnnoTagger* component. The figure further demonstrates that the expert annotators request the annotatable pages using their unique credentials and annotate the pages using *AnnoTagger*. They are expected to perform *content-level annotation* and *page-level annotation*, as detailed in Algorithm 2 and Section 5.1.4. The system has been setup to flush annotation data to the ES index after every 10 annotation activities. The attributes of a typical annotation task incldue *Annotated Text, Source text, Tags, User, URI, AnnotationID and Annotation Date (creation and updated)* etc. Each annotation on a web page (or group of web pages) is then added up to the semantic index, defined in Chapters 1 and 4 as *SemDex*. The evaluation of results, based on such annotations, has been detailed in Chapter 8.

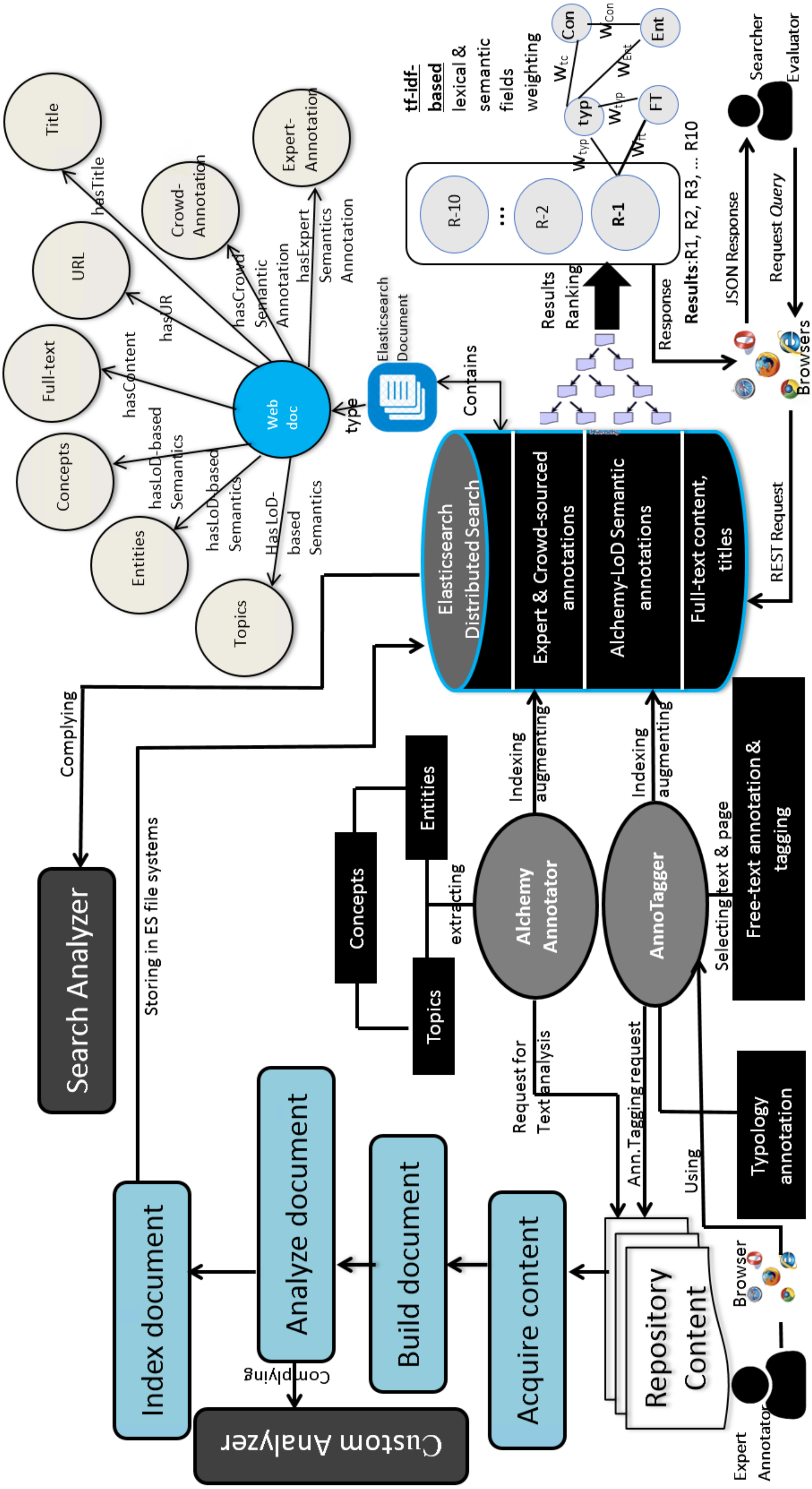


Figure 5.1: Implementation diagram for crowd-annotation based semantic annotation and tagging along with search results evaluation

5.1.2 Annotator: webpage-embedded manual annotation tool

The Annotator[1] tool has been embedded in the entire ReStore website as part of various crowd annotation experiments. The embedded tool enables the annotators to add free-text annotation to a web page by selecting different pieces of text and images inside the respective page. Algorithm 2 in Appendix C describes the deployment of *Annotator* and the three stages of crowd-annotation (content level and web page-level) i.e. *annotator authentication, content level and page level annotations, adding/updating of annotation to ES index* in a step by step process, required for the implementation of *AnnoTagger*. Furthermore, algorithm 4 describes the *content-level free-text annotation* and *entire webpage-level tagging* in Appendix (C).

5.1.3 Validation of tags assigned by search results evaluators

Figure 5.1 outlines the roles of expert annotators and search results evaluators with one positioned at the annotation side and the other at the retrieval side. After the experts have augmented the index with content level and webpage-level annotation, as specified in Algorithm 2, it is necessary to ascertain the opinion of non-expert crowd regarding search results retrieved against natural language queries. Each query request fetches and displays search results, along with popular tags and typology items, clustered around search results, for evaluators to choose from. Besides, evaluating the relevance of search results against their query, they are expected to assign as many tags as necessary based on their understanding of the content of the web page being reviewed. Experiment C (6.6) in Chapter 6 further details this participant-based search results evaluation and cross-tags validation. As can be seen in Figure 5.1, search users submit their requests to the ES KB and various users with a unique user identification, evaluate search results and associate typology and/or free-text keywords to the top 10 results matching their queries. As part of the sustainable crowd-sourced annotation model, it might be an option that the experts periodically review those tags against the content of the web pages to approve/disapprove association on a random number of web pages. Once approved, all qualified tags are then permanently associated with the web documents in question and ES KB gets enriched with further crowd annotations ready to be exploited by the future search users. However, the appropriateness of crowd-sourced tags (typology or non-typology tags), associated to a piece of text or webpage is better assessed in the proposed SRR system, when queries are

[1] Annotator is an open-source JavaScript ibrary to add annotation feature easily into your applications. Available at http://annotatorjs.org/

searched against the semantic index of those pages. Experiment C therefore includes this aspect of tags association to search results and the author presents the outcome from the analysis in Section 8.6.5 in Chapter 8.

5.1.4 Anatomy of crowd annotation

Figure 5.2 shows an anatomy of various crowdsourcing-based annotation elements using graph data modelling techniques. The relationships between various nodes are established through properties and labels and all nodes have originated from the `Experts-Annotation` node. The *content level* and *page level* nodes sub divide the crowd annotation into two broader categories with *allTags* as a common node between the two. The typology or vocabulary-based annotations share common ground with allTags via *vocabulary-based* node which is the *type* of *allTags* node.

The purpose of showing and capitalising on such relationships at search time improves and sustains relevance in search results. For example, Elasticsearch query formulation allows refining the ranking of search results by *field boosting* factor which is akin to assigning more weights to crowd annotation fields compared to full-text fields e.g. content and topical keywords. Further examples have been given later in this chapter under Section 5.5 to elaborate the query formulation and processing in ES.

5.1.5 Search results retrieval (*SRR*) system

The online searching environment for the retrieval performance, crowd-annotation and tagging experiments required a robust search application which helps the users to trawl for their *annotatable* web pages based on their queries. A complete search application, as illustrated in Figure 5.1, was implemented using the Elastica client at the browser end and ES cluster mounted on the Linux VM (Virtual Machine). The JSON response from ES is processed using PHP client and other client-side scripting software libraries i.e. JQuery, Ajax and Javscript etc. The client-side scripting tools are also used for client side feedback collection during each search results evaluation and ranking session. Other JavaScript frameworks i.e. AngularJS[2] and TyepeheadJS[3] have been used to implement the typing-based

[2] AngularJS is a structural framework for dynamic web apps which allows you to use HTML as your template language and lets you extend HTML's syntax to express your application's components clearly and succinctly- Available at https://docs.angularjs.org/guide/introduction

[3] *typehead* is a flexible JavaScript library that provides a strong foundation for building robust typeaheads on a variety of data in the form of auto-completion- Available at https://twitter.github.io/typeahead.js/

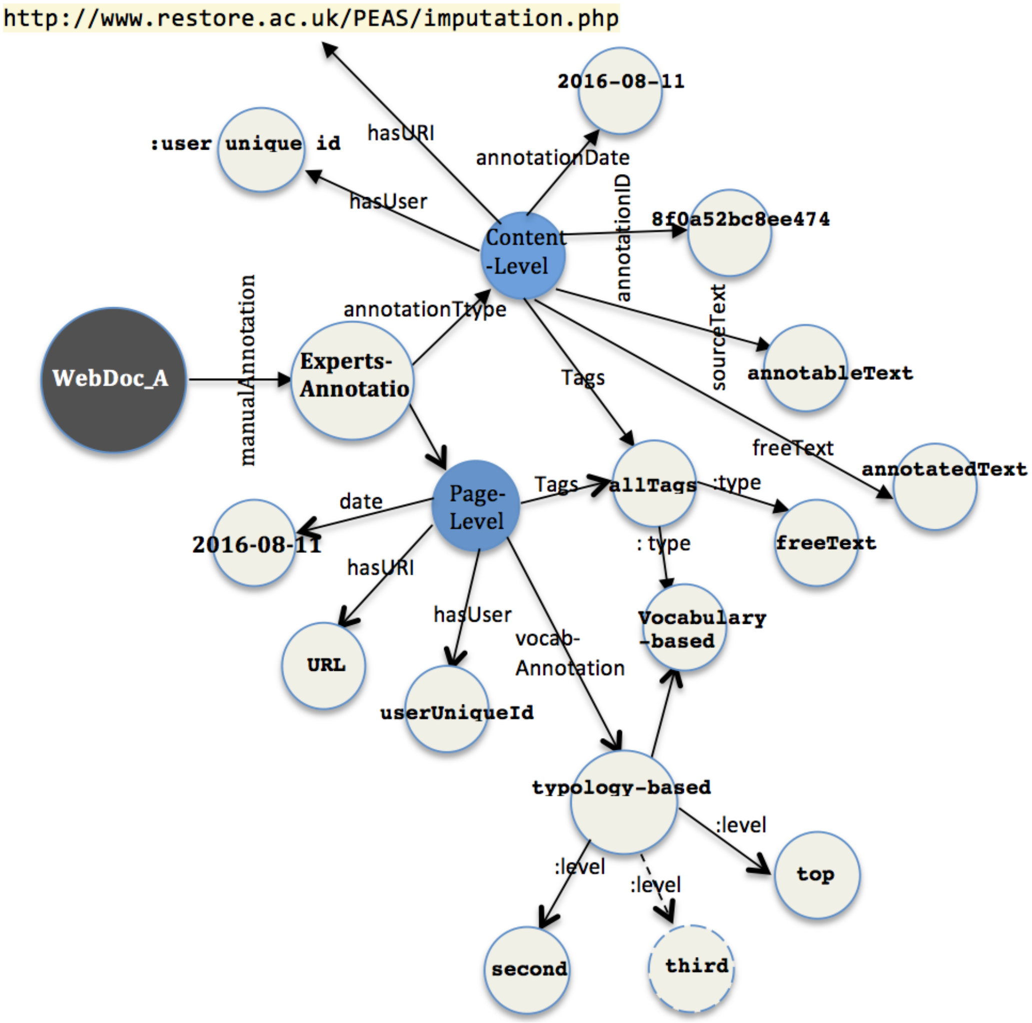
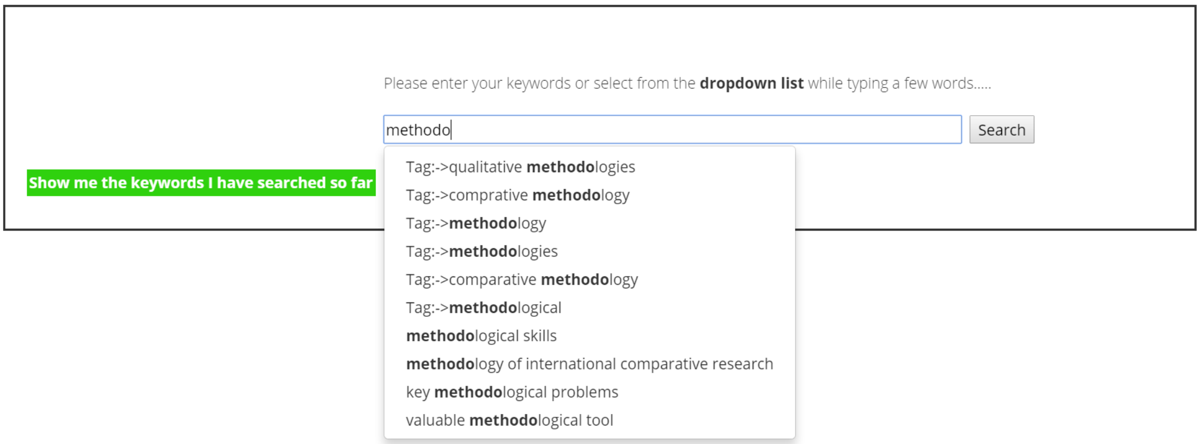


Figure 5.2: Representation of relationships among crowd annotation elements in a web page using graph data modelling

Query auto-complete on the online search box in a web page. The *auto-complete* feature remained a very popular query formulation tool in all of the three experiments (See Chapter 6), based on the author's observations in the focus group and non-focus-group experimentation sessions.

5.1.5.1 Experimental usage of Search application

The SRR system has been extensively used in the search results evaluation and crowd-annotation experiments. Moreover, it was critically important to ensure the availability of a search engine, which could retrieve results in several situations i.e. executing queries against full-text index only, semantic index (*SemDex*) and hybrid index (*SemCrowDex*). All participants in the 3 experiments (Sections6.4, 6.5, 6.6) made extensive usage of the search application for (firstly) retrieving *annotatable*



Please enter your keywords or select from the **dropdown list** while typing a few words.....

methodo

Tag:->qualitative **methodo**logies
Tag:->comprative **methodo**logy
Tag:->**methodo**logy
Tag:->**methodo**logies
Tag:->comparative **methodo**logy
Tag:->**methodo**logical
methodological skills
methodology of international comparative research
key **methodo**logical problems
valuable **methodo**logical tool

Figure 5.3: *Auto-complete*-based query formulation in the SRR system automatically fetching popular vocabulary and free-text tags from the index

pages as part of expert crowd-annotation and (secondly) retrieving search results against their queries for assessment and ranking purposes. Query auto-complete feature was consistently used in all experiments which was gradually enriched by the popular vocabulary and free-text tags to facilitate participants in formulating natural text queries.

5.2 Technology: customised implementation

Most of the technologies, aimed at implementing the proposed components of annotation, indexing and retrieval framework, are based on open source software. They include Elasticsearch stack as the search engine, Elastica library as the enabler of front-end search results application and RDBMS instances i.e. MySQL database to ensure access to dynamically populated web pages. PHP (Hypertext Processor), Javascript, AngularJS and MySQL have been used to render the user interface for taking users' input and fetching search results for evaluation purposes. A system has also been implemented on top of the client-server architecture which is both usable and scalable in both limited as well large scale environment without compromising the quality of annotations, efficiency of search, accuracy of search results and compatibility of system modules. The system architecture includes ReStore web server, ReStore database server containing two different MySQL Databases, which populate almost 2000 web pages in the ReStore website and the Elasticsearch dedicated search engine. Both types of users i.e. annotators and search results evaluators interact with the system in their browsers via RESTful interface using annotation and search APIs.

5.2.1 ElasticSearch: A knowledge management platform

The Elasticsearch server has been deployed on the current ReStore web server as a fully-fledged dedicated semantic and full text search engine which is available to the front end annotation and searching applications. Two instances of the ES process are mounted on a dedicated Linux VM (Virtual Machine), one for data storage (active) and one holding the backup (replica) in case of non-availability of the *primary* instance. Synonyms filter has also been defined to process disciplinary jargon and acronyms and assess the impact on relevance score calculation at the time of searching. *synonyms-descriptors* mapping has been demonstrated in Figure A.3 in Appendix A. As illustrated in Figure 5.1, ES has a flexible and adaptable custom analyzer that stems each word to its root form. Analyzes are discussed in more details in the forthcoming sub sections. ES provides horizontally scalable infrastructure for indexing and searching, which makes it an ideal choice for web-based repository content indexing and searching.

5.2.2 Embedded Apache Lucene

At the heart of ES, Lucene is a powerful search library from Apache which provides powerful indexing and searching capabilities for applications. The text analysis in indexing process forms the core of the entire indexing and searching exercise. Figure 5.1 illustrates the typical Lucene architecture aimed at indexing and searching starting from *content acquisition* to *document indexing*. Lucene enables Elasticsearch to index data in textual format and can be used with any data source as long as information can be extracted from the source in textual format. The structure, which is used by the Lucene is called *data inverted index* which is stored on the file system or memory as a set of index files. Before the data and further updates are added to the index, the text data is first processed by an analyzer (in this thesis *WhiteSpaceAnalyzer*) which takes into account token and stop filters defined at the time of creating the index. As illustrated in Figure 4.1 and Figure 5.1, Elasticsearch search platform uses the hybrid index to perform fast processing of queries against single and multiple searchable fields in one or multiple indices. The size of the current hybrid semantic index (full-text, LoD-based annotations and crowd-annotations) is 240MB with 8GB of physical memory for sharing with the ReStore web server. The standard or custom analyzers convert the textual content into units of searching called "terms" which are stored in the Lucene file system retrievable via queries against the concerned index.

Apache Lucene comes with various built-in analyzers such as *SimpleAnalyzer, StandardAnalyzer, StopAnalyzer, SNowballAnalyzer* and more. Table 5.1 illustrates various analyzers performing different operations during the indexing process. In the author's custom-built indexing analyzer, *WhiteSpaceAnalyzer* has been used with token, stop and synonyms filters to boost the precision in terms of retrieving relevant results against user queries. The *SearchAnalyzer* in Figure 5.1 is utilised by the ES search engine at the time of searching to match documents with user queries based on relevance. Figure 5.1 also shows the user's interaction with the search system where the author's system implements the PHP client called Elastica[4] in the front-end interface to display results in the user's browser against their natural text queries.

Analyzers	**Operations done in the text data**
WhitespaceAnalyzer	Splits tokens at whitespace
SImpleAnalyzer	Divides text at non-letter characters and puts text in lowercase
StopAnalyzer	Removes stop words and puts text in lower case (not good for searching)
StandardAnalyzer	Tokenizes text based on sophisticated grammar, puts text in lowercase and removes stop words.

Table 5.1: Widely used analyzers in Elasticsearch indexing and searching supported by Lucene

5.2.3 ES score computation

Figure 5.1 illustrates the overall scoring component employed by ES, taking into consideration semantic annotations and full-text content of documents. The ranking model capitalises on the *tf-idf*-based results ranking incorporating *fieldNorm* and boosting factors in ES. `R-1, R-2 and R-3` represent search results and the ranking of each search result involves calculation of the weights of the semantic entity nodes and their relationships with proximity nodes. $W_{ft}, W_{typ}, W_{tc}, W_{Con}, W_{Ent}$ represent the weight of full-text, typology, topical keywords (or topics), concepts and entities respectively in the document weights computation. Some scenarios are presented later in this chapter to highlight the role of *field-length-norm* and *synonyms-based* scoring in relation to the weight computation depicted in Figure 5.1.

[4] An open source PHP client for Elasticsearch - Available at http://elastica.io/

The basic or standard formula for score calculation (without manipulation) is given as follows:

$$\begin{aligned} Score(q,d) = queryNorm(q) * coord(q,d) * \\ \sum(tf(tind) * idf(t^2) * t.getBoost() * norm(t,d))(t \in q) \end{aligned} \tag{5.1}$$

where score (q,d) is the relevance score of document d for q, *queryNorm (q)* is the query normalization factor, *coord(q,d)* is the coordination factor, the sum of the weight for each term t in query q for document d is *tf.idf* and *t.getBoost()* is the boost factor applied to the query and *norm(t,d)* is the *field-length norm* which implies that the shorter the field length, the greater will be the weight of the term in it. The above equation (5.1) is the modified version of the standard score calculation formula based on *tf.idf*, which is given as:

$$Score(q,d) = \sum_{t \in q \cap d}^{N} tf_{t,d}.idf_{t,d} \tag{5.2}$$

In the ES-based custom-built search retrieval model, the equation (5.2), therefore, changes to equation (5.3) below.

$$Score_{q,d} = \sum_{t \in q \cap d\{k,c,e,\forall(annotations)\}}^{N} 1 + log\, tf_{t,d}\ .\, log\,(\frac{N}{df_t}) \tag{5.3}$$

where $idf_t = log_{10}\frac{N}{df_t}$ and idf is the measure of informativeness of the term and term t represents all terms in q and d including all annotations terms in document d. The score is 0 if none of the query terms is present in the document. The log weighted weights calculation thus leads to trading off between long and short documents vectors (after new annotations were added by the crowd annotators or automatic APIs) having comparable weights. ES will obviously give more weight to rare terms, as they are more important than frequent terms hence the increased IDF in the above equation. The IDF component increases with the rarity of the term in the *document collection* but it also increases with the number of *occurrences* within a document thereby increasing the length of that document vector.

Elasticsearch analyzers first analyze all the content belonging to each document via JSON-formatted URLs and relevant scores are stored against keywords, entities and concepts. Each document $\mathrm{D_j}$ represents a vector space model in the following manner:

$$D_j = (t_k, t_e, t_c..., t_k ec)$$

Where $\mathrm{t_k}$, $\mathrm{t_e}$, $\mathrm{t_c}$, $\mathrm{t_k}$ are the keywords (k), entities (e) and concepts (c) terms. With this representation, each document is a vector having the above elements for influencing the ranking of search results. The scoring computation is based upon statistical and NLP techniques employed by Elasticsearch.

5.2.4 Elastica

In terms of the search results retrieval interface, Elastica[5] library has been used along with JQuery to embed the annotation and tagging tool into the ReStore repository website to facilitate an intuitive user interface to human annotators. Elastica has also been used for rendering a complete web-based search application (with PHP client) to analyse search results based on automatic semantic annotation as well as crowd-annotation. The reasons why Elastica is an ideal choice to be used as a client include:

- Compatibility with PHP SDK, which is the AlchemyAPI's client for indexing documents
- Availability of up to date documentation for the implementation of best possible client application
- Class, sub-class, methods and properties modularity for implementing application in Object Oriented Programming environments.

5.2.5 Front-end of search application

A standard level natural query-based searching interface, has been developed for both search results evaluators and expert crowd-annotators. As illustrated in Figure 5.1, participants in both groups have extensively used the application to evaluate search results and annotate web pages respectively. The availability of annotatable pages to the expert participants based on their *typed keywords* was important to make optimal use of their time and improve the overall experience of annotation. The Elastica-based web application, developed using PHP as a client, displays search results in a web page using JQuery, Javascript and CSS along with showing related semantic tags around each search result. A multi-faceted search application was initially implemented for experimentation, however, the evaluation of top 10 search results was performed using natural queries in a search box without tracking *facets-clicks*. The reason being, the time factor for

[5] Elastica is the documentation of Elastica, a PHP client for Elastic search available at http://elastica.io/

each individual experiment session increased drastically, when *keywords search* and *facet-clicks* were included in the pilot sessions, as compared to the targeted session time, for the actual search result evaluation experiments.

5.3 Creating the semantic index

In order to implement a hybrid semantic indexing-based information retrieval, it is important to describe in detail the index schema, synonyms, mapping, index and search level analyzers. The index creation stage is important in its own right in that a fully-fledged schema for the index has been defined, which comprises of selecting appropriate *nGram tokenizers* at the parsing stage (Custom analyzer in Figure 5.1) and *standard analyzer* (white space) at the searching stage. The annotation and indexing components of the system address the search results retrieval from the outset by custom-tokenizing the text available from different sources. The *custom analyzer* in Figure 5.1 tokenizes the text using `textWhiteSpaceAnalyzer` which takes into account token and stop filters defined at the time of creating the index. A screenshot, showing the semantically extracted *Keywords* and *Concepts* is given in Figure A.5 in Appendix A. Term selection, stop words removals and term weighting have also been defined at the time of creating each index to enhance relevance and precision in search results while removing stop words from becoming part of the scoring criteria. Such terms are sometime referred to as normalization, ignorable character removal, dehyphenation and stop terms removal etc. (Gal et al., 2003). Moreover, specific settings have also been defined for all of the indexes before indexing the actual data, which has already shown promising results in terms of accuracy in search results. Figure A.4 in Appendix A shows mapping of the *searchable fields* in the semantic index.

Such mapping is flexible (unlike RDBMS-based mapping for storage and retrieval) and further fields can be added in the future without causing any conflict at the time of indexing or retrieval. *Stopwords* and *synonyms* filters were created to map domain and scientific-discipline-specific acronyms (89 in total) to actual text, in order to further improve upon relevance between query and top 10 search results. The purpose of mapping synonyms to obsolete domain-specific terms is to address the issue of concepts obsolescence (partially if not entirely) in a scientific discipline. A number of terms in the NCRM Typology has been mapped on to contemporary descriptors with a view to improve *query-document* similarity at the time of retrieval. Appendix A.3 shows a brief version of the schema, incorporated by both *SemDex* and *SemCrowDex*. A detailed description (screen shots of the original code) of initialising Alchemy API, annotating the *description*, *title* fields in a typical web document, extracting *keywords, concepts, entities* is given in Appendix B Figure B.1, B.2. The annotated document along with all semantic meta-data and properties is then indexed/stored in the ES index as shown in Appendix B B.3.

5.3.1 Document annotation

Based on the feedback, received from experts in pre-experimental pilot sessions, it became evident that they were more interested in web pages annotation instead of non-webpage documents. The expert-annotation of documents and search results evaluation were therefore restricted to web pages only in Exp.B (6.5) and Exp.C(6.6) respectively. As elaborated in Section 4.7, *crawlers* and custom-built scripts were used to extract text from web documents and prepare it for ingesting into semantic indexes. In all of these cases, the author has preserved the full-text fields e.g. title, content of each document, and enriched the indexed document with semantic entities, concepts and typology terms as part of document annotation. The *title* and *content* fields are the full-text fields whereas the *topics, entities, concepts, expert annotations and crowd-annotations* are semantic fields. The ranking score computation has been performed in Section 5.6.1 to further highlight the level of importance of each field.

5.4 Scoring algorithms based on semantic enrichment

Algorithm 1 Adding data to the index for searching and retrieval purpose

```
Input: Input
Cosine Similarity Score (q, d)
Float Score[n]=0
        Foreach term t in query vector
        Do calculate $Score[d] = queryNorm * coord(q, d)$ *
        $\sum(tf(tind) * idf(t)^2 * t.getBoost() * norm(t, d))(t \in q)$
        Read the array Length
        $[d \in D_{tokens \in fulltext,concepts,entities,keywords,Crowd-Annotations}]$
            foreach d
            Do $Score[d] = Score[d]/Length[d]$
            Return top 10 documents with highest $Score[d_1, d_2, \ldots d_{10}]$
            end for
        end for
Output: new Document added successfully to:myIndexName
```

Algorithm 1 presents the algorithm, being used to calculate the score for the top 10 documents against a user query. The array *Score* holds the scores for each of the documents and the Length array holds the normalized length of each document in the result set. It is important to keep in mind that the length of documents changes with every new crowd annotation added into the specified Elasticsearch document. The change is then reflected in the search results in terms of the rank of the respective document based on the new score assigned to that document in the Score array. The *queryNorm* is the normalization factor aimed at normalizing a query so that the results from one query may be compared with the results of another. The coordination factor `coord` in the above figure represents a weight to reward documents that contain a higher percentage of the query terms. The more query terms that appear in the document, the greater the chance that the document is a good match for the query.

5.5 Concepts and entities-based retrieval

The kind of web-based search application, the author is going to evaluate results with, involves semantically tagged concepts and those added by the crowd as part of the crowdsourcing element of this research. In order to further highlight and showcase the relationships that exist between documents, based on mutually related concepts, four web documents have been shown alongside their automatically extracted semantic concepts and URLs in Figure 5.4.

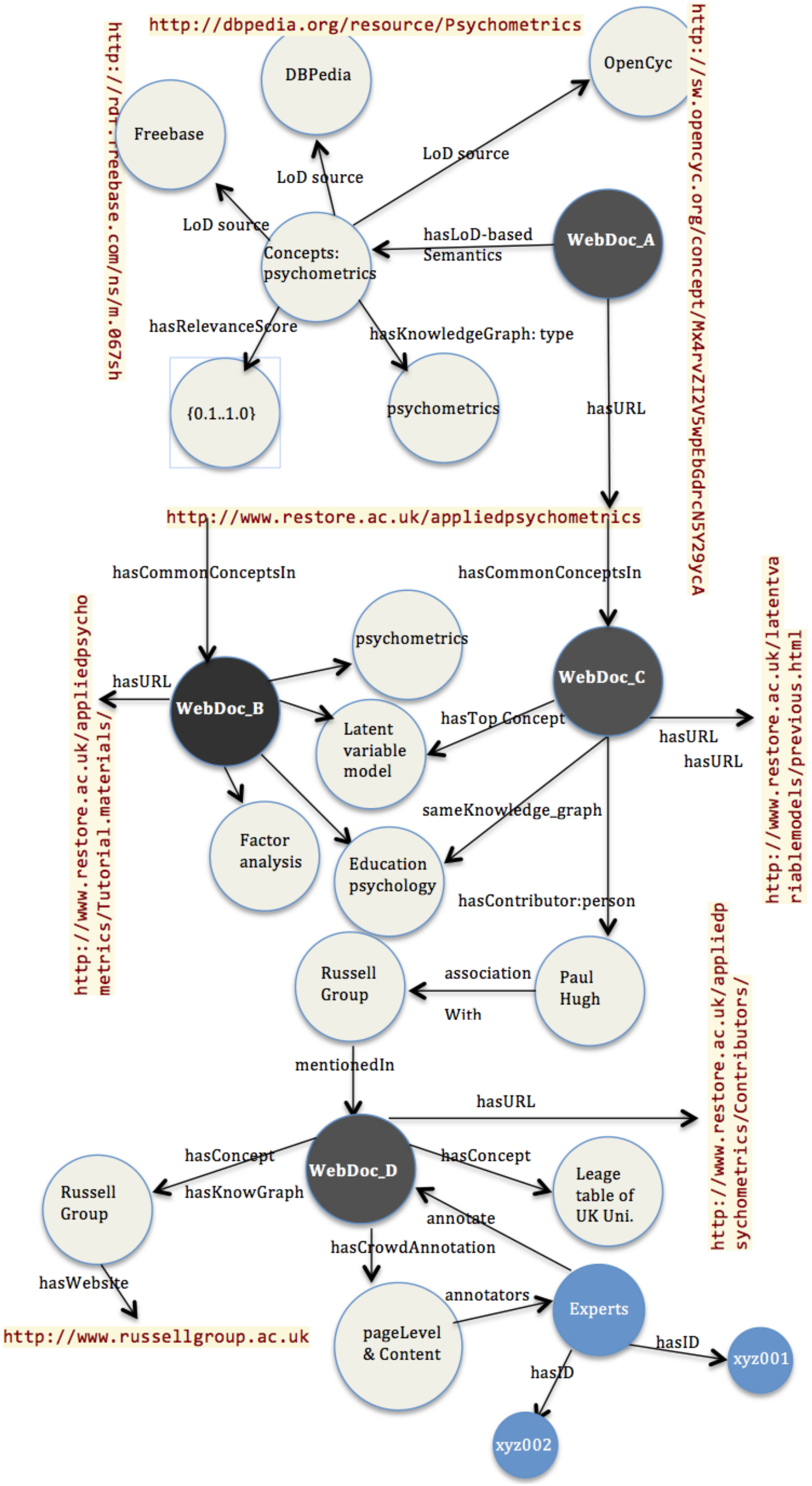


Figure 5.4: Using Graph data modelling to elicit the relationship between semantically annotated documents in our ES KB

The nodes and labels exhibit the potential of retrieving relevant web resources based on common concepts and entities using carefully crafted Elasticsearch queries. The capacity and potential of ES KB enables us to exploit that relationship at the time of searching, taking into account the scoring algorithm employed by the

ES scoring algorithms. For instance, let us assume that a user is interested in retrieving documents having instances of "*psychometrics*" concepts and all the related concepts and entities associated with psychometrics. The search results in that case will include documents having not only *psychometrics* instances but also keywords, as well as those documents having instances and entities related to the original "*psychometrics*" concepts in ReStore repository. Examples of queries retrieving results based on *entity type* and *concepts* have been given in Section C.2 , Appendix C. Chapters 6 and 8 explain this further in more detail.

5.6 Results ranking: *fields-based* score computation

This section details the computation of score when a query is searched against various *searchable* fields in ES hybrid index. Four different scenarios have been presented in the rest of this section to explain *fields*-based *tf-idf* score computation aimed at relevant search results retrieval. In each scenario, *tf-idf* score is summed up for each query $tf(t, q)$ term followed by document terms $tf(t, d_{d1,d2,d3...dn \in D})$. The methods of calculating *tf-idf* in these scenarios are slightly different from the traditional methods of calculation in that the *tf* for term t in document d is the *square root* of the number of times the term appears in the document. This calculation is more in line with the Elasticsearch *Explain API*[6] which has been used to understand the ranking score calculation in each scenario. TF is therefore given by: $tf(t, d) = \sqrt{n_{td}}$ and IDF is calculated as $idf(t, D) = 1 + log * (\frac{n_d}{n_{dt}+1})$ where n_{td} is number of occurrences of the term in the document, n_d is a number of elements in D (document space) and n_{dt} is a number of elements of D containing the term t. The entire *tf.idf* is therefore given as:

$$tf.idf(t, d, D) = tf(t, d)idf(t, D)$$

5.6.1 Scenario 1: *tf-idf* score computation

In this scenario, a query *population survey analysis* is searched against two fields i.e. *title* and full-text *content* in four documents (Doc1, Doc2, Doc3, Doc4). The *tf-idf* score is subsequently calculated along with total normalized score for each document, which determines the ranking of all four documents.

[6] Explain API computes a score explanation for a query and a specific document in the results list. Available at https://www.elastic.co/guide/en/elasticsearch/reference/current/search-explain.html

Tables 5.2 and 5.3 demonstrate the terms computation and documents ranking without taking into account any semantically enriched *searchable* field. The remaining three scenarios elaborate the impact of adding semantic fields and synonyms in the *tf-idf*-based results ranking process. The semantic enrichment ultimately leads to better ranked results retrieval based on *query-document* relevance.

Table 5.2: *Query-documents* ranking score calculation based on document fields using components of TF/IDF algorithm and query: *population survey analysis*

Doc	*field*	$tf(t,d) = \sqrt{n_{td}}$	$idf(t,D)$	$tf.idf$	$normalizedtf - idf = tf.idf * \frac{1}{\sqrt{n_{tfd}}}$
			population		
1	Title	1	1.30	0.57	1.30 * $\frac{1}{\sqrt{3}}$ = 0.75= 0.741
	Content	1.73	1	1.73	1.73 * $\frac{1}{\sqrt{981}}$=0.055
2	Title	0	1.30	0	0
	Content	0	1	0	0
3	Title	0	1.30	0	0
	Content	0	1	0	0
4	Title	0	1.30	0	0
	Content	2	1	2	2 * $\frac{1}{\sqrt{713}}$ =0.074
			survey		
1	Title	0	1.30	0	0
	Content	1.73	0.90	1.557	1.557 * $\frac{1}{\sqrt{981}}$=0.049
2	Title	0	1.30	0	0
	Content	1.73	0.90	1.557	$\frac{1}{\sqrt{2223}}$ 0.033
3	Title	1	1.30	1.30	1.30 * $\frac{1}{\sqrt{8}}$= 0.455
	Content	2.23	0.90	2.89	2.89 * $\frac{1}{\sqrt{1123}}$= 0.086
4	Title	0	1.30	0	0
	Content	1.41	0.90	1.83	1.83 * $\frac{1}{\sqrt{713}}$ =0.067
			analysis		
1	Title	0	1.30	0	0
	Content	1	1	1	1 * $\frac{1}{\sqrt{981}}$=0.032
2	Title	1	1.30	1.30	1.30 * $\frac{1}{\sqrt{12}}$=0.3757
	Content	2	1.30	2.60	2.60 * $\frac{1}{\sqrt{2223}}$ =0.055
3	Title	1	1.30	1.30	1.30 * $\frac{1}{\sqrt{8}}$= 0.455
	Content	2.23	1.30	2.89	2.89 * $\frac{1}{\sqrt{1123}}$= 0.083
4	Title	0	1.30	0	0
	Content	1.41	1.30	1.83	1.83 * $\frac{1}{\sqrt{713}}$ =0.067

The term-based grand computation, based on *title* and *content* fields, for each document in Table 5.3 clearly shows that Doc3 is the highest ranked document followed by Doc4, Doc2 and Doc1 respectively.

Table 5.3: Cumulative ranking score calculation by taking into account multiple searchable fields in the semantic index

ranked docs	Title (normalized)	Content (normalized)	Total (normalized)
Doc3	0 + 1.30 + 1.30= 2.60 (0.90)	0 + 2.89 + 2.89= 5.78 (0.169)	8.38 (1.06)
Doc4	0 + 0 + 0= 0 (0)	2 + 1.83 + 1.83= 5.66 (0.208)	5.66 (0.208)
Doc2	0 + 0 + 1.30= 1.30 (0.375)	0 + 1.557 + 2.60= 4.15 (0.088)	5.45 (0.463)
Doc1	0.57 + 0 + 0=0.57 (0.741)	1.73 + 1.557 + 1= 4.28 (0.136)	4.857 (0.877)

The *normalized* score computation has been demonstrated in Table 5.3; in order to show in the following section that document ranking changes (for the sake of *document-query* relevance improvement) when full-text fields (title, content) are augmented with semantic annotations and synonyms assignment. The ranking of documents in Table 5.3 therefore, changes when the semantic fields are included in the equation in addition to the full-text fields as elicited in Listings 5.1 and C.4. *GET* API is called against *SemDex*, which has been aliased to *idx_ restore_ v2*, and the query *cross-national comparison* is executed against *title, keywords, concepts* and *entity* fields

```
GET idx_restore_v2/annotations/_search?pretty&size=10
{"explain": true,  "query": {"multi_match": {
      "query": "cross-national comparison",
      "fields": [ "title","allkeywords.keywords", "
   allconcepts.concepts", "allentities.entity"]
   }}
}
```

Listing 5.1: Queries against full-text and Semantic annotation fields

Similarly, another ES query, given in Appendix C (Listing C.4), retrieves results based on cumulative score, computed after searching the fields in Query 5.1 plus crowd-added fields i.e. `crowdAnotation` and `vocabularyAnnotation`.

5.6.1.1 Discussion

The query listings 5.1 and C.4 produce ranked lists of top 10 search results, which vary in terms of relevance and *query-document* match. The ranking score is the *normalised tf-idf* score computed against two types of searchable fields i.e. 1. full-text and semantic field and 2. crowd-specific field. ES query in listing 5.1

fetches results based on the user's query (*crosss-national comparison*), using *term-matching* in multiple fields (in *SemDex*) i.e. *title, keywords, concepts, entities.* The score computation is purely done by the ES engine using Equation 5.1 and incorporating custom analyzer in Appendix A.3. However, it is by all means possible to incorporate the Alchemy API-generated score in the query in case one wants to filter out low-scoring terms in the indexed field. In addition, a filter can be applied on all indexed fields based on their *types* and/or *score* in order to maximise the degree of satisfaction in terms of achieving precision in search results retrieval as discussed in Section 5.5 and Figure 5.4. This discussion continues in Chapter 8 in Section 8.1.1 and Section 8.3.2, where the dimension changes from document ranking to relevance maximisation based on the inclusion/exclusion of certain searchable fields in ES queries.

5.6.2 Scenario 2: *tf-idf* score computation of full-text, semantic and crowd-annotated fields

Here, this scenario presents a query, executed against *SemDex* and *SemCrowDex* fields, and the resulting ranking of the search results. The scenario also elaborates the impact on the overall results ranking score by changing the searchable fields against which a particular query is executed.

```
GET index_SemCrowDex/_search?pretty
    from": 0,"size": 10,
      "query": {"bool": {
      "should": [
        /*****querying fulltext  and automatically
   created semantic annotation fields****/
        {"query_string": {"fields":
           ["title", "topical_keyworrds",  "sem_entities"
   , "sem_concepts" \par
          ],"query": "socio-economic policy framework"
          }
        },
        / *******querying crowd-annotation fields
   alongside the above******/
        {"nested": {
            "path": "crowdAnnotation",
            "score_mode": "max",
                 "query": {"query_string": {
   "fields": ["crowdAnnotation.sourceText", "
   crowdAnnotation.annotatedText"],
      "query": "socio-economic policy framework"
          } } } } ]
    }},"sort": [ {
        "_score": {
           "order": "desc"
        }
     }
```

Listing 5.2: Query against full-text, semantic annotation and crowd annotation fields combined

ES query in listing 5.2 comprises of two components i.e. {full-text, semantic fields} and {crowd-annotated fields}. Both fields are searched by the ES engine for query terms *socio-economic policy framework* to produce a list of the top 10 ranked search results. The integrated ranking score is calculated using the normalised *tf-idf* score as shown in Table 5.4, taking into account full-text and annotation fields as shown in the above query listings.

Table 5.4: *query-documents* relevance score calculation using searchable fields including vs. excluding crowd-annotation fields

	Fields	Normalized score (with Crowd element) {Doc 1}	Normalized score (no crowd element) {Doc 1}
	query=Socio-economic policy framework		
1	Title	0	1.81
2	Topical Keywords	0.57	0.35
3	Entities	0	0
4	Concepts	1.14	1.11
5	Crowd-sourceText	2.32	0
6	Crowd-annotatedText		
7	Crowd-popularTags		
Total normalized score		4.03	3.27

Table 5.4 also highlights the potential and impact of semantically populated fields in the ranking of documents following a query submission. *Doc 1* ranks higher when query *socio-economic policy framework* is searched against *SemCrowDex* and it gets demoted when the same query is searched against all field except the {crowd-annotated fields}.

5.6.2.1 Ranking of results: With vs. without crowd-annotation fields

Table 5.5 demonstrates the computationally ranked results against a *query: socio-economic policy framework* in two formats, one having the crowd-annotation weights taken into account in the overall scoring and ranking, and the other having no crowd-annotation elements. It also shows how worthy a ranked result is to the search results evaluators in terms of the number of stars, with 5 stars being the perfect match and 1 being the worst match.

Table 5.5: Comparison of ranked search results (documents) against a query by including /excluding crowd-annotation elements in the semantic index

Results based on ES Normalized Score (ENS)	**Results including crowd fields**			**Results excluding crowd fields**		
	Ranked Result ID	score	Stars rating by results evaluators	Ranked Result ID	score	Stars rating by results evaluators
query=Socio-economic policy framework						
(Result1)	2013	3.75	4 stars	2018	3.20	3 stars
(Result2)	2014	1.65	4 stars	2015	2.34	2 stars
(Result3)	330	1.19	3 stars	2014	2.05	2 stars
(Result4)	322	1.18	4 stars	2013	1.99	3 stars
(Result5)	2018	1.15	3 stars	191	1.48	1 stars
(Result6)	2016	1.07	3 stars	99	1.38	2 stars
(Result7)	2015	0.84	2 stars	2016	1.22	1 stars
(Result8)	324	0.83	2 stars	2017	1.18	1 stars

Table 5.5 highlights the impact of searching a query (*query:Socio-economic policy framework*) against multiple fields in *SemCrowDex* and *SemDex*. The table also shows the users' rating of particular results in terms of *stars* (5 stars being the best and 1 star being the worst match). The ranking changes, as does the rating, when the searchable fields include the crowd-populated fields. This phenomenon is due to the rarity of crowd-terms in the entire document corpus, length of fields being used for annotation and contemporarity of crowd and Alchemy-assigned annotations (of text and web pages) in the ReStore repository. The *star-rating* of search results has been detailed in Chapter 6 under Section 6.4 and 6.6.

5.6.3 Scenario 3: *Synonyms-based* ranking score computation

In this scenario the variation in ES normalized score computation is outlined when it takes into account the *synonyms* and their interpretation and representation in the queries.

Table 5.6: Computation of TF/IDF normalized score using certain searchable fields against a *query* by including synonyms filter in the search analyzer

Fields		Normalized score (with synonyms element) { Doc 1}
	query=rdi	
1	Title	0
2	Content	0.36
5	Synonyms	2.83
Total normalized score		Sim (q,d)=dotProduct=3.19/1.23=2.58

Table 5.6 describes the normalised *tf-idf* score computation for a potential result (Doc 1), retrieved after a query search (*query:rdi (researchers development initiative)*) is performed against the hybrid semantic index *SemCrowDex*. The *total normalised score* shows the total ranking score, which Doc 1 (result) has posted against the above query. The score, assigned to *synonyms* field among the 3 fields (*title, content and synonyms*), highlights the role of synonyms in the document-query similarity (dotProduct) score computation.

Applying this criteria, the ranked list of results can be understood in Table 5.7. Three dimensions have been highlighted in the tabular computation i.e. 1. *weight of crowd element*, 2. *weight of actual term* and 3. *weight of synonymous term.* The final score column in Table 5.7 includes elements from all of these 3 dimensions which determine the ranking of the top 10 search results. The popularity of the ranked results has also been highlighted in *star-rating column* in terms of the number of stars assigned to each result by the search results evaluators. The star-based ranking further corroborates the relevance and ranked order of these results.

Table 5.7: Comparison of ranked search results using the synonyms elements along with crowd-annotation vs. non-crowd-annotation fields

List of retrieved results	**Results including crowd fields**			**Synonym/actual score comparison**	
	Ranked Result ID	score	Stars rating by results evaluators	actual term weight	synonym weight
			query=rdi		
Result 1	2641	2.58	5 stars	*rdi=0.36*	2.22
Result 2	2516	2.22	3 stars	*rdi=0.31*	1.91
Result 3	2491	2.137	4 stars	*rdi=0.27*	1.86
Result 4	2522	2.130	4 stars	*rdi=0.29*	1.84
Result 5	2642	2.130	3 stars	*rdi =0.288*	1.84

5.6.4 Scenario 4: *Faceted data visualisation*

In this scenario, further possibilities have been highlighted aimed at formulating *NoSQL* queries[7] for retrieving potential search results to appreciate the potential of *SRR* system in terms of faceted data visualisation. The faceted aspect of search results retrieval and presentation addresses the *berry picking* phenomenon explained in Section 2.2.2 in Chapter 2. Retrieving search results based on the above-mentioned scoring algorithms in a faceted way increases the possibility of maximising the *query-document* relevance score.

Listing C.5 in Appendix C shows a *NoSQL* query having two dimensions i.e. 1. filtering out results having concepts type *General Household Survey* and 2. retrieving all significant concepts in the form of facets. Similarly, the faceted query in Listing 5.3 demonstrates another dimension of search results retrieval based on the indexed data in the hybrid index *SemCrowDex*. Although this aspect has not been included in the participants-based evaluation, it is surely interesting enough to fall under the remit of future research work.

Similarly, a *filtered* NoSQL query aimed at retrieving results based on *NCRM Typology* (see Figure 2.6 in Chapter 2) can be written as:

```
#Retrieve all results based on specific typology level1
   and level2 items
GET repository_index1,repository_index2.../_search?pretty
{"query": {"filtered": {
   "query": {"match_all_fields": {}},
   "filter": {"and": {
      "terms_of_filter_one": {
      "typology_level2.terms": [
         "Sampling",
         "Participant Recruitment",
         "Survey and Questionnaire Design"
      ] }},
      {"terms_of_filter_two": {
            "typology_level1.terms": [
            "Data Collection" ]}}
    }} }}
```

Listing 5.3: Filtered search results retrieval using NCRM Typology terms i.e. level2 and level1

[7] Elasticsearch processes NoSQL queries or non-relational queries and output results in JSON format

Note that there are 3 levels in the NCRM typology (classification system) with *level 1* being the top level followed by *level2* (the middle level) and the descriptor level i.e. *level3*. It can also be seen in Listing 5.3 that the query is being executed against multiple indexes and multiple repositories which addresses one of the author's research claims stating that the proposed annotation and searching framework are not limited to only one repository and its content. The real time extensibility of the annotation and searching framework is therefore an essential feature which necessitates the annotation & retrieval system to be adaptable and extensible as and when required.

5.7 Summary

Based on the analysis in this chapter and author's understanding, it is crucial to have a robust and scalable annotation and retrieval framework to ascertain the efficiency and effectiveness of a search system. When it comes to semantic search in repositories of heterogeneous content, pre-processing of content and establishing linkages between documents, based on context, meaning and association, take centre stage, which is evident from the methodology explained in this and the preceding chapter. It has also been discovered that by involving the online users in search results efficacy verification, the performance of an online search application can be continually improved in a scientific repository. This way, new concepts and terms in a scientific discipline can be analysed and evaluated by the users while they evolve with changes in technology, culture, socio-economic and life styles etc. On top of that, it is a source of encouragement for the author to know that by involving the relevant research community in designing the annotation and tagging environments, the overall performance of a search system can be sufficiently improved in terms of relevant and precise search results retrieval. Various experiments and evaluation analysis have been conducted in Chapter 6, Chapter 7 and Chapter 8 to test the performance of the implemented system, focusing on research questions, given in Chapter 1.

Chapter 6

Experimental design: *Semantic annotation and searching*

This chapter explains and discusses, the need for conducting a range of annotation and search results retrieval experiments. It investigate through these experiments, the role of LoD-based annotation and crowd-supported annotation and tagging in maximising relevance in the ever increasing number of search results against natural text queries. Two types of experiments (3 in total) have been conducted, one involving expert crowd annotators and two experiments focusing on search results evaluators. The three different types of experiments are elaborated, which were conducted to annotate several sections of the ReStore repository website, followed by the search results evaluation experiments. The SRR is based on different benchmark queries (see section 6.2 below), selected from all experiments, each involving diverse user groups comprising of research students, academics, librarians, research fellows and other volunteers. As outlined in Chapter 4, these experiments are labelled as Experiment A (Exp.A), Experiment B (Exp.B) and Experiment C (Exp.C). The complete search results evaluation from Exp.A is performed in Chapter 7 and the crowd-annotation supported search results retrieval is evaluated in Chapter 8.

6.1 The need for experimentation and evaluation

The RQs (1.3) in this thesis, dictate that a dynamic experimental platform needs to be setup for knowledge representation and retrieval using two different means of semantic annotation i.e. automatic LoD-based and crowdsourcing-based annotation & tagging. Figure 1.2 clearly demonstrated that need, from the outset in Chapter 1, along with the proposed components. The experimental setup needs to investigate the plausibility of accessing all types of content in a repository, followed by mass semantic annotation of all content for accurate knowledge representation, storage and ranked results retrieval in a *client-server* environment. In general terms, the aim is to assess the relationships between the changing information need and the degree of fulfilment of that need; with and without the intervention of the contemporary research community. Another claim, to be established through these experiments, is that improvement in search results (as a result of semantic enrichment) is not a singleton event in any way. Moreover, it is also established that such experiments and the likes of these can be repetitive in nature and may continually be conducted until the desired results are achieved.

In order to investigate and understand the role of API-based automatic semantic annotation, augmented by the expert crowd annotation in web repositories, the author aims to conduct various online annotation and searching experiments.

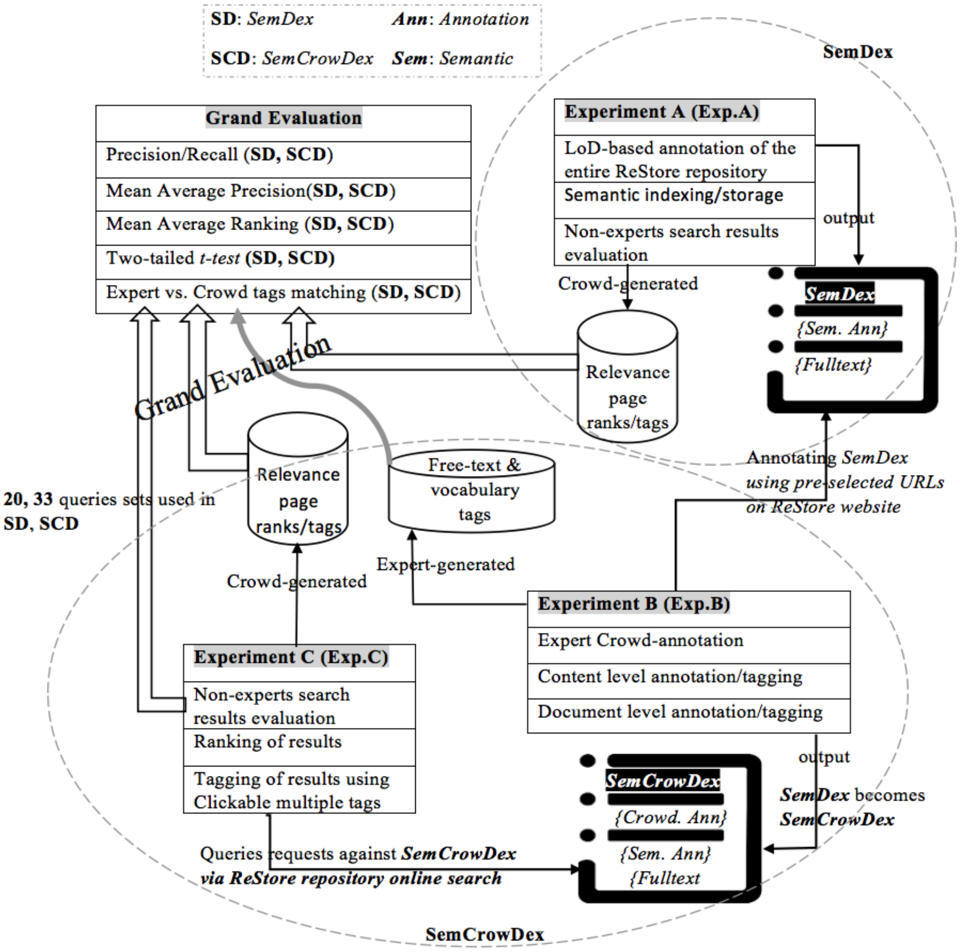

SD: SemDex
Ann: Annotation
SCD: SemCrowDex
Sem: Semantic
SemDex
Grand Evaluation
Precision/Recall (SD, SCD)
Mean Average Precision(SD, SCD)
Mean Average Ranking (SD, SCD)
Two-tailed t-test (SD, SCD)
Expert vs. Crowd tags matching (SD, SCD)
Experiment A (Exp.A)
LoD-based annotation of the entire ReStore repository
Semantic indexing/storage
Non-experts search results evaluation
output
Crowd-generated
SemDex
{Sem. Ann}
{Fulltext}
Relevance page ranks/tags
Grand Evaluation
Relevance page ranks/tags
Free-text & vocabulary tags
Annotating SemDex using pre-selected URLs on ReStore website
20, 33 queries sets used in SD, SCD
Crowd-generated
Expert-generated
Experiment B (Exp.B)
Expert Crowd-annotation
Content level annotation/tagging
Document level annotation/tagging
Experiment C (Exp.C)
Non-experts search results evaluation
Ranking of results
Tagging of results using Clickable multiple tags
output
SemCrowDex
{Crowd. Ann}
{Sem. Ann}
{Fulltext
SemDex becomes SemCrowDex
Queries requests against SemCrowDex via ReStore repository online search
SemCrowDex

Figure 6.1: Anatomy of experimental design showing SemDex, SemCrowDex, getting enriched as a result of Exp.A (6.4), Exp.B (6.5) & Exp.C (6.6) with grand evaluation component

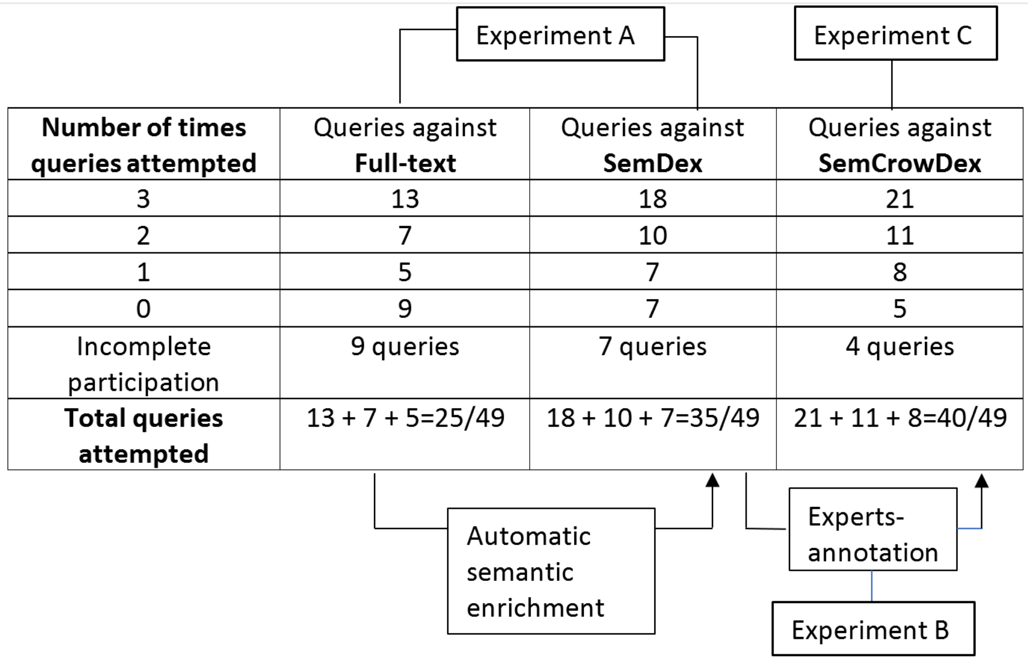
Exp.A (6.4), Exp.B (6.5), Exp.C (6.6) have been conducted; to evaluate search results retrieval in two systems and assess the feasibility and sustainability of automatic and crowd-sourced annotation in web repositories. Figure 6.1 illustrates the artefacts of 3 experiments and their relationship with other components e.g. search indices and evaluation processes. *SemDex* and *SemCrowDex* are the two indices, the author has used for participants' based search results retrieval (using natural language queries) and evaluation. Figure 6.1 further explains the experimental aspect of Figure 4.1, and establishes connection between the three experiments using various evaluation metrics. The following sections will frequently refer to Figure 6.1, while detailing various components in the figure.

6.2 Preparing a set of benchmark queries

A total of 50 queries were initially selected from the top 100 most searched queries on Google Analytic (GA), searched in 3 months' time, immediately before Exp.A. These queries represent the typical information needs of a particular research community i.e. Social Sciences, Web & Internet Sciences, etc. The criteria for selecting these queries included the number of clicks they had generated and the number of users they had brought to the ReStore site via Google search. In addition, the users' feedback to the participation advert was followed by an email, from the author, with a request to the participants, to supply at least 3 areas of research, which represented their research interests. Participant's research interests further helped out in refining the list of 50 queries intended for evaluation in Exp.A. Queries having single term, multiple terms and mixture of keywords and concepts were also considered to reduce bias in all the queries. The reason for selecting 50 queries was based on the following factors:

1. At least 3 times repetition of each query (with each session lasting for an hour) which means every query has to be attempted 3 times by participants.
2. 25 to 30 participants target was set in order to achieve at least 3 times repetition
3. Attrition bias in both experiments resulted in choosing different samples

Furthermore, two sets of benchmark queries were prepared; one for Exp.A and the other for Exp.C., based on the criteria, laid out in Figure 6.2. It is important to note that by "attempted" queries we mean those queries which were entered in the search box and the results were consequently retrieved, rated and tagged. In Exp.A, we sampled out 20 queries for evaluation purposes because search results



Number of times queries attempted	Queries against **Full-text**	Queries against **SemDex**	Queries against **SemCrowDex**
3	13	18	21
2	7	10	11
1	5	7	8
0	9	7	5
Incomplete participation	9 queries	7 queries	4 queries
Total queries attempted	13 + 7 + 5=25/49	18 + 10 + 7=35/49	21 + 11 + 8=40/49

Figure 6.2: Criteria for benchmark queries selection in Exp.A (6.4) & Exp.C (6.6)

evaluation was performed based on queries, which were attempted at least twice. That criteria led the author to choose 20 queries from *SemDex*, *Full-text*; and to analyse the performance based on precision/recall score and individual star-ranking of search results. *Full-text* and *SemDex* have 20 queries in common which were attempted at least twice in Exp.A. The criteria in Exp.C, however changed after we conducted Exp.B, because several web pages, which were annotated by the expert annotators, could be retrieved only by the one-time attempted queries. That phenomenon caused the addition of further queries (33 in total) in the analysis base, in order to capture web pages which were tagged and annotated using the typology and non-typology terms. Table 6.1 further explains the 20 and 33 sample queries selection for Exp.A and Exp.C respectively. The number of queries to be attempted remained the same i.e. 7 queries per participant. The number of participants in Exp.C were greater than all previous experiments, hence, the number of queries attempted 3 , 2 and 1 times were greater than previous experiments. However, the reason, for choosing 33 queries from both *SemDex* and *SemCrowDex*, was the number of *retrievable* results, which were ranked and tagged against these queries in both Exp.A and Exp.C. For example, let a *queryA* be searched in Exp.A against *SemDex*, and a participant ranks and tags 7 out of 10 results. The same query is then searched against *SemCrowDex* in Exp.C and another participant ranks and tags all 10 results. That query in question is qualified enough to be included in the grand evaluation analysis, as illustrated in Figure 6.1. The selection of this query, now, is not only based on the number of times it is attempted and

Queries	**Experiment A**	**Experiment C**
	FT vs. SemDex	SemDex vs SemCrowDex
Multivariate logistic regression analysis, Design effects in statistics, Online survey disadvantages, Randomized control trials, Media analysis, Finite population correction, Indirect geo-referencing, Cohort sequential design, Macrodata guide, Evaluating interaction effects, Primary sampling unit, Reasoning, Qualitative benchmarking, UK address example, Logistic regression in SPSS, Data collection skills, Case study on non-verbal communication, Ethnic group, Forecasting, Exploratory factor analysis, Social influence, thematic analysis	20 queries Attempted 2 or more time and no crowd-annotation instance	17 queries Attempted 1 or more time against SemDex and retrieved results have instances of experts' annotations and tagging
Components of research proposal, Stages of a systematic review, Natural experiments in Social Sciences, random sample enumeration, critical thinking, Mixture model, Paradata in survey research, Online research methods, Trust and respect in a team, latent class analysis mplus, qualitative research skills, sample enumeration, Using imputation for missing values, factor analysis, Stages of systematic review, sociolinguistics	0	16 queries Attempted 1 or more,time against SemCrowDex, and, retrieved results have instances, of experts' annotations and tagging
Total queries for each experiment	**20**	**33**

Table 6.1: Distribution of benchmark queries for Experiment A and C

ranked but whether the results (from that query) have enough tags, to be analysed for assessing the *crowd vs.experts*'s agreement on tagging. In addition, the author wanted to consider those queries for evaluation analysis, which produced results, tagged by expert annotators in Exp.B. That way, the improvement in relevant search results retrieval is attributed to the expert annotation; when *SemDex* and *SemCrowDex* are evaluated based on Precision/Recall score. It was observed that some queries were repeated more than five times against *SemDex*, but only equal number of attempted queries were averaged in Exp.A and Exp.C. Figure 6.2 explains that 18 queries in *SemDex* and 21 in *SemCrowDex* were attempted 3 times respectively. Moreover, 11 queries were attempted twice against *SemCrowDex* and 10 against *SemDex*. Furthermore, 7 queries against *SemDex* and 5 against *SemCrowDex* were not attempted by various participants for unknown reasons. They however attempted other queries against *SemCrowDex* but having no matching query in *SemDex*, evaluation was not performed on those queries.

6.2.1 Distribution of queries among participants

Table 6.2 explains the distribution of 49 queries among participants of both experiments (Exp.A and Exp.C). 1 out of 50 queries was removed because it retrieved many non-webpage results e.g. PDF documents. A set of queries comprises of 7 queries and each participant from 1-7 gets 7 queries each with no repeated queries.

The next group of participants (8-14) receives a set of repeated 7 queries, which ensures the equal number of repetition for each query until all participants receive their list of queries. The target for query repetition was set at 3 at the start of first experiment, which meant that each query would be repeatedly attempted by participants in order to reduce bias in Precision/Recall analysis.

Queries Participants	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Total
1-7	N	N	N	N	N	N	N	49
8-14	R	R	R	R	R	R	R	49
15-21	R	R	R	R	R	R	R	49
22-25 (-3)	R	R	R	R	R	R	R	28
N=New, R= Repeated								

Table 6.2: Query distribution for search results evaluation experiments

Table 6.2 demonstrates the distribution of queries among 25 participants, who participated in Exp.A. It is noted that some queries were not repeated 3 times (22-25 in the last row) but that bias has been considered in the evaluation analysis. Similar criteria was applied on 30 participants who participated in Exp.C, which is being discussed in Section 6.6. *N* in the table represents *New* query which is not repeated yet. *R*, on the other hand, shows the distribution of *Repeated* queries after the first cycle of 7 participants is completed.

6.2.2 Potential reasons for incomplete participation

One of the reasons for not evaluating 9 queries (including repeated) in Exp.A was the non-retrieval of relevant information against *Full-text* index, which might have led participants to enter a new query from the provided list of queries or entering an *out-of-the-list* query. The low number of unattempted queries in *SemDex* and *SemCrowDex* is due to the retrieval of more relevant search results, which led to the completion of all 10 search results evaluation against the respective queries. Distributing a pre-selected list of queries (in Table 6.2) consolidated the comparison and evaluation process around the 49 queries. However, due to the non-retrieval of relevant results against the provided queries, the participants used other queries from the full list of queries, harvested from Google Analytic[1]. In a focus group observation, it was noted that a particular participant received a query (among 7 queries), which was less representative of his/her research interests; hence it was either not attempted or attempted but not rated. The reward model might also

[1] Link to keywords is available at http://goo.gl/QpFmpf

have resulted in leaving some queries unattempted. In Exp.A, the reward was a lottery-based gift hamper and in Exp.C, participation was rewarded with a confirmed payment. Such variations in query selection led to shrinking the size of comparable queries, which were assessed against full-text and *SemDex* indices.

6.3 Recruiting Participants

Participants were recruited for all experiments by displaying adverts in the academic Schools' foyers (Education, Social Sciences, Geography, Statistics, Psychology, Computer Sciences) etc. Emails were written directly to the module leaders in the Faculty of Social & Human Sciences in Southampton and module leaders in Edinburgh, Cambridge, Cardiff, Manchester, Loughborough, Warwick, Kent, Portsmouth Universities in the UK requesting them to disseminate the study advert to post-doc researchers and PhD students in their respective departments. Other potential participants e.g. research fellows, web resource authors of content in the ReStore repository and professionals were directly approached via their connection with the University of Southampton e.g. UK Data Service[2], Language & Computation research group in University of Essex [3] requesting them for their participation. Despite the enormity and novelty of the task i.e. annotating text and tagging web pages using both free text and vocabulary annotations, the author was successful in getting sufficient number of participants who were both curious and motivated in participating in the study. This approach helped in filtering out unwanted annotation and tags from the outset. However, some participants needed more details before agreeing to participate in the experiments e.g. through frequent phone calls, email exchange, watching introductory video (http://www.restore.ac.uk/focusgroup) etc. A screen shot of the above webpage has been given in Figure B.7 Appendix B. The author also mass-emailed PhD students and Master students at the School of Social Sciences in Southampton [4] seeking their participation with options to either participate in focus group annotation/tagging study or attempt independent annotation, following guidance materials sent out by emails. A copy of the invitation email, sent out to potential participants, has been shown in Appendix B in Figure B.14. A web page, containing information about the study and joining details aimed at PhD students, was created on the ReStore website at http://www.restore.ac.uk/focusgroup. Training materials[5] were provided before the experiment and telephone calls were made and emails written when participants requested for more information.

[2] https://www.ukdataservice.ac.uk

[3] http://lac.essex.ac.uk

[4] http://www.southampton.ac.uk/socsci

[5] Available at http://goo.gl/QpFmpf

6.3.1 Defining crowdsourcing-based tasks

The following points have been addressed in order to ensure that the optimal standard is met when it comes to crowdsourcing-based annotation of web documents.

- Clearly defining the type of crowd required for the experiments
- Each crowd-annotation session was comprised of tasks with a clear goal defined and documented
- Crowd participants knew from the start the reward for participation in the experiment. Some participants voluntarily refused to be paid for the tasks while others accepted the quoted reward.
- The purpose, scope and description of the research and the goal to be achieved were clearly outlined in the invitation emails as well as web-based annotation guidance materials.
- It was clearly outlined from the outset that the tasks had to be performed online (using the lab or home computer). The lab sessions also helped in enhancing the usability and overall functionality of the system (through focus group data collection techniques) in addition to the performance of core crowdsourcing tasks.
- The training materials remained available online and were also sent to the participants, ahead of the actual participation experiment.

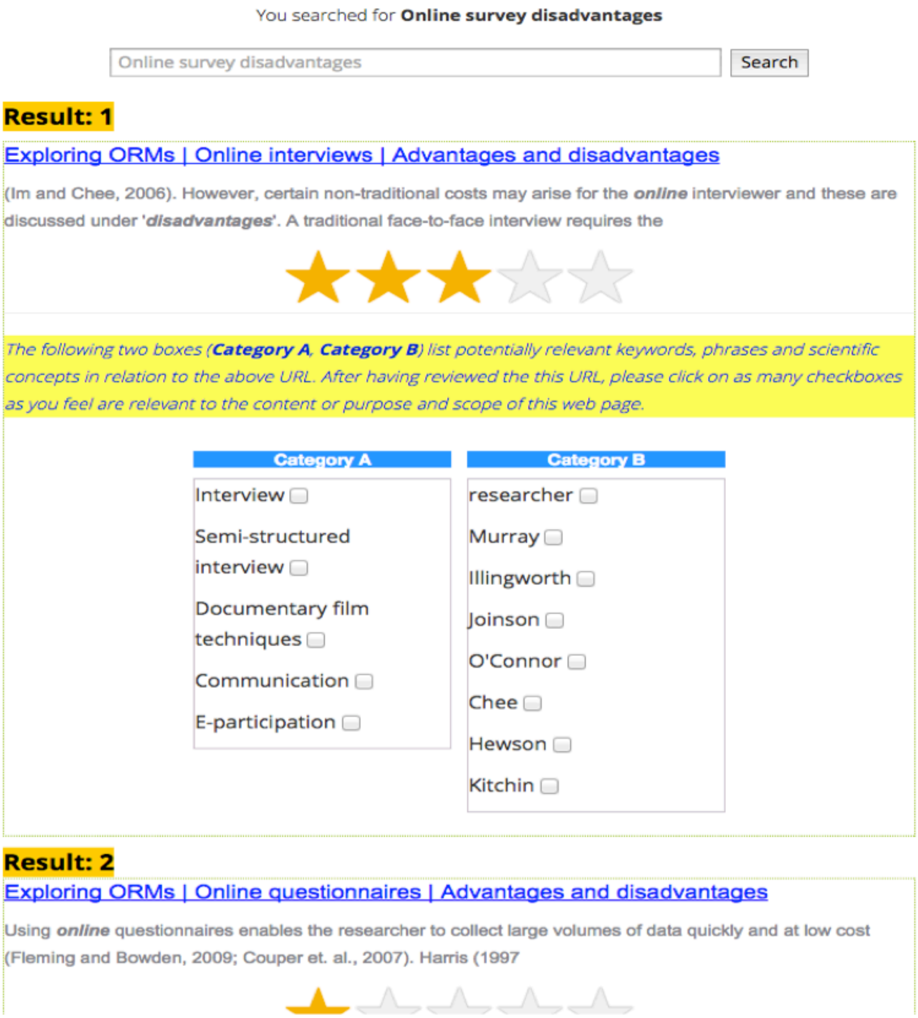
6.4 Experiment A: Lexical vs. Semantic search

This experiment aimed to investigate whether automatic semantic indexing on top of full-text index helps in improving relevance in search results. As shown in Figure. 6.1, this was the first investigation by (Khan et al., 2015) after having annotated the entire ReStore website using the LoD-supported Alchemy API. After having added up topical keywords, concepts and entities to 3400+ documents and had them stored along with the relevant API-generated score for each Elasticsearch document, precision and recall analysis (based on 25 participants) was conducted as part of the search results evaluation. The total number of concepts identified in 3400+ documents was 13629 along with 71, 613 keywords and 18, 770 entities. Only those concepts, entities and keywords were identified and indexed whose relevance score was more than 0.3 and the sentiment value was positive or neutral. The rationale behind that cut off (0.3/1.0), was to filter out *border-line* keywords, entities and concepts (in terms of relevance) and focus on more relevant extracted entities. A total of 25 participants (i.e. librarians, Social Science academics, Social Science research fellows) and 8 advanced PhD students (coming from various disciplines e.g. Social Sciences, Education, Geography & Environment and Statistics) participated in the search results evaluation. The evaluation experiment followed the automatic semantic annotation of content in the ReStore repository. The search results evaluators were tasked to use the pre-selected set of queries for search results retrieval. Their assessment included whether (a) a search result is relevant or not after viewing the content of the results by clicking on the link; and (b) ranking the result in terms of the number of *stars* corresponding to each results; and (c) validating potential concepts/entities from the list having association with each result. A copy of the invitation email, sent out to potential participants for this experiment, has been shown in Appendix B in Figure B.9.

After entering individual queries in the search box, a participant was expected to classify a result as either relevant or irrelevant i.e. 1 for relevant and 0 for irrelevant web documents. The star-rating of results and tagging relevant concepts and entities, retrieved along with individual results, will be used for measuring average ranking across the set of queries in Chapter 7. Each participant was provided with 7 queries from the benchmark query collection which required them to analyse 70 documents (10 per query) and rank each search result.Their searching, ranking and semantic tags corroborating activities were recorded in the database against a unique user session. A total of 886 documents were evaluated and 2,555 semantic concepts and entity tags were added up by the search results evaluators. The search results page showed 10 results to users with a summary for each highlighted matched words in the query, which helped users make a quick sense of the result before clicking on the link. Furthermore, in star-rating a result, one star means not relevant or a worst match and 2 *stars* means barely relevant. 5 *stars* means the most relevant results in terms of relevance to the query in question. Each participant was then required to assign at least 2 tags form each category i.e. Category A, Category B, which were relevant, as per their understanding, to the content of the web page. Category A listed up top 10 tags (representative of each result), which comprised of topical keywords that explicitly existed in the actual content. Category B, on the other hand, listed up LoD-based automatically generated concepts, entities, representing each result (10 results in total), as shown in Figure 6.3.

It is important to note here that some queries were evaluated by more than one evaluator (several queries were evaluated 5 times against *SemDex* and *SemCrowDex*) in which case the ranking and precision were averaged before being used in the evaluation analysis. On the basis of their assessment, MAP (Mean Average

Precision) was computed and the TREC[6] 11-points Precision/Recall curves were constructed in the next Chapter 7 to show that enhanced semantic metadata attachment to the actual content clearly improves precision in search results with maximum recall.



You searched for **Online survey disadvantages**

Online survey disadvantages

Result: 1

Exploring ORMs | Online interviews | Advantages and disadvantages

(Im and Chee, 2006). However, certain non-traditional costs may arise for the ***online*** interviewer and these are discussed under '***disadvantages***'. A traditional face-to-face interview requires the

★★★☆☆

*The following two boxes (**Category A**, **Category B**) list potentially relevant keywords, phrases and scientific concepts in relation to the above URL. After having reviewed the this URL, please click on as many checkboxes as you feel are relevant to the content or purpose and scope of this web page.*

Category A

- Interview □
- Semi-structured interview □
- Documentary film techniques □
- Communication □
- E-participation □

Category B

- researcher □
- Murray □
- Illingworth □
- Joinson □
- O'Connor □
- Chee □
- Hewson □
- Kitchin □

Result: 2

Exploring ORMs | Online questionnaires | Advantages and disadvantages

Using ***online*** questionnaires enables the researcher to collect large volumes of data quickly and at low cost (Fleming and Bowden, 2009; Couper et. al., 2007). Harris (1997

★☆☆☆☆

Figure 6.3: Search results evaluation interface for full-text and semantic index-based searching

6.4.1 Semantic entities and concepts tagging during evaluation

Alongside actual results evaluation and ranking of each result, participants also tagged relevant concepts and entities, presented to them by the system next to each result. The tagging has been of help in understanding participants' decision element for ranking a particular result. For example when *"forecasting"* is searched, one of the concepts that was suggested to users for tagging in a few

[6] Text Retrieval Conference is actually a set of relevance benchmarks organised by the National Institute of Standards and Technology (NIST)-http://trec.nist.gov/overview.html

results was *"prediction"*, *"decision theory"*, *"Bayesian inference"* and *"statistical inference"*. Those relevant concepts had already been identified by the automatic semantic annotator but participants' tagging enabled the author to re-validate the system's accuracy, which is reflected in assessing the degree of user happiness or MAR. Computation of MAR has been detailed in Chapter 7 under Section 7.4. In contrast, when *forecasting* was searched in the existing online search facility of ReStore, most of the results in top 10 results were retrieved because of the mention of the word *forecasting.* Likewise, when *forecasting* was searched by multiple users as part of our evaluation, they highly ranked a result, which was at no 3 position in the top 10 list and it had no mention of *forecasting* but the content were about a research tool used to predict housing, income and education situations of participants taking part in a case study. Google's top 10 results, however, included those results which defined *forecasting, Meteorological office forecasting* and *baseball game forecasting* thus making a different contrast to the above-mentioned results. The purpose of highlighting these examples is to highlight the variation in retrieved results and potential responses from the evaluators in terms of ranking and tagging.

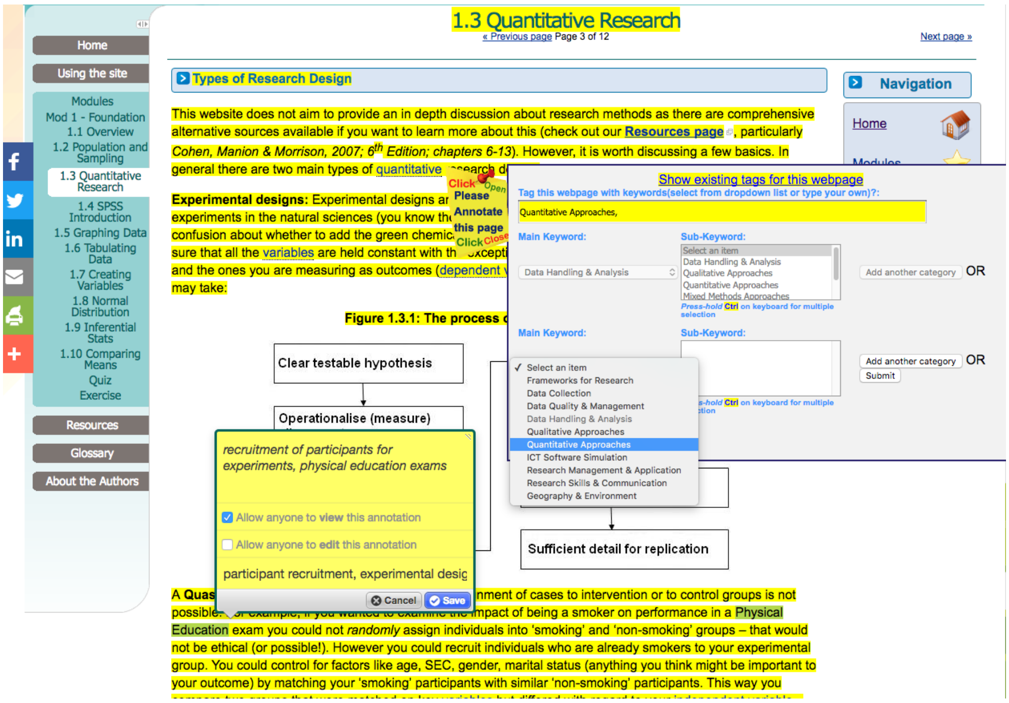
6.4.1.1 ***Discussion*: supplementary-tagging and ranking**

Tagging at the time of searching, influences the ranking criteria in SRR system. For example, after having searched a query "*mixture model*", the SRR system determines that the terms don't explicitly exist in the document vector space under the full-text "topical keywords" sub-index, but it has a high score under the "concepts" sub-index. Similarly, terms of another query "*Randomized control trials*" explicitly exist in the vector space under "keywords" sub-index, but with a low score; and a high score under the "concepts" sub-index. Thus, when searched, the documents having instances of such concepts, with a high score, will be presented to the user as highly relevant documents. Likewise, when "*reasoning*" is searched, a document containing "*critical thinking*" concept comes up first in top 10 search results; and it is subsequently considered as "relevant" by the evaluators. This trend of better performance in semantic search results evaluation has been demonstrated in Chapter 7 in Figure 7.3 and Figure 7.4.

Participants-based semantic tagging, at the time of search results evaluation, will be analysed in Exp.C (6.6), where suggested tags are presented to search results evaluators from two sources i.e. automatically generated and manually generated (crowd-annotation-based). Such assessment has been done to ascertain the proximity of agreement, on tagging web resources, between human and machine annotators.

6.5 Experiment B: Crowd-annotation & tagging

In this experiment, as depicted in Figure 6.1, the research community was involved in the expert annotation process by using the custom-built annotation tool (*AnnoTagger*) embedded in the entire ReStore repository website. As part of this experiment, participants were required to (a) access a web page, representative of their research interests: either using online search or using a pre-selected list of URLs (sent out in the email from the author), (b) scan the web page and annotate the inside text or content (*in-page annotation*), and (c) assign at least 3 keywords using the *page-level slider* (as illustrated in Figure 6.4). Participants for this annotation experiment were recruited using various sources of communication as explained in Section 6.3. Figure 6.4 shows a screen shot of the web page, which has been annotated by the expert annotators using the *AnnoTagger* (As explained in Chapter 4 under Section 4.6.5 and 4.3). The *AnnoTagger* tool[7], as shown in



1.3 Quantitative Research

« Previous page Page 3 of 12 Next page »

Types of Research Design

This website does not aim to provide an in depth discussion about research methods as there are comprehensive alternative sources available if you want to learn more about this (check out our Resources page, particularly *Cohen, Manion & Morrison, 2007; 6th Edition; chapters 6-13*). However, it is worth discussing a few basics. In general there are two main types of quantitative research d

Experimental designs: Experimental designs ar experiments in the natural sciences (you know th confusion about whether to add the green chemic sure that all the variables are held constant with th xcepti and the ones you are measuring as outcomes (dependent v may take:

Figure 1.3.1: The process

Show existing tags for this webpage

Tag this webpage with keywords(select from dropdown list or type your own)?:

Quantitative Approaches,

recruitment of participants for experiments, physical education exams

☑ Allow anyone to **view** this annotation

□ Allow anyone to **edit** this annotation

participant recruitment, experimental desig

A Quas nment of cases to intervention or to control groups is not possible. impact of being a smoker on performance in a Physical Education exam you could not *randomly* assign individuals into 'smoking' and 'non-smoking' groups – that would not be ethical (or possible!). However you could recruit individuals who are already smokers to your experimental group. You could control for factors like age, SEC, gender, marital status (anything you think might be important to your outcome) by matching your 'smoking' participants with similar 'non-smoking' participants. This way you

Figure 6.4: Screen shot of annotation and tagging environment *AnnoTagger* inside a web page

Figure 6.4, becomes available to each participant after having logged in to the system using the *login page* and anonymised credentials. A list of pre-selected URLs was provided to each expert annotator (27 in total) based on their research interests and their preference for certain social science topics. The need for generating a list of pre-selected URLs, based on experts research interests, was arisen in pilot experiments (prior to actual experiments). It was observed in those sessions,

[7] A demo page https://goo.gl/de5AkC showing the annotation tools in action.

that the experts spent time, first, on locating an annotable page before annotating and tagging the content. The total time, spent on searching and annotating the page, was therefore inconsistent with the goal and objectives, set out for this experiment i.e. number of pages to be annotated with in a stipulated time period. A worksheet, containing pre-selected keywords, harvested from Google Analytics, was also sent out, in case a participants may want to use meaningful keywords, for locating an *annotable* web page. It was noted at the end of the experiment, that majority of the participants simply clicked on the pre-selected links and annotated the page; while some preferred to use the online searching facility (in combination with pre-selected URLs)to track down web pages for annotation and tagging. A total of 27 expert annotators annotated 450 web pages, with 640 comments made on the content of these pages. They also provided 1670 typology or vocabulary tags and 350 free-text tags. The typology-based tags comprise of two levels: a broader level called `vocabularyAnnotation.level_1.level1` and a narrower level called `vocabularyAnnotation.level_2.level2` in query formulation. Annotators made use of 17 different broader level typology tags in 298 instances while 66 different narrower level typology tags were used in tagging web pages 318 times. The *allinOne* field in the *tagging* slider plugin offers the *autocomplete* feature to annotators based on typology terms as well as popular tags. A tag became a popular tag when it was used by participants at least 3 times.

More than 400 broader and narrower typology terms were offered through the AnnoTagger tool (using custom-built auto-complete) to enable annotators to assign at least 3 different tags to each web page being annotated. The auto-complete not only facilitated existing word selection but also influenced new keyword formulation which led to establishing new relationships between documents at the time of searching. For example, *"sociolinguistics"* term was used for content related to *socio-demographic* while *"economy, society and space"* terms were used for content having the general theme of *economy*. Similarly, *"critical discourse analysis of text"* term was used for *discourse analysis*, and *"corpus linguistics"* term was assigned to a paragraph describing *corpus & documentary analysis* and so on.

Some individual experiments were conducted in focus group sessions, which helped the author discover the benefit of offering popular and typology tags through an up to date *auto-complete search box*. It also helped participants formulate new keywords which were the source of enhanced linkage between documents at the time of search retrieval in the next phase. Annotators took 90 minutes on average to complete the task of annotating/tagging 15-20 web pages but they had the freedom of attempting it at their own convenience by logging on to the system. This approach was adopted to distribute the participants in two groups i.e.

focus groups for local participants, to understand their behaviour to annotation & tagging and improve the system on the fly; and those intending to complete at the place of their choice in multiple attempts as per their convenience. The outcome and findings from this experiment were published by Khan et al. (2017), arguing that crowd-sourced annotation improves the performance of SRR systems in web repositories.

6.5.1 Questionnaire and participants feedback

To provide for the basic usability components i.e. *learnability, efficiency, memorability* and *satisfaction* (Gabrilovich and Markovitch, 2007) and measure them in each case, expert participants' feedback was received at the end of each individual annotation exercise. The short questionnaire included questions such as how desirable was it to annotate a piece of text or tag an entire web page, the usability of both annotation and tagging tools, the suitability of typology terms for associating with web pages, and their willingness to assign their own keywords for tagging a set of web pages. A 5-point Likert scale for interpreting participants' feedback was used i.e. *Strongly Agree, Agree, Neither agree nor disagree, Disagree* and *Strongly Disagree.* It was quite encouraging to note that almost 80% answered *Strongly agreed* to questions on usability and ease of use. 85% answered *strongly agreed* to finding text relevant to their research topics in a web page and were able to annotate the content. Only 30% *agreed* that they felt the need for re-annotating already annotated content in a set of web pages. 55% answered *agreed* to question on formulating their own keywords when the existing popularity-based and vocabulary tags exhausted in the auto-complete drop-down on the first few words typed into the text box.

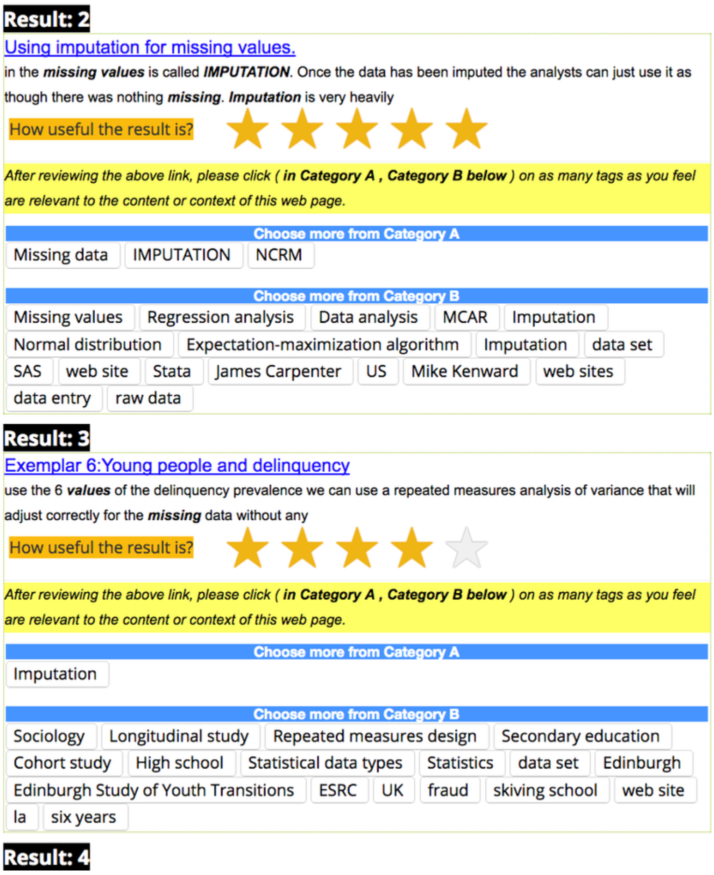
The outcome of the crowd-annotators' feedback is presented and discussed in Figure 8.9 in Chapter 8. Questionnaire, consent forms and invoice for crowd-annotation have been given in Appendix B in Figure B.4, Figure B.5 and Figure B.8. This experiment was aimed at collecting experts' collective intelligence in the form of free-text and vocabulary annotations and tags. However, no evaluation based on search results retrieval was performed to assess the retrieval performance against a benchmark queries set. The author therefore elaborates the search results retrieval performance in Experiment C (6.6) using two indices i.e. *SemDex* and *SemCrowdex* (see Figure 4.1) to ascertain the validity of the claim that crowd-sourcing-based expert annotation in web repositories improves and sustains relevance in search results.

6.6 Experiment C: Search results retrieval evaluation

Following on from the crowd-annotation experiment, a search results evaluation study was needed to ascertain the influence of expert crowd annotation over the semantic index created in Exp.A (6.4). This experiment, Exp.C, was highly significant because the search results retrieval was completed by more than 30 participants (non-expert multi-disciplinary students), who were paid for their participation, as shown in Figure 6.1. Participants were largely students in the Faculty of Social & Human Sciences, Web & Internet Sciences research group and various other interdisciplinary researchers from across the UK Universities. Some participants were locally available for a focus group session while others preferred to attempt the tasks online after having gone through the instructions and guidance materials. A copy of the invitation email, sent out to potential participants for this experiment, has been shown in Appendix B in Figure B.14. As detailed in Section 6.2, a list of 7 queries was provided to each participant, making sure that each query is attempted along with results evaluation and ranking. However, in the grand evaluation, some queries were found to have been repeated more than 3 times, in which case, precision has been averaged for all retrieved results, both against *SemDex* and *SemCrowDex*. Payments of 10 British Pounds/participation were made to all participants except some researchers who voluntarily preferred not to be paid for.

The evaluation included: (a) whether a search result is relevant or not after viewing the content of the results by clicking on the link, (b) ranking the result in terms of the number of *stars* corresponding to each result, and (c) intellectually verifying potential concepts/entities from the list having association with each result. The total number of search results retrieved were 1572 against 101 distinct queries. The number of distinct web pages evaluated by the participants were 435. The total number of distinct semantic concepts associated by the participants to all web pages was 442. All these concepts were initially generated as part of the LoD-based semantic annotation in Exp.A. In this experiment, those concepts were retrieved by the ES search application against users' queries as "relevant complementary concepts". All of them appear as tag clouds around search results in Figure 6.5 for evaluators to click on (as many as relevant), based on their semantic relevance with the respective search result.

After having completed 30 participants, a grand evaluation of search results retrieval was performed (as illustrated in Figure 6.1), based on participants' search results evaluation in Exp.A and Exp.C. Only those queries were selected which were used by participants in Exp.A (6.4) as well as participants in Exp.C (6.6) in order to evaluate the performance of two indexed systems i.e. *SemDex* and *SemCrowDex*.



Result: 2

Using imputation for missing values.

in the ***missing values*** is called ***IMPUTATION***. Once the data has been imputed the analysts can just use it as though there was nothing ***missing***. ***Imputation*** is very heavily

How useful the result is? ★★★★★

*After reviewing the above link, please click (**in Category A , Category B below**) on as many tags as you feel are relevant to the content or context of this web page.*

Choose more from Category A

Missing data | IMPUTATION | NCRM

Choose more from Category B

Missing values | Regression analysis | Data analysis | MCAR | Imputation | Normal distribution | Expectation-maximization algorithm | Imputation | data set | SAS | web site | Stata | James Carpenter | US | Mike Kenward | web sites | data entry | raw data

Result: 3

Exemplar 6:Young people and delinquency

use the 6 ***values*** of the delinquency prevalence we can use a repeated measures analysis of variance that will adjust correctly for the ***missing*** data without any

How useful the result is? ★★★★☆

*After reviewing the above link, please click (**in Category A , Category B below**) on as many tags as you feel are relevant to the content or context of this web page.*

Choose more from Category A

Imputation

Choose more from Category B

Sociology | Longitudinal study | Repeated measures design | Secondary education | Cohort study | High school | Statistical data types | Statistics | data set | Edinburgh | Edinburgh Study of Youth Transitions | ESRC | UK | fraud | skiving school | web site | la | six years

Result: 4

Figure 6.5: A typical search results page showing numbered results (10 in total) with tags clusters in Category A, B

6.6.1 Discussion

As shown in Figure 6.5, a search result page typically lists top 10 numbered search results against each query along with related tags, comprising of semantic concepts, entities, free-text and vocabulary tags. Each participant was then required to assign at least 2 tags form each category i.e. Category A, Category B.

Category A, however, in Exp.C (unlike Exp.A) listed up top 10 tags, which comprised of typology annotations and free-text keywords and Category B listed up LoD-based automatically generated concepts, entities and popular keywords, representing each result. In some cases, participants didn't star-rate (considered not relevant) a particular search result retrieved against their query; but they still did assign semantic tags from the two categories, as potentially relevant to the content of that particular result. Using Category B, the number of distinct(not including the repetition) entities assigned by the evaluators to search results, was 456. Similarly, using Category A, the total number of free-text tags, associated to search results, by the participants in Exp.C, was 369. The number of NCRM Typology tags association was totalled as 745. Such disparity in the number of total associated tags, for two distinct groups of tags, shows the variation in participants' subjective decisions, based on their understanding and preferences. The fact that the participants didn't know the type of tags (typology or free-text), they were associating each result with, further vindicates the author's claim that typology-based crowd-annotation and free-text tags can be extensively used; to address the issues of content heterogeneity, obsolescence of concepts in web repositories at the time of searching. The search application used by the participants can be accessed at https://goo.gl/od7WHc

6.6.1.1 Order of search results annotation & tagging

In the focus group sessions, it was observed that participants were not following one universal pattern of search results ranking and tagging. Higher star-rating of results was usually followed by several tags association and vice versa. Some participants preferred to associate tags to search results of their choice before even rating the results against the submitted query. In the guidance materials, however, it was emphasised that the search results rating was the most important activity following the submission of a query to the search engine. Having said that however, the percentage of pages star-rated, tagged remained at 95% and in some cases, more than the expected queries were submitted and results evaluated by the participants. Among the 5% non-evaluated results, most of the results were tagged but weren't assigned any ranking star. In the Precision/Recall analysis in Chapter 8 such pages have been classified as relevant where the number of associated tags were 3 or more.

6.6.1.2 Expandability and reproduceability of the platform

In addition to measuring relevance of results against queries, one of the purposes of carrying out a range of annotation and ranking experiments was to assess the system's *extensibility* and platform *reproduceability*. The characteristics, which the author wanted to take into consideration while conducting these experiments included: (a) the amount of technical efforts required for the reproduction of the experimental platform, and (b) the robustness of the system in the face of increasing volume and types of content. Since the author has developed and operationalised the entire system in a *client-server* environment, it would be perfectly valid to reproduce the entire infrastructure in a different domain of interest or academic discipline etc. Point (b), however, has been fulfilled partially in that in Exp.A, the entire content of the ReStore website were semantically annotated (using Alchemy API) but the crowd-annotators only annotated the web pages (static and dynamic) in the ReSore repository. The reason being, was the pilot annotation exercises, which the author had conducted before the actual experiments, it became quite clear that most of the pilot participants were opening web pages or clicking on web pages instead of *PDF* or *Word* documents for potential annotation and tagging. However, given the extra amount of time dedicated to each individual session and the availability of non-web-page document inside similar container for annotation (e.g. PDF doc opening inside a web page ready to be annotated), the experimental system can be further expanded to include more heterogeneous objects of data from multiple repositories.

6.7 Summary

In this chapter, a number of experiments aimed at *lexical-semantic* and *lexical-semantic-crowdsourced* annotation of content were presented for better search results retrieval in web repositories. The core of the experimental analysis of annotation-based search results retrieval lies in the real time impact of semantic annotation on search results retrieval. Moreover, this chapter also presented the parameters and variables of each experiment and outlined the recruitment criteria for expert crowd-annotators and search results evaluators along with benchmark queries preparation. Furthermore, this chapter also highlighted the *reproduce-ability* and *expandability* elements of the experimental platform, which would be crucial when it comes to annotating, collating, indexing and crowd-annotating the content of multiple web repositories.

The next Chapter 7 details the performance metrics being used to assess the quality of relevant information retrieval in web repositories. Chapter 7 also presents the outcome of Exp.A (6.4) and discusses improvement in the retrieval of relevant search results attributed to the LoD-based semantic annotation of content in web repositories.

Chapter 7

Lexical-semantic search results retrieval & evaluation

In this chapter, the author puts to test the two searchable indices; one containing full-text content of the ReStore repository website, called FT index, and the other containing full-text and automatically generated semantic annotations i.e. *SemDex*. The performance of the search results retrieval system has been analysed in terms of search results relevance, precision and recall in two different categories i.e. (1) searching on the basis of keywords and actual content (2) searching on the basis of topic keywords, semantic concepts and entities extracted by Alchemy APIs. In the benchmark document collection, 3,400 documents have been annotated and a semantic index has been populated, ready to be exploited by the SRR system.

7.1 Evaluation of SRR system

In the evaluation analysis, it is assumed that every web user would typically want every result on the first page to be relevant (high precision), but have little interest in knowing, let alone looking at every document that is relevant. To address this phenomenon, precision/recall measures have been used to determine the system performance in terms of assessing evaluators' responses to each result following the submission of a query. Average Precision (AP) has been calculated on query level (20 queries in total) using TREC's 11-points average recall/precision rates. List of benchmark queries has been given in Appendix A (Table A.3), which shows all 20 queries, used in Exp.A. Moreover, comparison has been made between the keywords-based full-text (FT) index and *semDex* index; based on the SRR system's performance, in terms of precise search results retrieval, from the two indices. To properly quantify the level of relevance in both scenarios, a combined measure

(Mean Average Precision-MAP) has been used that assesses the precision/recall trade-offs. Furthermore, MAR (Mean Average Ranking) has also been computed to ascertain the degree of relevance in terms of users' satisfaction based on users' ranking in individual results evaluation sessions.

7.1.1 Top 10 search results evaluation

The evaluation analysis has to determine how many relevant pages r=r^1,r^2, . . .r^n} could be retrieved in the top 10 pages, which were produced by the SRR system, against each query (a) from the keywords index Q(k) = {k^1, k^2, k^3...k^7} and (b) the semantic index Q(s) = {s^1, s^2, s^3...s^7}. In other words, how best could the SRR system interpret keywords in users' queries, turn them into topical keywords, concepts and entities, and retrieve a ranked list of documents, based on the score at the time of searching (using the pre-defined *Search analyzer*). Precision at *k* documents has been our assumption throughout the experimentation process, which implies that the best set of search results appears on the first page of results set and the total number of best results is 10. This approach is usually appropriate as most users scan only the first few (e.g. 10) hyperlinks that are displayed on the search results page. Similarly recall at *k* documents is based on our assumption that the relevant documents at the time of submitting each query will remain 10 documents. This approach has also been adopted based on the nature of web searching in the ReStore repository which hosts archived content and most users are, by and large, interested in the first 10 results to maximise their satisfaction in terms of finding relevant results.

The author recognises that there may be a need to go beyond the 10-pages mark, but given the observations made during the three experiments involving participants, almost 95% clicked on one or more top 10 retrieved results for ranking and tagging of results. In addition, based on our analysis of Google-based queries (Google Analytics), it was discovered that the first 100 popular queries in Google search listed results from ReStore (amongst other website links) in the top 10 results. In other words, almost every result from the ReStore was among top 10 results (not in *k+1*) which were retrieved against those 100 popular queries in the Google search.

7.1.1.1 Keywords-based full-text searching

The criteria, used to assess the system and actual performance is explained in Table 7.1. The table shows the Keywords-based (Natural Language) Precision/Recall contingency table explaining the criteria for retrieving relevant web pages against users' keywords. Furthermore, the retrieval of search results, based on these contingency tables scenarios, is explained in Chapter 8, where these criteria have been modified: by including the crowd component on top of the lexical keywords and semantic keywords, concepts and entities.

	Truth	
Our System	**Relevant**	**Not Relevant**
Set of web pages selected based on keywords annotation and were considered relevant by the system	True Positive-TP	False Positive-FP
Set of web pages not selected based on keywords annotation and were considered irrelevant by the system	False Negative -FN	True Negative-TN

Table 7.1: Keywords Precision/Recall contingency table

7.1.1.2 Semantic-index based searching

Table 7.2 shows the Semantic annotations-based Precision/Recall contingency table explaining the criteria for retrieving relevant web pages against users' keywords.

	Truth	
Our System	**Relevant**	**Not Relevant**
Set of web documents selected for having semantic concepts, entities, keywords, crowd-produced annotations and considered relevant by the system [**RETRIEVED**]	True Positive-TP	False Positive-FP
Set of web pages not selected based on concepts, entities, keywords annotation and considered to be irrelevant [**NOT RETRIEVED**]	False Negative -FN	True Negative-TN

Table 7.2: Precision/Recall contingency table based on semantic concepts/entities/tags annotations

The precision and accuracy of the system in identifying the relevant results against a submitted query is based on the ranking score computation, as elaborated in Section 5.6.1 (Chapter 5). In this, and in the forthcoming sections, the system's efficiency has been put to real time testing (as part of Exp.A (6.4)) and Precision/Recall, MAP, MAR analyses have been performed as part of investigation.

7.2 Precision/Recall/Average Precision metrics

As mentioned earlier, the first benchmark result set contains 10 documents which the system has retrieved based on users' information needs. The evaluators have judged the documents and have classified the documents on the basis of a "Relevant" and "Not Relevant" scale, as given below.

Precision is P(relevant | retrieved) and Recall is P(retrieved | relevant) and, as per contingency tables above, it could be further simplified in the following:

$$P = \frac{tp}{(tp + fp)}$$

$$R = \frac{tp}{(tp + fn)}$$

Table 7.3 and Table 7.4 show Precision/Recall, Average Precision and Interpolated Precision for a single query, searched against FT index and semantic index *SemDex* respectively. A query-based (*finite population correction*) analysis has been presented as an example; in order to highlight the overall precision/recall computation, required for the evaluation of two systems. The tables explain various measures, which have been used for the performance evaluation of SRR systems. Individual *precision/recall* values have been obtained against every result R (top 10 results) along with *ranking* values. The results have also been expressed in terms of Average Precision (AP) and Average Ranking (AR) for the query *finite population correction* out of the 20 benchmark queries. The variation in figures has been depicted in Figure 7.1 and Figure 7.2. The data in Table 7.3 and Table 7.4 clearly indicate that the relevance of search results, retrieved from semantic index (SemDex) is greater than the full-text index.

Query Results Measures	Finite population correction (searched against FT index)									
Result	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
P	1	1	1	0.75	0.6	0.66	0.57	0.62	0.55	0.5
Int. P	1	1	1	0.75	0.66	0.66	0.62	0.62	0.55	0.5
Ranking	4	3	2	0	0	2	0	2	0	0
AP	Average Precision									0.72
AR	Average User Ranking									1.3

Table 7.3: Computing Precision/Recall/Interpolated Precision, Average Precision and Average users' ranking for one query out of 20 benchmark queries, when semantic annotations are not included in the searching index.

Query Results Measures	Finite population correction (searched against SemDex index)									
Result	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
P	1	1	1	1	0.8	0.83	0.85	0.875	0.88	0.8
Int. P	1	1	1	1	0.88	0.88	0.88	0.88	0.88	0.8
Ranking	5	4	3	4	0	3	3	2	2	0
AP	Average Precision									0.91
AR	Average User Ranking									2.6

Table 7.4: Computing Precision/Recall/Interpolated Precision, Average Precision and Average users' ranking for one query out of 20 pre-selected batch of queries when semantic annotations are included in the searching index.

Figure 7.1 and Figure 7.2 show the Precision/Recall graph in two situations i.e. Keywords-based full-text search and semantic index-based search; after a query "*Finite population correction*" was entered in the search box. Table A.2 in Appendix A shows all variables needed to collect data against searching for one single query in a single results evaluation task.

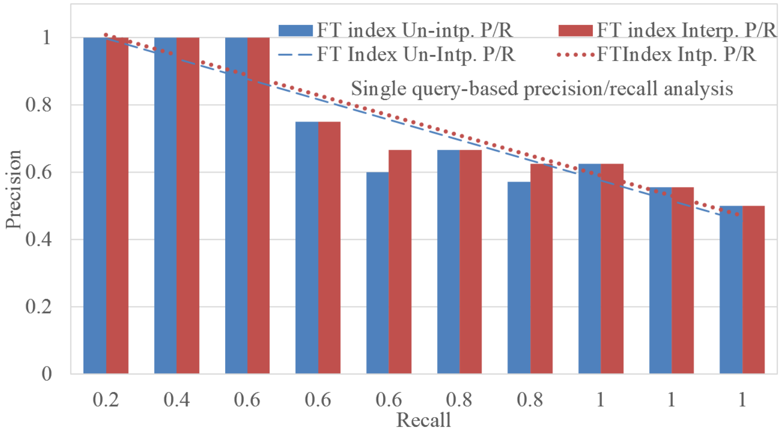


Figure 7.1: Precision/Recall graph based on a single query against FT index only

Figure 7.1 performs very well in the first 3 results against the query, by retrieving first 3 relevant results. However, the following two results (4th, 5th) were classified as irrelevant by the participants, which decreased the precision against the increasing recall. On the contrary, participants considered first 4 results, retrieved against the *SemDex* in Figure 7.2, as relevant, which maintained a steady precision, despite the decreasing recall. The overall trendline shows better performance for *SemDex* as there were only 2 irrelevant results. However, five out of ten results, were relevant when the query was searched against FT index, which explains the lowest precision (0.5) in Figure 7.1 .

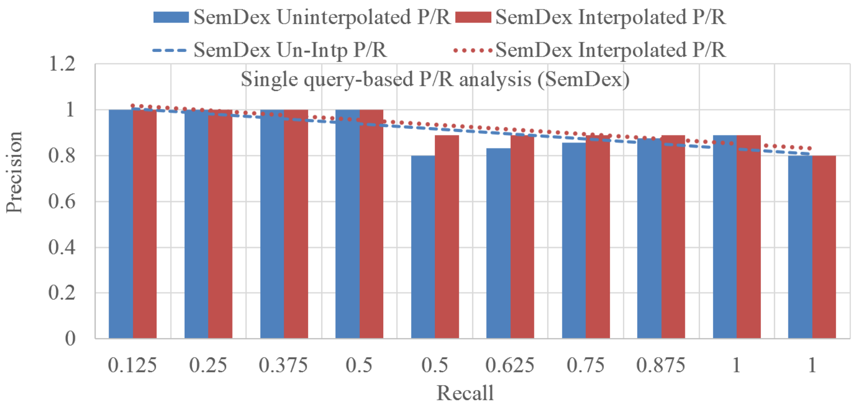


Figure 7.2: Precision/Recall graph based on a single query against *SemDex* using query *Finite population correction*

7.3 Interpolated AP vs. Uninterpolated AP measurement

It has been assumed, during the evaluation of search results, that a user will examine a fixed number of retrieved results and precision will be calculated at that rank and interpolated rank. Hence the fixed average precision is given by: $P(n) = \sum_{n=1}^{N} \frac{r(n)}{n}$ where r(n) is the number of relevant items (at cut off k relevant document) retrieved in the top n which in our case is 10 documents (N) at each level of individual information needs in the form of user queries. However, if (k + 1)th retrieved document is not relevant, precision will drop but recall will remain the same. Similarly if (k+1)th document is relevant both precision/recall increase. Therefore these measures have to be extended by using ranked retrieval results, which is a standard with search engines. Interpolated precision is therefore given by:

$$P_{11-pts} = \frac{1}{11}\sum_{j=0}^{10}\frac{1}{N}\sum_{i=1}^{N} P_i(r_j) \tag{7.1}$$

where $P(r_j)$ is the precision at the recall points but $P(r_j)$ does not coincide with measurable data point r if number of relevant documents per query is not divisible by 10 in which case, the interpolated precision is given by: $P_{interpolated}(r) = max\{P_i : r_i \geq r\}\ where(P_i, r_i)$ are raw values obtained against different queries or information needs. So, the new average interpolated precision is given by:

$$P_{11-pts-interpolated} = \frac{1}{11}\sum_{r\in\{0,0.1,0.2,.......1.0\}} P_{interp}(r) \tag{7.2}$$

In other words, interpolated precision shows the maximum of future precision values for current recall points. An increase in both precision and recall means that the user is willing to look at more results. All of this tells us about the expected precision/recall values for another set of results (k + 1,2,3...n). Since $P(n)$ ignores the rank position of relevant documents retrieved above cut off (i.e. 10 + 1), interpolated precision has been calculated at each query level to assess the system performance at n + 1th documents, which is beyond the existing cut off point. Figure 7.3 shows the TREC-11 points ranked retrieval precision/recall curve which is representative of the two systems i.e. topical keywords or full-text search vs. semantic index-based searching (*SemDex*). Furthermore, the Interpolated Precision and Recall curve shows SRR system performance i.e. keywords (FT) searching vs. *SemDex*-based searching over the entire queries batch (See Table *A.3* in Appendix A). Tables A.4 and Table A.5 in Appendix A show the actual data used to construct Figure 7.3. A consistent outlook, illustrated in Figure 7.3, is representative of the SRR system, which retrieved results from FT index and *SemDex* using 20 benchmark queries. The rate of increase in recall, causes the precision to fall more rapidly in the case of FT-index. On the contrary the fall in precision is way more slower in *SemDex* case, which shows the strength and accuracy of the SRR system, while using the latter index. However, given the nature of search engine performance, it is important to keep into account the corpus size e.g. what happens after more heterogeneous documents are annotated and indexed. Our system has been tried in one repository; so it would be a good future work to include another repository and carry out further experiments to validate the system performance.

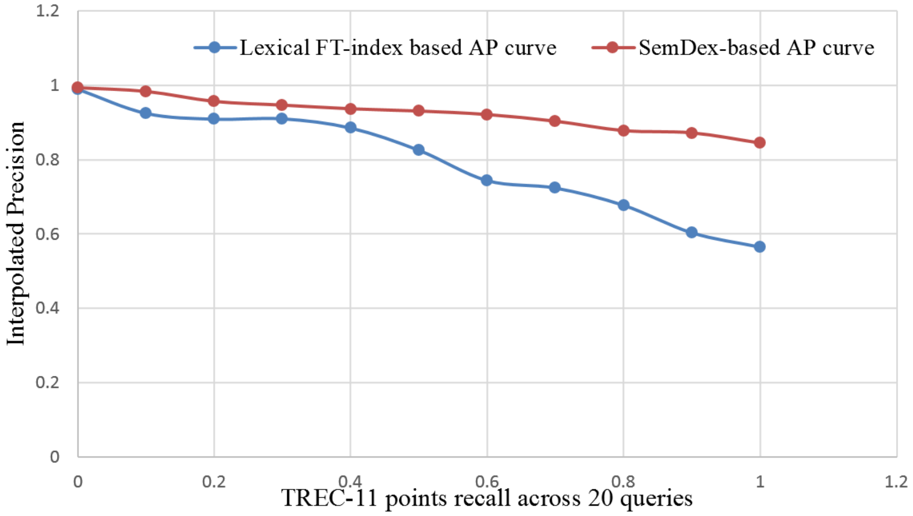


Figure 7.3: TREC 11 points Interpolated Precision and Recall curve showing system performance i.e. keywords (full-text) searching vs.*SemDex*-based searching

7.3.1 Fixed uninterpolated AP

Uninterpolated average precision was also calculated at each query level which is given by:

$$P(r) = \sum_{i=1}^{N_q} \frac{P_i(r)}{N_q} \tag{7.3}$$

where P(r) is the average precision at Recall level r and N_q is the number of queries. $P_i(r)$ is the precision at Recall level r for the i-th query. The uninterpolated Precision/Recall graph demonstrates a comparatively different but consistent trend in Figure 7.4, based on uninterpolated AP values. Table A.6 in Appendix A shows the actual data used to construct Figure 7.4.

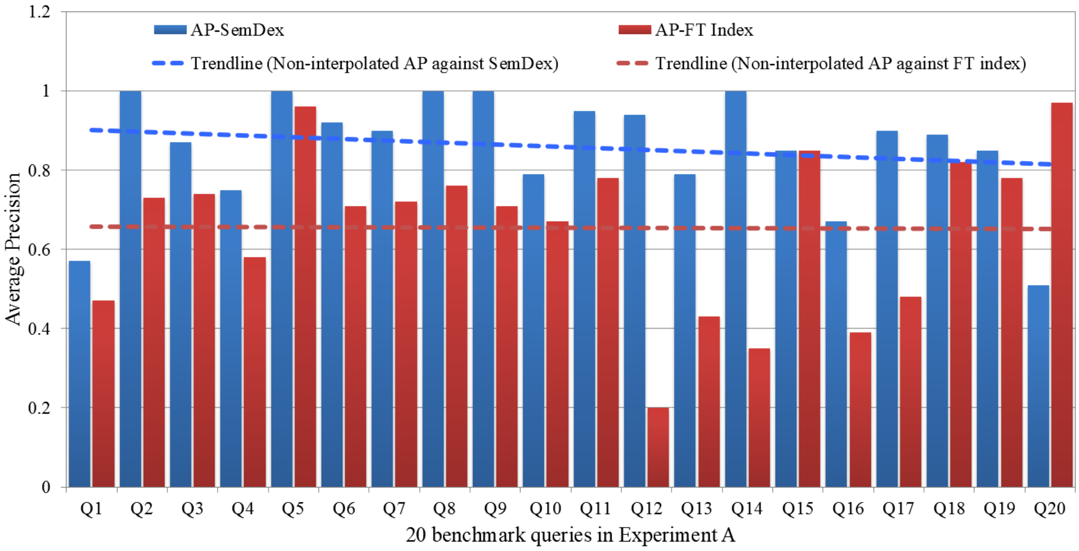


Figure 7.4: TREC 11 points uninterpolated Average Precision (AP) curve showing precision/recall performance over the entire set of queries. *SemDex*-based curve performs better.

The trendline in Figure7.4 demonstrates that the uninterpolated full-text (FT) AP fluctuates between 67 and 70. On the other hand, the linear trendline, representing the *SemDex* AP, shows precision performance between 80% and 90% across the entire 20 benchmark queries. As highlighted in Chapter 6, the precision for each benchmark query was averaged (assigned by two evaluators), before reporting the overall performance of SRR system against FT and *SemDex* indices. Figure 7.4 clearly shows better performance in terms of uninterpolated P/R when queries were searched against *SemDex*.

However, there are exceptions of a few cases, where queries, searched against FT index, perform better i.e. Q18, Q20. Moreover, SRR system posted the lowest AP score for Q12 (*reasoning*), when searched against FT index, compared to a very high score, when the query was searched against *SemDex*. This anomaly, is due to five irrelevant results out of 10 results, retrieved from FT index, against this query. On the contrary, eight out of 10 search results were considered relevant by the participants, which explains the disparity in columns in Figure 7.4, for the query in question. While analysing the performance of Q18 (*ethnic groups*), against both indices, it is noted that the number of relevant results against FT index is nine out of 10 and eight out of 10 against *SemDex*. In this case, the relevance has regressed apparently due to the *synonyms*-based association (during automatic semantic annotation using custom analyzer), defined at the time of index creation. It is also noted that the first 3 results, in both cases, were considered relevant, which explains that, the association of NEs, concepts and synonyms didn't negatively affect the relevant score of top 10 results. However, it is difficult to conclusively assume and expect such outcomes, based on the averaged performance of 2 queries. Looking at Q20 (exploratory factor analysis), however, reveals that the high score, attributed to the FT index, is due to the more relevant results (nine out of 10), compared to eight out of 10 relevant results in *SemDex*. The noticeable difference in the two columns, however, is due to the first two irrelevant results, retrieved against *SemDex*. This phenomenon, is however, offsetted by the interpolated precision-recall curve in Figure 7.3; where the semantic-index-based performance remains consistent, across the 20 queries.

7.3.2 Mean Average Precision (MAP) calculation

MAP has provided an insight into whether enhanced semantic annotations have improved search results using the purpose built search application. Average precision has been computed at each standard recall level across all queries, with the conclusion that MAP in *SemDex*-based searching performs better than the full-text index. Using MAP, fixed recall levels are not chosen and there is no interpolation. MAP has been used to assess the overall precision across the 20 benchmark queries, for both keywords-based and semantic searching. MAP is given as: $MAP = (\sum_{i=1}^{Q} AP_i)/Q$ where Q = number of queries in a batch. MAP ensures that equal weight is given to all queries i.e. those containing rare and common terms with different recalls. The calculated MAP measures for search results, retrieved against FT and *SemDex*, are 66% and 86% respectively, which highlights the precision in relevant results retrieval using *SemDex*.

7.4 Mean Average Ranking

Precision and Recall curves do not allow for the degree of relevancy as such when it comes to retrieving precise and relevant documents. This is partly because these measures are based on binary classification and individuals' perception. What is relevant to one person may not be relevant to another. To address this issue, the author has averaged the ranking of all relevant documents against each query in the set of baseline queries. The averaging is done to calculate the combined Mean Average Ranking (MAR) figure which represents the author's ranking model, applied in Figure 7.5. MAR is computed here as:

$$MAR = (\sum_{j=1}^{Q} AR_j)/Q \tag{7.4}$$

where Q = number of queries in a batch and AR_j is the average ranking of results against each query. The combined MAR for the entire 20 queries, is, therefore, 1.42 , when the users searched against only lexical *full-text* index. On the other hand, MAR for *SemDex* is 2.70.

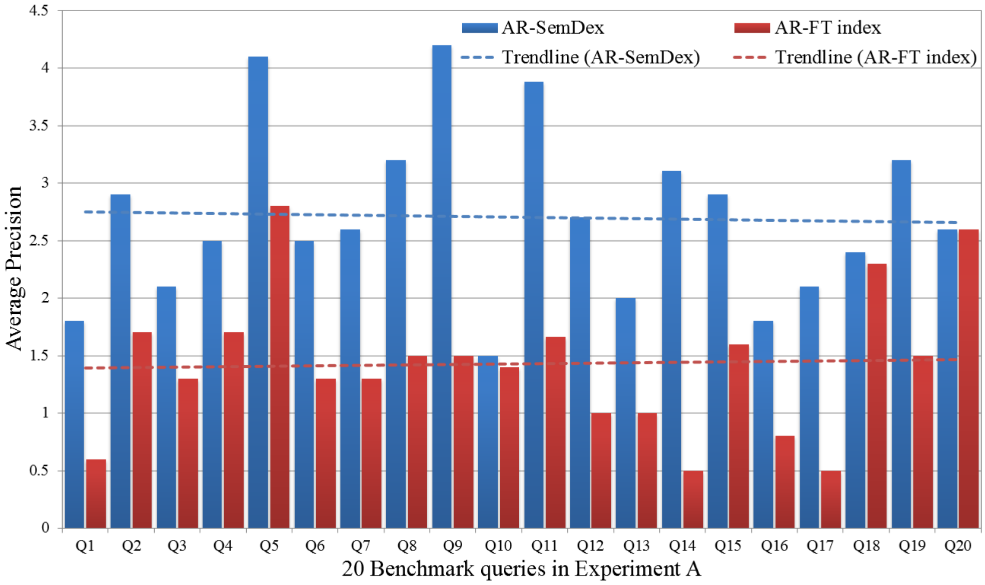


Figure 7.5: Average ranking curve showing average ranking at each information need or query level across the set of benchmark queries

Figure 7.5 also highlights, some interesting cases, where AR for two types of indexes, remained the same, competitive or different to each other i.e. Q10, Q17, Q18 and Q20. Searching Q10 (*Evaluating interaction effects*) against full-text index, 5 out of 10 results were classified as relevant, and subsequently ranked (in terms of stars) by the participants in Exp.A. Similarly 6 out of 10 results, against

SemDex were classified relevant and subsequently ranked by the participants. The slight difference in ranking is due to the number of relevant results (5 and 6), which indicates that the semantic concepts and entities, didn't make a huge difference to the ranking for this particular result. Likewise, the ranking score for Q20, is the same, when the query was searched against full-text index and *SemDex*. The higher number of relevant results in both cases, indicate that the query-document relevance score, based on actual full-text content, is sufficiently high and semantic indexing has barely enriched the content. In the case of Q17, it is noted that only 2 results were deemed relevant, when searched against full-text index and 8 results were considered relevant against *SemDex*. Furthermore, 5 results were ranked 3-stars and 3 results got ranked 2-stars, after their retrieval from *SemDex*. On the contrary, both relevant results, retrieved from full-text index, were ranked 2-stars. Furthermore, the trend in Figure 7.5 supports the claim, that users achieve greater satisfaction, while searching their queries against *SemDex* using the SRR system. Such disparities in ranking, for some queries, explain the importance of full-text indexing, which justifies the author's hybrid indexing approach towards heterogeneous documents indexing, as explained in the ealrier chapters.

7.4.1 Summary

In this chapter, the author has introduced the evaluation measures and presented the performance of SRR system using FT and *SemDex* indices. The author has also discussed, in detail, the query representation in both systems in terms of precision/recall graphs; and demonstrated better performance, achieved by the *SemDex*-based SRR system. The ranking of search results, along with the relevant/non-relevant classification, helped the author to assess the impact of semantic annotation in web repositories on sustainable basis.

Having said that however, the author still understands that human intervention in the content classification and knowledge organisation in web repositories will further improve relevance in search results; despite changes in communities, language, culture and disciplinary transitions. To address the "human intervention" or *crowdsourcing* component, the author looks into the *crowd* element of the semantic annotation in Chapter 8. The *crowd-annotation*-based semantic annotation and tagging will add another dimension to this research. It will demonstrate that using controlled and un-controlled vocabularies, expert crowd-annotators can play a pivotal role in preserving the performance of search engines over time.

Chapter 8

Lexical-semantic & Crowdsourcing-based search results retrieval and evaluation

The author has evaluated the performance of the search engine in Chapter 7 based on data obtained from users' participation in experiment Exp.A (6.4). It has also been investigated that by structuring the heterogeneous content of web repositories, using automatic semantic annotators, the performance of retrieval systems can both be improved and sustained. In this chapter, the author is adding the crowd component to the semantic annotation framework, to ascertain the impact of expert crowd annotation on search results retrieval. This chapter details the crowd-supported IR in ES and ranking score computation when a particular document is annotated and is supposed to be retrieved against a query (subject to *query-document* relevance). The evaluation in this chapter includes the evaluation of two experiments i.e. Exp.B (6.5) and Exp.C (6.6). The former deals with the manual annotation of content in the ReStore repository, and the later assesses the impact of those annotations/tags on relevant search results retrieval. In other words, the evaluation is performed by comparing the ranking & tagging feedback of participant evaluators from Exp.A and Exp.C. Furthermore, the evaluation includes a benchmark set of queries carefully selected from Exp.A and Exp.C aimed at comparing performance in terms of accurate search results retrieval from two indices: one having crowd-annotation element *SemCrowDex* and the other having no crowd component *SemDex*. The author also looks at the sustainability element of the entire framework in order to answer one of the research questions (1.3) in Chapter 1. In addition, this chapter also analyses the association of tags to search results by the expert crowd and non-expert results evaluators in Exp.B and Exp.C respectively to assess the degree of agreement between the two different groups of crowd.

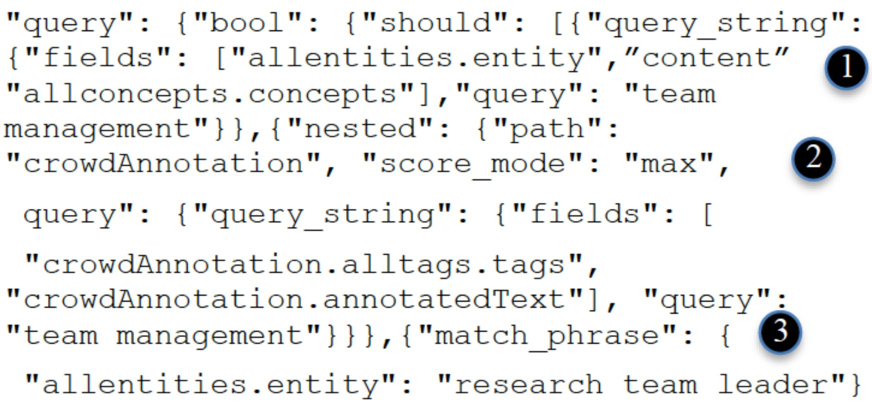
8.1 Crowd-supported IR in Elasticsearch

In order to practically benefit from the crowd annotation and tagging, the SRR system facilitates the evaluation of relevance in search results, retrieved against *SemDex* and *SemCrowDex* indices. Furthermore, in order to link whatever the users type in the search box with the available popular tags, a fully-fledged auto-complete feature has been mounted on the search box. This was another dimension, added to the investigation in Exp.C (6.6), in order to understand users' preference for existing searchable terms (free-text or vocabulary tags) or newly typed keywords at the time of submitting queries.

As outlined in the preceding chapters, following the submission of a query, a typical search engine matches the query terms with indexed tokens to gather all matching documents. The retrieval system then ranks the documents using the scoring criteria, before showing the top results to a user. It is important to note that search engines and *document matchers* are not focused on classification of new documents; rather their primary goal is to retrieve documents. In this chapter, as part of the search results evaluation, the Elasticsearch-based SRR system will determine the relevant pages r={r^1,r^2,...r^n} retrieved as part of the top 10 results against each query searched in semantic index (*SemDex*) and Hybrid semantic index (*SemCrowDex*). The query representation in semantic index is expressed as Q(s) = {s^1, s^2, s^3...s^7} and in hybrid semantic index it can be expressed as Q(c) = {c^1, c^2, c^3...c^7}. Let x be a set of all annotation elements x={*full-text, keywords, entities, concepts, crowd-annotations*} such that $Q(c) \in x \bigcap Q(s) \in x$. Each participant has searched for 7 queries in the hybrid semantic index; each one, then, has evaluated 10 results (70 results at minimum) per query, followed by *star-rating* of each result and associating relevant representative tags to each result, as explained in Exp.C (6.6). In the next section, the author talks about VSM, query-document relationship using CSM in order to highlight how best the system interprets keywords, entities, crowd-annotations at the time of SRR.

8.1.1 Manipulation of weights for Relevance Maximization (RM)

The hybrid semantic indexing and searching platform offers the flexibility of term score manipulation at query time i.e. retrieving those results having a specific named entity with the maximum score in addition to the matched query's terms in other fields. The ES scoring algorithms and potential scenarios have been elaborated in Chapter 5 under Section 5.2.3 and Section 5.6.1. Here, a few examples are described where RM for the top 10 results can be achieved by fields-based manipulation in the crowd-annotated semantic index *SemCrowDex*. For example, the following query fetches results from the Elasticsearch KB based on users' query *team management*. In simple terms, the user wants to get all relevant documents having content on *"team management"* and the fields to be searched for the query include *content, allconcepts, allentities, annotatedText and sourceText*. The *content* field is a full-text field and the rest are semantic fields in the hybrid semantic index. The filter being applied for maximum relevance is `allentities.entity` field, which must match those documents, which have entity of type "*research team leader*".



```
"query": {"bool": {"should": [{"query_string":
{"fields": ["allentities.entity",”content”
"allconcepts.concepts"],"query": "team
management"}},{"nested": {"path":
"crowdAnnotation", "score_mode": "max",
 query": {"query_string": {"fields": [
 "crowdAnnotation.alltags.tags",
"crowdAnnotation.annotatedText"], "query":
"team management"}}},{"match_phrase": {
 "allentities.entity": "research team leader"}
```

Figure 8.1: Elasticsearch query for data management keywords with a filter on specific Entity

The above query in Figure 8.1 retrieves results based on a cumulative score, which is 13.39 for the first result. The lowest score is 0.026, which shows variation in the maximum and minimum scores for a given query as above. The most important aspect of the above query is that the relevance score is calculated based on the 3 components, labelled 1, 2, 3 in the figure.

By executing the above query, 252 results are retrieved with the top most result having a score of 13.39. However, when component 2 (`crowdAnnotation`) is removed from the query and the query is re-executed, a similar number of results are retrieved but they are sorted based on their different score computation. The maximum score for result no 1 is now 9.19 but it is evident that the score has been calculated purely based on lexical and semantic content in the index with no weight manipulation caused by the expert annotations.
The next filter component (3) in the query (Figure 8.1) is the type of entity, which could be specified by the user at the time of search after the first set of results is retrieved (based on component 1, 2) against a given query. For example, in

the above query, *team management* partially matched with `annotatedText` as well as `sourceText` fields, but only one of the two words matched with the actual content of the page. However, since every result has to conform to the 3rd component i.e. a result should have an entity of type "JobTitle" and label value "*Research team leader*", the relevance in top 10 results increases greatly. Moreover, conformance of results to component 3 is not a must (due to loose filter *should* on line 1 of the query); as it is preferred to get filtered results, but it is left to the SRR system to calculate the score based on the combination of components.

In another scenario, the keywords are replaced in the above query (Figure 8.1) with *multilevel modelling* and entity of type *Person* having label value of *Patrick Sturgis*, the total number of results produced by the search engine is 50 with a maximum score of 1.75. Furthermore, observation of the top most result in the top 10 results reveals that *multilevel* and *multilevel modelling* exist in many fields including the crowd-annotation field. However, the 3rd component does not conform to the name of the entity of type *Person* i.e. *Patrick Sturgis* but the search engine has still listed the page as top of the 10 results. That is because the filter criteria is based on *should* instead of *must* which implies that ideally a result should conform to the *type* filter but not always. Removing the 2nd component (2), however, from the query produces 40 results, with 1st results conforming to component 3 but with no presence of any crowd activity on the page whatsoever. In this case, Elasticsearch has applied the standard TF-IDF scoring algorithm to retrieving relevant search results but the page popularity, in terms of crowd annotations and tags, have influenced the ranking of the page in top 10 search results.

8.2 The angle of document-query relevance

In Figure 8.2, a user searches for "*Multilevel modelling*" but wants to filter out results based on association of content with various semantic entities and discipline-specific vocabulary. As evident in the figure, a query vector representation shows weight 5 for *Multilevel* and 2 for *modelling*. Doc1 is closer in terms of a smaller angle, but the closest document to the query is doc3, based on other factors (a_w), in addition to the mere presence of words, in those documents as illustrated in Figure 8.2. Figure 8.2 also highlights that the association of vocabulary tags to documents by the crowd-annotators (a_w) also prove to be a source of retrieval (based on closer document-query angle) of these documents among the top 10 search results.

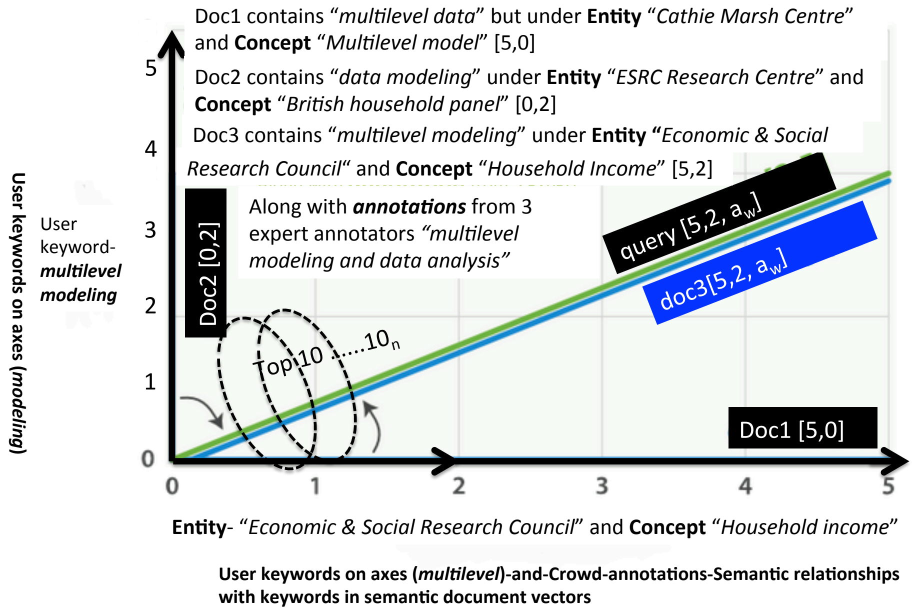

Doc1 contains "multilevel data" but under Entity "Cathie Marsh Centre" and Concept "Multilevel model" [5,0]
Doc2 contains "data modeling" under Entity "ESRC Research Centre" and Concept "British household panel" [0,2]
Doc3 contains "multilevel modeling" under Entity "Economic & Social Research Council" and Concept "Household Income" [5,2]
Along with annotations from 3 expert annotators "multilevel modeling and data analysis"
Doc2 [0,2]
query [5,2, a_w]
doc3[5,2, a_w]
Top 10 .....10_n
Doc1 [5,0]
User keyword- multilevel modeling
User keywords on axes (modeling)
0
1
2
3
4
5
0
1
2
3
4
5
Entity- "Economic & Social Research Council" and Concept "Household income"
User keywords on axes (multilevel)-and-Crowd-annotations-Semantic relationships with keywords in semantic document vectors

Figure 8.2: Two-dimensional representation of query vector in Vector Space Model(VSM)

8.2.1 Assignment of score to documents

A mechanism for assigning a score to a *query-document* pair has been outlined in Chapter 5 along with presentation of the most relevant search results to users. This section further builds on that mechanism in the context of adaptable *query-document* vector space. In the semantic search and in the case of enhanced crowd-annotations, the emphasis has to be on the context of a term rather than the occurrence of lexical, semantic or crowd-annotated terms in a document or the collection of documents in a hybrid vector space. Along with "how many times" the query term occurs in the document, the interest should be in "where" the term or word occurs and "how important" it is to be considered worth placing in the top 10 search results. The document length, however, affects the score computation following the automatic and manual semantic annotation of documents. In order to measure similarity between the query and document, normalisation of length takes priority in order to measure the proximity of documents in an adaptable VSM space. In other words, a document vector can be length-normalized by dividing each of its components by its length i.e.

$$|V| = \sqrt{\sum_{i=1}^{N(i.c.e.a)} v_i^2} \tag{8.1}$$

where components i,c,e,a represent additional layers of annotation in a document vector which ES will use to calculate the score for ranked documents retrieval.

In order, to visualize a semantic document vector in a |V| dimensional vector space, the user's query can be thought of as a query vector. Document terms lie on the axes of the vector space and document vectors are points, which will be multi-dimensional in a *query-document* vector space. All document vectors having close proximity to query vectors in the space will be ranked higher. In terms of the searchable document vector, an elongated document vector is indexed in an ES KB. The document vector is adaptable in terms of content (changed or modified content with potential semantic annotation) and ranking score depending on requirements in the future. In terms of the query vector, based on the data obtained from Google analytics, user queries are not abnormally lengthy but they are not single term either, which will be a benefit when calculating IDF later in this section as part of the Cosine similarity calculation. Most measures of vector similarity are based on the Dot product, which is given in Equation 2.5.

8.2.2 Adaptation of classical VSM using crowdsourcing-based annotation

Adaptation of the classical VSM is exploited fully when expert crowd-annotations are semantically indexed, ready to be considered in the ranked retrieval score computation. Figure 8.3 shows this phenomenon in a 3-dimensional VSM where a 3-terms query (*participatory longitudinal research*) is being searched in a hybrid space of lexical, semantic and crowd-annotated documents.

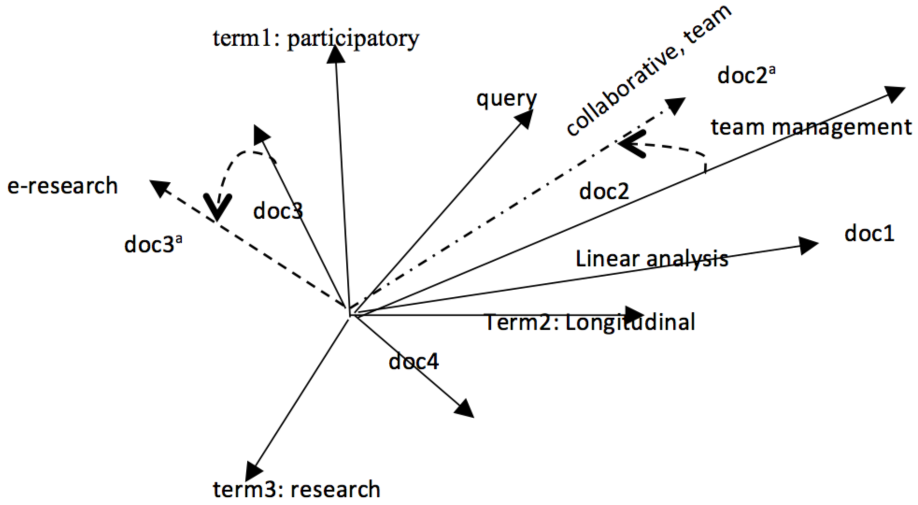


Figure 8.3: 3-dimensional Vector Space Model with a query, 4 documents and 3 terms

The length normalisation of each document and query vector by the ES is obtained by dividing each component of a single vector by its length. The relative distribution of terms is thus offset in a set of documents and proximity and relevance is computed against the query vector in question. In other words, long and short documents' vectors now have comparable weights after new annotations were first added automatically (in Exp.A); and then by the crowd (in Exp.B and Exp.C). New and contemporary tags assignment by the expert annotators thus becomes significant in terms of establishing relationships between the two documents and their ranking in the top 10 search results. Figure 8.3 clearly illustrates the tilt in the angle of `doc2` and `doc3` towards the axis of query terms. The more enriched the hybrid vector space is with the contemporary scientific terms and vocabulary terms, the more accurate search results ranking will be, as explained in the various Precision/Recall figurative analysis in this chapter.

8.3 Incorporating the crowd element in SRR framework

Online users typically express their information needs in the form of a query, which comprises of a set of keywords submitted to a search application. The application retrieves relevant information in the form of documents, which the search algorithms assess to be relevant to users' information needs. Relevance here represents the similarity between the *selected* and *suggested* results. A screen shot of a document having actual content and annotation components, stored in ES KB has been given in Appendix C Figure C.1. It is important to look at the SRR system from the perspective of crowd-annotation and tagging in order to ascertain whether this layer of semantic annotation can further reduce the angle between documents and query vectors, on a sustainable basis, in the hybrid vector space. For example, a search is made for *"social research"*, against the hybrid semantic index (*SemCrowDex*), with a filter of semantic Named Entity (NE) containing the term *"research methods"* of type "PrintMedia". One of the results in the top 10 results shows an entity *"International Social research methods case studies"* of type "Print Media". On a closer inspection of the indexed document, it is discovered that the full-text keywords list also contained *social research methods* as a top keyword due to its high relevance score. But in the annotation components of the document index, the top annotation (free text annotation typed by an annotator) is *"research methods bank"* and the source text (the text that has been selected for annotation in a web page) contains *social research methods case studies*. So the scoring was performed based on the *annotated term, sourced text, Entity mention and full-text keywords* respectively. In comparison to the first result, when the second result is observed in the top 10-result set, it is evident that there are more annotations containing the word *"research"* in them e.g. *"mobile research"*, *"e-research"*, *"online research links"*, *"research framework"* and the DBPedia concept *"Research Methods"* but with a low score of 0.59, which is not enough for the search engine to flag this result up at no 1 position. That is largely due to the comparatively larger similarity angles between the query terms and documents elements compared to the first result. Another interesting element in the first result is that the *keywords* and *concepts* both list *"Social research methods"* and *"social research"* as top keywords respectively in their token list, which is a cross of the original query *"social research"*, and entity filter *"research methods"*. This kind of heterogeneous query building (based on post-query-submission in our search application) proves to be an effective tool in retrieving most relevant search results. The *field norm* characteristics widely used in Elasticsearch in documents ranking, gives an extra weight to the number of times a web document has been annotated (the fields-norm scoring computation has been detailed in Chapter5)

8.3.1 Annotations-based relationships between repository documents

After having submitted the following query, newly relevant results in the top 10 search results, can be quickly discovered based on crowd tagging and annotation. *User to user, user to web documents, experts-tagged web resources to automatically annotated web resources* are a few to name when the search application extend the scoring criteria from full-text index to more meaningful elements of a document index. The query in Listing 8.1, highlights the possibilities to relate various web resources based on experts' annotations, expert' research interests, web page tagging or even the source text which is the text they select inside the web page to attach their annotation to.

```
GET index_SemCrowDex/_search?pretty&size=10
"query": {"nested": {
      "path": "crowdAnnotation",
        "query": {"filtered":  [#comment:tag-specific filter]
         {"query": {
       "match_phrase":{#comment:actual keyword terms
            "crowdAnnotation.alltags.tags": "methodological
   innovation"}},
             "filter":{"bool":  [#comment:annotator-specific filter]
      \{"must": [ [#comment:boolean condition with "must" attribute]
                {"term": \{"crowdAnnotation.user": "user_xyz" }}
                ]}}}
       }
```

Listing 8.1: Elasticsearch query for retrieving annotation/tagging based results

Given that most of the participants were experts in their fields and they attempted the annotation tasks very earnestly, with genuine interest in the content, it is quite encouraging to see the high number of related web pages annotated by multiple users and retrieved by the SRR system.

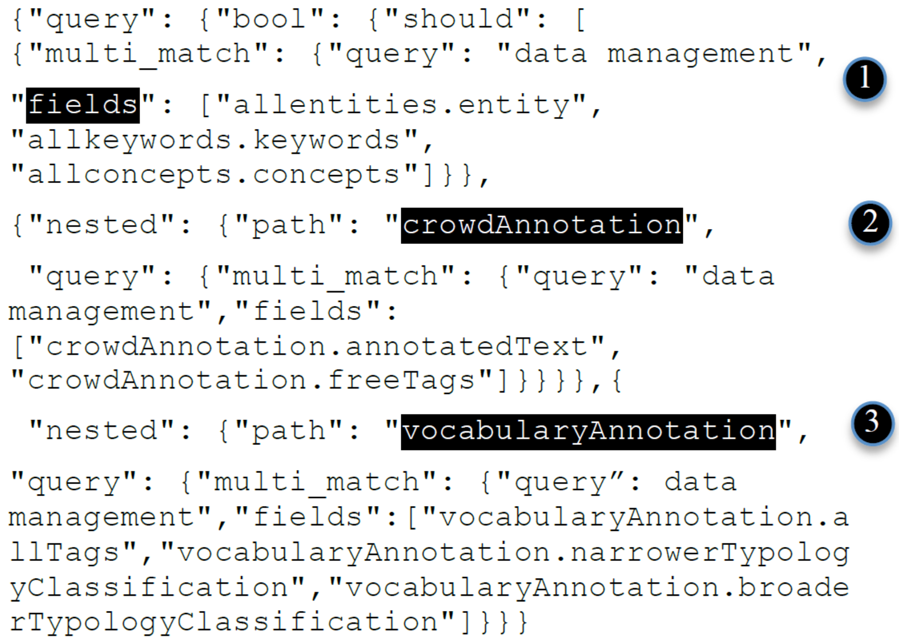
8.3.2 Annotations are more summative than semantics and *topical keywords* combined

When the query (*data management*) in Listing 8.2 is executed against the *SemCrowDex* index, the most relevant result in the top 10 retrieved results, is the one having been annotated and tagged with phrases "*data management*", "*data quality & management*" and "*data quality & data management*" by 3 different expert annotators. Interestingly, the list of keywords associated with the same document include "*classification variables*", "*large datasets*", "*smaller units*", "*conventions*" where as the concepts include "*critical thinking*", "*want*", "*need*" etc. Nowhere in the full text, has the document suggested *data management* as an activity except the title of the page where the closest phrase is "*managing your analysis*".

```
GET index_SemCrowDex/_search?pretty&size=10
"query": {"bool": {"should": [ #comment:boolean condition with "
   should" attribute
     {
      "query_string": {"fields": [ #comment:specific searchable
   fields
      "allkeywords.keywords", "allentities.entity", "allconcepts.
   concepts"],
       "query": "data management"}}, {"nested":{
         "path": "crowdAnnotation", #comment:nested field objects in
   an ES document
                 "score_mode": "max",
                  "query": {"query_string": {
                  "fields":[ #comment:specific searchable nested
   fields
                  "crowdAnnotation.alltags.tags", "crowdAnnotation.
   annotatedText"],
                   "query": "data management" #comment:actual keyword
    terms
                   }}
```

Listing 8.2: Elasticsearch query for *data management* keywords combining LoD-based semantic index with crowd-annotated index

What the above query in Listing 8.2 lacks is the connection with Typology-based annotations of the web pages. The ranking of retrieved search results will change when that query is changed to the following query in Figure 8.4



```
{"query": {"bool": {"should": [
{"multi_match": {"query": "data management",
"fields": ["allentities.entity",
"allkeywords.keywords",
"allconcepts.concepts"]}},
{"nested": {"path": "crowdAnnotation",
 "query": {"multi_match": {"query": "data
management","fields":
["crowdAnnotation.annotatedText",
"crowdAnnotation.freeTags"]}}}},{
 "nested": {"path": "vocabularyAnnotation",
"query": {"multi_match": {"query”: data
management","fields":["vocabularyAnnotation.a
llTags","vocabularyAnnotation.narrowerTypolog
yClassification","vocabularyAnnotation.broade
rTypologyClassification"]}}}
```

Figure 8.4: Elasticsearch query for *data management* keywords combining LoD-based semantic index with crowd-annotated index (including vocabulary annotation)

Now the query in Figure 8.4 searches for terms against selected fields in non-typology and typology-based annotations (2,3) along with semantic concepts, entities and topical keywords (1). The fact that all crowd-annotation fields have less data (due to shorter and meaningful annotations) in them as compared to the full-text content and title fields, they impact the retrieval scoring to a greater extent. The *field length norm* feature of the Elasticsearch scoring algorithm (see Section 5.2.1 in Chapter 5) measures smaller field by giving them higher weighting except those modified by the *boost factor*. As evident in the above query, the `vocabularyAnnotations.allinOne.tags`, `crowdAnnotation.annotatedText` and `crowdAnnotation.freeTags` are the fields, which have been filled up by users' annotation and tagging activity, hence they carry more weight when it comes to score calculation using the Elasticsearch standard scoring algorithms. Same is the case with the typology-based fields i.e. those having a prefix of `vocabularyAnnotation`, which is considered by the scoring algorithm when ranking the top 10 search results. A screen shot of a document having actual content and annotation components, stored in ES KB has been given in Appendix C Figure C.1.

8.4 Ranked retrieval in hybrid semantic Vector Space

As outlined earlier, to measure how well documents and query match, one has to look into the lengths of document vectors and get them normalized before computing cosine similarity of queries and documents vectors. For example, a document vector with lexical and semantic annotation components will have longer lengths than those having none. However, the importance and rarity of terms will still remain important elements of ranking at the time of retrieval as Elasticsearch uses TF.IDF weighting distributions to compute relevance and ranks of document in top 10 search results. IDF of the vocabulary tags (*rarity of informativeness*), added by the crowd-annotators and those added up by the Alchemy annotators, played a role in ranking those documents higher on the scale. For example, *British Sociological Association of Ethical Practice* is a statement in one of the many web pages but a certain *web page A* becomes more relevant when it was annotated by two expert annotators with "*BSA guide*" and a URL to the guide. Similarly another annotator annotated the text with a free tag "*Ethical research practices*". The length of the document vector increases with the addition of these terms and phrases and tags but the rarity of the web page has also increased for these reasons: (a) it was annotated and tagged hence IDF increases, which in turn increases the overall score; (b) more contemporary data was added linking the document with more similar documents hence in the range of small cosine angles clusters in the semantic vector space; (c) words are likely to be important, based on expert annotations and vocabulary-based annotations. Throughout this analysis, IDF has been considered as a measure of informativeness of the term and the fact that IDF affects the ranking of documents for queries with at least two terms. For example in the above query, IDF weighting makes occurrences of "*BSA*" counts far more in the document ranking than occurrences of "*guide*" for its being a common term. Also since VSM does not consider the ordering of term tokens in a document so in the crowd-annotation model and LoD-based semantic annotation, the order of words inside the vector stack does not matter. Rather the context, place, rarity and importance of the token will matter regardless of whether the token was generated from the full text, semantic annotation or crowd-annotation inside a single document vector.

8.5 Relevant search results retrieval using cosine similarity

This section presents a *query-documents* similarity score computation scenario, in a hyper semantic vector space, using one of the queries, evaluated in both Exp.A and Exp.C. A comparison of weighted document vectors has been made including non-semantic document vectors and semantic document vectors. Furthermore, cosine similarity between the query vector and each semantic and non-semantic document vectors has been computed to determine the *cosine angle* of similarity between query and documents i.e. *Doc1, Doc2, Doc3.* Log frequency TF.IDF weights of semantic document vectors and non-semantic document vectors have been given in Appendix A (Table A.12 and Table A.13). The vector representation of query "*randomized control trials*", non-semantic vectors of *Doc1, Doc2, Doc3* along with cosine similarity scores have been demonstrated in Table 8.1.

Query {*N=3400*}				**Doc1**		**Doc2**		**Doc3**	
Query terms	$W.tf_{t,d}$	DF	IDF	NW	DP	NW	DP	NW	DP
randomized	1	2	3.23	0.22	0.71	0.11	0.35	0	0
control	1	1	3.53	0	0	0.21	0.74	0	0
trials	1	1	3.53	0	0	0.47	1.65	0.29	1.02
Final Similarity Score b/w *q* & *d*				**0.71**		**.35 + 0.74+1.65=2.74**		**1.02**	
NW=Normalized weight, DP=Dot product, W.=Weighted									

Table 8.1: Cosine similarity computation between *query-documents* in a non-semantic vector space model

Table 8.1 shows that the similarity score between doc2 and query is 2.74, which suggests that, the document is closely related to the query terms after normalizing the length of the document. Doc3 and Doc1 are ranked second and third respectively. To compute the similarity score based on semantic document vectors, all the three documents have now been annotated with relevant phrases in which some terms include *control, trials,* with the exception of *randomized* term. Table 8.2, now shows, that *control* term has been added to Doc1 and Doc3, *trials* to Doc 1 and Doc 3, which have changed the normalized TF-IDF weighted score of semantic document vectors. The terms that have been added are directly augmenting the meaning of the query terms but these terms could have been one of the synonyms, defined at the time of index creation. For example web pages containing terms like *hypothesis testing research* or *experimental design* or *experimental research* would have been listed up by the retrieval system in the search results against this query based on *synonymical* relationships. The contextual proximity of these synonyms with the actual query terms *randomized control trials* is determined at the time of creating synonyms filters as part of semantic index creation (See section 5.3). After having computed, the cosine similarity between query and document vectors, different scores have been obtained in Table 8.2.

Query $\{N=3400\}$				**Doc1**		**Doc2**		**Doc3**	
Query terms	$W.tf_{t,d}$	DF	IDF	NW	DP	NW	DP	NW	DP
randomized	1	2	3.23	0.20	0.64	0.11	0.35	0	0
control	1	3	3.05	0.18	0.54	0.21	0.64	0.20	0.61
trials	1	3	3.05	0.28	0.85	0.35	1.06	0.36	1.09
Final Similarity Score b/w *q* & *d*				**.64+.85+.54=2.03**		**.35 +.64+1.06=2.05**		**.61+1.09=1.7**	
NW=Normalized weight, DP=Dot product, W.=Weighted									

Table 8.2: Showing revised weights after new terms were added through semantic or crowd-annotation

It is quite evident that Doc1 has gone further higher in the ranking score but Doc2 has regressed due to the low score for *trials* term. Similarly, Doc3 has slightly improved but due to the low cumulative score for all terms, it is now ranked 3rd. Such phenomenon impacts the overall score more sharply if the annotation was a vocabulary term instead of a free text word, which may contain more *stopwords* or repeated words. The IDF in the case of vocabulary terms/tags in annotations will increase on the basis of rarity of terms in the collection of documents; thus making the document or set of documents more relevant against a given query.

8.6 Participants-based SRR evaluation

This chapter has so far presented crowd-supported IR, weight and score manipulation, similarity of query and document vectors with and without crowd annotation elements, adaptation of classical VSM (from classical to hybrid semantic VSM) to justify the incorporation of the crowd elements in the SRR system. Further discussions were made on the possibilities and potential of query formulation to exploit the hybrid semantic index and optimise the SRR system in terms of retrieving very relevant search results. Furthermore, the ranked retrieval of search results was described using normalized TF.IDF with and without crowd-sourced annotation.

Now, in order to assess the usefulness of the SRR system in terms of accurate and relevant search results retrieval against a set of benchmark queries, performance evaluation is carried out using *SemDex* and *SemCrowDex* search indices. In this context, the performance of SRR system has already been evaluated based on a set of benchmark queries against full-text and semantic indices in Chapter 7.

Now, in this section, the enriched index *SemCrowDex* is put to test to ascertain the improvement in relevance, based on participants' evaluation in Exp.C. It is now important to analyse the systems' performance, based on how search results evaluators (participants) interacted with the system (as part of Exp.C 6.6) in terms of retrieving relevant search results; and substantiating (or not) the crowd-generated tags and annotations originality, created by the expert crowd in Exp.B.

8.6.1 Methodology:*Performance metrics*

The performance of SRR systems in this research has been evaluated in terms of very relevant top 10 results by using TREC-11 Recall-Interpolated precision, MAP (Mean Average Precision), two-tailed-t tests and Mean Average Ranking (MAR). In addition, this analysis shows that results evaluators, who are not domain-specific experts, can perform just as well as expert crowd annotators by assigning the correct and relevant semantic tags to search results. In this evaluative analysis, the focus remains on the three components of a retrieval model i.e. the set of document D, the set of information need representations or queries Q and the relevance function R $(d,\ q)$. The relevance function associates a real number between 0 and 1 for a document $d \in D$ and query $q \in Q$ to define an ordering for the documents in D with respect to the query q. In other words, a document is retrieved by the system if $R(d,\ q) = \begin{cases} 1 if\ d \rightarrow q \\ 0\ Otherwise \end{cases}$

Given a query q_i, let the documents, in the hybrid index, relevant to query q_i, be denoted by the set $\{d_i1, d_i2, d_i3, ..., d_im\}$ where $d \in D_{NEs,keywords,crowd_annotations}$ in a single Elasticsearch document. Let the ranks of the above documents be retrieved by the set $\{r_i1, r_i2, r_i3, ..., r_im\}$, where the rank r_ik corresponds to the document d_ik relevant to query q_i in the collection of documents in the hybrid ES index. Furthermore, this section also explains that the expert annotators' feedback can improve documents ranking by modifying term weights in documents against users' queries in real time (as elaborated in the earlier sections of this chapter). The author has also prepared and uploaded all the data files, used in the search results evaluation, to a web page which can be can be accessed at https://goo.gl/IZ3XQT

8.6.2 Query-document relevance

In the *query-document* relevance framework, annotation of expert crowd is employed as relevance feedback aimed at improving the ranking of documents by giving better term weights in documents against users' queries. The change in documents is achieved by crowd and semantic annotation and change in the query term is caused by external factors as highlighted in the introduction section. In other words, a document d remains relevant if its current terms $d = (x_1, x_2, ..., x_n)$ are relevant to query $q = (y_1, y_2, ..., y_n)$ where $x_i = \begin{cases} 1 if\ document\ d\ is\ indexed\ by\ term\ t_{i \in T} \\ 0\ Otherwise \end{cases}$ and

$y_i = \begin{cases} 1 if\ query\ q\ contains\ term\ t_{i \in Q} \\ 0\ Otherwise \end{cases}$ Where T represents the terms in document d comprising of full-text, NE, Concepts, Keywords and Crowd annotation while Q is a single query out of the 33 benchmark queries set. Table A.7 in Appendix A shows the list of 33 queries, evaluated in this chapter. The values of the terms in both cases, are always subject to change depending on when the content in the documents is annotated by the crowd-annotators; and then subsequently indexed in the hybrid semantic index for search results retrieval against future queries. So the probability of a document being relevant against a query P$(r|d|q)$ and the probability of the same document no longer relevant P$(\neg r|d|q)$ depends upon the presence/absence of terms in the hybrid semantic index.

8.6.3 Precision and Recall analysis

A Precision and Recall analysis has been carried out on the data collected from two experiments i.e. Exp.A (6.4) and Exp.C (6.6) in the form of (a) search results classification(relevant, not-relevant), (b) star-ranking of results and (c) associating tags from supplementary tag clouds (free-text and vocabulary tags) around each search results against a given query. Average Precision (AP) was calculated for both groups of search results evaluators in Exp.A (6.4) and Exp.C (6.6), based on interpolated and uninterpolated precision against 33 benchmark queries to ascertain the efficacy and impact of crowd-annotation on the overall relevant search results retrieval. AP score against *SemDex* for some queries vary in Exp.A and Exp.C due to the number of times a query has been evaluated and subsequently averaged. As outlined in Chapter 6 that the 33 benchmark queries set was carefully selected from both Exp.A and Exp.C in order to evolve two systems with one having the *crowd-annotated searchable fields* (part of Exp.C) and another having no crowd-annotation fields (part of Exp.A). The *precision/recall* analysis enabled the author to assess the performance of two SRR systems in terms of precise search results without compromising heavily on recall.

8.6.3.1 Fixed (FP) vs. Interpolated precision (IP) analysis

As described in Chapter 7 that in *precision/recall* evaluation analysis in this thesis, each search results evaluator has examined a fixed number of retrieved results and precision has been calculated at that rank and interpolated rank. Similar *precision/recall* computation measures have been used here as presented in Section 7.3 in Chapter 7 but in a relatively different context. As per the FP/IP analysis in Section 7.3, we fit the average IP on the data obtained from Exp.A (6.4) and Exp.C (6.6) which is given by:

$$P_{11-pts-interpolated} = \frac{1}{11} \sum_{r \in \{0,0.1,0.2,.......1.0\}} P_{interp}(r) \tag{8.2}$$

It can be observed in Equation 8.2 that IP is capable of measuring future precision values for current recall points as shown in Figure 8.5. In contrast, normal P/R curve reacts differently to such variations as illustrated in Figure 8.6. It is evident that an increase in both precision and recall means, that the users are willing to look at more results beyond the top 10 retrieved results. It is therefore derived that the (k+1,2,3..n) set of results will also remain relevant and users will continue exploring the results. $P_{11-pts-interpolated}$ therefore gets a clear edge over $P_{11-pts-non-interpolated}$ in terms of reliable results and search systems' performance.

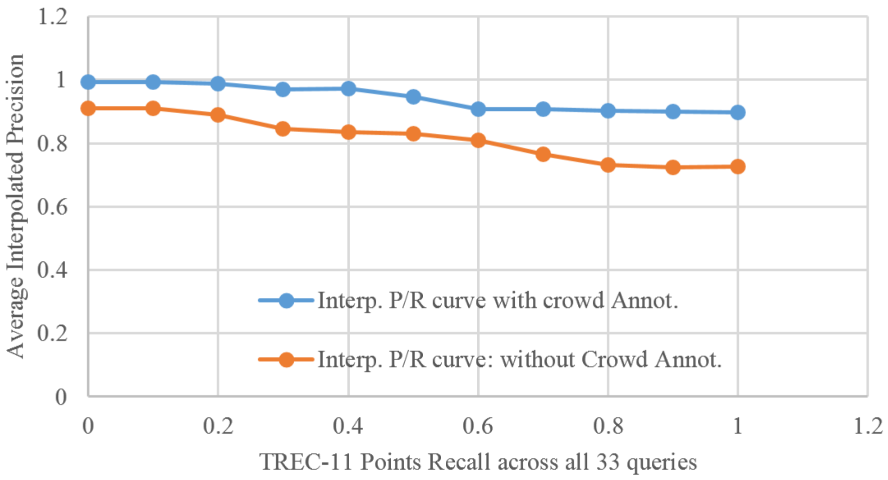


Figure 8.5: TREC-11 points Interpolated Average Precision (AP) curve showing search results retrieval performance over the entire set of 33 benchmark queries (See Table A.7). Tables A.8 and A.9 show the actual data in Appendix A.

Figure 8.5 illustrates the TREC-11 points ranked retrieval precision/recall curve, which is representative of the two indices *SemDex* and *SemCrowDex*; and the performance of SRR system has been assessed against these two indices. The figure demonstrates that the behaviour of the curve shows the non crowd-sourced hybrid tends to have slid downwards from 30% recall and 89% precision to 77% at 100% recall respectively. Contrary to that, the tendency of the curve representing the Crowdsourced hybrid semantic index, remains consistent until 40% recall at 98% precision and then tends to move downward to almost 90% of precision.

This system-level performance (33 queries) is indicative of a better search results retrieval by the SRR system against *SemCrowDex*.

8.6.3.2 Average (AP) vs. Uninterpolated precision MAP analysis

After interpolated analysis above, it is important to look at the results with a different perspective and conduct a uninterpolated analysis of the data obtained in Exp.A and Exp.C. Fitting the Non-Int-AP $P(r) = \sum_{i=1}^{N_q} \frac{P_i(r)}{N_q}$ is the right measure to conduct the uninterpolated precision analysis where P(r) is the average precision at Recall level r and N_q is the number of queries. Figure 8.6 shows average uninterpolated precision across the the 33 queries benchmark showing system performance in terms of precise search results when queries were searched against *SemDex* and *SemCrowDex*.

Performance in both situations suggests that Crowd-annotation-based graph performs better. The crowd-annotation AP trendline in Figure 8.6 also indicates better performance of the system, with the exception of a few queries where *SemDeX*-based searching performs better but the fluctuations in crowd-sourced-based searching remains minimal. The performance of *SemDex* against Q13, in Figure8.6, remains dismal, with only 2 out of 10 results were deemed relevant. *SemCrowDex*, on the other hand, performs better against the same query with only 2 out 10 results were evaluated as irrelevant. Similarly, Q15 (*what is media analysis*) was evaluated against *SemCrowDex* and the comparison was made to the query variant of Q15 (*media analysis*) against *SemDex*. The high AP score in *SemCrowDex* for Q15 indicates that the expert-annotated tags like *content analysis software, ICT software simulation, qualitative software etc.* did influence the score in favour of the index. This is further clarified in Q29 *media analysis*, which clearly shows better search results retrieval for *SemCrowDex* against *SemDex*. Similarly Q20 has posted a low score against *SemCrowDex* compared to *SemDex*. After having analysed the evaluation data, it is evident that the high score was down to the full-text keywords and LOD-based annotation; and the crowd-annotation didn't impact the relevance score for this particular query. Such findings further supports the methodology, set out in this thesis, that including the full-text alongside semantic annotation is essential part of the SRR system. Moreover, the convergence of bars in the case of Q31, is due to the fact that the first result out of 10 results against *SemCrowDex* was classified as irrelevant. In Precision/Recall analysis, the first result getting irrelevant, has ripple effects on the rest of evaluation of the remaining 9 results. *SemDex*, on the other hand performed well as the first 2 results were classified relevant followed by 2 irrelevant results. All this led to causing crowd-annotation AP bar trailing behind non-crowd-annotation AP bar.

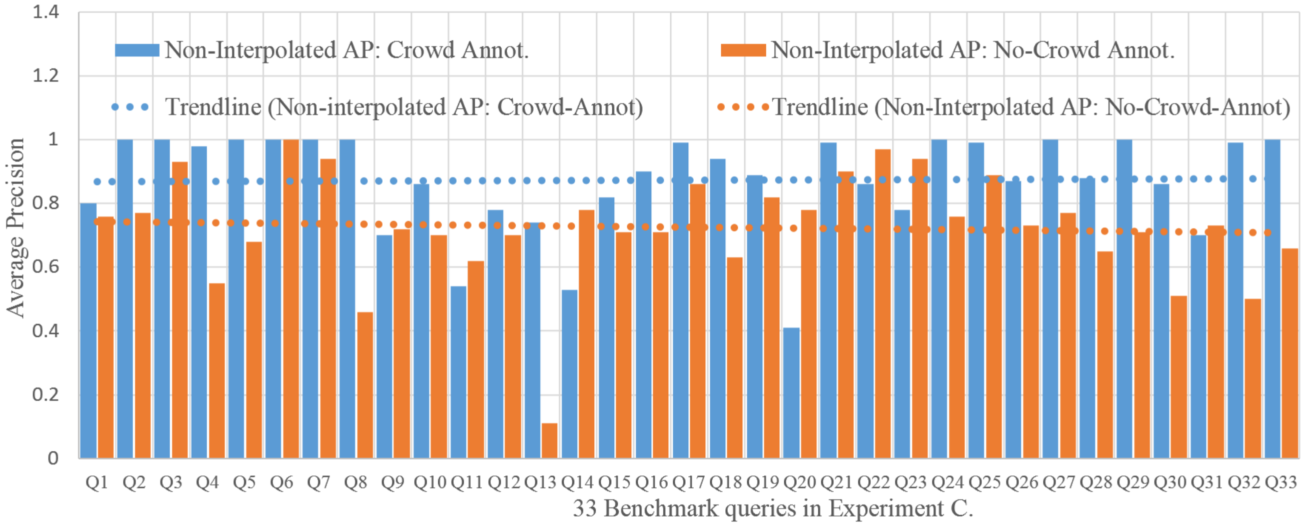


Figure 8.6: Uninterpolated Average Precision (AP) graph, showing retrieval performance over 33 benchmark queries. Table A.10 shows the actual data in Appendix A.

In terms of the single numeric measure MAP (Mean Average Precision), uninterpolated weights are used to construct Figure 8.6. MAP calculates average precision (AP) for each query and the average precision is calculated for each relevant document as: $AP = \sum_{i=1}^{R} \frac{i}{rank_i}/R\ where$ R=number of relevant docs for a particular query and $\frac{i}{rank_i} = 0$ if document was not retrieved. Hence, the overall MAP $=MAP = \frac{(\sum_{i=1}^{Q} AP_i)}{Q}$ where Q=number of queries in a batch. MAP ensures that equal weightage is given to all queries i.e. those containing rare and common terms with different recalls.

The MAP figures for two groups of evaluators (searching against *SemCrowDex* and *SemDex*), therefore, are 87% and 72% respectively. The percentage difference between the two systems resonate the interpolated difference in the Interpolated precision-recall curve (Figure 8.5) which further vindicates the author's claim that this improvement is needed on a sustainable basis in web repositories searching in order to offset the impact of change in web content with the passage of time.

8.6.4 Mean Average Ranking (MAR)

Figure 8.7 demonstrates that average ranking was calculated for each search result against each query and the individual ranking score was averaged, due to the fact that each query was used by more than one evaluator. That way the bias has been kept in check throughout the evaluation stage against the two indices i.e. *SemDex* and *SemCrowDex*. Figure 8.7 shows that the ranking of results against Q11, Q14, Q23 were higher, when searched against *SemDex*, as compared to *SemCrowDex*. On a closer look at the individual ranking of each results, it is noted that 2 out of 10 results were ranked 5-star and further 1 result was ranked 4-star, when Q11 was searched against *SemDex*. On the contrary, 2 out of 10 results were ranked 3-star and similar number of results were ranked 2-stars, when Q11 was searched against *SemCrowDex*. The primary reason for such low ranking against *SemCrowDex* is, the automatically extracted semantic *concepts* and *entities*, which provided a good match based on *query-document* similarity. The relationship between Q11 (*Mixture model*) and entity like *European Social Surveys-ESS* or concept like *multi-level models* in the document, appeared more favourable to the users for ranking the result high. Similarly, the average ranking is high for Q14 (*evaluating interaction effects*), when query was searched against *SemDex*. It is clear, that the users ranked the result higher when the query was searched against *SemDex*. In other words, the full-text and automatically extracted terms matched more favourably (e.g. stratified sampling, sample size etc.) with the query and the crowd-annotation didn't cause participants in Exp.C to rank it high, when Q14 was searched against *SemCrowDex*. Moreover, in the case of Q23 (*reasoning*), the ranking of results, retrieved against *SemDex*, remained high due to a good match between the query and full-text content as well as semantic entities. The participants in Exp.C, however didn't rank the set of results high enough against this query, due to the change in *query-document* relevance, caused by crowd-annotation and tagging. The trendline, showing the overall ranking for *SemCrowDex*, however, demonstrates marked improvement in the ranking-based user satisfaction. MAR has also been computed, which reflects the system's ranking in terms of degree of accuracy, results relevance and satisfaction level. MAR is 62% when search results were evaluated based on the*SemCrowDex*; and it is almost 47% when the search results retrieval was carried out against the

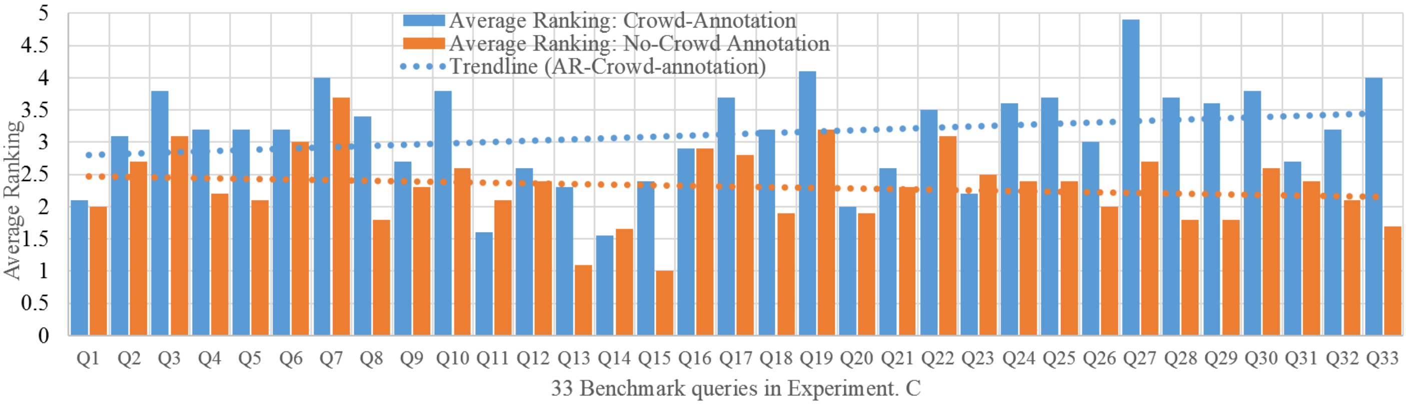


Figure 8.7: Average ranking curve shows averages ranking over the entire set of the 33 queries. Table A.10 shows the actual data in Appendix A.

SemDex. This figure is different from the MAR figure, we obtained in Chapter 7, because, we included results of queries, repeated more than three times, against *SemDex*.

8.6.5 Matching expert's tags with non-expert evaluators

Another interesting evaluation of tags, assigned to search results by expert crowd-annotators and non-expert evaluators, was performed based on data from Exp.A and Exp.C. An evaluation analysis was needed to compare the top tags (free-text and vocabulary), assigned by the expert annotators to each web page, with those assigned by the two groups of search results evaluators in both experiments. The minimum criteria specified for this task was based on the following two points:

1. Clicking on at least 3 tags from Category A and B in Exp.A (6.4)
2. Clicking on at least 3 tags from Category A and B in Exp.C (6.6). The sources of tags in category A and B in both these points are LoD-based automatic annotation, expert crowd free-text annotation & tagging and typology-based tagging as explained in Exp.A and Exp.C in Chapter 6.

The above points sum up the distinctive difference between the two tagging tasks carried out by the search results evaluators. The participants in Exp.A had to choose tags from two categories all of which were generated by the LoD-based Alchemy semantic annotations as part of Exp.A. On the other hand, in point 2, the evaluators performed the same tasks in Exp.C but the sources of tags in both categories are different i.e. Category A tags were generated by the expert crowd annotators in Exp.B (using free-text and typology) and Category B tags were generated automatically by the LoD API-based semantic annotation in Exp.A.

As per the comparison figures which have been obtained across the 33 benchmark queries, it has been discovered that a substantial number of tags were verified by the non-expert search results evaluators. The ranking and positions of tags in both Category A and Category B were based on popularity level of the tags regardless of the type of tags whether NCRM Typology (vocabulary tag), or free-text tag. That approach was adopted because the author wanted to know the reasons for *star-rating* a particular result highly or lowly, alongside the subjective ranking of each result. It helped the author to further develop confidence in the truthfulness of search results evaluators, with considerably low disagreement and disambiguation element. It also addresses the sustainable relevance point once more by verifying semantic tags of experts with the non-expert evaluators. This analysis also suggests that crowd-sourcing the task of tagging and annotating content in web repositories to online users will eventually strengthen the online search systems to retrieve relevant results, despite various changes, taking place in literature and discourse over time.

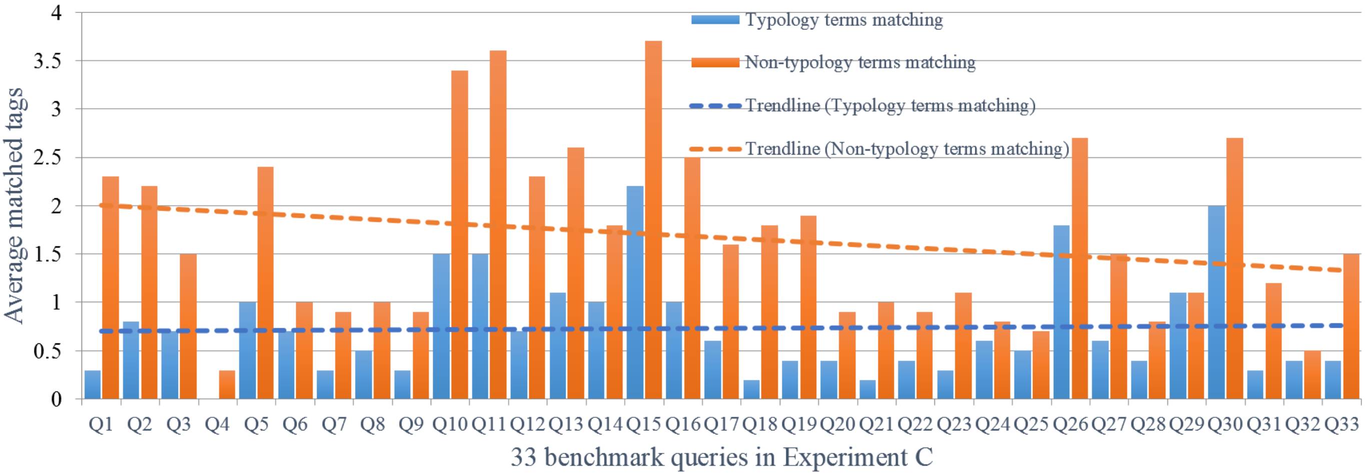


Figure 8.8: Matching of controlled/uncontrolled tags from expert annotators and non-expert results evaluators, across 33 queries

Figure 8.8 shows the average values from the two categories (typology & non-typology tags) calculated across the 33 benchmark queries for each top 10 result set. The number of non-typology terms clearly have got an edge over the typology terms across the queries, which suggests that search results evaluators tended to select non-typology keywords from the two categories, presented to them on the search results page. That also implies that the search results evaluators agreed more with the expert crowd-annotators when they used non-typology term. The trend in Figure 8.8 further suggests that offering both vocabulary and uncontrolled keywords for annotation to experts greatly increases the likelihood of reaching an agreement with the ordinary online users in the multidisciplinary repositories and reduces the over-reliance on having a controlled vocabulary for annotation and tagging of content. Furthermore, the typology-based matching trendline in Figure 8.8 highlights the consistent approach, adopted by the search results evaluators, towards selecting typology terms across the queries.

However, the author also recognises that the popularity-based tags suggestion, in the auto-complete of (*AnnoTagger*), might have contributed, to a certain extent, to this disparity between the twos i.e. typology and non-typology tags. It was observed during the focus group experiments, while the experts attempted the annotation tasks and also through online experiments, that the experts would usually annotate the text inside a web page; and would pay less attention to annotating the entire webpage as an entity. It is understood that splitting the two tasks (*in-page text-level* and *webpage-level annotation*) amongst two distinct expert crowd groups might have produced different results, but the availability of both types of annotation ensured invoking multiple annotations and tags from experts for a single webpage. Moreover, the amount of automatically generated tags, listed in Category A, B in Exp.A and Category B in Exp.C, provided a greater variety of tags (tags abundance) to the evaluators in Exp.A and Exp.C, to chose from. As per the number of annotations, given in Exp.A (6.4), the automatically generated entities, concepts and keywords are far greater than those of typology tags (*abundant vs. scattered*) (See Exp.B (6.5)). This number-based disparity might have also influenced the experts, in Exp.B, to assign non-typology tags, using the popularity-based *auto-complete*. In other words, *one-word free-text term* was popped up sooner (as typing started) than the *multiple-words typology term*, by the auto-complete, due to number-based popularity. As a result typology-based matching trendline in Figure 8.8 remained lower but steady, as compared to the non-typology matching trendline. This trend can be improved by trading off the weights of the typology and non-typology terms in the auto-complete component of *annoTagger*. Another, approach would be to synchronously map the automatically extracted synonyms and descriptors to the typology items to reduce the weight disparity between the twos at the time of expert crowd-annotation.

8.6.6 Expert annotators' feedback questionnaire

In this section, the outcome of the individually collected feedback questionnaire has been summed up, which was completed by the experts at the end of their respective annotation sessions in Exp.B (6.6). Figure 8.9 illustrates the overall attitude of expert crowd-annotators towards using the annotation and tagging tool *AnnoTagger*. Table B.1, the data source of Figure 8.9, shows the data collected from the feedback questionnaire in Appendix B. Screenshots of the questionnaires are given in Figure B.5 and Figure B.6 in Appendix B. There are 20 questions in the questionnaire form and the score was calculated using the 5-points *Likert scale* i.e. *Strongly agree, Agree, Neither agree nor disagree (Neutral), Disagree and Strongly disagree*. The questionnaire has been divided into 4 categories each having sub questions in the form of Q1.1, Q1.2 ...Q4.4 etc. This has been done to give equal weightage to all the questions on the scale against the 5-points. The blue curve shows the tendency of the annotators towards *Strongly agree, Agree and Neutral* points on the scale. The points of disagreement (*Strongly disagree, disagree*) were rarely chosen against those questions which were meant to assess the edition of existing annotations, difficulty in posting comments, difficulty in finding relevant descriptive tags etc.

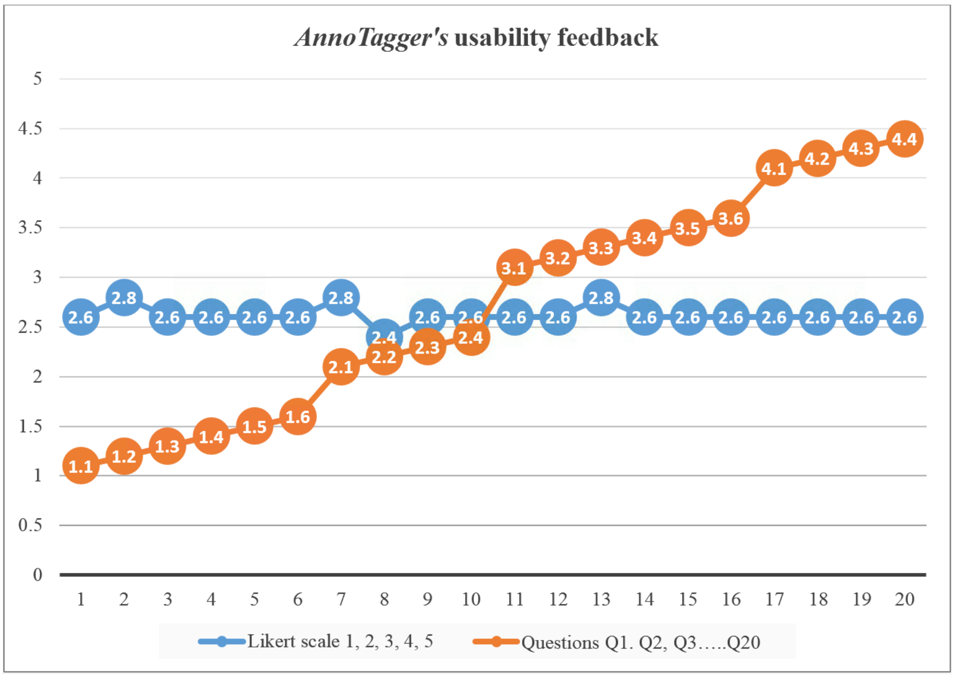


Figure 8.9: Questionnaire feedback of experts annotators on 1-5 Likert scale

8.6.7 Two-tailed t test

In order to compare the two systems i.e. *SemDex* and *SemCrowDex* in terms of standard error of the sample means, the two-tailed t test metric has been used. In this analysis, two t-values i.e. *calculated t-value* and *critical t-value* are needed to accept or reject the null H_0 hypothesis. If the *calculated t-value* is greater than *critical t-value* then the null hypothesis is rejected and vice versa.
Independent t-test formula is given by:

$t_{calculated} = \frac{M_1 - M_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$ *where* M_1 and M_2 are two means, S_1^2, S_2^2 are two variances from two samples of baseline queries searched against two different indices i.e. *SemDex* and *SemCrowDex*. n_1, n_2 are the sample sizes which remain the same in both groups i.e. 33 queries. The denominator is always the kind of engine room of t-test in that it represents the standard error of the difference between means i.e. the lesser the denominator the better.

8.6.7.1 Calculated t-value

So the null hypothesis here implies that significance of search results in terms of relevance is the same whether or not expert crowd annotators annotate the web resources i.e. $H_0 : \mu_1 = \mu_2$. In other words, the ranking of documents based on relevance in top 10 search results does not change even if content in web repositories are annotated by the expert annotators. The contrary to that would be $H_1 : \mu_1 \neq \mu_2$ which states that it increases relevance and changes the ranking of documents at the time of searching.

Now two values from each group are calculated i.e. Mean M_1, M_2 and Variance,S_1^2, S_2^2 depending on the two sample sizes in the test. The data, used for calculating *t-value*, is given in Appendix A in Table A.11.

After having calculated the means, variance and sample sizes based on the AP values obtained in section 8.6.3.2, the calculated t-value can be calculated[1] using the above formulae as: $t_{calculated} = \frac{0.8727-0.6927}{\sqrt{\frac{0.024}{33}+\frac{0.042}{33}}} = 4.09$.

After having obtained $t_{calculated}$ value, critical $t_{critical}$ value is now needed as it might have happened by chance due to the fact that the sample size is not very big.

8.6.7.2 $t_{critical}$ value

Before getting the $t_{critical}$, it is required to calculate degrees of freedom (df) and specify the alpha level (α) which is almost always 0.05. The α level of 0.05 indicates that there is 5% chance of fooling oneself into thinking there is a difference at the sample level when in fact there is no difference. Now, with df and α, critical t-value can be identified in the *t-table*. As the *calculated t-value* has already been obtained based on two means from two groups (Participants in experiments Exp.A and Exp.C), the comparison between the calculated and critical will now lead to either rejecting or accepting the null hypothesis H_0.

Df here is calculated as: $df = n_1 + n_2 - 2 = 33 + 33 - 2 = 64$. With df=64 and α= 0.05 , the critical t-value has been identified in the t-table as 1.9970 or |1.9970|. The value has been enclosed into *pipes* to show that this is an absolute value and that may include positive and negative values. *t table* does not have negative values.

The author understands that the $t_{calculated}$ *value* (4) is greater than $t_{critical}$ *value* which leads to rejecting the null hypothesis H_0. This implies that the null hypothesis of equal μs can be rejected , $p<0.05$. In literal terms, the significance of crowd-annotation to improve search results retrieval based on 33 baseline queries is greater than not including any crowd-annotation-based searchable fields against which the queries can be searched. Therefore the alternative hypothesis of unequal μs $H_1 : \mu_1 \neq \mu_2$ can be accepted, $p<0.05$. In other words, crowd-annotation based semantic searching in System B produces results at greater precision level,

[1] The complete data file is available at https://goo.gl/IZ3XQT

t(64)=4, p<0.05. In general terms, the bigger nominator value is what should be desired as compared to the denominator which shows variance or standard error of difference between two means of the two distinct systems *SemDex* and *SemCrowDex*. In summary, the author concludes that the AP-based t-test calculation of the two systems prefers *SemCrowDex*; and that there is a difference in terms of performance between the two SRR systems when it comes to search results retrieval.

8.7 Summary

The author has looked deeper into the incorporation of crowd-annotation of web content in the SRR framework with a view to assessing further improvement in *query-document* relevance and expert-non-expert agreement on tagging and annotation. The amalgamation of output from two experiments (Exp.A and Exp.C) against the baseline 33 benchmark queries gave the author a leverage to analyse the SRR systems' performance from multiple angles. The author also observed participants' interaction with the annotation and retrieval systems in the labs while executing very laborious parallel *semantic indexing* and *searching* processes. Such observations-based performance evaluation was performed to ascertain the robustness of the system and strengthen the author's confidence in the search results retrieval system. The evaluation results, in terms of different performance metrics, clearly show that given the support of disciplinary community, expert intervention in reclassifying the content of web repositories improves search engines' performance. On top of that, the total control over the usability and implementation of crowdsourcing-based semantic annotation and SRR in a web-based environment, makes it possible without the need for contracting it out to third party services and experts.

Chapter 9

Conclusions & Future Work

This thesis has identified the problem of insufficient information representation in multi-disciplinary web repositories and has advocated for real time, sustainable and verifiable semantic representation and annotation methods. The author has demonstrated, through rigorous annotation and searching experiments, that such methods can be implemented and widely adopted in web-based repositories and archives with minimal technical complication. The author recognises that this area of research continues to evolve with the fast-paced information revolution. The author also concludes that there is a greater need for the community of online users to aid the IR systems in determining the degree of relevance between *information need* and *results*, at the time of searching.

Willingness to contribute in terms of annotating and tagging content in a web repository, alongside the consumption of information, will go a long way in terms of relevant and precise information retrieval. This chapter formally draws this thesis to a close by (a) summarising the work, undertaken by the author and (b) highlighting the need for future work. Furthermore, It is also important to highlight the issues and limitations, which have the potential of influencing the findings of this research. Section 9.1 critically summarises contribution and limitations of this research, in the light of research questions:

RQ1: Can accuracy and relevance in search results be improved by employing automatic semantic annotations?
RQ2: Can the relevance of search results be further improved by periodically adding contemporary semantic annotations?
RQ3. Can the relevance of search results be further improved by expert and non-expert crowd-sourced semantic annotations & tagging?
In Section 9.3, the need for further research work is highlighted in this field with

a special focus on interoperability and *RDFisation* of semantic annotation data that is *interpretable* by cross-platform search results retrieval systems.

9.1 Annotation and searching framework

There is a great potential in undertaking and innovating the semantic annotation of heterogeneous content of web repositories, using LoD-based semantic annotators and vocabulary-supported manual annotation. Analysis of various textual resources, in this thesis, has led the author to discover a range of methods and techniques for knowledge interpretation, organisation and discovery. The objective was not only the accurate semantic entity and concept extraction from text but it was equally important to verify the accuracy and consistency of quality by involving human judgement. The annotation, searching and experimental framework in this thesis have shown consistent performance (based on output from various experiments) when tested on a variety of content. The research work in this thesis was completed in the following phases:

1. Development of a system to ascertain improvement, sustainable accuracy and relevance in search results in web repositories by employing automatic semantic annotation methods.
2. Extension of the system in (1) to investigate if expert crowd-sourced semantic annotation can be exploited (on a sustainable basis) on top of the automatic semantic annotation, to address the context obsolescence issues, affecting the ranking of relevant search results in a domain-specific web repository.
3. Further diversification of the system to examine the impact of expert and non-expert tagging, at the time of searching, on the overall relevant search results retrieval. Moreover, the investigation also included assessment of the degree of agreement and consensus among the expert crowd annotators and non-expert search results evaluators on assigning contemporary scientific terms to online resources.

9.1.1 Development of a sustainable annotation and searching system

The need for a full- fledged system, which would fully take into account the nature of web repositories, their content and full-text searching issues, was identified from the outset. Annotation and searching frameworks were piloted to create the building blocks for assessing various systems on the scale of precision, accuracy and relevance in search results. The system development strategy draws from the kind of work, related to our research work in Chapter 3, which advocate "build ontology first" before annotation and domain-specific semantic searching. The divergence from this typical narrative, towards using the existing LoD-enabled semantic annotators, helped the author explore some novel paths in the semantic annotation and knowledge representation domains.

The author wanted to build, extend and sustain an annotation system, which would perform the annotation tasks based on a well defined evaluation criteria, adopted for improved search results retrieval. The functional criteria generally included content diversity, multi-level annotation, simplicity in terms of usability, dynamic customizability and extensibility. Such arrangements, led to the search results evaluation, firmly based on well established IR performance metrics, as well as participants-based evaluation for system verification and quality assurance. Marked improvement in search results, validated the objectives of setting up a full-fledged annotation and searching system. However, in order to further build on that improvement, the author wanted to diversify the system by including the expert and non-expert crowd-annotation element, as discussed in the following section.

9.1.2 Diversification of the annotation and searching framework

Chapter 5 extensively detailed the implementation of the author's entire semantic indexing, searching, and evaluation framework. The emphasis has been on the utilisation of two sources of annotation: automatically generated semantic annotation and crowdsourcing-based annotation. The expert crowd element is of substantial importance, as it aims to augment the *searchable* knowledge base with contemporary domain-specific terms and refresh the SRR system for better and up-to-date search results retrieval. Several pilot exercises were conducted before implementing the above, followed by participants-based experimentation to establish performance evaluation metrics for the follow-up experiments. Those exercises also helped the author to streamline the infrastructure (hardware and software), required for parallel annotation and retrieval processes. Typology and free-text terms tagging, have been the primary sources of contemporary semantic annotation, which further enriched the KB for sufficiently improved search results retrieval. The improvement in SRR's performance was verified by search research evaluators as explained in the next section. However, it was observed in Exp.B that annotating content inside a page (*content-level*) took 70% time and the *page-level* annotation was completed in 30% of the total time, required for completing one single result evaluation. In addition to the issues, raised in Section 8.6.5, it is also important to highlight the need for diversification of the searching framework to include typology-based *facet-searching*. It remains to be seen, however, whether *facet searching*, may bring further improvement to the SRR system performance in the long run, based on continual typology-based tagging.

9.1.3 Optimal SRR performance using expert and non-expert crowd

The development of a 3-tier experimental framework streamlined both the in-house processes as well as external participants-based activities. The usability components for 3 different experiments have been a challenge, but due to the collaboration of the designated research community (e.g. social sciences, statistics, geography, Web & internet sciences etc.), several pilot exercises were organised before running the actual online and focus group experiments. The custom-built crowdsourcing-based annotation environment has been instrumental in expert annotation and tagging, based on the participants feedback and our observations during the focus group sessions. Evaluation and tagging of search results, by non-expert crowd, have also been a great source of performance validation of the two systems i.e. *SemDex* and *SemCrowDex*. Disparity in the number of participants and the number of results, evaluated in Exp.A and Exp.C, however, influenced the preparation of benchmark queries sets. Furthermore, it was observed in Exp.A and Exp.C, that participants would ignore ranking a particular page after reading the content but would attach tags from both categories. Such practice was more common in Exp.A than Exp. C, which suggests that star-rating should have appeared only after the ranking of a page, in order to get the ranking-related feedback first-thing from a participant. Moreover, the amount of work, involved in evaluating a single query, needs to be further curtailed in order to enable the participant spend more time on evaluating a single result (out of 10 results) against a single query. Such amendments have the potential of attracting more participants, performing deeper semantic enrichment through *definite* ranking and *dense* tagging.

9.1.4 Assessment of the degree of agreement between experts and non-experts

One of the significant contributions of this research has been the analysis of both *expert* and *non-expert* annotators of online resources in web repositories. Forming the baseline of annotation comments, vocabulary and free-text tags have been part of evaluation analysis, in order to ascertain the level of consensus between experts and non-experts. The consensus evaluation in Section 8.6.5, Chapter 8, demonstrates promising results, with the exception of *low agreement* on typology-based tagging. It is noted that the mutual agreement on non-typology terms was greater than the typology terms. This disparity can be addressed by adjusting the weight of typology terms against non-typology terms, at the time of *in-page content-level* and *page-level* annotation. Nonetheless, an agreement between the two participant groups, over the selection of non-typology terms (compared to *lesser* typology term agreement) for similar search results, shows the trust of both groups in automatically extracted concepts and entities.

9.1.5 Contributions of this research

This research has contributed to the wider research by introducing a *workable approach* to semanticizing and searching the content of multi-disciplinary repositories or archives of scientific research data. The analysis of the current semantic representation and search retrieval tools and techniques in Chapter 3, revealed that many approaches required the building of a domain-specific ontology, the availability of ontology engineers and RDF-based triple store, before annotating content for better searching. This research, however has adopted a different but related approach as summed up in the following points:

1. A complete semantic knowledge representation and retrieval system, which has been built on an open source foundation, with minimal domain experts intervention.
2. Knowledge acquisition as a parallel activity, through expert crowd-annotation, from a multi-disciplinary research community at the time of browsing and searching.
3. Discipline-specific vocabulary (NCRM Typology) utilisation in the semantic annotation framework, as a knowledge organisation system, and setting up a complete working model to periodically upgrade the typology with the involvement of multi-disciplinary academic and subject experts
4. A comprehensive multi-layered evaluation model based on (a) algorithmically proven retrieval score computation and (b) experimentally validated (through users participation) SRR system

9.2 Opportunities, assumptions and limitations

The main aim of this research is to underscore the insufficiency of keywords-based full-text searching in multi-disciplinary web repositories. It has been established that the retrieval of relevant search results improved after semantically extracted metadata was added to the content. It was also demonstrated that domain experts could further influence the way knowledge is structured and searched in that system in a less technical and more practical ways. A substantial amount of research work has been done by the concerned research community to address the above-mentioned issue by using different approaches based on different semantic models, document formats and domains. It is, however, understood that no "fit for all" solution exists so far. Therefore, a domain-specific, lightweight and sustainable solution is needed to address the issue in web repositories and archives, without acquiring complex hardware and software infrastructure or services. This research presented an opportunity to design and implement such solution in a web repository of multi-disciplinary research data. However the following assumptions and limitations, restrict the scope and replication of the system in other environments.

- It is assumed that the types of content in web repositories are hosted by the respective organisations with full access control regardless of the types of content e.g. a static web page, dynamic web page, and other documents
- Expert crowd can only annotate web-page-only documents and only automatic annotators can annotate the entire corpus of documents
- It can also be argued that non-expert annotators may annotate the content both at *content-level* as well as *page-level* (like expert annotators in Exp.B. However, the assessment of *non-tagged* annotation i.e. comments, notes etc. requires the involvement of domain-specific academic experts to critically evaluate the association of comments with a web page or its content, which would be impractical in many cases.
- It is assumed that there is enough awareness and recognition of the need for a contemporary semantic search systems among the technical professionals and domain experts for better dissemination of research outputs
- Shallow corpus of heterogeneous documents with no comparison to be made with other domain-specific repository, based on different time periods e.g. past 10 years.

9.3 Future work

Although the author has developed and tested the annotation, retrieval and evaluation system for sustainable search results retrieval in web repositories, there are several ways this work can be further advanced. So far the author has experimented with the content of the ReStore repository[1]. However, expanding the

[1] www.restore.ac.uk

corpus base would further strengthen the author's confidence in both automatic and manual semantic annotation for future research. It is also recognised that there is a great potential for further crowdsourcing-based annotation experiments based on *multi-faceted* searches, led by typology-based tagging. Online searching is a vast area and although a substantial majority of users do type their search queries into the *search box*, a faceted search in web repositories and archives is a potential area of further research, worth evaluating for further improvement in search results retrieval. It is also important to mention here that most of the *ontology-based* searching either do not facilitate faceted browsing or filter results based on fixed facets for all searches(Butt et al., 2015).
Another important future work could be the integration of our framework with the RDF-Compliant eco-system of repositories and KBs. The author describes, the potential of using a new relationships-based technology called GraphDB [2], which enables interoperability and RDF-compliance on the existing KBs. It is also important to note that, after the retirement of Alchemy API in 2016, IBM NLU (Natural Language Understanding) service, the upgraded version of Alchemy API, now offers entity extraction configuration (e.g. classifier training, custom modelling) (Dale, 2018). It is, by all means, a step forward and a future opportunity to extend the current framework for the automatic classifier-based identification of new concepts/entities in web repositories using LoD platform.

9.3.1 ES KB as a *GraphDB-Compliant* KB

Elasticsearch offers impressive performance when it comes to full-text search or aggregations, but they are aggregate oriented databases and therefore have limitations when it comes to *relationships-based* connected data. Considering that RDF is a building block of the Semantic Web, one would ideally want the current SRR system to be compatible with RDF speaking clients. In other words, a connector or linkage is needed to store and present the existing data as a relationship graph made up of nodes and vertices, connected via relationships or edges, to form a mesh of information, the like of which has been shown and explained in Chapter 5. Querying graphs instead of documents should therefore be investigated to ascertain the performance of the SRR when integrated with another graph-based KB. Querying or traversing graphs using an RDF-Compliant language i.e. SPARQL, ensures cross-platform access to knowledge bases (with edges and nodes), which would make the author's ES-based KB compatible with

[2] In computing, a graph database is a database that uses graph structures for semantic queries with nodes, edges and properties to represent and store data. Available at https://en.wikipedia.org/wiki/Graph_database

the *graph-compliant* eco-system of KBs. GraphDB (former OWLIM[3]) is therefore the most scalable semantic repository which includes triple store, inference engine and SPARQL query engine. In order to further extend this work by ensuring the retrieval in RDF-compliant format with SPARQL endpoint, the author has explored various possibilities. The most promising, workable and sustainable solution is to use the GraphDB connector which has been the latest discovery as of 2018 by Ontotext[4]. GraphDB is one of the RDF stores which can infer and serve new knowledge from existing facts in our ES KB. Having an interoperable, searchable knowledge base will ensure cross-platform searching on a comparatively larger scale with a focus on enhanced search results retrieval.

9.3.2 *RDFisation* of the entire annotation & indexing system

Following on from the discussion in the preceding section, it is understood that currently, the GraphDB Connector has the potential of serving the Elasticsearch-indexed data in triples. In other words, GraphDB connector provides the means of converting a graph-query into an ES query which may include *INSERT/DELETE/UPDATE* statements. Figure 9.1 clearly illustrates the proposed integration of the ES KB (*SemCrowDex* index) with the GraphDB-based graph index, interoperable with each other through the GraphDB Connector. The different components in Figure 9.1 demonstrates that there is promising potential for integration of the author's framework with an RDF-Compliant triple store, ready to be evaluated for search results retrieval. By implementing such platform the author of this thesis assumes and concludes that:

- The semantic annotation and SRR framework does not stay inside a silo and is ready to be queried by SPARQL-speaking semantic web repositories
- The core indexing and domain searching framework remain the same and only querying and retrieval is extended which makes it part of the bigger semantic web-based RDF KBs eco-system
- The distributed , open source and horizontal scaling feature of ES makes it an ideal choice for higher education institutions and research organisations. On top of that, the framework presented in this thesis performs better when it comes to sustainable and cost effective solutions, to organising knowledge in

[3] OWLIM is a family of semantic repositories, or RDF database management systems, with compliance of RDFS, OWL,SPARQL

[4] Ontotext is a Bulgarian software company headquartered in Sofia whose main domain of activity is the development of software products and solutions based on the Semantic Web languages and standards, in particular RDF, OWL and SPARQL

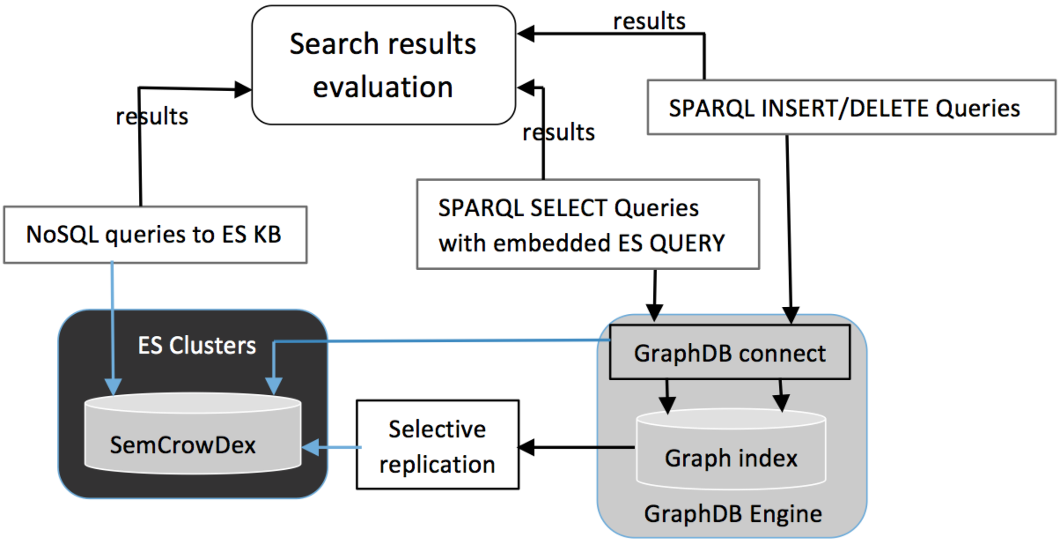


Figure 9.1: Proposed integration of search results retrieval from ES KB and GraphDB KB using NoSQL and SPARQL queries respectively

the ever burgeoning repositories of data published by the government, higher education and research organisations.

The author has also demonstrated the blueprint of the proposed *RDF interoperability model* in Figure D.1, where various domain-specific concepts have been presented in the form of RDF triples ready to be investigated for storage and retrieval. Furthermore, a detailed RDF representation has also been illustrated in Figures D.2 and D.3, showing *entity, relationships-based* triples, representing a particular webpage in a web repository. However, the formal representation of RDF triples does not yet include the manually generated annotation and tags, associated with the webpage in question. Moreover, computation of ranked results against users' queries in an *RDF-SPARQL*-compliant KB have to be thoroughly examined (using the ranking methodology set out in Chapter 5) to ensure that cross-platforms search results retrieval does not affect the retrieval of relevant search results despite the challenges, outlined in Chapter 1 of this thesis.

Appendix A

Background, semantic indexing, search results retrieval and evaluation

A.1 ESRC's research initiatives in social science research

Initiatives	Title	Online resource location	Funding ended/active
RMP	Research Methods Programme	ReStore repository	2008
RDI	Researchers Development Initiative	ReStore repository	2010
QMI	Quantitative Methods Initiative	ReStore repository	2013
NCRM	National Centre for Research Methods	NCRM Repository	2019
ADRN	Administrative Data Research Network	ADRN Repository	2018

Table A.1: A list of various past and current initiatives funded by the Economic and Social Research Council (ESRC) to further Social Science research across the UK.

A.2 KIM semantic publishing architecture

Figure A.1 demonstrates different components in the KIM's (Knowledge and Information Management) semantic publishing architecture showing content storage, text processing, rules writing and relation extraction etc (Georgiev et al., 2013).

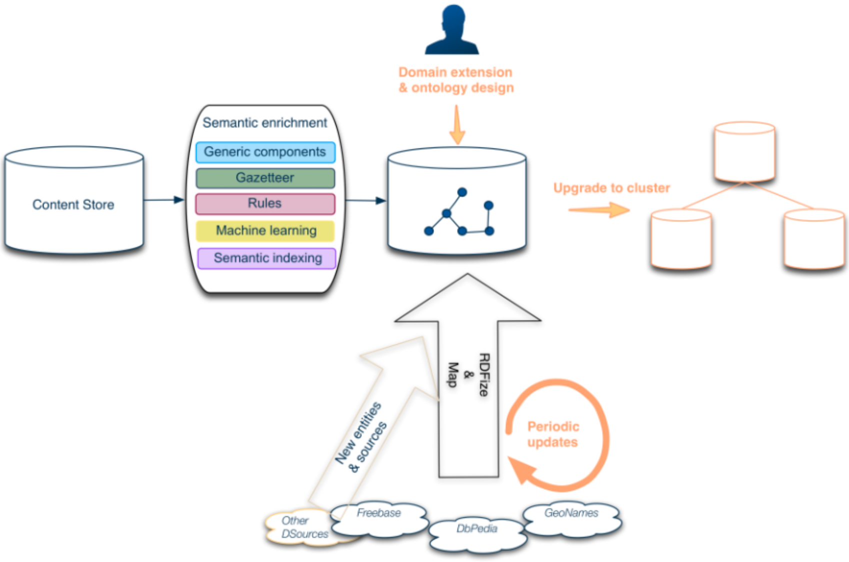


Figure A.1: KIM- Generalised Semantic Publishing Architecture

KIM's approach has been described as one of the related works in Chapter 3 in Section 3.2.1.

A.3 Extensible and Scalable Elasticsearch cluster

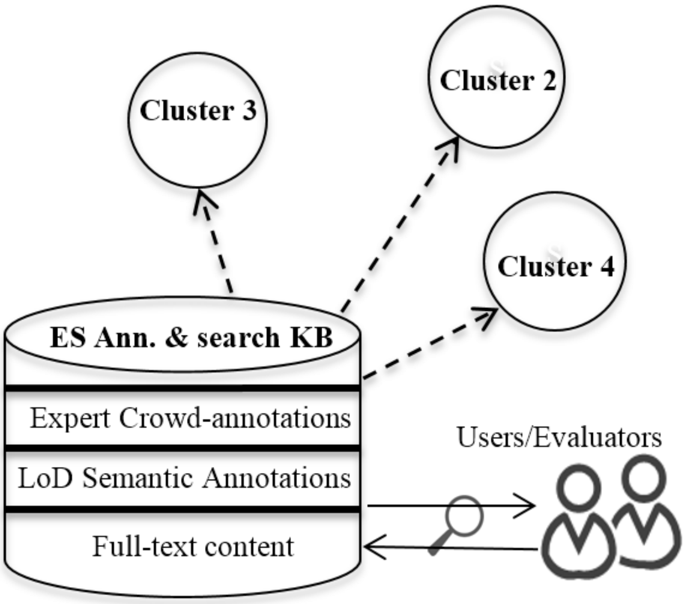


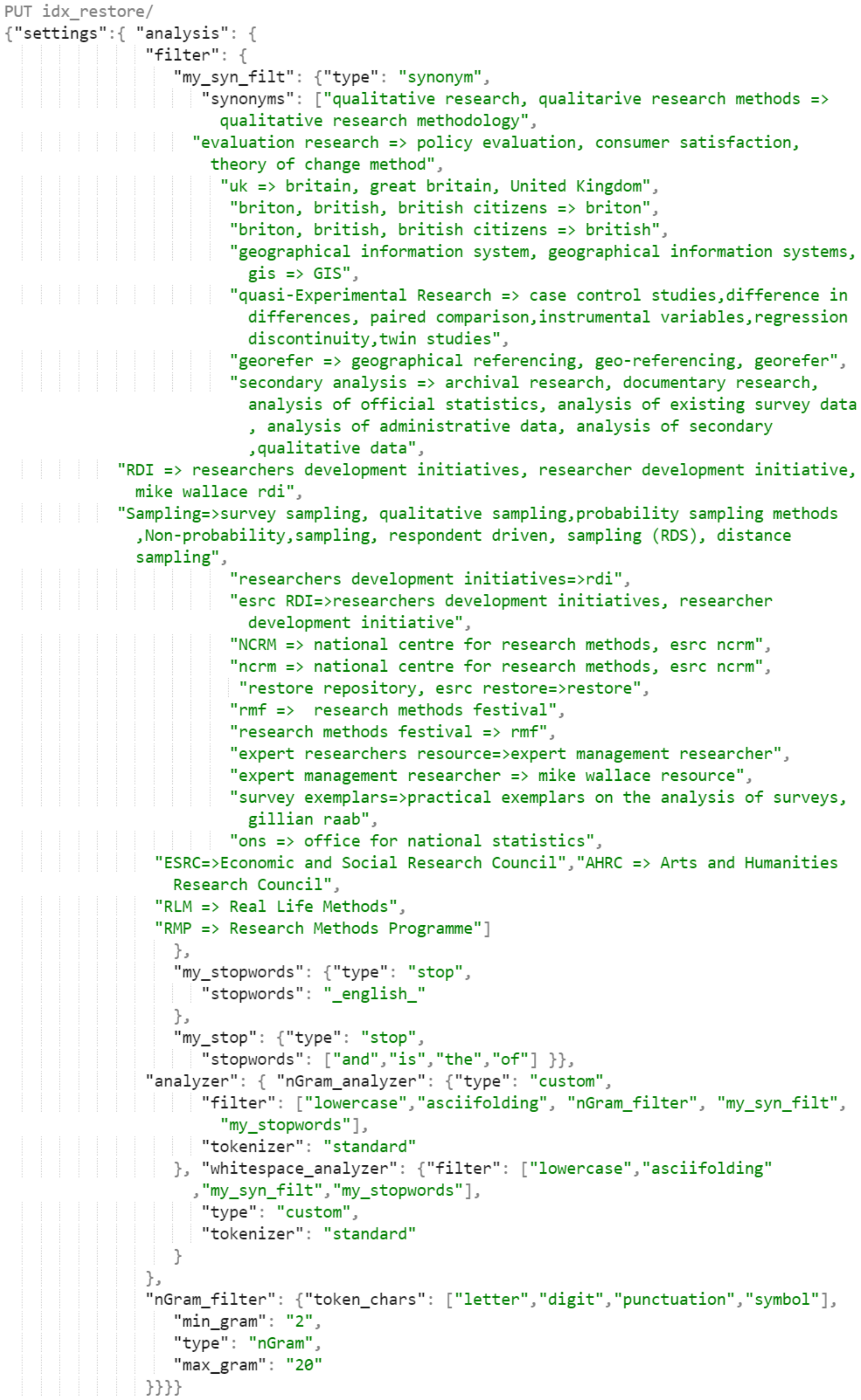
Figure A.2: Extensible and scalable semantic indexing, annotation and search framework

Figure A.2 demonstrates that one Elasticsearch cluster can be combined into multiple clusters thus making it a domain-independent, multi-disciplinary annotation and search platform accessed by the users via a universal user interface.

Figure A.4(a) shows a screenshot of the semantic index mapping without crowd-sourced annotation and Figure A.4(b) depicts a screenshot of the hybrid semantic index mapping having the crowd-annotation & tagging elements.

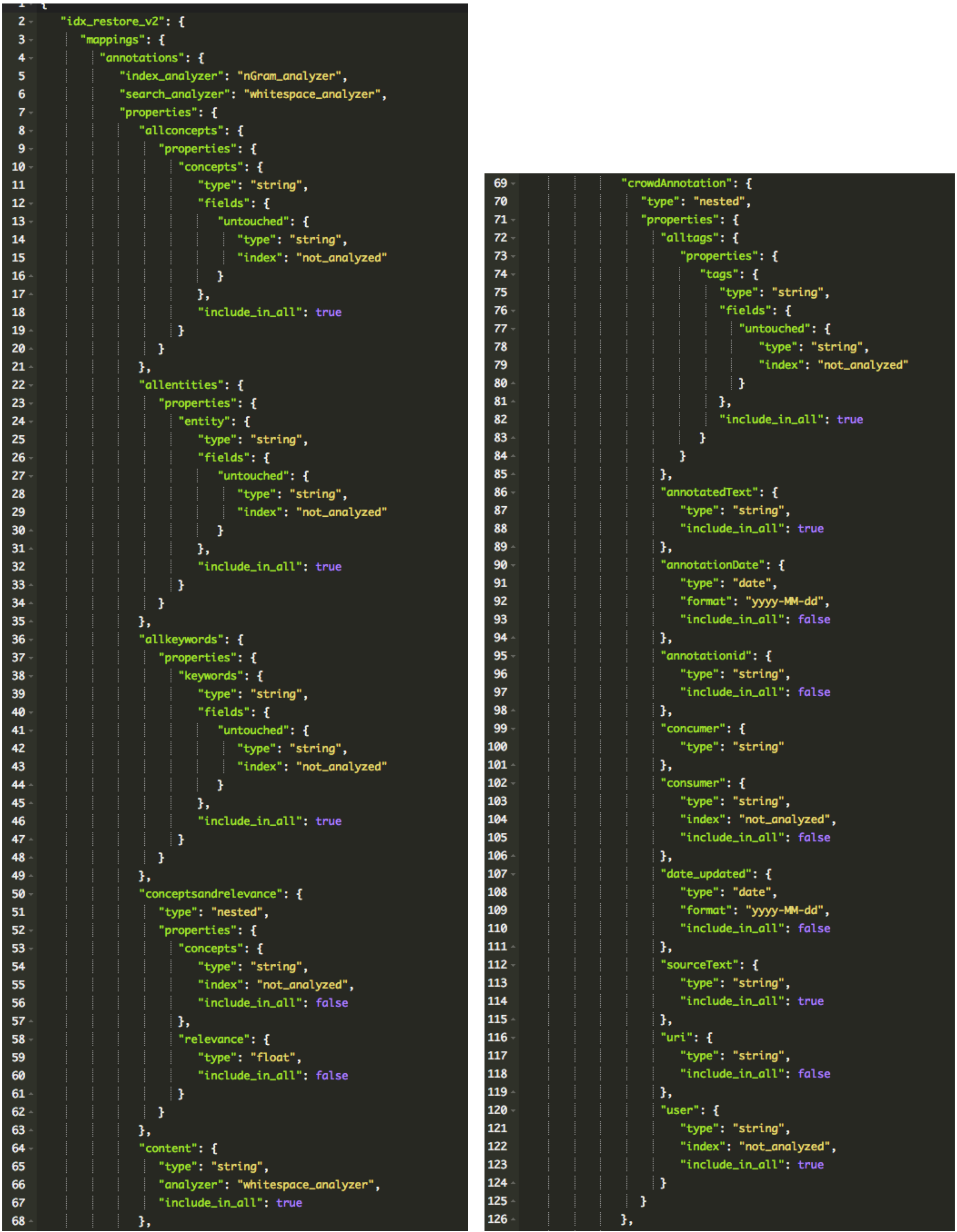
A.4 Index settings, schema mapping and fields definition

Figure A.3 demonstrates the brief version of the custom analyser, used by the Elasticsearch to analyse and index web documents. The relationship between *synonyms* and *descriptors* has been illustrated as $(synonyms) => (descriptors)$.



```
PUT idx_restore/
{"settings":{ "analysis": {
               "filter": {
                  "my_syn_filt": {"type": "synonym",
                     "synonyms": ["qualitative research, qualitarive research methods =>
                       qualitative research methodology",
                    "evaluation research => policy evaluation, consumer satisfaction,
                      theory of change method",
                       "uk => britain, great britain, United Kingdom",
                        "briton, british, british citizens => briton",
                        "briton, british, british citizens => british",
                        "geographical information system, geographical information systems,
                          gis => GIS",
                        "quasi-Experimental Research => case control studies,difference in
                          differences, paired comparison,instrumental variables,regression
                          discontinuity,twin studies",
                        "georefer => geographical referencing, geo-referencing, georefer",
                        "secondary analysis => archival research, documentary research,
                          analysis of official statistics, analysis of existing survey data
                          , analysis of administrative data, analysis of secondary
                          ,qualitative data",
            "RDI => researchers development initiatives, researcher development initiative,
              mike wallace rdi",
            "Sampling=>survey sampling, qualitative sampling,probability sampling methods
              ,Non-probability,sampling, respondent driven, sampling (RDS), distance
              sampling",
                        "researchers development initiatives=>rdi",
                        "esrc RDI=>researchers development initiatives, researcher
                          development initiative",
                        "NCRM => national centre for research methods, esrc ncrm",
                        "ncrm => national centre for research methods, esrc ncrm",
                         "restore repository, esrc restore=>restore",
                        "rmf =>  research methods festival",
                        "research methods festival => rmf",
                        "expert researchers resource=>expert management researcher",
                        "expert management researcher => mike wallace resource",
                        "survey exemplars=>practical exemplars on the analysis of surveys,
                          gillian raab",
                        "ons => office for national statistics",
                "ESRC=>Economic and Social Research Council","AHRC => Arts and Humanities
                  Research Council",
                "RLM => Real Life Methods",
                "RMP => Research Methods Programme"]
                  },
                  "my_stopwords": {"type": "stop",
                     "stopwords": "_english_"
                  },
                  "my_stop": {"type": "stop",
                     "stopwords": ["and","is","the","of"] }},
               "analyzer": { "nGram_analyzer": {"type": "custom",
                     "filter": ["lowercase","asciifolding", "nGram_filter", "my_syn_filt",
                       "my_stopwords"],
                     "tokenizer": "standard"
                  }, "whitespace_analyzer": {"filter": ["lowercase","asciifolding"
                    ,"my_syn_filt","my_stopwords"],
                     "type": "custom",
                     "tokenizer": "standard"
                  }
               },
               "nGram_filter": {"token_chars": ["letter","digit","punctuation","symbol"],
                  "min_gram": "2",
                  "type": "nGram",
                  "max_gram": "20"
               }}}}
```

Figure A.3: Designing semantic index schema, fields and synonyms in an Elasticsearch cluster



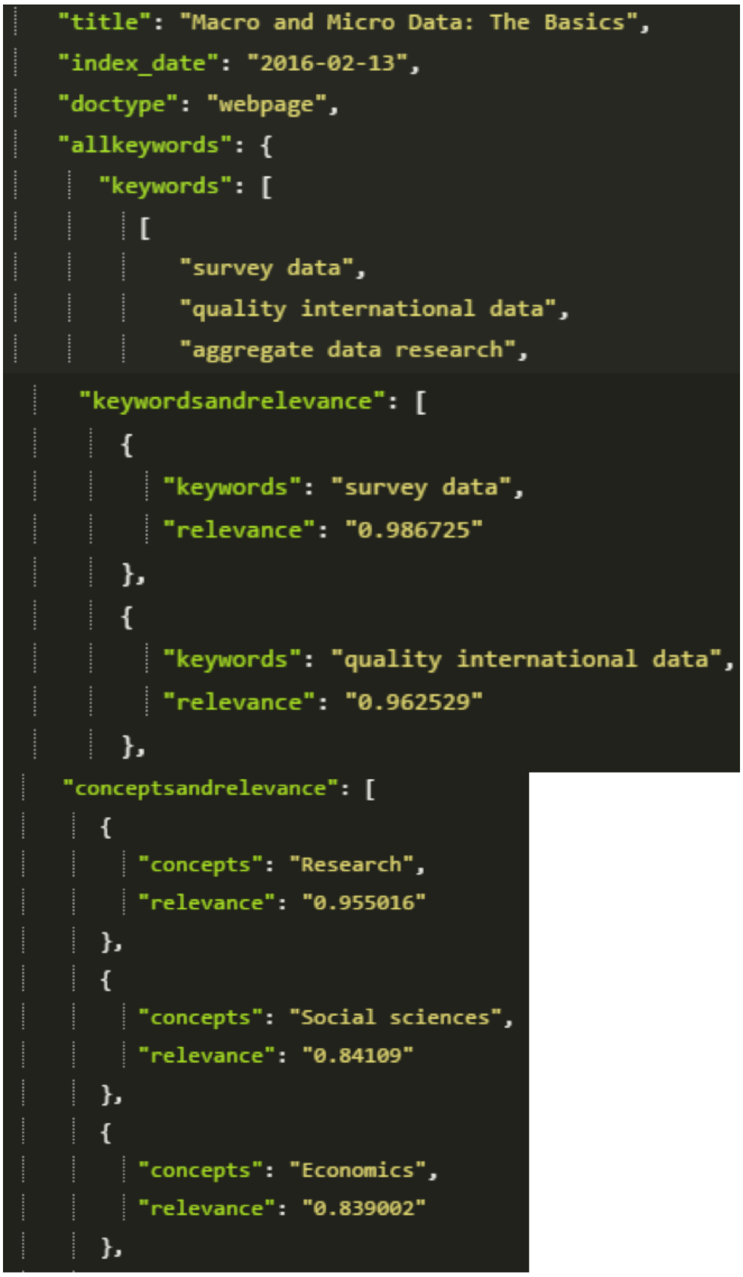
```
{
  "idx_restore_v2": {
    "mappings": {
      "annotations": {
        "index_analyzer": "nGram_analyzer",
        "search_analyzer": "whitespace_analyzer",
        "properties": {
          "allconcepts": {
            "properties": {
              "concepts": {
                "type": "string",
                "fields": {
                  "untouched": {
                    "type": "string",
                    "index": "not_analyzed"
                  }
                },
                "include_in_all": true
              }
            }
          },
          "allentities": {
            "properties": {
              "entity": {
                "type": "string",
                "fields": {
                  "untouched": {
                    "type": "string",
                    "index": "not_analyzed"
                  }
                },
                "include_in_all": true
              }
            }
          },
          "allkeywords": {
            "properties": {
              "keywords": {
                "type": "string",
                "fields": {
                  "untouched": {
                    "type": "string",
                    "index": "not_analyzed"
                  }
                },
                "include_in_all": true
              }
            }
          },
          "conceptsandrelevance": {
            "type": "nested",
            "properties": {
              "concepts": {
                "type": "string",
                "index": "not_analyzed",
                "include_in_all": false
              },
              "relevance": {
                "type": "float",
                "include_in_all": false
              }
            }
          },
          "content": {
            "type": "string",
            "analyzer": "whitespace_analyzer",
            "include_in_all": true
          },
```

(a) Schema mapping of the keywords, concepts, entities semantic index with no crowd-sourced elements

```
          "crowdAnnotation": {
            "type": "nested",
            "properties": {
              "alltags": {
                "properties": {
                  "tags": {
                    "type": "string",
                    "fields": {
                      "untouched": {
                        "type": "string",
                        "index": "not_analyzed"
                      }
                    },
                    "include_in_all": true
                  }
                }
              },
              "annotatedText": {
                "type": "string",
                "include_in_all": true
              },
              "annotationDate": {
                "type": "date",
                "format": "yyyy-MM-dd",
                "include_in_all": false
              },
              "annotationid": {
                "type": "string",
                "include_in_all": false
              },
              "concumer": {
                "type": "string"
              },
              "consumer": {
                "type": "string",
                "index": "not_analyzed",
                "include_in_all": false
              },
              "date_updated": {
                "type": "date",
                "format": "yyyy-MM-dd",
                "include_in_all": false
              },
              "sourceText": {
                "type": "string",
                "include_in_all": true
              },
              "uri": {
                "type": "string",
                "include_in_all": false
              },
              "user": {
                "type": "string",
                "index": "not_analyzed",
                "include_in_all": true
              }
            }
          },
```

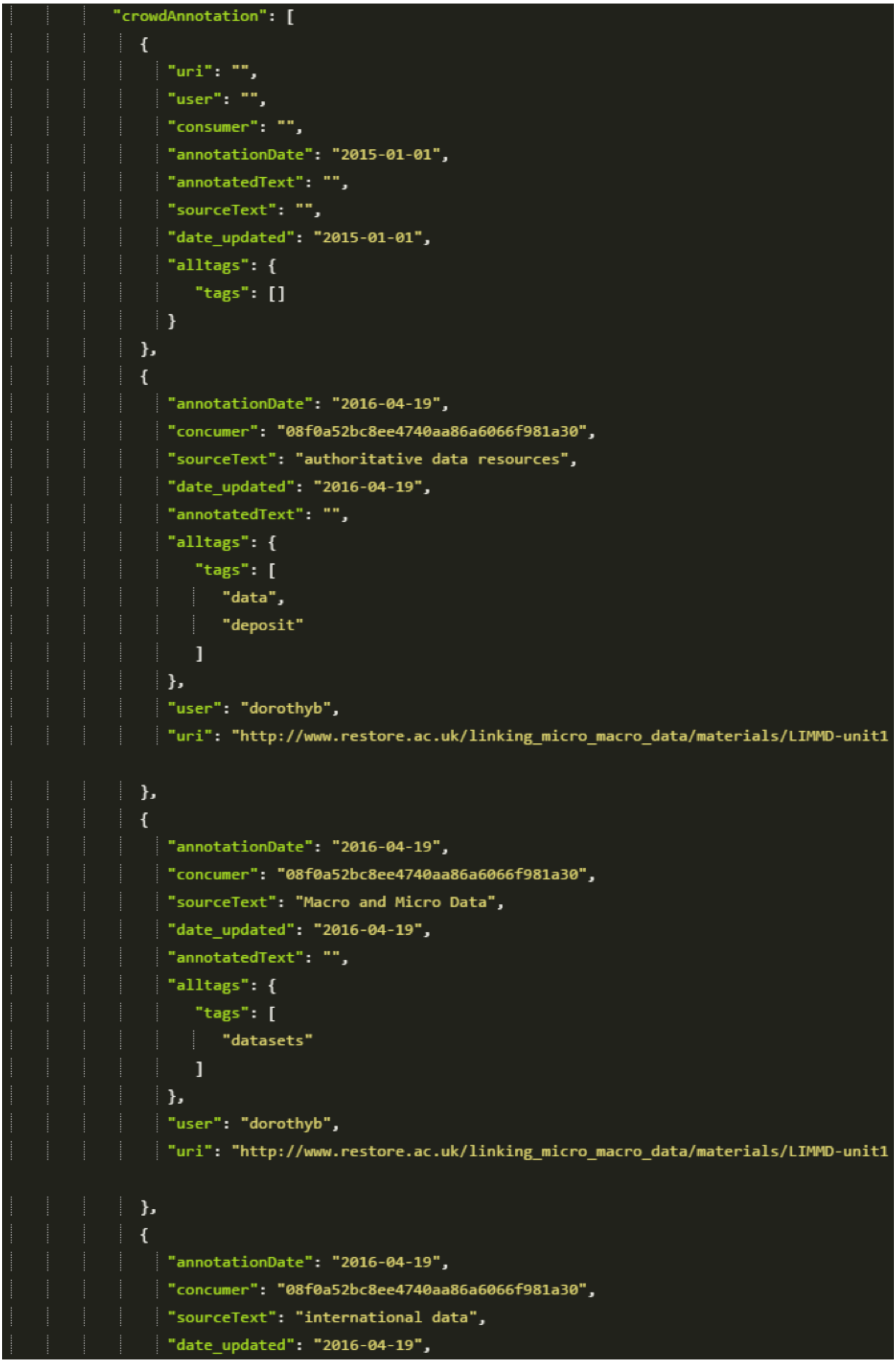
(b) Modified scheme mapping (from left) with the addition of crowd-sourced annotations element

Figure A.4: Screen shots of schema mapping of the hybrid semantic index incorporating automatically generated keywords, concepts, entities along with crowd-sourced annotations and tags using free text and vocabulary terms



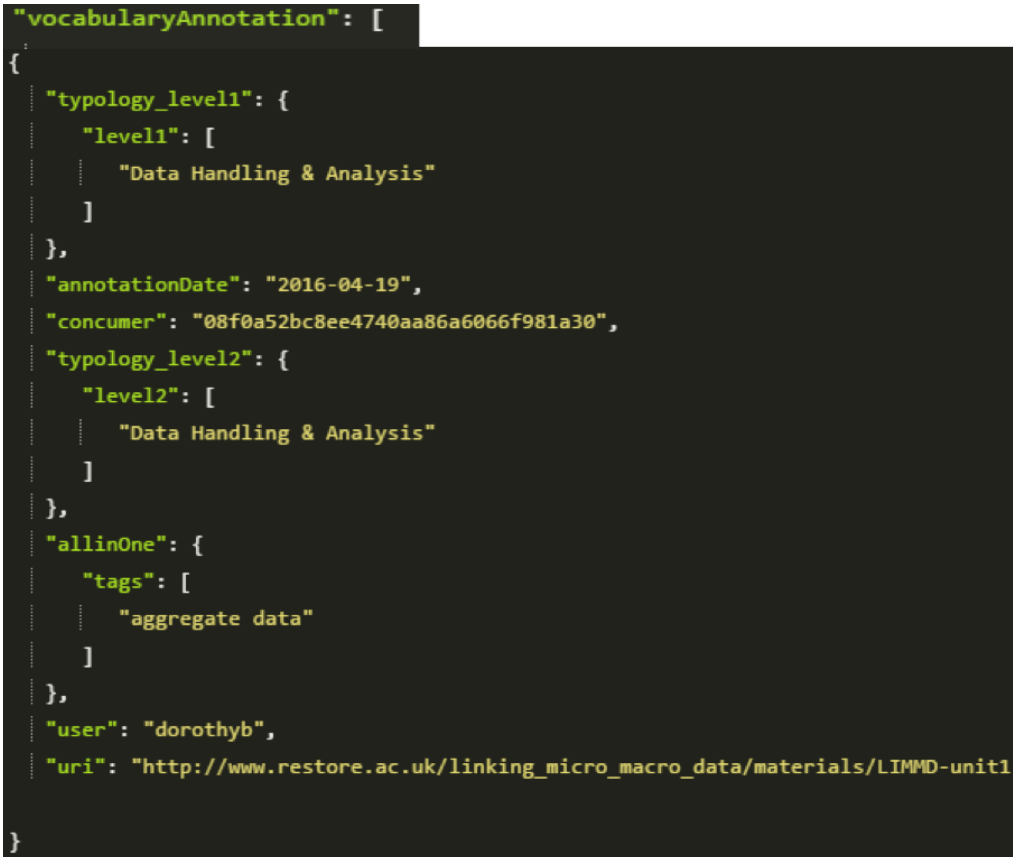
```
    "title": "Macro and Micro Data: The Basics",
    "index_date": "2016-02-13",
    "doctype": "webpage",
    "allkeywords": {
        "keywords": [
            [
                "survey data",
                "quality international data",
                "aggregate data research",
    "keywordsandrelevance": [
        {
            "keywords": "survey data",
            "relevance": "0.986725"
        },
        {
            "keywords": "quality international data",
            "relevance": "0.962529"
        },
    "conceptsandrelevance": [
        {
            "concepts": "Research",
            "relevance": "0.955016"
        },
        {
            "concepts": "Social sciences",
            "relevance": "0.84109"
        },
        {
            "concepts": "Economics",
            "relevance": "0.839002"
        },
```

Figure A.5: A screenshot showing topical keywords and ranked concepts as part of a single ES document along with individual Alchemy API-generated score in ES search results console produced against a query



```
            "crowdAnnotation": [
                {
                    "uri": "",
                    "user": "",
                    "consumer": "",
                    "annotationDate": "2015-01-01",
                    "annotatedText": "",
                    "sourceText": "",
                    "date_updated": "2015-01-01",
                    "alltags": {
                        "tags": []
                    }
                },
                {
                    "annotationDate": "2016-04-19",
                    "concumer": "08f0a52bc8ee4740aa86a6066f981a30",
                    "sourceText": "authoritative data resources",
                    "date_updated": "2016-04-19",
                    "annotatedText": "",
                    "alltags": {
                        "tags": [
                            "data",
                            "deposit"
                        ]
                    },
                    "user": "dorothyb",
                    "uri": "http://www.restore.ac.uk/linking_micro_macro_data/materials/LIMMD-unit1

                },
                {
                    "annotationDate": "2016-04-19",
                    "concumer": "08f0a52bc8ee4740aa86a6066f981a30",
                    "sourceText": "Macro and Micro Data",
                    "date_updated": "2016-04-19",
                    "annotatedText": "",
                    "alltags": {
                        "tags": [
                            "datasets"
                        ]
                    },
                    "user": "dorothyb",
                    "uri": "http://www.restore.ac.uk/linking_micro_macro_data/materials/LIMMD-unit1

                },
                {
                    "annotationDate": "2016-04-19",
                    "concumer": "08f0a52bc8ee4740aa86a6066f981a30",
                    "sourceText": "international data",
                    "date_updated": "2016-04-19",
```

Figure A.6: A screenshot showing Crowd-annotation fields and data added by the expert crowd-annotators as part of a single ES document in ES search results console produced against a query



```
"vocabularyAnnotation": [
{
  "typology_level1": {
    "level1": [
      "Data Handling & Analysis"
    ]
  },
  "annotationDate": "2016-04-19",
  "concumer": "08f0a52bc8ee4740aa86a6066f981a30",
  "typology_level2": {
    "level2": [
      "Data Handling & Analysis"
    ]
  },
  "allinOne": {
    "tags": [
      "aggregate data"
    ]
  },
  "user": "dorothyb",
  "uri": "http://www.restore.ac.uk/linking_micro_macro_data/materials/LIMMD-unit1

}
```

Figure A.7: A screenshot of vocabulary or typology annotation fields as part of a single ES document in ES search results console produced against a query

A.5 Single query precision/recall ranking calculation: Lexical vs. Semantic

Table A.2 also shows the lexical and semantic representation of one of 20 queries assessed in Exp.A(6.4) using the TREC's 11-points recall scale to construct Figure 7.1 and Figure 7.2 . The data also shows the computation of Average Precision (AP) in both lexical and semantic cases along with Average User Ranking (AR) which have been collectively used to construct Figure 7.5. Moreover, Table A.2 is split into two halves one showing precision(P), interpolated precision (IP), user ranking for a query (*query:finite population correction*) on the basis of *lexical* and *semantic* representation of data. The R/N column shows whether a result against this query is relevant (1) or not relevant (0).

	query: Finite Population Correction						*query: Finite Population Correction*					
	Lexical representation						*Semantic representation*					
Results	**R|N**	**$\sum$RD@K**	**P**	**R**	**IP**	**User ranking**	**R|N**	**$\sum$RD@K**	**P**	**R**	**IP**	**User Ranking**
1	1	1	1	0.1	1	4	1	1	1	0.1	1	5
2	1	2	1	0.2	1	3	1	2	1	0.2	1	4
3	1	3	1	0.3	1	2	1	3	1	0.3	1	3
4	0	3	.75	0.3	0.75	0	1	4	1	0.4	1	4
5	0	3	.60	0.3	0.66	0	0	4	0.8	0.4	0.88	0
6	1	4	.66	0.4	0.66	2	1	5	0.833	0.5	0.88	3
7	0	4	.57	0.4	0.625	0	1	6	0.857	0.6	0.88	3
8	1	5	.63	0.5	0.625	2	1	7	0.875	0.7	0.88	2
9	0	5	.55	0.5	0.55	0	1	8	0.88	0.8	0.88	2
10	0	5	.50	0.5	0.5	0	0	8	0.8	0.8	0.8	0

Table A.2: Table showing the lexical and semantic representation of one of 20 queries assessed in Exp.A(6.4)

A.6 20 Benchmark queries used in Experiment A (6.4)

Queries evaluated in Experiment A	
Query No.	**Query**
1	cohort sequential design
2	multivariate logistic regression analysis
3	design effects in statistics
4	online survey disadvantages
5	randomized control trials
6	media analysis
7	finite population correction
8	indirect geo-referencing
9	macrodata guide
10	evaluating interaction effects
11	primary sampling unit
12	reasoning
13	qualitative benchmarking
14	UK address example
15	logistic regression in SPSS
16	data collection skills
17	case study on non-verbal communication
18	ethnic group
19	forecasting
20	exploratory factor analysis

Table A.3: List of 20 benchmark queries, searched against FT index and SemDex in Experiment A

A.7 TREC-11 Points Interpolated Precision/Recall across queries

Table A.4 shows the carefully calculated semantic query-wise scores for the entire set of queries

TREC	Queries																				
R	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**AP**
0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.87	1	1	1	1	0.9935
0.1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.87	1	1	1	0.8	0.9835
0.2	0.66	1	1	1	1	1	1	1	1	1	1	1	0.8	1	1	0.87	1	1	1	0.8	0.9565
0.3	0.66	1	1	0.8	1	1	1	1	1	1	1	1	0.8	1	1	0.87	1	1	1	0.8	0.9465
0.4	0.66	1	1	0.8	1	1	1	1	1	1	1	1	0.8	1	1	0.87	1	0.8	1	0.8	0.9365
0.5	0.66	1	1	0.8	1	1	0.8	1	1	1	1	1	0.88	1	1	0.87	1	0.8	1	0.8	0.9305
0.6	0.6	1	1	0.71	1	0.88	0.88	1	1	0.8	1	1	0.88	1	1	0.87	1	1	1	0.8	0.921
0.7	0.6	1	1	0.71	1	0.88	0.88	1	1	0.71	0.88	1	0.88	1	1	0.87	0.85	1	1	0.8	0.903
0.8	0.6	1	0.77	0.7	1	0.88	0.88	1	1	0.71	0.88	0.88	0.88	1	0.9	0.87	0.8	1	1	0.8	0.8775
0.9	0.66	1	0.77	0.7	1	0.88	0.88	1	1	0.6	0.88	0.88	0.88	1	0.9	0.8	0.8	1	1	0.8	0.8715
1.0	0.6	1	0.7	0.7	1	0.8	0.8	1	1	0.6	0.88	0.8	0.8	1	0.9	0.7	0.8	1	1	0.8	0.844

Table A.4: This table shows the semantic index-based interpolated precision/recall score for each query in order to construct the Figure 7.3

Table A.5 shows the carefully calculated lexical or keywords-based query-wise scores for the entire set of queries

TREC	Queries																				
R	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**AP**
0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.8	0.99
0.1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0.5	1	1	1	1	0.925
0.2	1	1	1	1	1	1	1	1	1	1	1	0.25	0.44	1	1	0.5	1	1	1	1	0.9095
0.3	1	0.77	1	1	1	1	1	1	1	1	1	0.5	0.44	1	1	0.5	1	1	1	1	0.9105
0.4	1	0.77	1	1	1	0.8	1	1	1	1	1	0.3	0.44	1	1	0.5	1	0.9	1	1	0.8855
0.5	1	0.77	1	0.5	1	0.8	1	1	1	1	1	0.3	0.44	0.5	1	0.5	1	0.9	0.8	1	0.8255
0.6	1	0.77	0.75	0.55	1	0.8	0.66	0.8	0.62	0.75	1	0.3	0.44	0.33	1	0.5	1	0.9	0.71	1	0.744
0.7	1	0.77	0.71	0.55	0.9	0.8	0.6	0.8	0.62	0.5	1	0.3	0.44	0.28	1	0.6	1	0.9	0.71	1	0.724
0.8	1	0.77	0.71	0.55	0.9	0.6	0.62	0.71	0.62	0.5	1	0.3	0.44	0.28	0.83	0.6	0.66	0.9	0.66	0.9	0.6775
0.9	1	0.7	0.62	0.55	0.9	0.55	0.55	0.62	0.62	0.5	0.55	0.3	0.44	0.25	0.66	0.4	0.4	0.9	0.66	0.9	0.6035
1.0	1	0.7	0.5	0.5	0.9	0.5	0.5	0.5	0.5	0.5	0.55	0.4	0.4	0.2	0.6	0.4	0.25	0.9	0.6	0.9	0.565

Table A.5: This table shows the keywords or lexical index-based interpolated precision/recall score for each query in order to construct Figure 7.3

Table A.6 shows the Average Precision scores for all 20 queries searched against lexical, semantic indices in order to construct the Figure 7.4. The data in the table also show AR (Average Rankings) used to construct MAR Figure 7.5 curve in Chapter 7.

TREC	**Queries**																				
Q	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**MAP/MAR**
AP-S	.57	1	.87	.75	1	.92	.9	1	1	.79	.95	.94	.79	1	.85	.67	.9	.82	0.85	.51	0.86 [MAP-S]
AP-L	.47	.73	.74	.58	.96	.71	.72	.76	.71	.67	.78	.2	.43	.35	.85	.39	.48	.89	.78	.97	0.658 [MAP-L]
AR-S	1.8	2.9	2.1	2.5	4.1	2.5	2.6	3.2	4.2	1.5	3.88	2.7	2	3.11	2.9	1.8	2.1	2.4	3.2	2.6	2.70 [MAR-S]
AR-L	.6	1.7	1.3	1.7	2.8	1.3	1.3	1.5	1.5	1.4	1.66	1	1	0.5	1.6	0.8	0.5	2.3	1.5	2.6	1.428 [MAR-L]

Table A.6: This table shows the Average Precision scores for all 20 queries searched against lexical, semantic indices in order to construct Figure 7.4. The data also shows AP used to construct MAR Figure 7.5 curve in Chapter 7

.

A.8 33 Benchmark queries used in Experiment C (6.6)

Queries evaluated in Experiment C	
Query No.	**Query**
1	interpreting logistic regression in SPSS
2	components of research proposal
3	primary sampling unit
4	stages of a systematic review
5	using imputation for missing values
6	indirect geo-referencing
7	randomized control trials
8	critical thinking
9	online survey disadvantages
10	multivariate logistic regression analysis
11	mixture model
12	trust and respect in a team
13	paradata in survey research
14	evaluating interaction effects
15	what is media analysis
16	online research methods
17	logistic regression in SPSS
18	factor analysis
19	forecasting
20	case study on non-verbal communication skills
21	finite population correction
22	UK address example
23	reasoning
24	random sample enumeration
25	ethnic group
26	design effects in statistics
27	qualitative research skills
28	data collection skills
29	media analysis
30	exploratory factor analysis
31	critical thinking
32	sample enumeration
33	natural experiments in social sciences

Table A.7: List of 33 benchmark queries, searched against SemDex and SemCrowDex in Experiment C

TREC	Queries [1-20]																			
R	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
0	1	1	1	1	1	1	1	1	0.90	1	1	1	1	1	1	1	1	1	1	1
0.1	1	1	1	1	1	1	1	1	0.90	1	1	1	1	1	1	1	1	1	1	1
0.2	1	1	1	1	1	1	1	1	0.90	1	1	1	0.8	1	1	1	1	1	1	1
0.3	1	1	1	1	1	1	1	1	0.90	0.90	1	1	0.8	0.67	1	1	1	1	1	1
0.4	1	1	1	1	1	1	1	1	0.90	1	1	1	0.8	0.67	1	0.90	1	1	1	1
0.5	0.80	1	1	1	1	1	1	1	0.90	1	1	0.75	0.8	0.67	1	0.90	1	1	0.90	1
0.6	0.80	1	1	1	1	1	1	1	0.90	1	1	0.75	0.8	0.50	0.71	0.90	1	1	0.90	0.43
0.7	0.80	1	1	1	1	1	1	1	0.90	1	1	0.75	0.8	0.50	0.71	0.90	1	1	0.90	0.43
0.8	0.80	1	1	1	1	1	1	1	0.90	1	1	0.75	0.8	0.50	0.71	0.90	1	0.90	0.90	0.43
0.9	0.80	1	1	0.90	1	1	1	1	0.90	1	1	0.75	0.8	0.50	0.71	0.90	1	0.90	0.90	0.43
1	0.80	1	1	0.90	1	1	1	1	0.90	1	1	0.70	0.8	0.50	0.71	0.90	1	0.90	0.90	0.43

	Queries [21-33]													
R	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	*AP*
0	1	1	1	1	1	1	1	1	1	1	0.90	1	1	*0.993*
0.1	1	1	1	1	1	1	1	1	1	1	0.90	1	1	*0.993*
0.2	1	1	1	1	1	1	1	1	1	1	0.90	1	1	*0.988*
0.3	1	0.90	1	1	1	1	1	1	1	0.90	0.90	1	1	*0.969*
0.4	1	1	1	1	1	1	1	1	1	0.90	0.90	1	1	*0.972*
0.5	1	1	0.75	1	1	1	1	1	1	0.90	0.90	1	1	*0.0.947*
0.6	1	1	0.75	1	1	0.83	1	0.86	1	0.90	0.90	1	1	*0.907*
0.7	1	1	0.75	1	1	0.83	1	0.86	1	0.90	0.90	1	1	*0.907*
0.8	1	1	0.75	1	1	0.80	1	0.86	1	0.90	0.90	1	1	*0.903*
0.9	1	1	0.75	1	1	0.80	1	0.86	1	0.90	0.90	1	1	*0.900*
1.0	1	1	0.75	1	1	0.80	1	0.80	1	0.90	0.90	1	1	*0.896*

Table A.8: Interpolated precision/recall score (AP) computed for each query (executed against *SemCrowDex* having crowd annotation searchable fields) as part of Exp.C (33 queries in total) in order to construct the Figure 8.5

TREC	Queries [1-20]																			
R	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
0	1	1	1	0.78	0.88	1	1	0.75	1	0.9	1	1	0.25	1	0	0.9	1	1	1	1
0.1	1	1	1	0.78	0.88	1	1	0.75	1.0	0.9	1	1	0.25	1	0	0.9	1	1	1	1
0.2	1	0.85	1	0.78	0.88	1	0.8	0.75	1	0.9	1	0.8	0.25	1	0	0.9	1	1	1	1
0.3	1	0.86	1	0.78	0.88	1	0.8	0.75	1	0.90	1	0.8	0.25	1	0	0.9	1	1	1	1
0.4	1	0.86	1	0.78	0.88	1	0.8	0.75	1	0.90	1	0.8	0.25	1	0	0.9	1	1	1	1
0.5	1	0.86	1	0.78	0.88	1	0.8	0.75	1	0.90	1	0.8	0.20	1	0	0.9	1	1	1	1
0.6	1	0.86	0.9	0.78	0.80	1	0.8	0.75	0.75	0.90	1	0.8	0.20	1	0	0.9	0.83	1	1	0.8
0.7	0.8	0.86	0.9	0.78	0.80	1	0.6	0.75	0.75	0.90	1	0.7	0.25	1	0	0.9	0.8	0.75	0.71	0.71
0.8	0.80	0.8	0.94	0.78	0.8	1	0.6	0.75	0.75	0.90	1	0.7	0.2	0.7	0	0.90	0.8	0.7	0.6	0.71
0.9	0.7	0.8	0.9	0.78	0.8	1	0.6	0.75	0.75	0.9	1	0.7	0.2	0.65	0	0.9	0.8	0.7	0.6	0.62
1	0.71	0.8	0.94	0.78	0.8	1	0.6	0.75	0.75	0.9	1	0.7	0.25	0.67	0	0.9	0.8	0.71	0.6	0.67
	Queries [21-33]																			
R	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	*AP*						
0	1	1	1	1	1	1	1	1	1	0.8	1	0.75	1	*0.909*						
0.1	1	1	1	1	1	1	1	1	1	0.8	1	0.75	1	*0.909*						
0.2	1	1	1	1	1	1	0.86	1	1	0.8	1	0.75	1	*0.888*						
0.3	0.86	1	1	0.8	1	0.78	0.86	0.75	1	0.8	0.6	0.75	0.8	*0.0.846*						
0.4	0.86	1	1	0.8	0.9	0.78	0.86	0.75	1	0.8	0.6	0.75	0.8	*0.835*						
0.5	0.86	1	1	0.8	0.9	0.78	0.86	0.71	0.8	0.8	0.67	0.75	0.8	*0.0.828*						
0.6	0.86	1	1	0.71	0.9	0.78	0.86	0.71	0.8	0.8	0.67	0.75	0.8	*0.0.809*						
0.7	0.86	1	1	0.71	0.9	0.78	0.86	0.71	0.8	0.8	0.71	0.56	0.6	*0.765*						
0.8	0.8	0.9	0.8	0.7	0.9	0.7	0.8	0.7	0.8	0.8	0.7	0.5	0.6	*0.731*						
0.9	0.8	0.9	0.8	0.7	0.9	0.75	0.8	0.7	0.61	0.8	0.78	0.56	0.6	*0.0.722*						
1.0	0.8	0.9	0.8	0.7	0.9	0.78	0.8	0.71	0.63	0.8	0.7	0.56	0.6	*0.727*						

Table A.9: Interpolated precision/recall score (AP) computed for each query (executed against *SemDex* having no crowd annotation searchable fields) as part of Exp.C (33 queries in total) in order to construct the Figure 8.5

Queries	AP-Crowd	AP-[No.Crowd]	AR-CROWD	AR-[No.Crowd]
Q1	0.8	0.76	2.1	2
Q2	1	0.77	3.1	2.7
Q3	1	0.93	3.8	3.1
Q4	0.98	0.55	3.2	2.2
Q5	1	0.68	3.2	2.1
Q6	1	1	3.2	3
Q7	1	0.94	4	3.7
Q8	1	0.46	3.4	1.8
Q9	0.7	0.72	2.7	2.3
Q10	0.86	0.7	3.8	2.6
Q11	0.54	0.62	1.6	2.1
Q12	0.78	0.7	2.6	2.4
Q13	0.74	0.11	2.3	1.1
Q14	0.53	0.78	1.55	1.66
Q15	0.82	0.71	2.4	1
Q16	0.9	0.71	2.9	2.9
Q17	0.99	0.86	3.7	2.8
Q18	0.94	0.63	3.2	1.9
Q19	0.89	0.82	4.1	3.2
Q20	0.41	0.78	2	1.9
Q21	0.99	0.82	2.6	2.3
Q22	0.86	0.97	3.5	3.1
Q23	0.78	0.94	2.2	2.5
Q24	1	0.76	3.6	2.4
Q25	0.99	0.89	3.7	2.4
Q26	0.87	0.73	3	2
Q27	1	0.77	4.9	2.7
Q28	0.88	0.65	3.7	1.8
Q29	1	0.71	3.6	1.8
Q30	0.86	0.51	3.8	2.8
Q31	0.7	0.73	2.7	2.4
Q32	0.99	0.5	3.2	2.1
Q33	1	0.66	4	1.7
Average	0.872727	0.7257	3.13181	2.30
	MAP-CROWD	MAP-NO.CROWD	MAR-CROWD	MAR-NO.CROWD

Table A.10: Non-interpolated AP scores for all 33 queries searched against semantic crowd index (*SemCrowDex*) as depicted in Figure 8.6. The Table also shows Average Ranking (AR) score depicted in Figure 8.7 with and without the Crowd-annotation element.

Table A.10 shows the non-interpolated AP scores for all 33 queries searched against semantic crowd index (*SemCrowDex*) in order to construct Non-IP Crowd and no-crowd AP chart in Figure 8.6. The Table also shows data which was used to construct Average Ranking (AR) chart in Figure 8.7 with and without the Crowd-annotation element.

Queries	AP-Crowd	AP-[No.Crowd]	(X-0.8727)^2	(X-0.6927)^2
Q1	0.8	0.76	0.005	.0045
Q2	1	0.77	0.016	.0059
Q3	1	0.93	0.016	.056
Q4	0.98	0.55	0.012	.0203
Q5	1	0.68	.016	.0001
Q6	1	1	.016	.0944
Q7	1	0.66	.016	.001
Q8	1	0.46	.016	.054
Q9	0.7	0.72	.030	.0007
Q10	0.86	0.7	.000	5.29E-05
Q11	0.54	0.62	.111	.0052
Q12	0.78	0.7	.009	5.29E-05
Q13	0.74	0.11	.018	.3395
Q14	0.53	0.78	.117	.007
Q15	0.82	0	.003	.4798
Q16	0.9	0.71	.001	.0002
Q17	0.99	0.86	.014	.0279
Q18	0.94	0.63	.005	.0039
Q19	0.89	0.8	.000	.0115
Q20	0.41	0.78	.214	.0076
Q21	0.99	0.82	.014	.0161
Q22	0.86	0.97	.000	.0768
Q23	0.78	0.94	.009	.0611
Q24	1	0.76	.016	.0045
Q25	0.99	0.89	.014	.0389
Q26	0.87	0.73	.000	.0013
Q27	1	0.77	.016	.0059
Q28	0.88	0.65	.000	.0018
Q29	1	0.71	.016	.0002
Q30	0.86	0.51	.000	.0333
Q31	0.7	0.73	.030	.0013
Q32	0.99	0.5	.014	.0371
Q33	1	0.66	.016	.0010
Average	$\overline{X}_{1[Crowd]}$=0.872727	$\overline{X}_{2[No-Crowd]}$ =0.692727	$\sum = 0.779$	$\sum = 1.402$
			$S^2_{crowd} = 0.024$	$S^2_{no-crowd} = 0.0438$

Table A.11: This table shows the Average Precision scores for all 33 queries and *t-test* calculation in Section 8.6.7 in Chapter 8 using the formula $t_{calculated} = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s^2_{[Crowd]}}{n_1} + \frac{s^2_{[No-Crowd]}}{n_2}}}$

A.9 Document-query vector normalisation

This section presents the normalization of weights in 3 documents against a query, searched in a hyper semantic vector space. The normalization takes into account all annotation elements as well full-text content in documents and query vectors. Table A.12 shows a vector representation of three documents in terms of real-valued vector of TF-IDF weights $\in \mathbb{R}^{|V|}$ calculated using Equation 2.4. In order to compute cosine similarity between query and document vectors, using Equation 2.4, cosine similarity will be calculated as below:

Query terms	**Doc1** $\{UN \rightarrow N\}$	**Doc2** $\{UN \rightarrow N\}$	**Doc3**$\{UN \rightarrow N\}$
design	0.33	0.20	0.34
experimental	0.62	0.48	0.35
ethical	0.38	0.46	0.30
random	0.41	0.24	0
childhood	0.11	0.57	0.43
control	0	0.21	0
randomised	0.22	0.11	0
evaluation	0	0.23	0.14
hypothesis	0.10	0.24	0.26
trials	0	0.47	0.29

Table A.12: Log frequency TF.IDF weights of non-Semantic Document Vectors $\{UN \rightarrow N\} \equiv$ {Un-normalized to Normalized} using Equation 2.3 for document vectors.

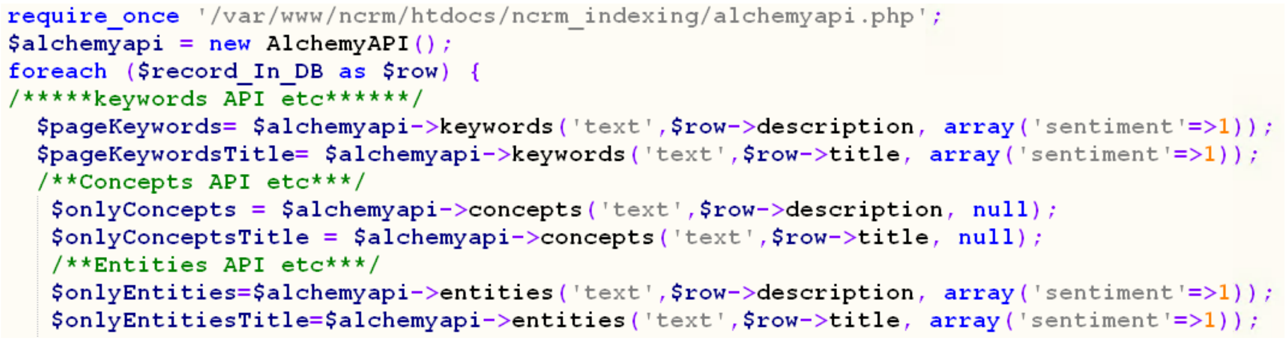
Here is the modified Table A.12, now called, Table A.13. annotated with further terms *control* and *trials*.

Query terms	**Doc1** $\{UN \rightarrow N\}$	**Doc2** $\{UN \rightarrow N\}$	**Doc3**$\{UN \rightarrow N\}$
experimental	0.51	0.37	0.62
ethical	0.58	0.55	0.47
random	0.38	0.46	0.29
childhood	0.41	0.24	0
trials	0.28	0.28	0.36
control	0.18	0.21	0.20
randomised	0.20	0.11	0
evaluation	0	0.17	0.20
hypothesis	0.10	0.24	0.25
design	0.10	0.56	0.39

Table A.13: Log frequency TF.IDF weights of Semantic Document Vectors $\{UN \rightarrow N\} \equiv$ {Un-normalized to Normalized} using Equation 2.3 for document vectors.

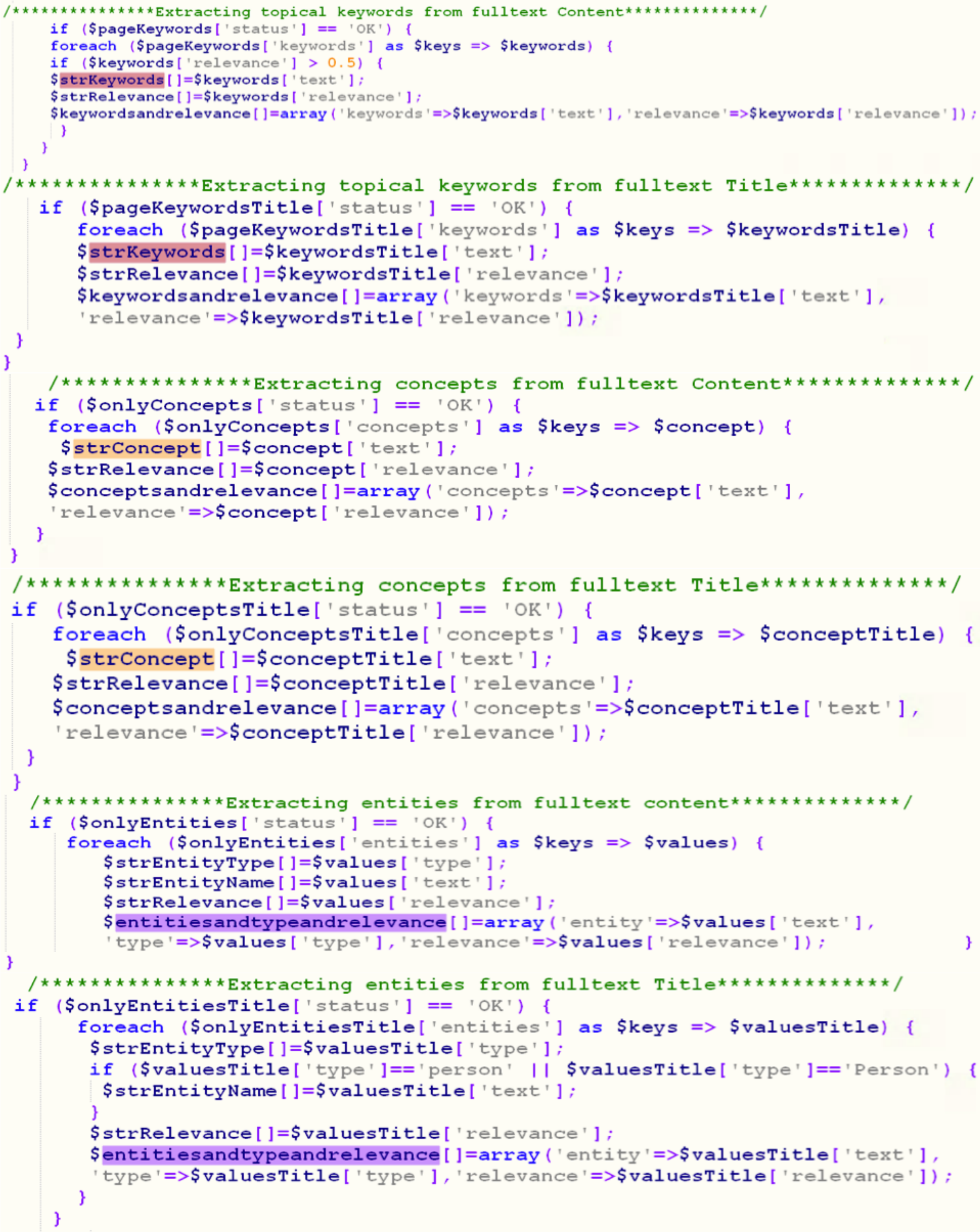
Appendix B

Experimentation, questionnaires and recruitment advertisements



```
require_once '/var/www/ncrm/htdocs/ncrm_indexing/alchemyapi.php';
$alchemyapi = new AlchemyAPI();
foreach ($record_In_DB as $row) {
/*****keywords API etc******/
  $pageKeywords= $alchemyapi->keywords('text',$row->description, array('sentiment'=>1));
  $pageKeywordsTitle= $alchemyapi->keywords('text',$row->title, array('sentiment'=>1));
  /**Concepts API etc***/
   $onlyConcepts = $alchemyapi->concepts('text',$row->description, null);
   $onlyConceptsTitle = $alchemyapi->concepts('text',$row->title, null);
   /**Entities API etc***/
   $onlyEntities=$alchemyapi->entities('text',$row->description, array('sentiment'=>1));
   $onlyEntitiesTitle=$alchemyapi->entities('text',$row->title, array('sentiment'=>1));
```

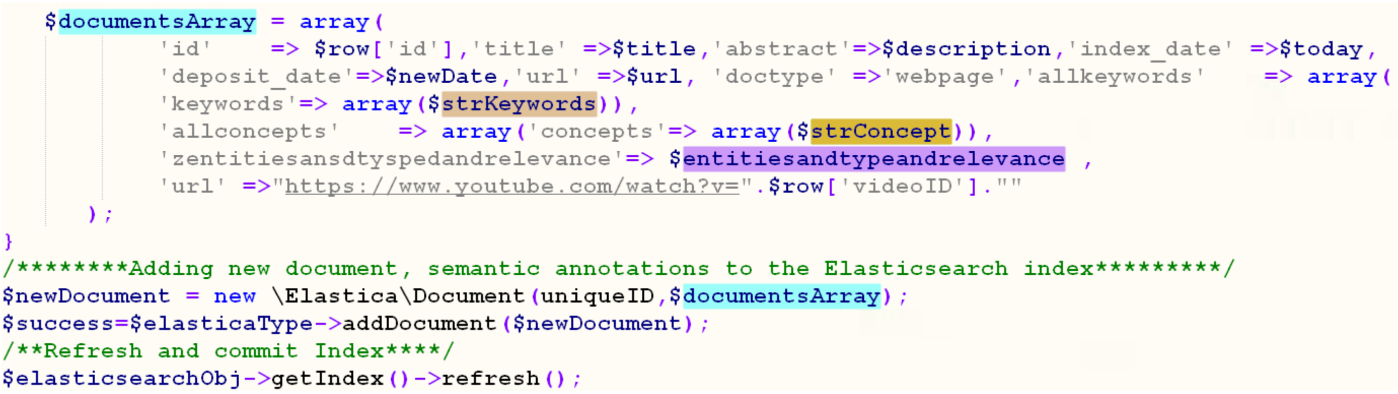
Figure B.1: Importing Alchemy framework and creating Keywords, Concepts, Entities array objects to store the respective instances in each iteration



```
/**************Extracting topical keywords from fulltext Content**************/
     if ($pageKeywords['status'] == 'OK') {
     foreach ($pageKeywords['keywords'] as $keys => $keywords) {
     if ($keywords['relevance'] > 0.5) {
     $strKeywords[]=$keywords['text'];
     $strRelevance[]=$keywords['relevance'];
     $keywordsandrelevance[]=array('keywords'=>$keywords['text'],'relevance'=>$keywords['relevance']);
      }
    }
  }
/***************Extracting topical keywords from fulltext Title**************/
  if ($pageKeywordsTitle['status'] == 'OK') {
     foreach ($pageKeywordsTitle['keywords'] as $keys => $keywordsTitle) {
     $strKeywords[]=$keywordsTitle['text'];
     $strRelevance[]=$keywordsTitle['relevance'];
     $keywordsandrelevance[]=array('keywords'=>$keywordsTitle['text'],
     'relevance'=>$keywordsTitle['relevance']);
 }
}
   /***************Extracting concepts from fulltext Content**************/
  if ($onlyConcepts['status'] == 'OK') {
   foreach ($onlyConcepts['concepts'] as $keys => $concept) {
    $strConcept[]=$concept['text'];
   $strRelevance[]=$concept['relevance'];
   $conceptsandrelevance[]=array('concepts'=>$concept['text'],
   'relevance'=>$concept['relevance']);
  }
 }
 /**************Extracting concepts from fulltext Title**************/
 if ($onlyConceptsTitle['status'] == 'OK') {
    foreach ($onlyConceptsTitle['concepts'] as $keys => $conceptTitle) {
     $strConcept[]=$conceptTitle['text'];
    $strRelevance[]=$conceptTitle['relevance'];
    $conceptsandrelevance[]=array('concepts'=>$conceptTitle['text'],
    'relevance'=>$conceptTitle['relevance']);
  }
 }
  /**************Extracting entities from fulltext content**************/
  if ($onlyEntities['status'] == 'OK') {
     foreach ($onlyEntities['entities'] as $keys => $values) {
        $strEntityType[]=$values['type'];
        $strEntityName[]=$values['text'];
        $strRelevance[]=$values['relevance'];
        $entitiesandtypeandrelevance[]=array('entity'=>$values['text'],
        'type'=>$values['type'],'relevance'=>$values['relevance']);            }
}
  /**************Extracting entities from fulltext Title**************/
 if ($onlyEntitiesTitle['status'] == 'OK') {
      foreach ($onlyEntitiesTitle['entities'] as $keys => $valuesTitle) {
       $strEntityType[]=$valuesTitle['type'];
       if ($valuesTitle['type']=='person' || $valuesTitle['type']=='Person') {
        $strEntityName[]=$valuesTitle['text'];
       }
       $strRelevance[]=$valuesTitle['relevance'];
       $entitiesandtypeandrelevance[]=array('entity'=>$valuesTitle['text'],
       'type'=>$valuesTitle['type'],'relevance'=>$valuesTitle['relevance']);
      }
    }
```

Figure B.2: Alchemy API i.e. Concepts, Keywords and Entities are being called against each database record having fulltext description and title of the webpage

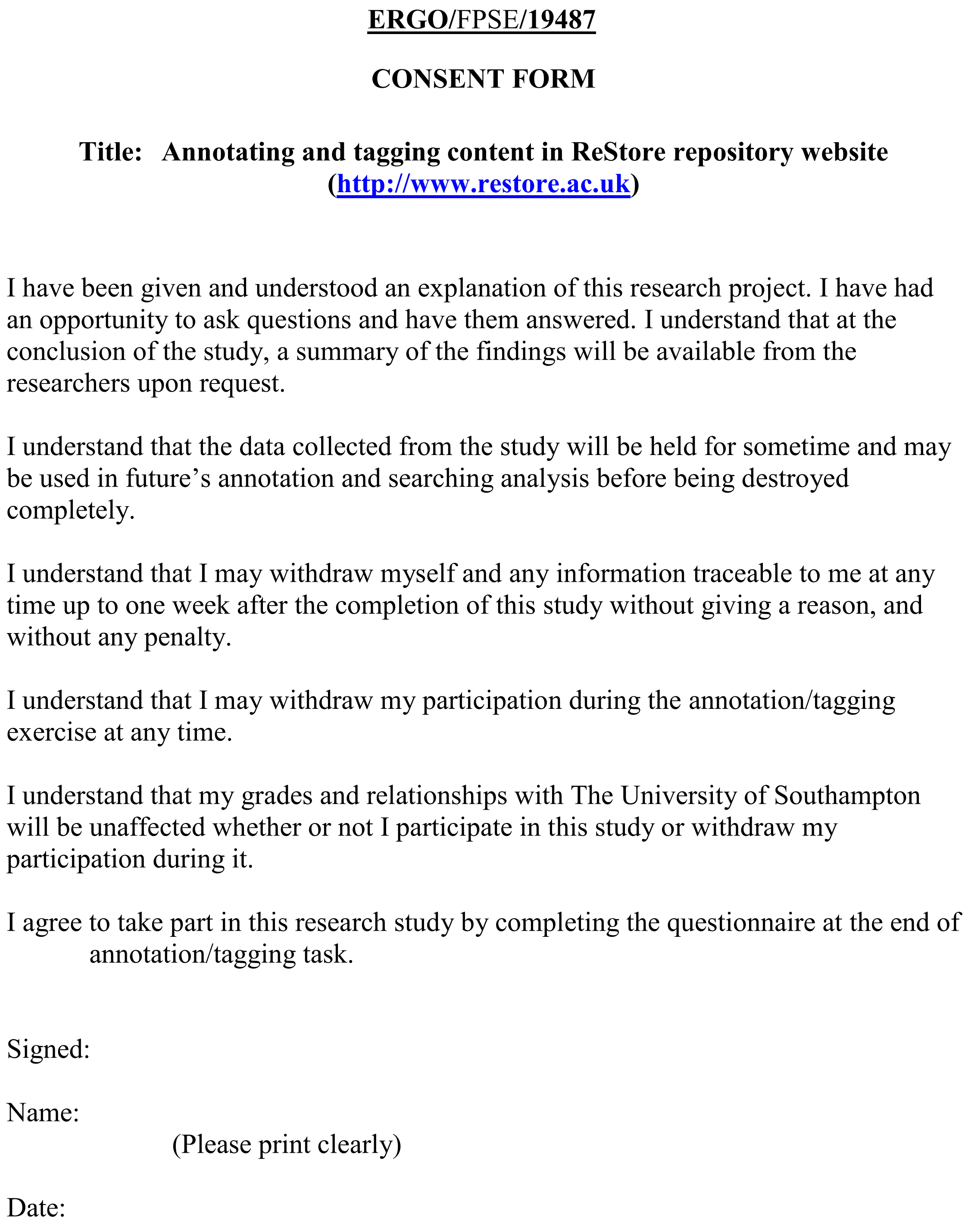
Figures B.1, B.2 and B.3 above, programmatically explain the entire process flow of importing Alchemy API service, document analysis for IE and annotated document



```
   $documentsArray = array(
            'id'    => $row['id'],'title' =>$title,'abstract'=>$description,'index_date' =>$today,
            'deposit_date'=>$newDate,'url' =>$url, 'doctype' =>'webpage','allkeywords'    => array(
            'keywords'=> array($strKeywords)),
            'allconcepts'    => array('concepts'=> array($strConcept)),
            'zentitiesansdtyspedandrelevance'=> $entitiesandtypeandrelevance ,
            'url' =>"https://www.youtube.com/watch?v=".$row['videoID']."
      );
}
/********Adding new document, semantic annotations to the Elasticsearch index*********/
$newDocument = new \Elastica\Document(uniqueID,$documentsArray);
$success=$elasticaType->addDocument($newDocument);
/**Refresh and commit Index****/
$elasticsearchObj->getIndex()->refresh();
```

Figure B.3: Document being added to the Elasticsearch index along with semantic annotations after text analysis is completed

addition to ES index as detailed in Section 5.2.2.



ERGO/FPSE/**19487**

CONSENT FORM

Title: Annotating and tagging content in ReStore repository website (http://www.restore.ac.uk)

I have been given and understood an explanation of this research project. I have had an opportunity to ask questions and have them answered. I understand that at the conclusion of the study, a summary of the findings will be available from the researchers upon request.

I understand that the data collected from the study will be held for sometime and may be used in future's annotation and searching analysis before being destroyed completely.

I understand that I may withdraw myself and any information traceable to me at any time up to one week after the completion of this study without giving a reason, and without any penalty.

I understand that I may withdraw my participation during the annotation/tagging exercise at any time.

I understand that my grades and relationships with The University of Southampton will be unaffected whether or not I participate in this study or withdraw my participation during it.

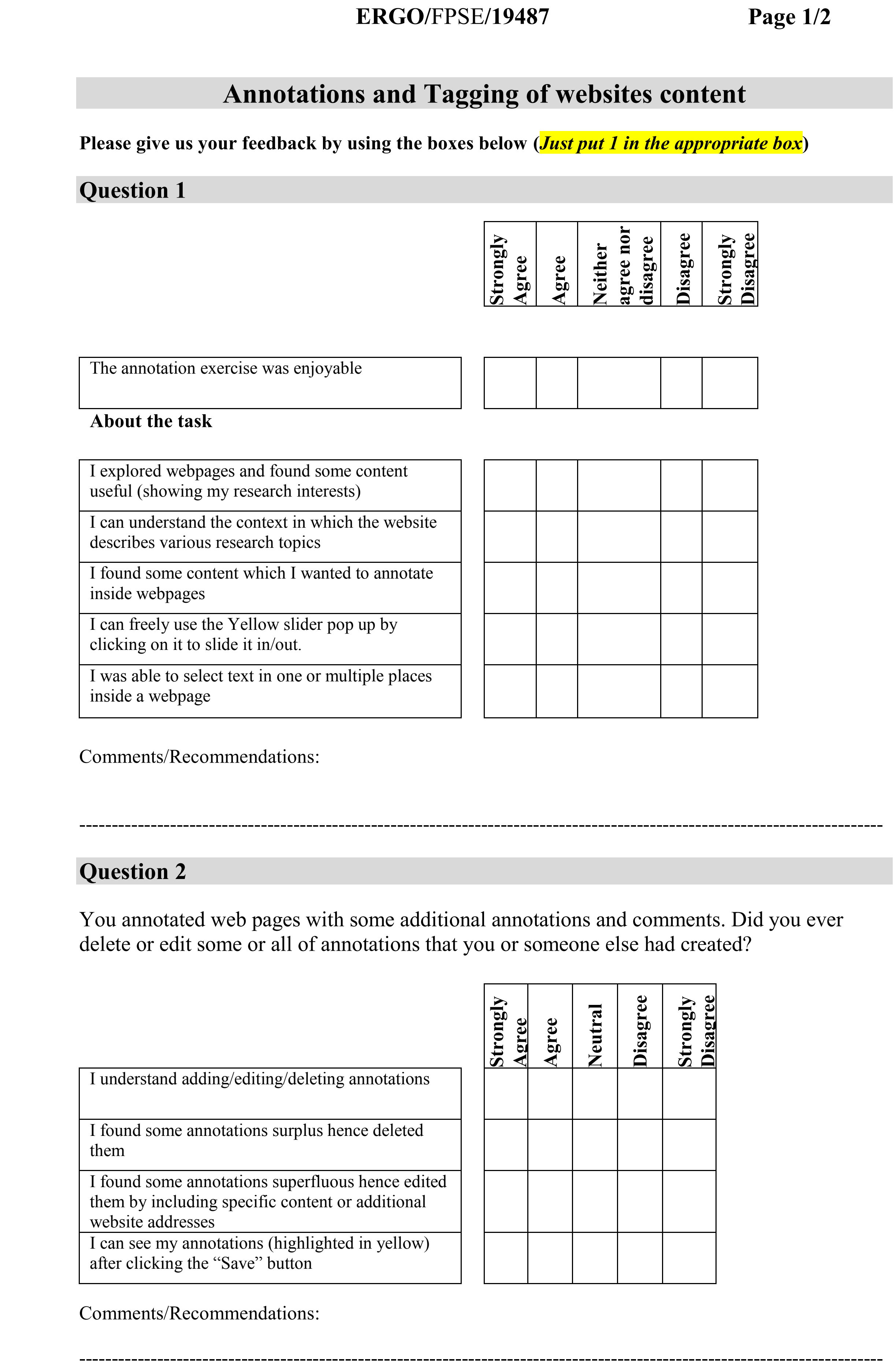
I agree to take part in this research study by completing the questionnaire at the end of annotation/tagging task.

Signed:

Name:
(Please print clearly)

Date:

Figure B.4: Consent form for the participants of crowd annotation and tagging experiment (Exp. B)



ERGO/FPSE/**19487** **Page 1/2**

Annotations and Tagging of websites content

Please give us your feedback by using the boxes below (*Just put 1 in the appropriate box*)

Question 1

	Strongly Agree	Agree	Neither agree nor disagree	Disagree	Strongly Disagree
The annotation exercise was enjoyable					

About the task

I explored webpages and found some content useful (showing my research interests)					
I can understand the context in which the website describes various research topics					
I found some content which I wanted to annotate inside webpages					
I can freely use the Yellow slider pop up by clicking on it to slide it in/out.					
I was able to select text in one or multiple places inside a webpage					

Comments/Recommendations:

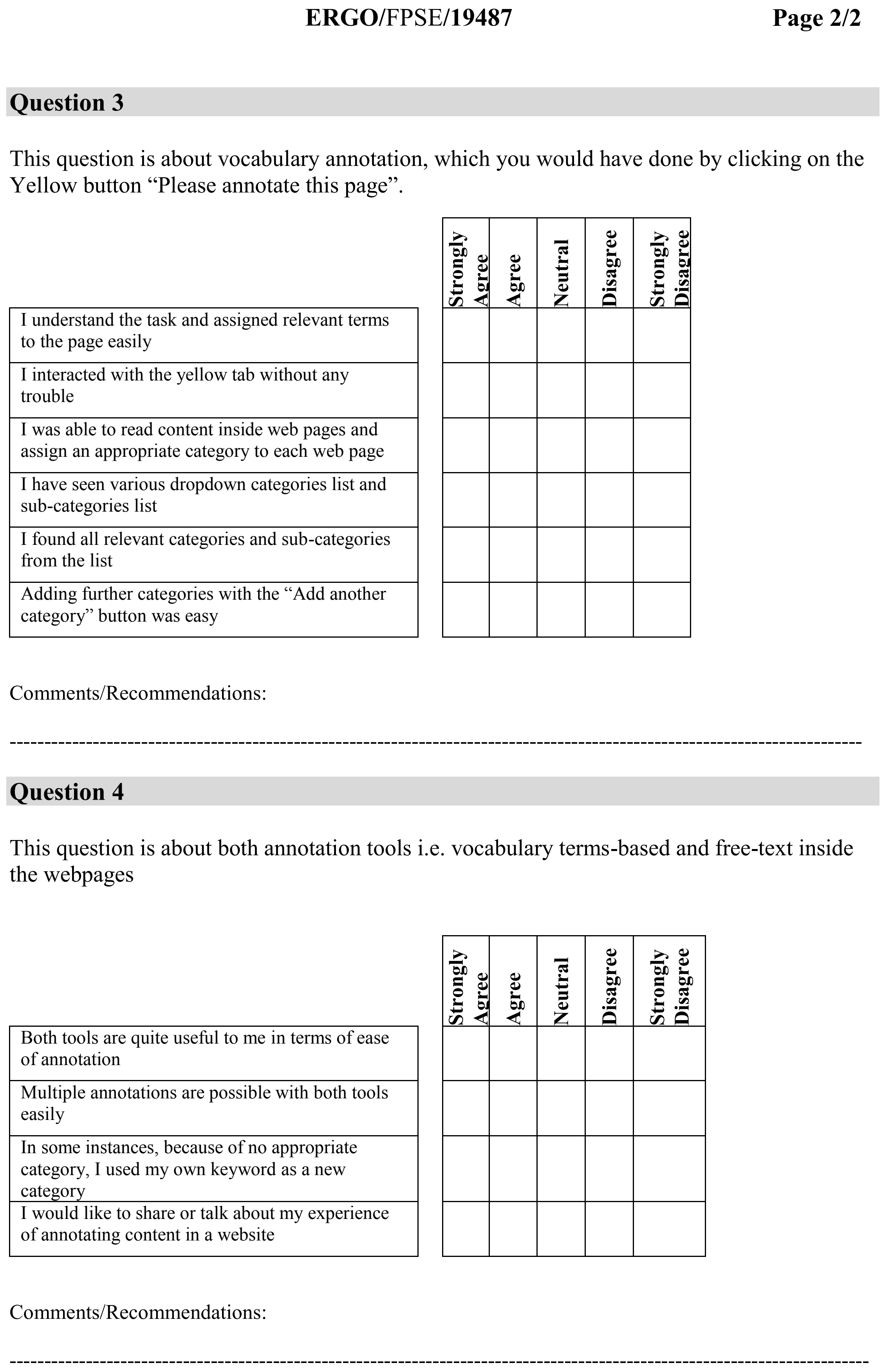
Question 2

You annotated web pages with some additional annotations and comments. Did you ever delete or edit some or all of annotations that you or someone else had created?

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
I understand adding/editing/deleting annotations					
I found some annotations surplus hence deleted them					
I found some annotations superfluous hence edited them by including specific content or additional website addresses					
I can see my annotations (highlighted in yellow) after clicking the "Save" button					

Comments/Recommendations:

Figure B.5: Page 1/2: Questionnaire form for the crowd annotators in Experiment B (Exp.B)



ERGO/FPSE/**19487** **Page 2/2**

Question 3

This question is about vocabulary annotation, which you would have done by clicking on the Yellow button "Please annotate this page".

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
I understand the task and assigned relevant terms to the page easily					
I interacted with the yellow tab without any trouble					
I was able to read content inside web pages and assign an appropriate category to each web page					
I have seen various dropdown categories list and sub-categories list					
I found all relevant categories and sub-categories from the list					
Adding further categories with the "Add another category" button was easy					

Comments/Recommendations:

Question 4

This question is about both annotation tools i.e. vocabulary terms-based and free-text inside the webpages

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
Both tools are quite useful to me in terms of ease of annotation					
Multiple annotations are possible with both tools easily					
In some instances, because of no appropriate category, I used my own keyword as a new category					
I would like to share or talk about my experience of annotating content in a website					

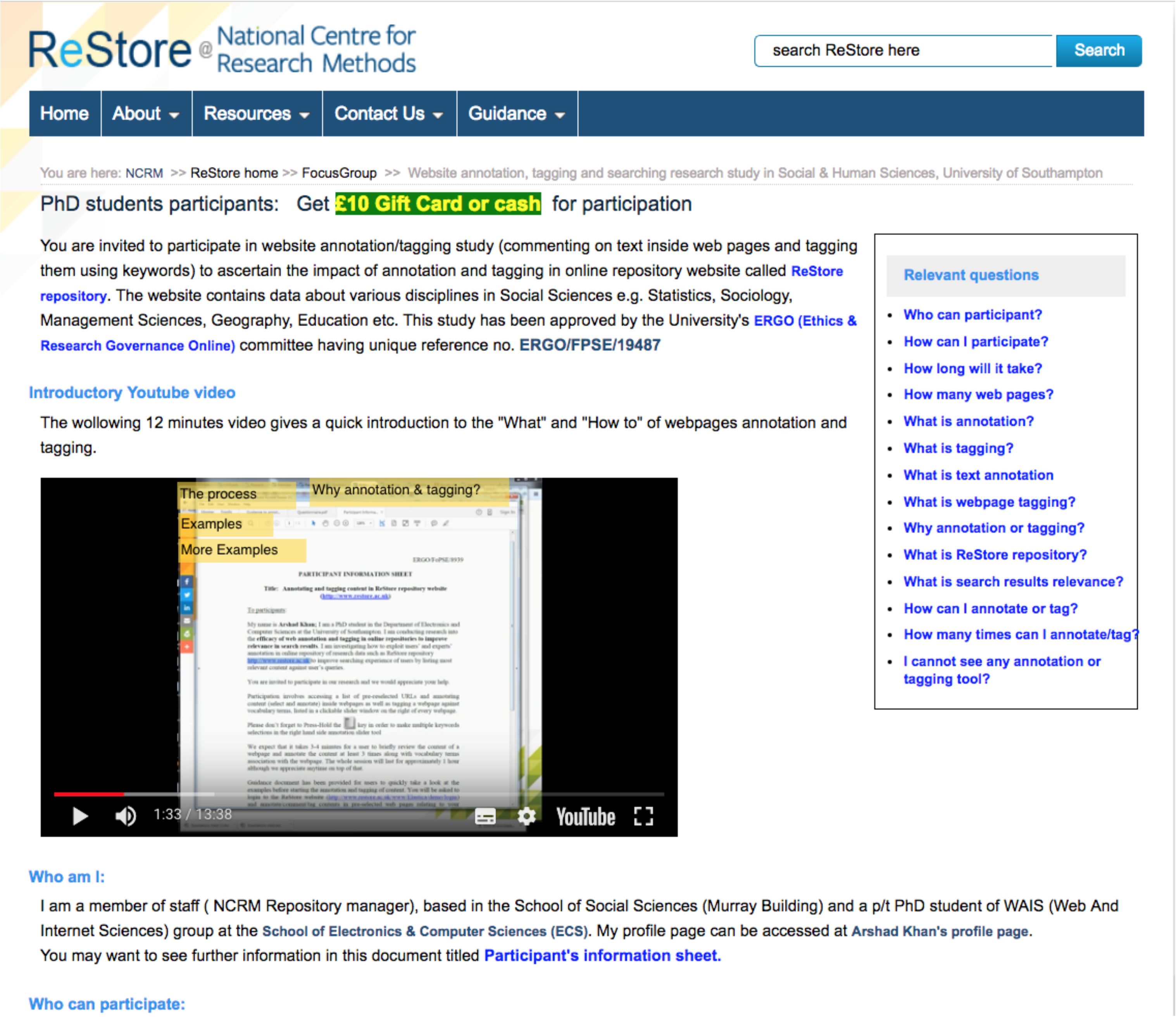
Comments/Recommendations:

Figure B.6: Page 2/2: Questionnaire form for the crowd annotators in Experiment B (Exp.B)

Table B.1: Results from the feedback questionnaire collected form the expert crowd-annotators at the end of crowd-annotation experiment

Questions	LS-1 Score	LS-2 Score	LS-3 Score	LS-4 Score	LS-5 Score	**Mean score**
Q 1.1	2	8	2	1	-	2.6
Q 1.2	4	7	1	2	-	2.8
Q 1.3	4	8	1	-	-	2.6
Q 1.4	4	8	-	1	-	2.6
Q 1.5	6	7	-	-	-	2.6
Q 1.6	10	3	-	-	-	2.6
Q 2.1	4	8	1	-	1	2.8
Q 2.2	-	-	1	7	4	2.4
Q 2.3	-	3	1	5	4	2.6
Q 2.4	8	5	-	-	-	2.6
Q 3.1	5	6	1	1	-	2.6
Q 3.2	7	4	2	-	-	2.6
Q 3.3	4	4	4	2	-	2.8
Q 3.4	6	7	-	-	-	2.6
Q 3.5	3	6	1	2	1	2.6
Q 3.6	5	3	3	2	-	2.6
Q 4.1	8	3	1	1	-	2.6
Q 4.2	7	5	1	-	-	2.6
Q 4.3	5	6	1	1	-	2.6
Q 4.4	-	4	6	3	-	2.6

Table B.1 shows the feedback from expert crowd-annotators (in terms of 5 LS=Likert Scale mean points against each question) collected through the feedback questionnaire given in Figure B.5 and Figure B.6. Figure 8.9 in Chapter 8 illustrates the overall attitude of expert crowd-annotators towards our annotation and tagging tool i.e. *AnnoTagger*.



You are here: NCRM >> ReStore home >> FocusGroup >> Website annotation, tagging and searching research study in Social & Human Sciences, University of Southampton

PhD students participants: Get £10 Gift Card or cash for participation

You are invited to participate in website annotation/tagging study (commenting on text inside web pages and tagging them using keywords) to ascertain the impact of annotation and tagging in online repository website called ReStore repository. The website contains data about various disciplines in Social Sciences e.g. Statistics, Sociology, Management Sciences, Geography, Education etc. This study has been approved by the University's ERGO (Ethics & Research Governance Online) committee having unique reference no. ERGO/FPSE/19487

Relevant questions

- Who can participant?
- How can I participate?
- How long will it take?
- How many web pages?
- What is annotation?
- What is tagging?
- What is text annotation
- What is webpage tagging?
- Why annotation or tagging?
- What is ReStore repository?
- What is search results relevance?
- How can I annotate or tag?
- How many times can I annotate/tag?
- I cannot see any annotation or tagging tool?

Introductory Youtube video

The wollowing 12 minutes video gives a quick introduction to the "What" and "How to" of webpages annotation and tagging.

ERGO/FoPSE/8939

PARTICIPANT INFORMATION SHEET

Title: Annotating and tagging content in ReStore repository website (http://www.restore.ac.uk)

To participants:

My name is Arshad Khan; I am a PhD student in the Department of Electronics and Computer Sciences at the University of Southampton. I am conducting research into the efficacy of web annotation and tagging in online repositories to improve relevance in search results. I am investigating how to exploit users' and experts' annotation in online repository of research data such as ReStore repository http://www.restore.ac.uk to improve searching experience of users by listing most relevant content against user's queries.

You are invited to participate in our research and we would appreciate your help.

Participation involves accessing a list of pre-reselected URLs and annotating content (select and annotate) inside webpages as well as tagging a webpage against vocabulary terms, listed in a clickable slider window on the right of every webpage.

Please don't forget to Press-Hold the key in order to make multiple keywords selections in the right hand side annotation slider tool

We expect that it takes 3-4 minutes for a user to briefly review the content of a webpage and annotate the content at least 3 times along with vocabulary terms association with the webpage. The whole session will last for approximately 1 hour although we appreciate anytime on top of that.

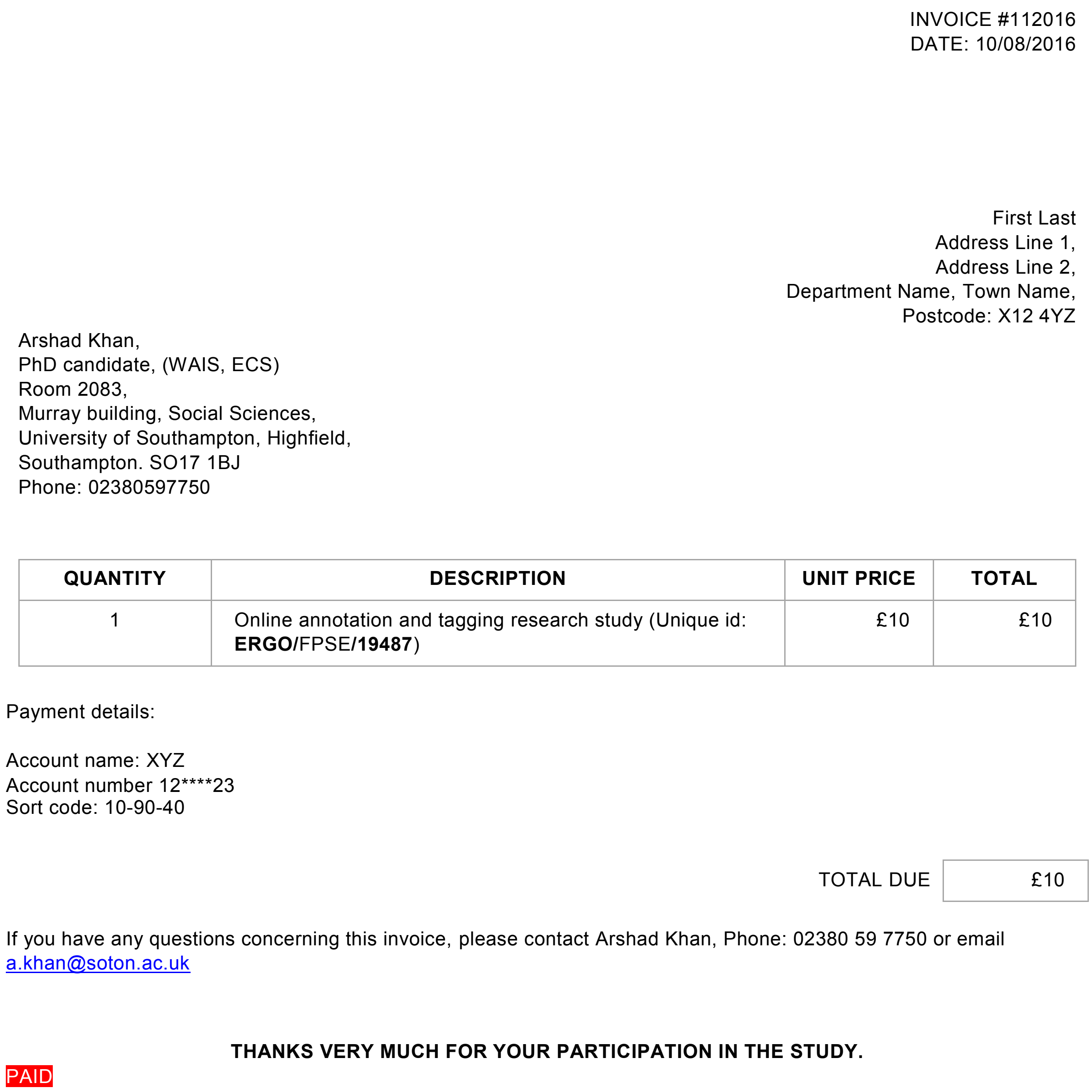
Guidance document has been provided for users to quickly take a look at the examples before starting the annotation and tagging of content. You will be asked to login to the ReStore website [illegible]

Who am I:

I am a member of staff (NCRM Repository manager), based in the School of Social Sciences (Murray Building) and a p/t PhD student of WAIS (Web And Internet Sciences) group at the School of Electronics & Computer Sciences (ECS). My profile page can be accessed at Arshad Khan's profile page.
You may want to see further information in this document titled Participant's information sheet.

Who can participate:

Figure B.7: A webpage in the ReStore website providing introductory video for crowd-annotation and tagging experiments along with step-by-step explanation aimed at participants' training before their participation. The webpage is available at http://www.restore.ac.uk/FocusGroup/



INVOICE #112016
DATE: 10/08/2016

First Last
Address Line 1,
Address Line 2,
Department Name, Town Name,
Postcode: X12 4YZ

Arshad Khan,
PhD candidate, (WAIS, ECS)
Room 2083,
Murray building, Social Sciences,
University of Southampton, Highfield,
Southampton. SO17 1BJ
Phone: 02380597750

QUANTITY	DESCRIPTION	UNIT PRICE	TOTAL
1	Online annotation and tagging research study (Unique id: **ERGO/**FPSE/**19487**)	£10	£10

Payment details:

Account name: XYZ
Account number 12****23
Sort code: 10-90-40

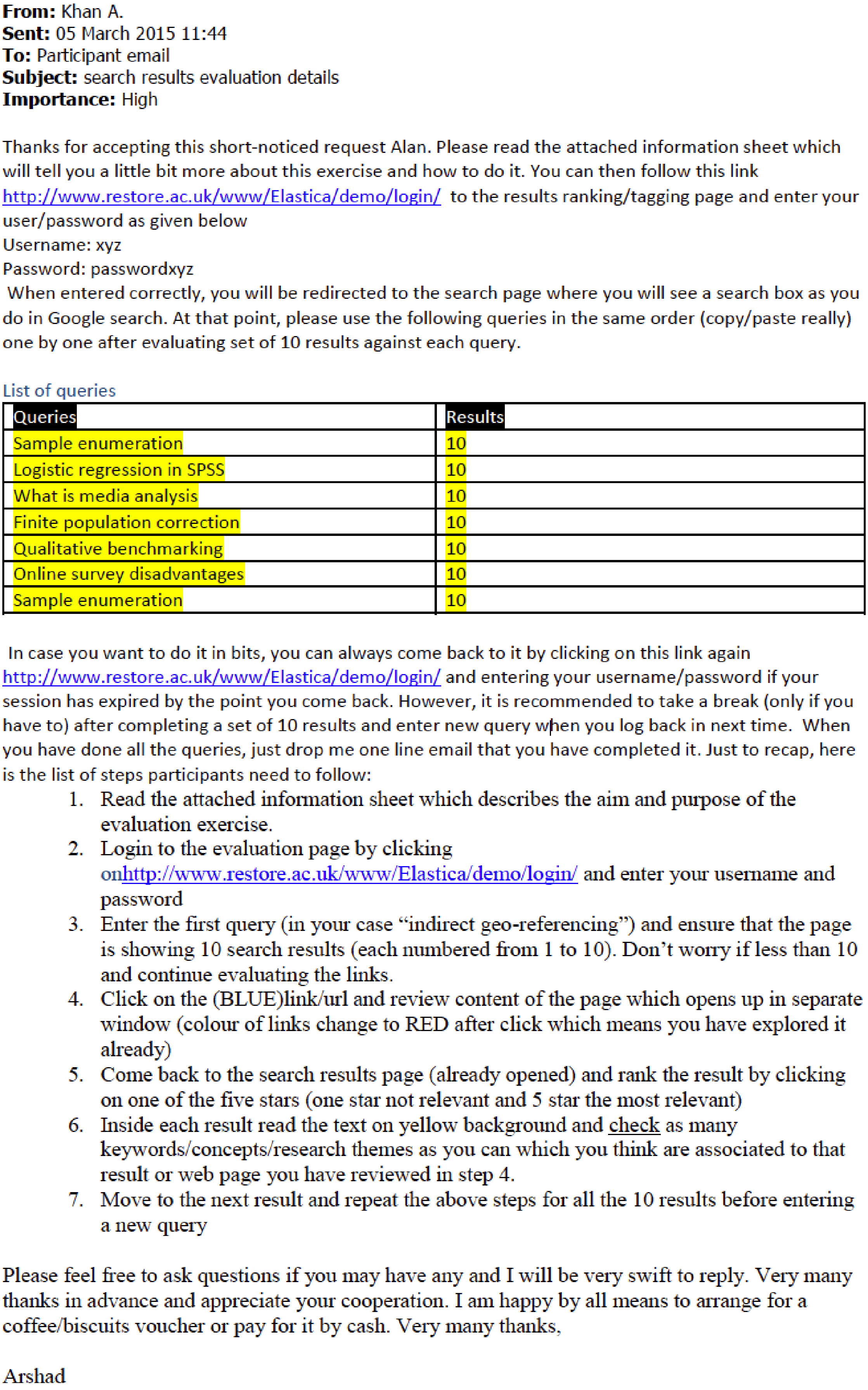
TOTAL DUE £10

If you have any questions concerning this invoice, please contact Arshad Khan, Phone: 02380 59 7750 or email a.khan@soton.ac.uk

THANKS VERY MUCH FOR YOUR PARTICIPATION IN THE STUDY.

PAID

Figure B.8: Anonymised paid invoice sent out for payment by crowd annotator and search results evaluators



From: Khan A.
Sent: 05 March 2015 11:44
To: Participant email
Subject: search results evaluation details
Importance: High

Thanks for accepting this short-noticed request Alan. Please read the attached information sheet which will tell you a little bit more about this exercise and how to do it. You can then follow this link http://www.restore.ac.uk/www/Elastica/demo/login/ to the results ranking/tagging page and enter your user/password as given below
Username: xyz
Password: passwordxyz
When entered correctly, you will be redirected to the search page where you will see a search box as you do in Google search. At that point, please use the following queries in the same order (copy/paste really) one by one after evaluating set of 10 results against each query.

List of queries

Queries	Results
Sample enumeration	10
Logistic regression in SPSS	10
What is media analysis	10
Finite population correction	10
Qualitative benchmarking	10
Online survey disadvantages	10
Sample enumeration	10

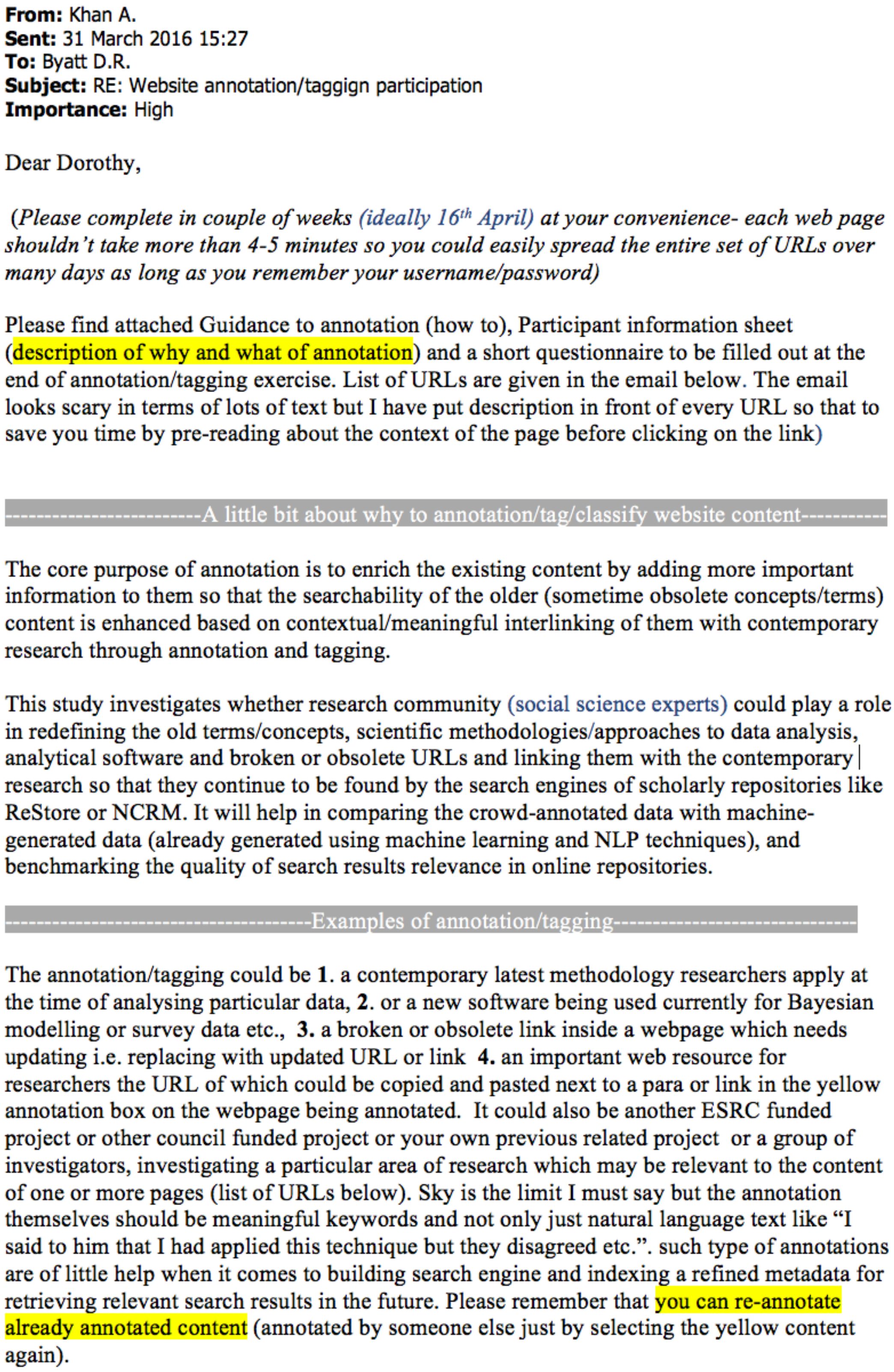
In case you want to do it in bits, you can always come back to it by clicking on this link again http://www.restore.ac.uk/www/Elastica/demo/login/ and entering your username/password if your session has expired by the point you come back. However, it is recommended to take a break (only if you have to) after completing a set of 10 results and enter new query when you log back in next time. When you have done all the queries, just drop me one line email that you have completed it. Just to recap, here is the list of steps participants need to follow:

1. Read the attached information sheet which describes the aim and purpose of the evaluation exercise.
2. Login to the evaluation page by clicking onhttp://www.restore.ac.uk/www/Elastica/demo/login/ and enter your username and password
3. Enter the first query (in your case "indirect geo-referencing") and ensure that the page is showing 10 search results (each numbered from 1 to 10). Don't worry if less than 10 and continue evaluating the links.
4. Click on the (BLUE)link/url and review content of the page which opens up in separate window (colour of links change to RED after click which means you have explored it already)
5. Come back to the search results page (already opened) and rank the result by clicking on one of the five stars (one star not relevant and 5 star the most relevant)
6. Inside each result read the text on yellow background and check as many keywords/concepts/research themes as you can which you think are associated to that result or web page you have reviewed in step 4.
7. Move to the next result and repeat the above steps for all the 10 results before entering a new query

Please feel free to ask questions if you may have any and I will be very swift to reply. Very many thanks in advance and appreciate your cooperation. I am happy by all means to arrange for a coffee/biscuits voucher or pay for it by cash. Very many thanks,

Arshad

Figure B.9: A typical email sent to the participants of Experiment A (6.4) for search results evaluation against a set of queries



From: Khan A.
Sent: 31 March 2016 15:27
To: Byatt D.R.
Subject: RE: Website annotation/taggign participation
Importance: High

Dear Dorothy,

(Please complete in couple of weeks (ideally 16th April) at your convenience- each web page shouldn't take more than 4-5 minutes so you could easily spread the entire set of URLs over many days as long as you remember your username/password)

Please find attached Guidance to annotation (how to), Participant information sheet (description of why and what of annotation) and a short questionnaire to be filled out at the end of annotation/tagging exercise. List of URLs are given in the email below. The email looks scary in terms of lots of text but I have put description in front of every URL so that to save you time by pre-reading about the context of the page before clicking on the link)

------------------------A little bit about why to annotation/tag/classify website content----------

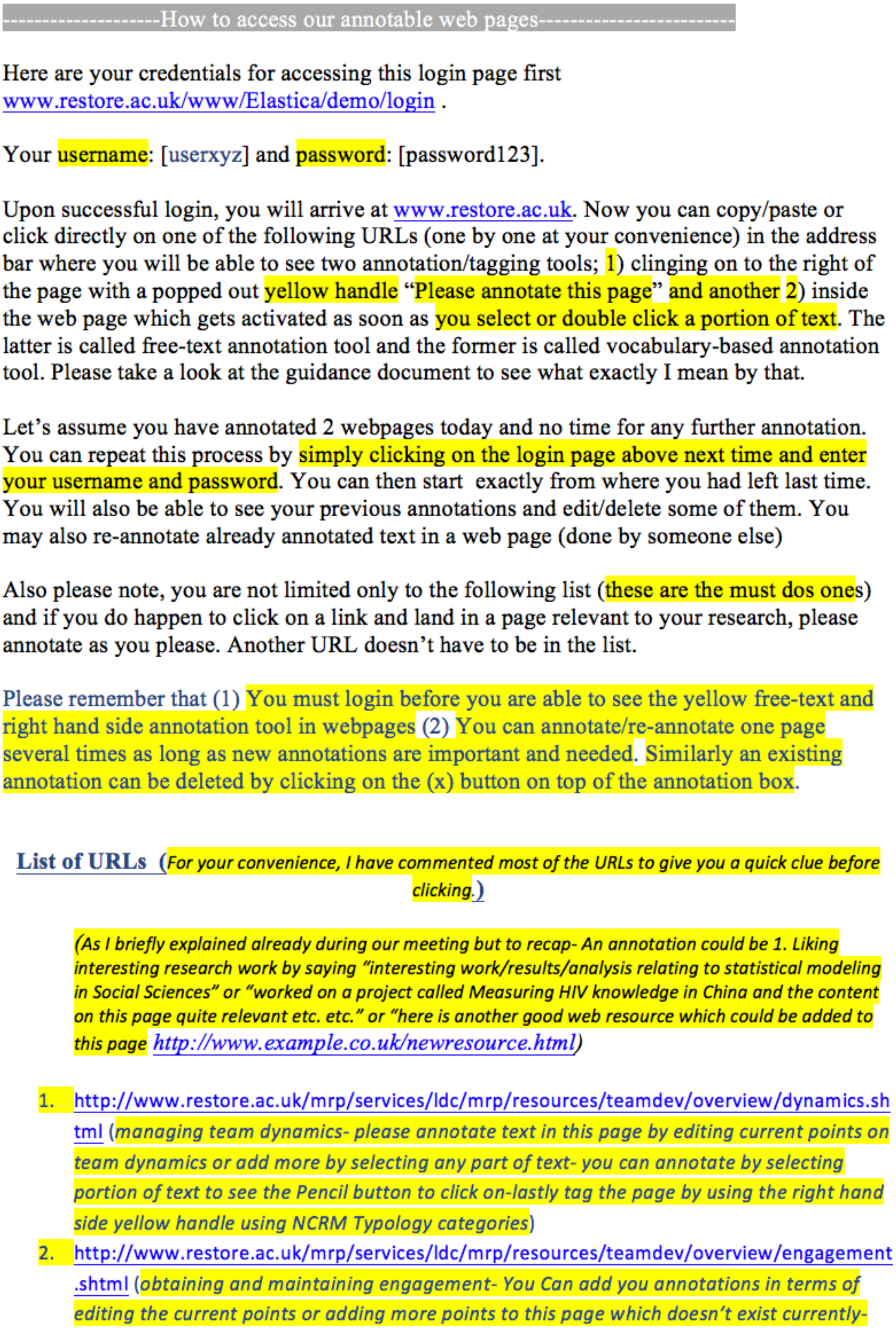
The core purpose of annotation is to enrich the existing content by adding more important information to them so that the searchability of the older (sometime obsolete concepts/terms) content is enhanced based on contextual/meaningful interlinking of them with contemporary research through annotation and tagging.

This study investigates whether research community (social science experts) could play a role in redefining the old terms/concepts, scientific methodologies/approaches to data analysis, analytical software and broken or obsolete URLs and linking them with the contemporary research so that they continue to be found by the search engines of scholarly repositories like ReStore or NCRM. It will help in comparing the crowd-annotated data with machine-generated data (already generated using machine learning and NLP techniques), and benchmarking the quality of search results relevance in online repositories.

--------------------------------------Examples of annotation/tagging------------------------------

The annotation/tagging could be **1**. a contemporary latest methodology researchers apply at the time of analysing particular data, **2**. or a new software being used currently for Bayesian modelling or survey data etc., **3.** a broken or obsolete link inside a webpage which needs updating i.e. replacing with updated URL or link **4.** an important web resource for researchers the URL of which could be copied and pasted next to a para or link in the yellow annotation box on the webpage being annotated. It could also be another ESRC funded project or other council funded project or your own previous related project or a group of investigators, investigating a particular area of research which may be relevant to the content of one or more pages (list of URLs below). Sky is the limit I must say but the annotation themselves should be meaningful keywords and not only just natural language text like "I said to him that I had applied this technique but they disagreed etc.". such type of annotations are of little help when it comes to building search engine and indexing a refined metadata for retrieving relevant search results in the future. Please remember that you can re-annotate already annotated content (annotated by someone else just by selecting the yellow content again).

Figure B.10: A typical crowd-annotation experiment participation email to disciplinary experts- Page: 1/4



-------------------How to access our annotable web pages-----------------------

Here are your credentials for accessing this login page first
www.restore.ac.uk/www/Elastica/demo/login .

Your username: [userxyz] and password: [password123].

Upon successful login, you will arrive at www.restore.ac.uk. Now you can copy/paste or click directly on one of the following URLs (one by one at your convenience) in the address bar where you will be able to see two annotation/tagging tools; 1) clinging on to the right of the page with a popped out yellow handle "Please annotate this page" and another 2) inside the web page which gets activated as soon as you select or double click a portion of text. The latter is called free-text annotation tool and the former is called vocabulary-based annotation tool. Please take a look at the guidance document to see what exactly I mean by that.

Let's assume you have annotated 2 webpages today and no time for any further annotation. You can repeat this process by simply clicking on the login page above next time and enter your username and password. You can then start exactly from where you had left last time. You will also be able to see your previous annotations and edit/delete some of them. You may also re-annotate already annotated text in a web page (done by someone else)

Also please note, you are not limited only to the following list (these are the must dos ones) and if you do happen to click on a link and land in a page relevant to your research, please annotate as you please. Another URL doesn't have to be in the list.

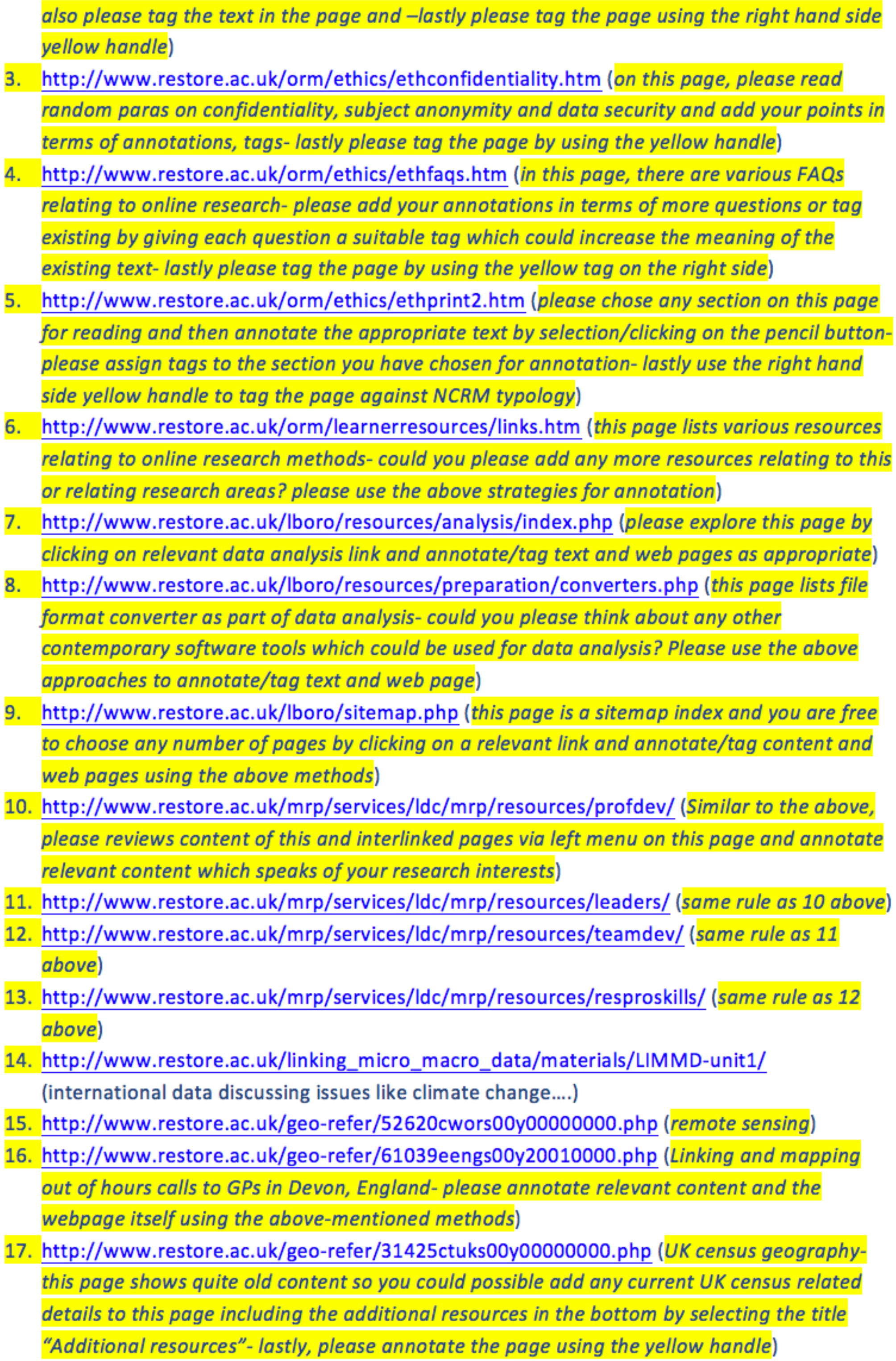
Please remember that (1) You must login before you are able to see the yellow free-text and right hand side annotation tool in webpages (2) You can annotate/re-annotate one page several times as long as new annotations are important and needed. Similarly an existing annotation can be deleted by clicking on the (x) button on top of the annotation box.

List of URLs (*For your convenience, I have commented most of the URLs to give you a quick clue before clicking.*)

(As I briefly explained already during our meeting but to recap- An annotation could be 1. Liking interesting research work by saying "interesting work/results/analysis relating to statistical modeling in Social Sciences" or "worked on a project called Measuring HIV knowledge in China and the content on this page quite relevant etc. etc." or "here is another good web resource which could be added to this page http://www.example.co.uk/newresource.html)

1. http://www.restore.ac.uk/mrp/services/ldc/mrp/resources/teamdev/overview/dynamics.shtml (*managing team dynamics- please annotate text in this page by editing current points on team dynamics or add more by selecting any part of text- you can annotate by selecting portion of text to see the Pencil button to click on-lastly tag the page by using the right hand side yellow handle using NCRM Typology categories*)
2. http://www.restore.ac.uk/mrp/services/ldc/mrp/resources/teamdev/overview/engagement.shtml (*obtaining and maintaining engagement- You Can add you annotations in terms of editing the current points or adding more points to this page which doesn't exist currently-*

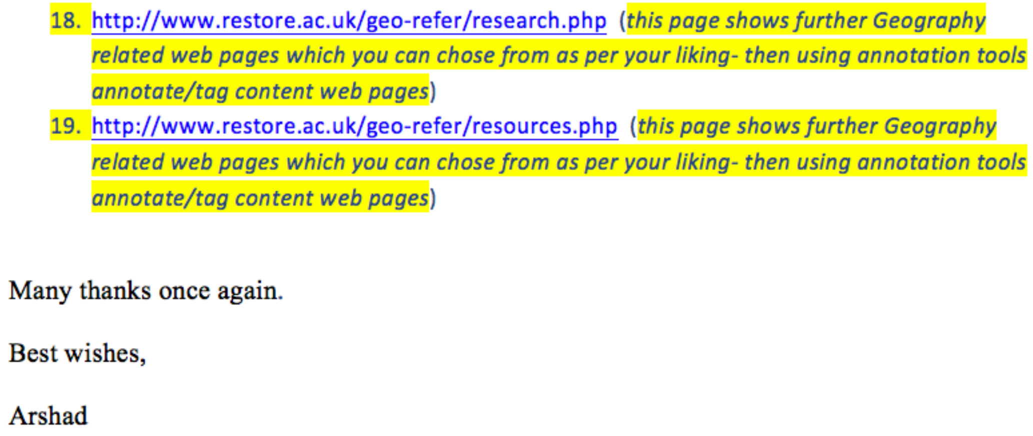
Figure B.11: A typical crowd-annotation experiment participation email to disciplinary experts- Page: 2/4



also please tag the text in the page and –lastly please tag the page using the right hand side yellow handle)

3. http://www.restore.ac.uk/orm/ethics/ethconfidentiality.htm (*on this page, please read random paras on confidentiality, subject anonymity and data security and add your points in terms of annotations, tags- lastly please tag the page by using the yellow handle*)
4. http://www.restore.ac.uk/orm/ethics/ethfaqs.htm (*in this page, there are various FAQs relating to online research- please add your annotations in terms of more questions or tag existing by giving each question a suitable tag which could increase the meaning of the existing text- lastly please tag the page by using the yellow tag on the right side*)
5. http://www.restore.ac.uk/orm/ethics/ethprint2.htm (*please chose any section on this page for reading and then annotate the appropriate text by selection/clicking on the pencil button- please assign tags to the section you have chosen for annotation- lastly use the right hand side yellow handle to tag the page against NCRM typology*)
6. http://www.restore.ac.uk/orm/learnerresources/links.htm (*this page lists various resources relating to online research methods- could you please add any more resources relating to this or relating research areas? please use the above strategies for annotation*)
7. http://www.restore.ac.uk/lboro/resources/analysis/index.php (*please explore this page by clicking on relevant data analysis link and annotate/tag text and web pages as appropriate*)
8. http://www.restore.ac.uk/lboro/resources/preparation/converters.php (*this page lists file format converter as part of data analysis- could you please think about any other contemporary software tools which could be used for data analysis? Please use the above approaches to annotate/tag text and web page*)
9. http://www.restore.ac.uk/lboro/sitemap.php (*this page is a sitemap index and you are free to choose any number of pages by clicking on a relevant link and annotate/tag content and web pages using the above methods*)
10. http://www.restore.ac.uk/mrp/services/ldc/mrp/resources/profdev/ (*Similar to the above, please reviews content of this and interlinked pages via left menu on this page and annotate relevant content which speaks of your research interests*)
11. http://www.restore.ac.uk/mrp/services/ldc/mrp/resources/leaders/ (*same rule as 10 above*)
12. http://www.restore.ac.uk/mrp/services/ldc/mrp/resources/teamdev/ (*same rule as 11 above*)
13. http://www.restore.ac.uk/mrp/services/ldc/mrp/resources/resproskills/ (*same rule as 12 above*)
14. http://www.restore.ac.uk/linking_micro_macro_data/materials/LIMMD-unit1/ (international data discussing issues like climate change….)
15. http://www.restore.ac.uk/geo-refer/52620cwors00y00000000.php (*remote sensing*)
16. http://www.restore.ac.uk/geo-refer/61039eengs00y20010000.php (*Linking and mapping out of hours calls to GPs in Devon, England- please annotate relevant content and the webpage itself using the above-mentioned methods*)
17. http://www.restore.ac.uk/geo-refer/31425ctuks00y00000000.php (*UK census geography- this page shows quite old content so you could possible add any current UK census related details to this page including the additional resources in the bottom by selecting the title "Additional resources"- lastly, please annotate the page using the yellow handle*)

Figure B.12: A typical crowd-annotation experiment participation email to disciplinary experts- Page: 3/4



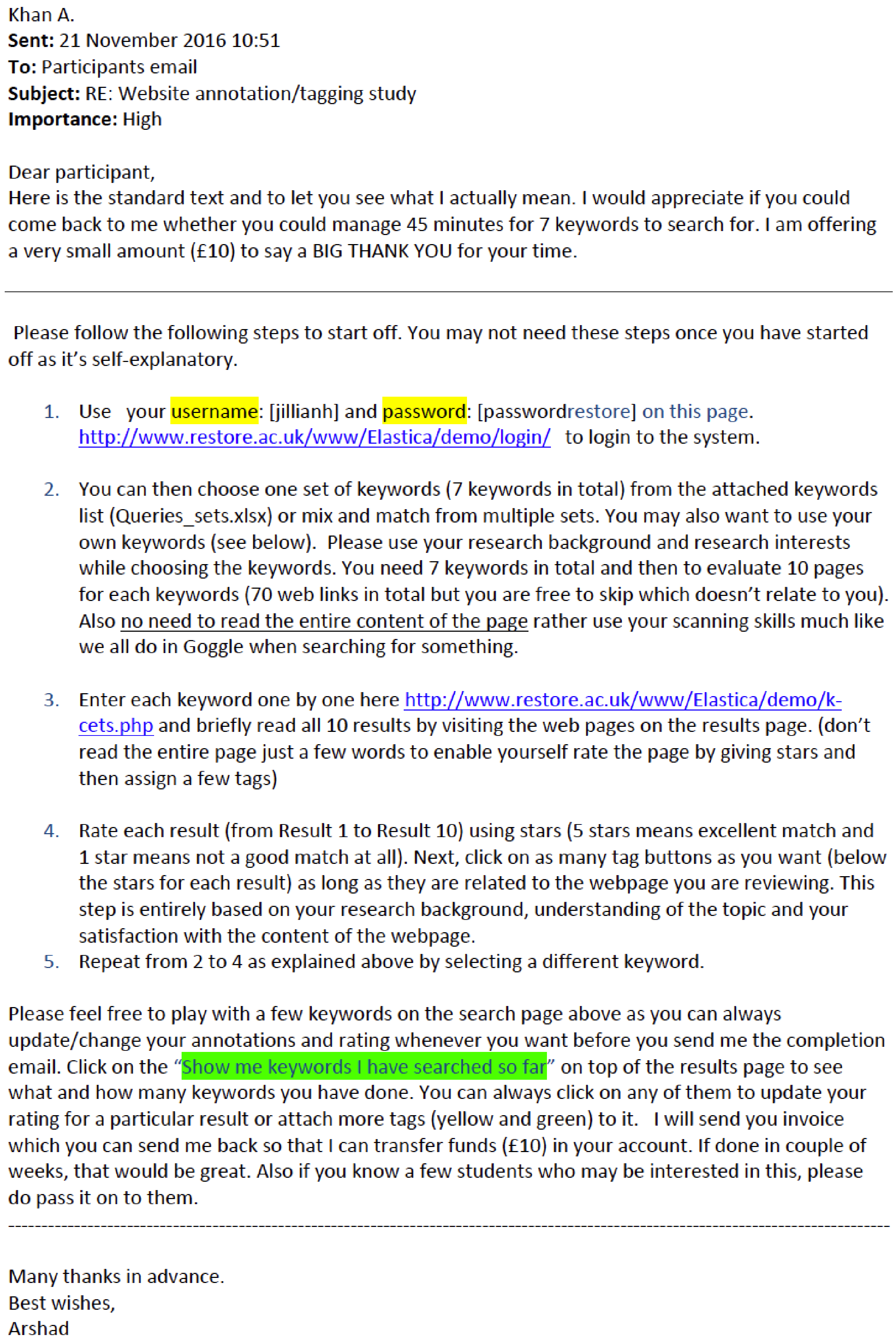
18. http://www.restore.ac.uk/geo-refer/research.php (*this page shows further Geography related web pages which you can chose from as per your liking- then using annotation tools annotate/tag content web pages*)
19. http://www.restore.ac.uk/geo-refer/resources.php (*this page shows further Geography related web pages which you can chose from as per your liking- then using annotation tools annotate/tag content web pages*)

Many thanks once again.

Best wishes,

Arshad

Figure B.13: A typical crowd-annotation experiment participation email to disciplinary experts- Page: 4/4



Khan A.
Sent: 21 November 2016 10:51
To: Participants email
Subject: RE: Website annotation/tagging study
Importance: High

Dear participant,
Here is the standard text and to let you see what I actually mean. I would appreciate if you could come back to me whether you could manage 45 minutes for 7 keywords to search for. I am offering a very small amount (£10) to say a BIG THANK YOU for your time.

Please follow the following steps to start off. You may not need these steps once you have started off as it's self-explanatory.

1. Use your username: [jillianh] and password: [passwordrestore] on this page. http://www.restore.ac.uk/www/Elastica/demo/login/ to login to the system.

2. You can then choose one set of keywords (7 keywords in total) from the attached keywords list (Queries_sets.xlsx) or mix and match from multiple sets. You may also want to use your own keywords (see below). Please use your research background and research interests while choosing the keywords. You need 7 keywords in total and then to evaluate 10 pages for each keywords (70 web links in total but you are free to skip which doesn't relate to you). Also no need to read the entire content of the page rather use your scanning skills much like we all do in Goggle when searching for something.

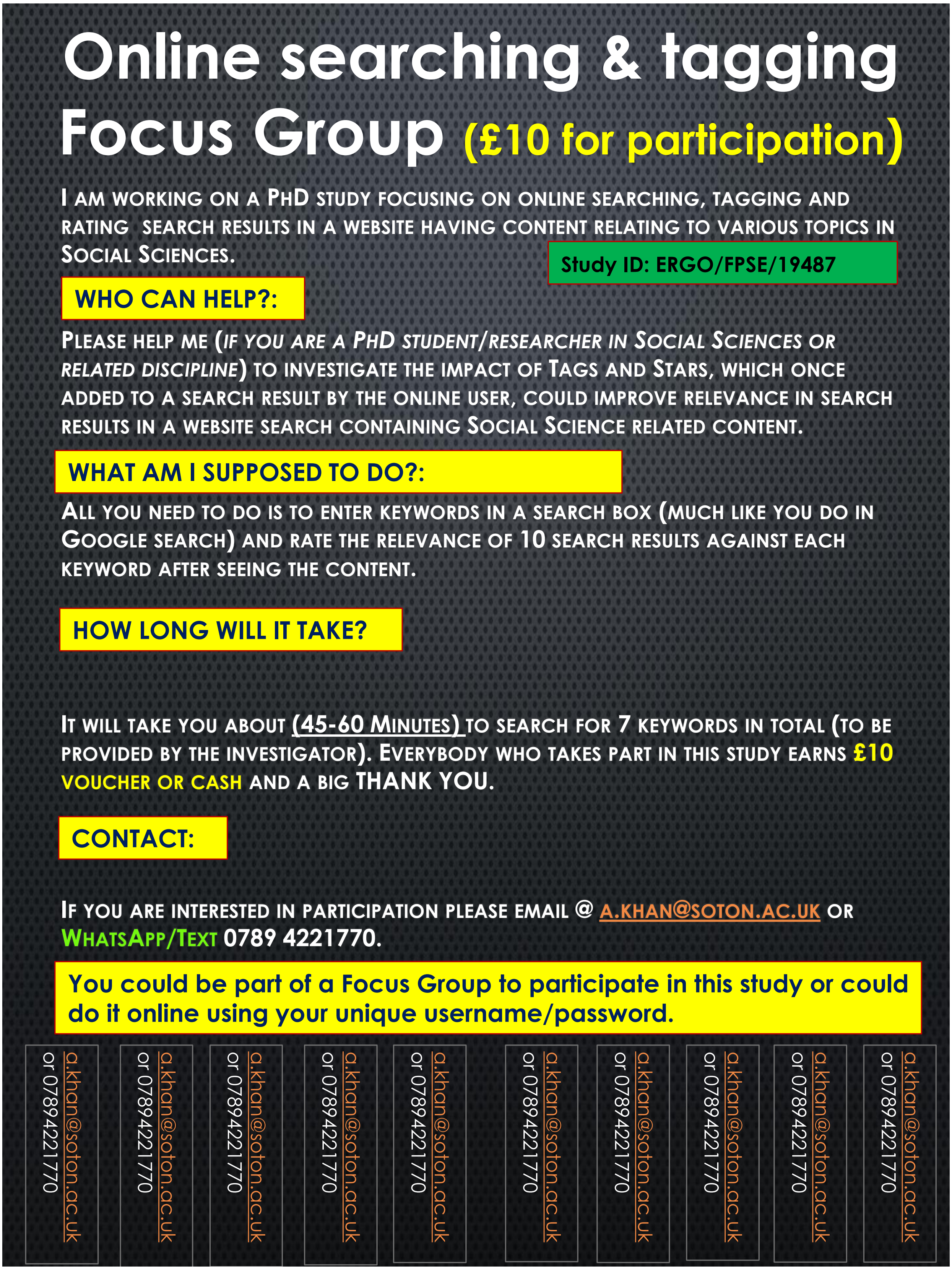
3. Enter each keyword one by one here http://www.restore.ac.uk/www/Elastica/demo/k-cets.php and briefly read all 10 results by visiting the web pages on the results page. (don't read the entire page just a few words to enable yourself rate the page by giving stars and then assign a few tags)

4. Rate each result (from Result 1 to Result 10) using stars (5 stars means excellent match and 1 star means not a good match at all). Next, click on as many tag buttons as you want (below the stars for each result) as long as they are related to the webpage you are reviewing. This step is entirely based on your research background, understanding of the topic and your satisfaction with the content of the webpage.
5. Repeat from 2 to 4 as explained above by selecting a different keyword.

Please feel free to play with a few keywords on the search page above as you can always update/change your annotations and rating whenever you want before you send me the completion email. Click on the "Show me keywords I have searched so far" on top of the results page to see what and how many keywords you have done. You can always click on any of them to update your rating for a particular result or attach more tags (yellow and green) to it. I will send you invoice which you can send me back so that I can transfer funds (£10) in your account. If done in couple of weeks, that would be great. Also if you know a few students who may be interested in this, please do pass it on to them.

Many thanks in advance.
Best wishes,
Arshad

Figure B.14: A typical email sent to the participants of final Experiment C (6.6) for search results evaluation against a set of 7 queries



Online searching & tagging Focus Group (£10 for participation)

I am working on a PhD study focusing on online searching, tagging and rating search results in a website having content relating to various topics in Social Sciences.

Study ID: ERGO/FPSE/19487

WHO CAN HELP?:

Please help me (*if you are a PhD student/researcher in Social Sciences or related discipline*) to investigate the impact of Tags and Stars, which once added to a search result by the online user, could improve relevance in search results in a website search containing Social Science related content.

WHAT AM I SUPPOSED TO DO?:

All you need to do is to enter keywords in a search box (much like you do in Google search) and rate the relevance of 10 search results against each keyword after seeing the content.

HOW LONG WILL IT TAKE?

It will take you about (45-60 Minutes) to search for 7 keywords in total (to be provided by the investigator). Everybody who takes part in this study earns £10 voucher or cash and a big THANK YOU.

CONTACT:

If you are interested in participation please email @ a.khan@soton.ac.uk or WhatsApp/Text 0789 4221770.

You could be part of a Focus Group to participate in this study or could do it online using your unique username/password.

a.khan@soton.ac.uk or 07894221770

a.khan@soton.ac.uk or 07894221770

a.khan@soton.ac.uk or 07894221770

a.khan@soton.ac.uk or 07894221770

a.khan@soton.ac.uk or 07894221770

a.khan@soton.ac.uk or 07894221770

a.khan@soton.ac.uk or 07894221770

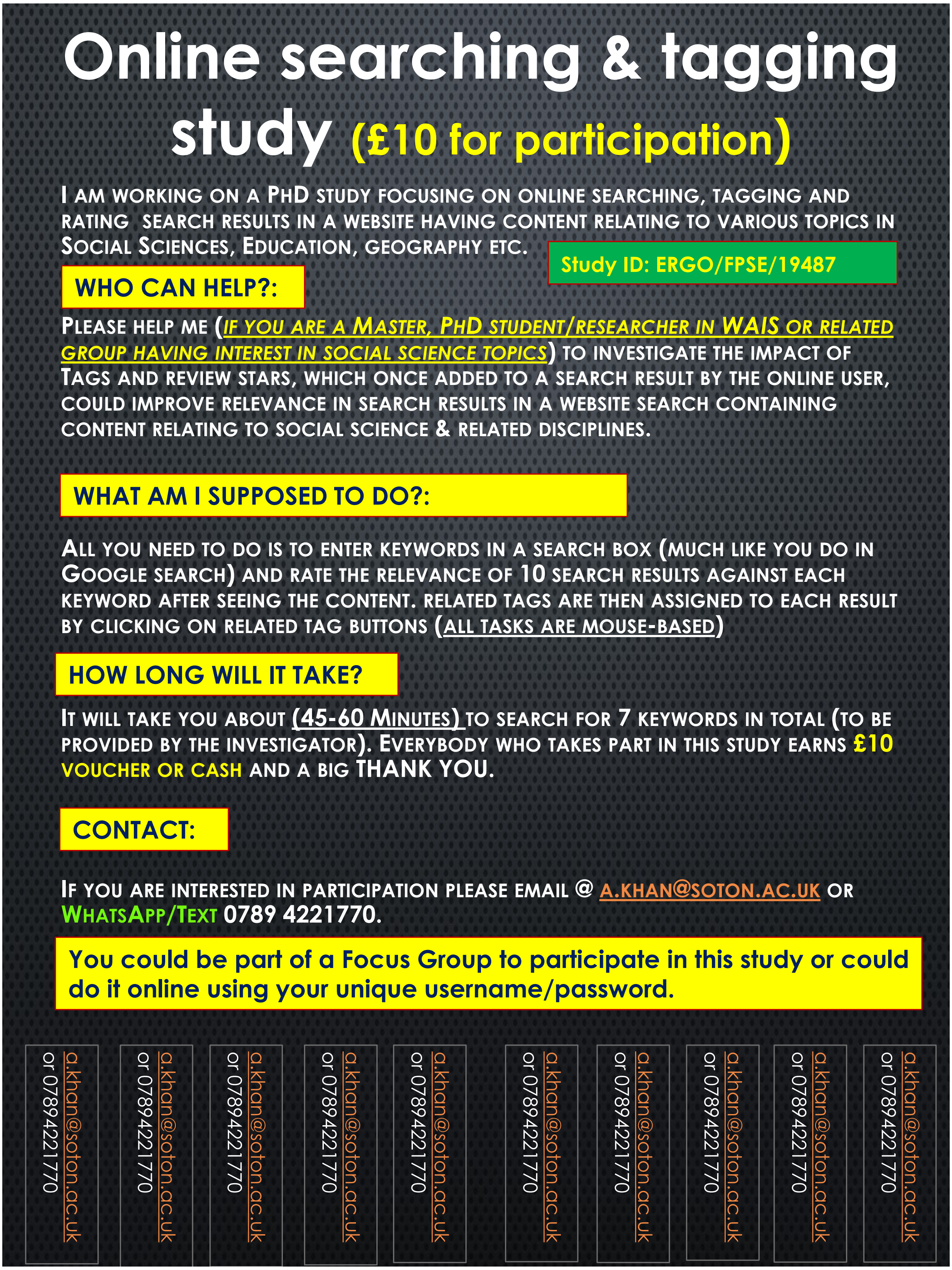
a.khan@soton.ac.uk or 07894221770

a.khan@soton.ac.uk or 07894221770

a.khan@soton.ac.uk or 07894221770

Figure B.15: Advertisement for the participation of search results evaluators in the focus group as part of final Experiment C (6.6)

Furthermore, a webpage with the embedded training video and description was created and shared with potential participants to describe the scope and purpose of annotation and tagging in web repositories. The webpage can be accessed at `http://www.restore.ac.uk/FocusGroup/`.



Online searching & tagging study (£10 for participation)

I am working on a PhD study focusing on online searching, tagging and rating search results in a website having content relating to various topics in Social Sciences, Education, geography etc.

Study ID: ERGO/FPSE/19487

WHO CAN HELP?:

Please help me (*If you are a Master, PhD student/researcher in WAIS or related group having interest in social science topics*) to investigate the impact of Tags and review stars, which once added to a search result by the online user, could improve relevance in search results in a website search containing content relating to social science & related disciplines.

WHAT AM I SUPPOSED TO DO?:

All you need to do is to enter keywords in a search box (much like you do in Google search) and rate the relevance of 10 search results against each keyword after seeing the content. Related tags are then assigned to each result by clicking on related tag buttons (all tasks are mouse-based)

HOW LONG WILL IT TAKE?

It will take you about (45-60 Minutes) to search for 7 keywords in total (to be provided by the investigator). Everybody who takes part in this study earns £10 voucher or cash and a big THANK YOU.

CONTACT:

If you are interested in participation please email @ a.khan@soton.ac.uk or WhatsApp/Text 0789 4221770.

You could be part of a Focus Group to participate in this study or could do it online using your unique username/password.

a.khan@soton.ac.uk
or 07894221770

a.khan@soton.ac.uk
or 07894221770

a.khan@soton.ac.uk
or 07894221770

a.khan@soton.ac.uk
or 07894221770

a.khan@soton.ac.uk
or 07894221770

a.khan@soton.ac.uk
or 07894221770

a.khan@soton.ac.uk
or 07894221770

a.khan@soton.ac.uk
or 07894221770

a.khan@soton.ac.uk
or 07894221770

a.khan@soton.ac.uk
or 07894221770

Figure B.16: Advertisement for the participation of search results evaluators in the online annotation/tagging study (non-focus group) as part of final Experiment C (6.6)

Appendix C

Documents annotation, tagging and indexing in Elasticsearch KB

C.1 Algorithms

C.1.1 Algorithm: Three stages of crowd-annotation

Algorithm 2 describes the three stages of crowd-annotation (content level and web page-level) i.e. *annotator authentication, content level and page level annotations, adding/updating of annotation to ES index.* Both content-level and page-level annotation have been explained in Algorithm 4

Algorithm 2 Updating the automatically indexed document with crowd-sourced annotation using Annotator API and Elasticsearch instance

```
Input: URL, AnnotatorID, AnnotatorConsumerKey, AnnotatorSecretKey
Select piece of text to annotation
annotatorID ← AnnotatorID
initialize Annotator API
objAnnotator= new Annotator(AnnotatorConsumerKey, AnnotatorSecretKey, AnnotatorID )
addAnnotation (objAnnotator, annotation, piceofSelectedText, freeTextTags, vocabularyTags)
....continue adding annotation on various other pages....

  Iterating through all annotations with respect to unique consumer key

  ObjArray[]= getAllannotation (AnnotatorConsumerKey)
Elastic myindex=Obj1.getIndex(myAnnotationIndex)
            foreach indexedDocument in myindex
                Do
                compare myindex['URL'] with ObjArray['URL']
                if matched (URL, URL)
                update myindex (ObjArray (array (annotation, annotator, vocabularyTags, free-
  textTags, sourceText)))
                total++
                while total < count (objArray)
            end for
Output: new Document added successfully to:myIndexA
```

C.1.2 Algorithm: Text analysis, entities extraction and indexing of data

The following algorithm (Algorithm 3) sums up the text analysis, entities extraction and indexing of data (from RDBMS (Refer to Figure 4.5) source in the ReStore repository) using Alchemy API and Elasticsearch.

Algorithm 3 Semantic entities extraction & storage using Alchemy API & Elasticsearch

```
Input: Content in DB for dynamic web pages
Load Elasticsearch (ES) libraries
ESObject instance
Set pointers to hyper index and type in ES
Establish connection with DBMS
Fetch selected fields for annotation
        Foreach selected fields as myFields
            cleanString(myFields)
            arrayKeywords()=AlchemyKeywordsAPI (myFields, sentiment(positive))
            arrayConcepts()=AlchemyConceptsAPI (myFields, sentiment(positive))
            arrayEntities()=AlchemyEntitiesAPI (myFields, sentiment(positive))
                foreach arrayKeywords as onlyKeywords
                keywordsText[]=onlyKeywords[text]
                relevance[]=onlyKeywords[relevance]
                end for
                foreach arrayConcepts as onlyConcepts
                conceptsText[]=onlyConcepts [text]
                relevance[]=onlyConcepts [relevance]
                end for
                foreach arrayEntities as onlyEntities
                entitiesText[]=onlyEntities [text]
                EntityType[]=onlyEntities [type]
                relevance[]=onlyEntities [relevance]
                end for
            End For
ESObject document[] = new document ()
document.addDocument ( documentID,myFields[title], myFields[fullText],
onlyKeywords[keywordsText],onlyKeywords[relevance],onlyConcepts[conceptsText], onlyConcepts[relevance],onlyEntities[entitiesText],onlyEntities[EntityType],onlyEntities[relevance],
index_date, version )

Output: Keywords, entities, concepts extracted and stored in Elasticsearch index
```

As shown in Algorithm 3, after having loaded the Elasticsearch instance (*ESObject*) and set pointers to the index, requests to the database are made iteratively for text analysis by the Alchemy API object. Keywords, Entities and Concepts APIs array objects store the JSON output from *Title* and *Content* fields in each textual document. All array objects are passed on to the Elasticsearch's *Document* object in order to be written to the index. The document is finally stored and indexed along with full text and semantic metadata ready to be searched via the web-based search application. Screen shots of the actual code have also been shown in Figures B.1, B.2 and B.3.

C.1.3 Algorithm: *Content-level free-text annotation* and *entire webpage-level tagging*

Algorithm 4 describes the entire annotation tasks in two steps. An expert crowd-sourced annotator logs into the annotation and tagging system and searches or browses for an annotatable page. After having read or scanned the content of the page partially or fully, he/she has to select a piece of text, caption of image or other objects in a typical web page which pops open the annotation/comment window (see Figure 6.4 in Chapter 6)to be filled in with potential annotation/tags. An annotation associated to a particular piece of text inside a web page can be edited or deleted by the creator annotator having a unique id. Similarly, the pop up window enables annotators to access popular typology tags and free-text tags in an auto-complete fashion. The typing-based auto-complete has been implemented in both the in-page textual annotation pop-up window as well as the entire webpage-level tagging slider.

Algorithm 4 *Crowd-annotation task*: Adding /Editing content level and webpage-level annotation and tags by expert annotators

```
Foreach (Webpages as URL)
    Input: URL, annotator
    FreeText annotation

 Content level freetext annotation

    annotatorID←currentAnnotator
  Step1: Add new annotation
    pieceofText= Select (pieceofText, WebPage)
    addAnnotation (pieceofText, annotatorID)
    popularTags= clusterofTags(typologyTags, freeTextTags)
    addPopularityTags (pieceofText, annotatorID, popularTags)
    commit()
  Step1A: Edit annotation
    pieceofText= Select (pieceofText, WebPage)
    editAnnotation (pieceofText, annotatorID)
    popularTags= clusterofTags(typologyTags, freeTextTags)
    editPopularityTags (pieceofText, annotatorID, popularTags)
    commit()

 Entire webpage-level tagging

  Step2: Add webpage-level Tags
    input←(URL, annotator)
selectTypology&FreeTextTags[ ]=FindrepresentativeTags(URL,typologyTags,freeTextTags)
    addAllTags (WebPage, selectTypology&FreeTextTags, annotatorID)
    commit()
End For
Output: New annotations and tags (content level and webpage-level) have been
        added/edited to :myIndexA
```

C.2 Examples:*entity type* , *crowd-annotation types* and *concepts* based search results retrieval

For example to find relevant documents having entities with a user-defined type (University of Newcastle of type Organization), the query would be formulated as:

```
GET index_SemDex/keywords /_search?pretty=true
{"query": {"nested" : {
       "path" : "entitiesandtypeandrelevance",
       "score_mode" : "avg",
       "query" : {"bool" : { [#comment:bool=boolean
   condition with a "must" attribute]
          "must" : [
            {"match" : {"entitiesandtypeandrelevance.
   entity" : "University of Newcastle"}
            },{
             "term" : {"entitiesandtypeandrelevance.type"
    : "Organization"}
              }
               ] } } } } }
```

Listing C.1: Named Entity-based retrieval using *type:organization*

Likewise, the NoSQL query for retrieving results based on entity of type *Person* is given as:

```
GET index_SemDex/_search?pretty
{"query": {"filtered": {
   "query": {"match_all_fields": {}},
    "filters": [
         {"filter1_term": {
      "entities.entity": "Paul Boyle"
   }},{"filter2_term": {
      "entities.type": "Person"
   }} ] } }
```

Listing C.2: Named Entity-based retrieval using *type:person*

Similarly to retrieve relevant web documents having concepts called “demography” and relevance score falling between 0.5 and 0.9, the query would be formulated as follows:

```
GET index_SemDex/_search?pretty=true
{"fields": [
   "conceptsandrelevance.concepts","conceptsandrelevance.
   relevance"
], "query": {"nested": {
   "path_name": "conceptsandrelevance",
       "score_mode" : "avg",
   "query": {"bool": {"must": [{
```

```
                        "match" : {"conceptsandrelevance.
   concepts" : "Demography"}
                    },
                   {
  "score_range" : {
    "conceptsandrelevance.relevance" : {
        "lte" : "0.9", [#comment:lte=less than equal to]
       "gte" : "0.5"  [#comment:gte=greater than equal to
   ]
                }} } ]}} }}}
```

Listing C.3: Concepts-based retrieval using disciplinary *concepts*

Example of ES query, which retrieves results based on relevance score computation incorporating 3 types of fields i.e. full-text (title, content), semantic (keywords, entities, concepts) and crowd-annotation (typology and non-typology fields). The query in Listing C.4 fetches the top 10 results using the indexed fields in *SemDex* as well as *SemCrowDex*. The crowd-annotated fields in this listing have a nested path accessible via *crowdAnnotation* and *vocabularyAnnotation* prefixes.

```
GET index_SemCrowDex/_search?pretty&size=50
{"explain": true,  "query": {
"bool": {"should": [ [#comment:bool=boolean condition
   with a "should" attribute]
   {"multi_match": {
      "query": "cross-national comparison",
      "fields": [ "title","allkeywords.keywords", "
   allconcepts.concepts", "allentities.entity"]
   }},{
 "nested": { [#comment:nested field objects in an ES
   document, is used as part of a condition]
   "path": "crowdAnnotation",
   "query": {"multi_match": {
      "query": "cross-national comparison",
      "fields": ["crowdAnnotation.annotatedText", "
   crowdAnnotation.alltags.tags"]
   } }}},
   {"nested": {
   "path": "vocabularyAnnotation",
   "query": {"multi_match": {
      "query": "cross-national comparison",
      "fields": ["vocabularyAnnotation.typology_level1.
   level1", "vocabularyAnnotation.typology_level1.level1",
    "vocabularyAnnotation.typology_level2.level2", "
   vocabularyAnnotation.allinOne.tags"]
    }
```

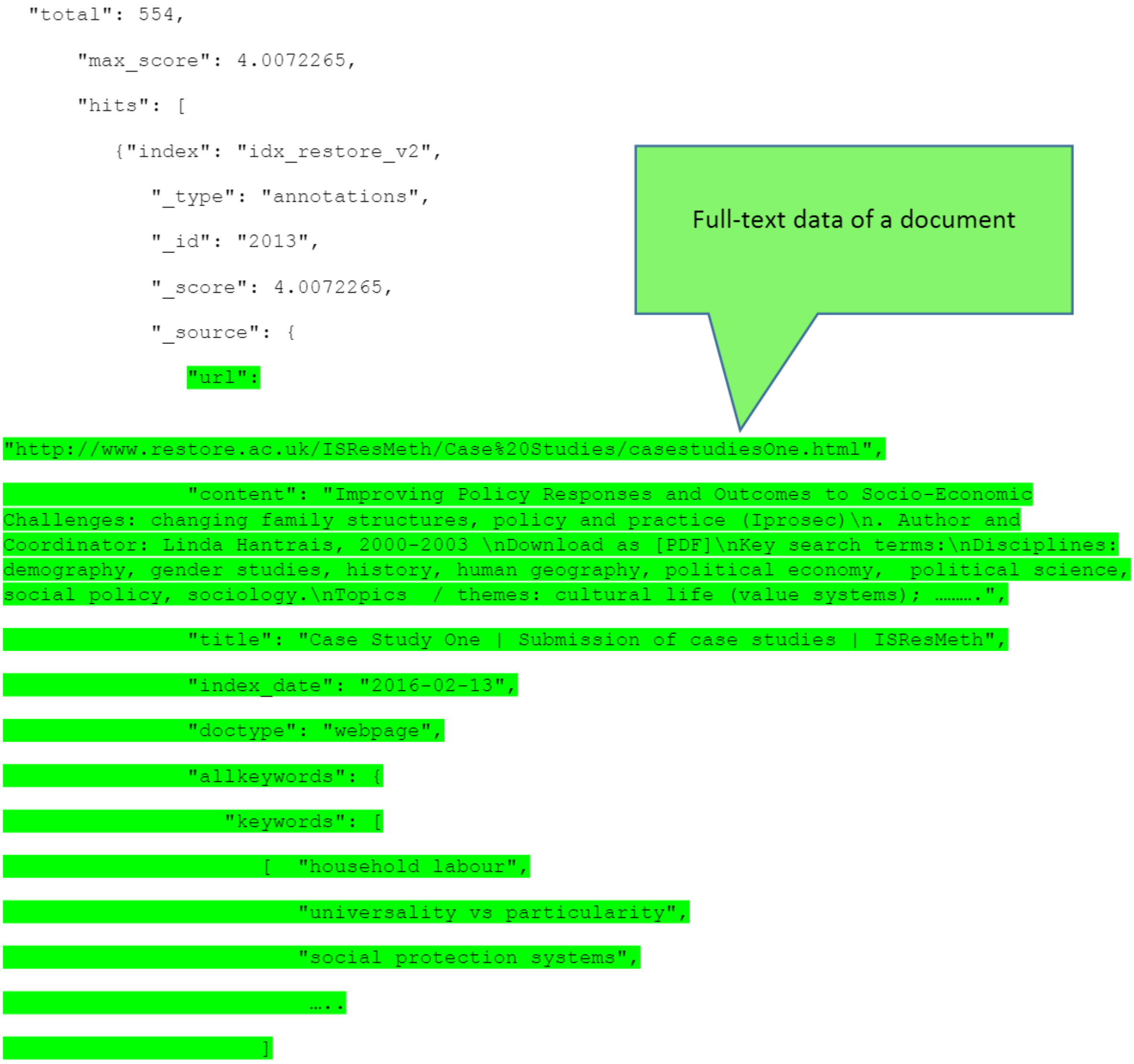
Listing C.4: Query against full-text, semantic annotation and crowd annotation fields combined

C.2.1 Faceted data visualisation and retrieval

```
#Retrieve all results where semantic concept is "General
   Household Survey" along with aggregated facet of top 10
    significant concepts
GET index_SemCrowDex/_search?pretty
{"query": {"match": {
   "concepts.type": "General Household Survey"
}}, "facets": {
   "types": {
      "terms": {
         "field": "allconcepts.facet",
         "size": 10
      } } } }
```

Listing C.5: Filtered search results retrieval using a particular semantic concept i.e. `General Household Survey`

C.3 Sample of ES document stored in ES hybrid index

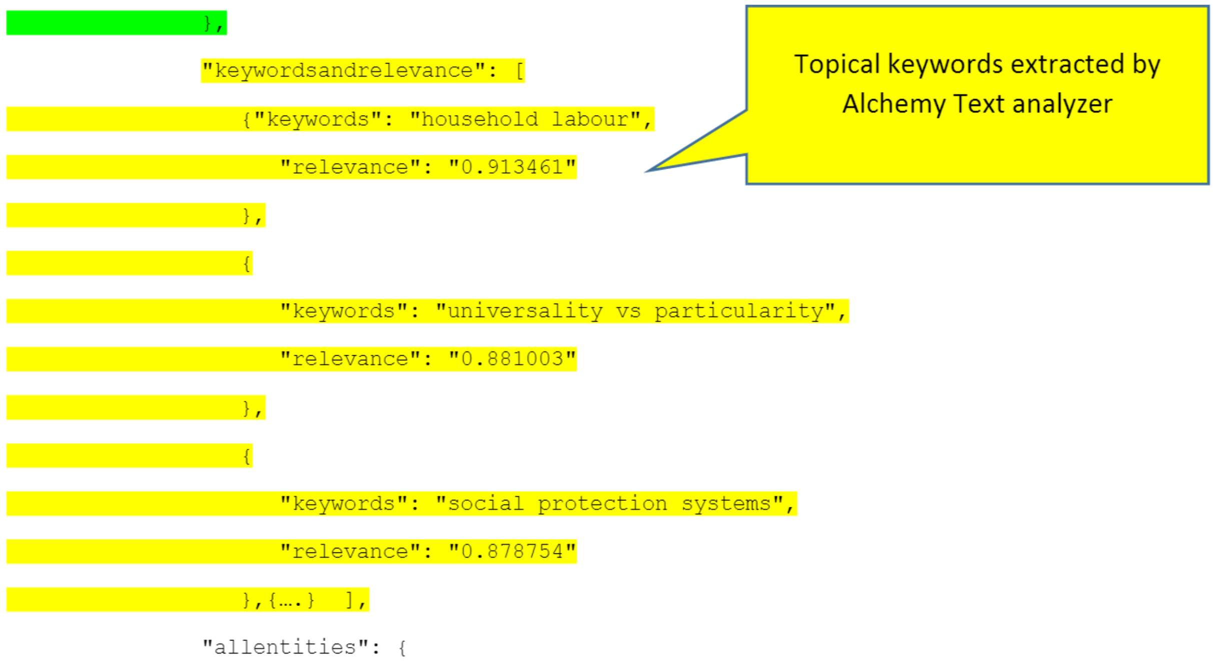


```
    "total": 554,

        "max_score": 4.0072265,

        "hits": [

           {"index": "idx_restore_v2",

              "_type": "annotations",

              "_id": "2013",

              "_score": 4.0072265,

              "_source": {

                 "url":

"http://www.restore.ac.uk/ISResMeth/Case%20Studies/casestudiesOne.html",

               "content": "Improving Policy Responses and Outcomes to Socio-Economic
Challenges: changing family structures, policy and practice (Iprosec)\n. Author and
Coordinator: Linda Hantrais, 2000-2003 \nDownload as [PDF]\nKey search terms:\nDisciplines:
demography, gender studies, history, human geography, political economy,  political science,
social policy, sociology.\nTopics  / themes: cultural life (value systems); ……….",

               "title": "Case Study One | Submission of case studies | ISResMeth",

               "index_date": "2016-02-13",

               "doctype": "webpage",

               "allkeywords": {

                  "keywords": [

                     [  "household labour",

                        "universality vs particularity",

                        "social protection systems",

                        …..

                     ]
```

Full-text data of a document

Figure C.1: Part 1/6 (Full-text part): A semantically indexed document in *SemCrowDex* and a part of hybrid vector space stored in ES KB retrievable by SRR system against a user's query

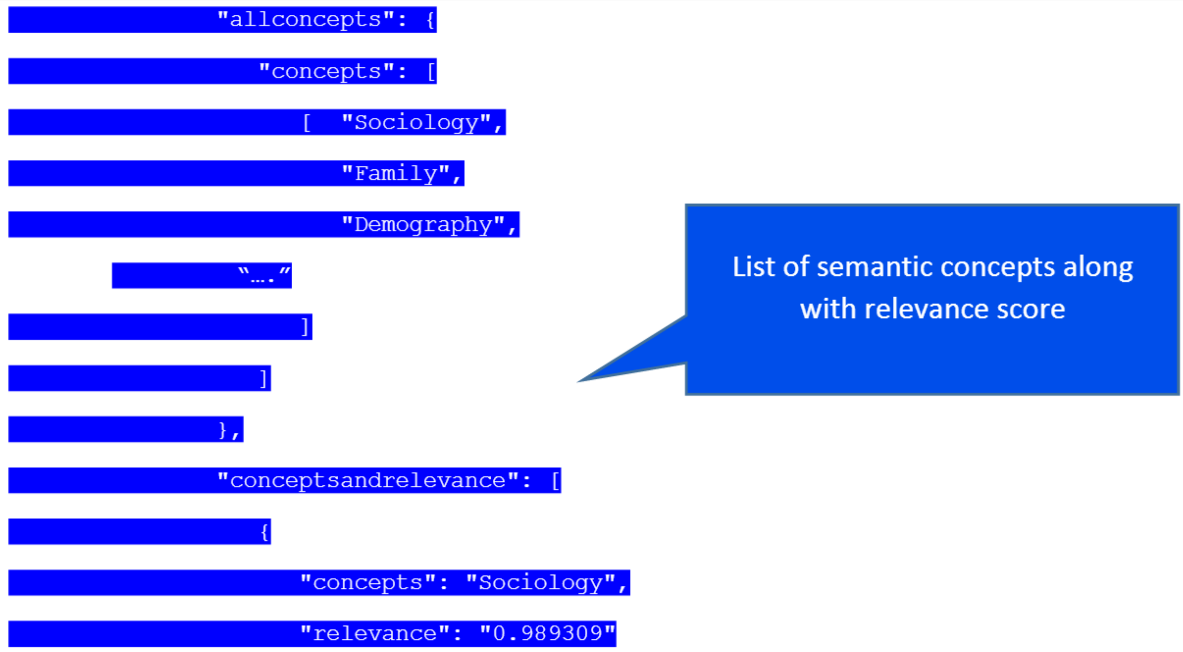
Figure C.1 shows a document in Elasticsearch KB having automatic semantic annotation and crowd-sourced semantic annotations along with individual relevance score. The document comprises of 6 parts shown in the following figures.



```
               },
               "keywordsandrelevance": [
                  {"keywords": "household labour",
                     "relevance": "0.913461"
                  },
                  {
                     "keywords": "universality vs particularity",
                     "relevance": "0.881003"
                  },
                  {
                     "keywords": "social protection systems",
                     "relevance": "0.878754"
                  },{….}  ],
               "allentities": {
```

Topical keywords extracted by Alchemy Text analyzer

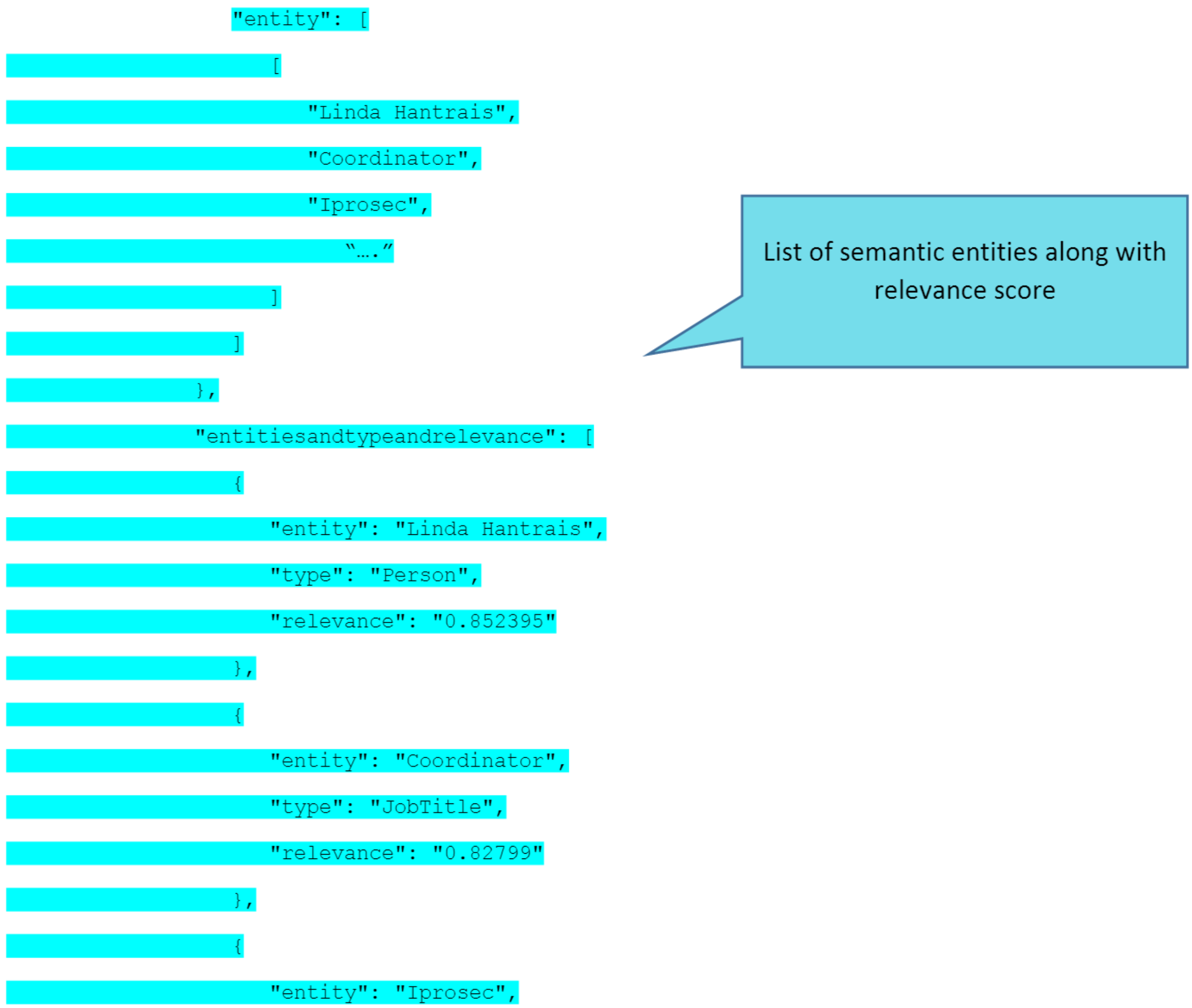
Figure C.2: Part 2/6 (Keywords part): A semantically indexed document in *SemCrowDex* and a part of hybrid vector space stored in Elasticsearch KB retrievable by SRR system against a user's query



```
"allconcepts": {
   "concepts": [
      [  "Sociology",
         "Family",
         "Demography",
 "….”
      ]
   ]
},
"conceptsandrelevance": [
   {
      "concepts": "Sociology",
      "relevance": "0.989309"
```

List of semantic concepts along with relevance score

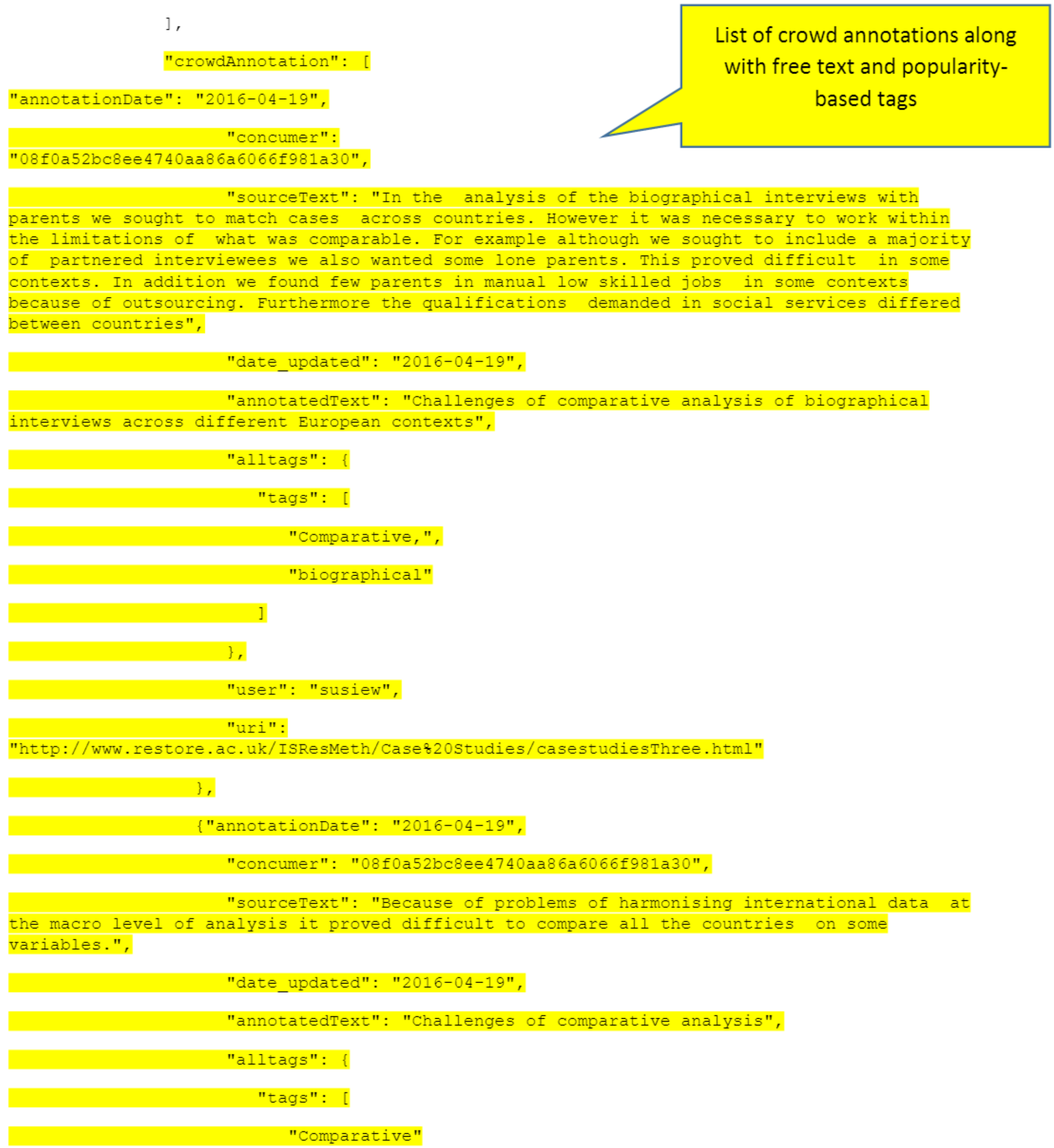
Figure C.3: Part 3/6 (Concepts part): A semantically indexed document in *SemCrowDex* and a part of hybrid vector space stored in Elasticsearch KB retrievable by SRR system against a user's query



```
   "entity": [
      [
         "Linda Hantrais",
         "Coordinator",
         "Iprosec",
            "…."
      ]
   ]
},
"entitiesandtypeandrelevance": [
   {
      "entity": "Linda Hantrais",
      "type": "Person",
      "relevance": "0.852395"
   },
   {
      "entity": "Coordinator",
      "type": "JobTitle",
      "relevance": "0.82799"
   },
   {
      "entity": "Iprosec",
```

List of semantic entities along with relevance score

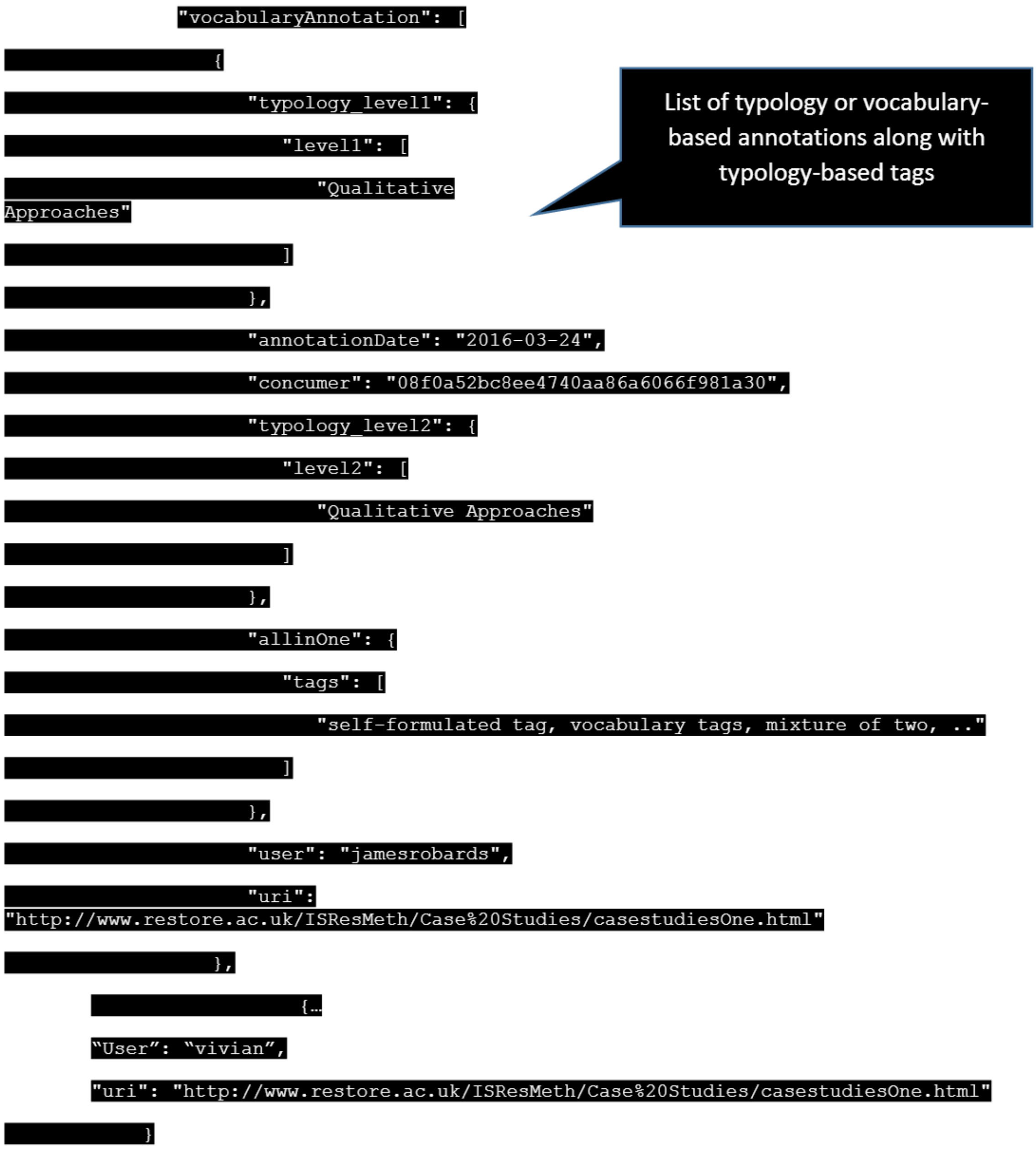
Figure C.4: Part 4/6 (Entities part): A semantically indexed document in *SemCrowDex* and a part of hybrid vector space stored in Elasticsearch KB retrievable by SRR system against a user's query



```
                ],
                "crowdAnnotation": [
"annotationDate": "2016-04-19",
                      "concumer":
"08f0a52bc8ee4740aa86a6066f981a30",
                      "sourceText": "In the  analysis of the biographical interviews with
parents we sought to match cases  across countries. However it was necessary to work within
the limitations of  what was comparable. For example although we sought to include a majority
of  partnered interviewees we also wanted some lone parents. This proved difficult  in some
contexts. In addition we found few parents in manual low skilled jobs  in some contexts
because of outsourcing. Furthermore the qualifications  demanded in social services differed
between countries",
                      "date_updated": "2016-04-19",
                      "annotatedText": "Challenges of comparative analysis of biographical
interviews across different European contexts",
                      "alltags": {
                         "tags": [
                            "Comparative,",
                            "biographical"
                         ]
                      },
                      "user": "susiew",
                      "uri":
"http://www.restore.ac.uk/ISResMeth/Case%20Studies/casestudiesThree.html"
                   },
                   {"annotationDate": "2016-04-19",
                      "concumer": "08f0a52bc8ee4740aa86a6066f981a30",
                      "sourceText": "Because of problems of harmonising international data  at
the macro level of analysis it proved difficult to compare all the countries  on some
variables.",
                      "date_updated": "2016-04-19",
                      "annotatedText": "Challenges of comparative analysis",
                      "alltags": {
                         "tags": [
                            "Comparative"
```

List of crowd annotations along with free text and popularity-based tags

Figure C.5: Part 5/6 (Crowd-annotation part): A semantically indexed document in *SemCrowDex* and a part of hybrid vector space stored in Elasticsearch KB retrievable by SRR system against a user's query



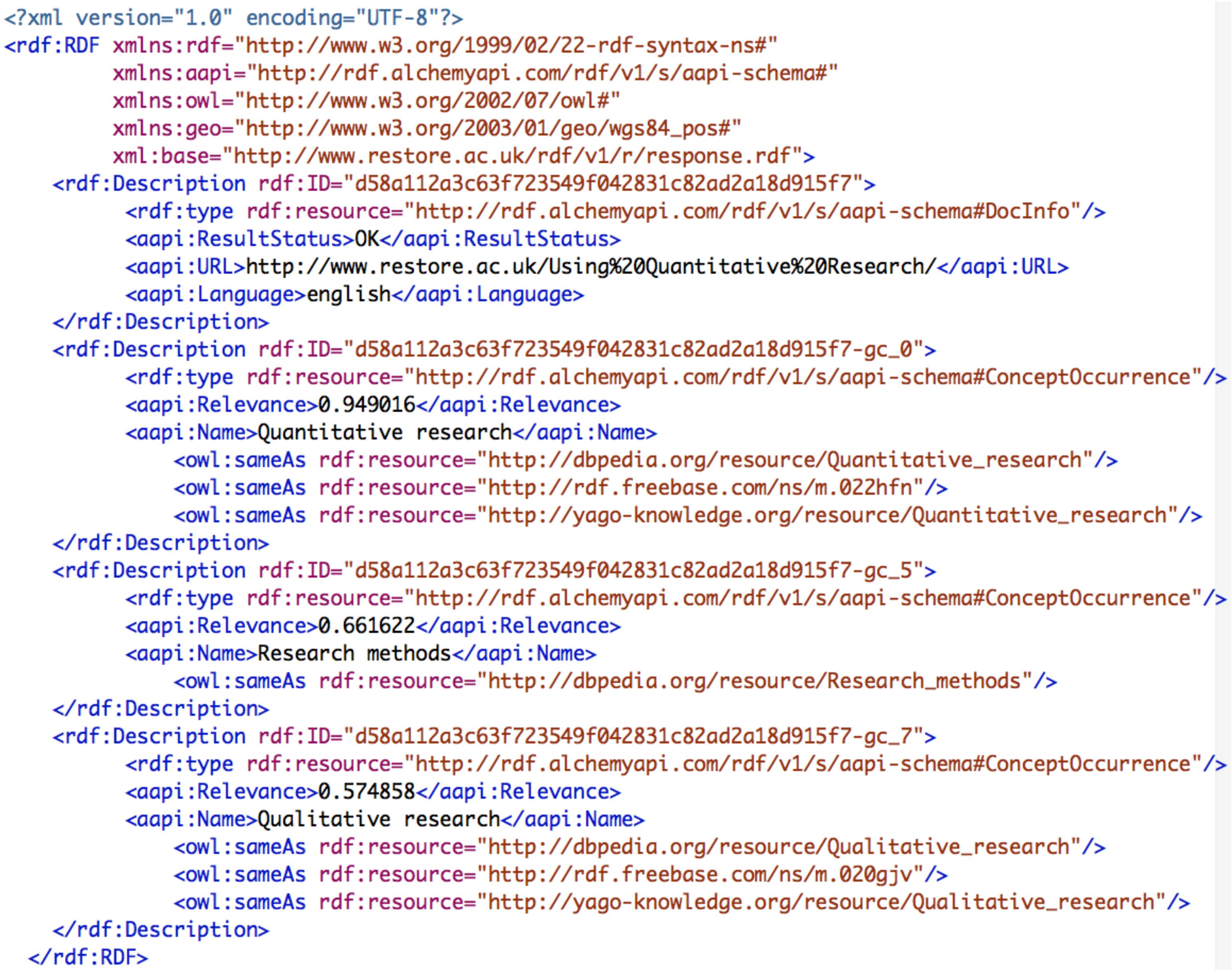
```
          "vocabularyAnnotation": [
          {
              "typology_level1": {
                "level1": [
                  "Qualitative
Approaches"
                ]
              },
              "annotationDate": "2016-03-24",
              "concumer": "08f0a52bc8ee4740aa86a6066f981a30",
              "typology_level2": {
                "level2": [
                  "Qualitative Approaches"
                ]
              },
              "allinOne": {
                "tags": [
                  "self-formulated tag, vocabulary tags, mixture of two, .."
                ]
              },
              "user": "jamesrobards",
              "uri":
"http://www.restore.ac.uk/ISResMeth/Case%20Studies/casestudiesOne.html"
          },
     {…
     “User”: “vivian”,
     "uri": "http://www.restore.ac.uk/ISResMeth/Case%20Studies/casestudiesOne.html"
        }
```

List of typology or vocabulary-based annotations along with typology-based tags

Figure C.6: Part 6/6 (Vocabulary/Typology part): A semantically indexed document in *SemCrowDex* and a part of hybrid vector space stored in Elasticsearch KB retrievable by SRR system against a user's query

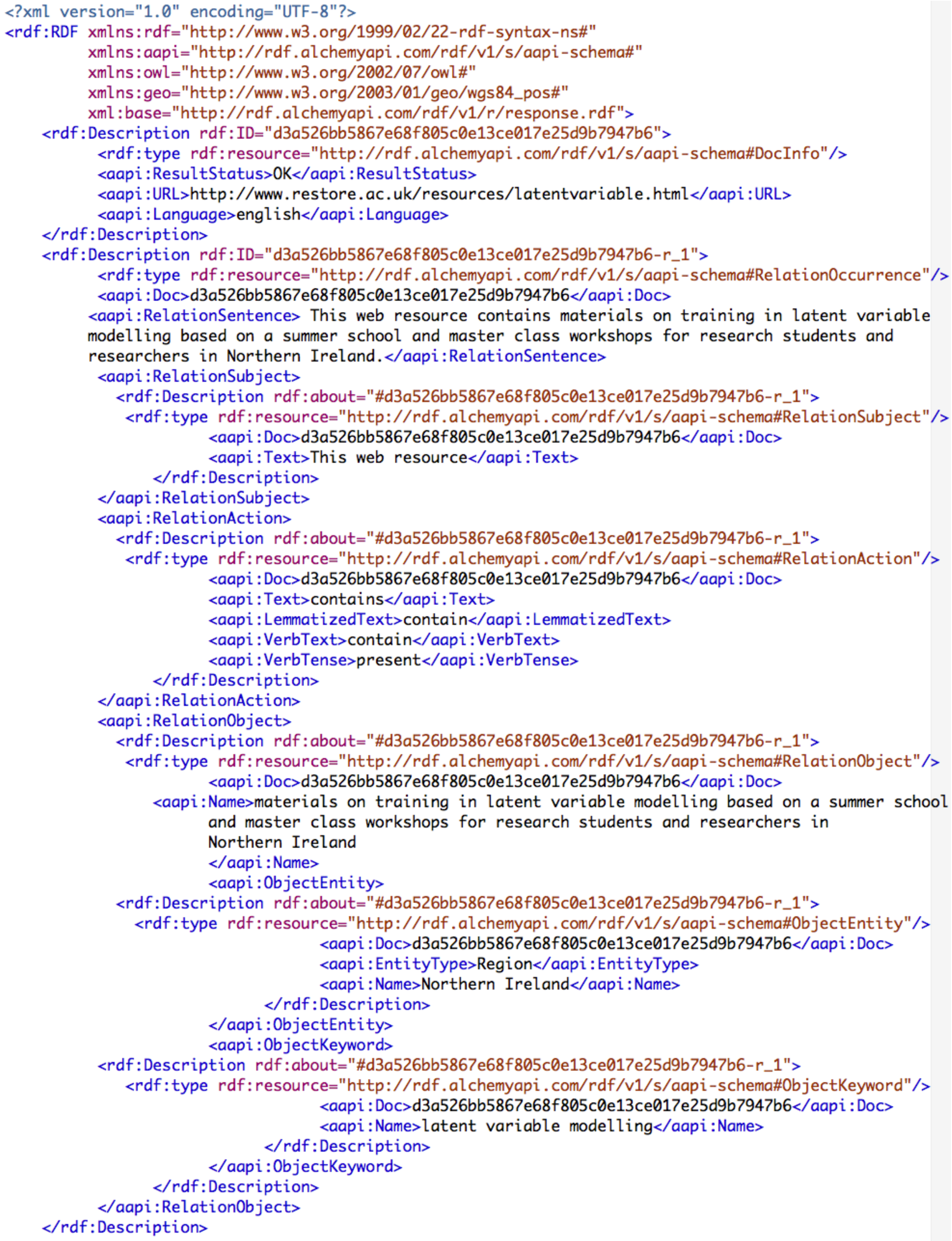
Appendix D

RDFisation of ES-indexed data



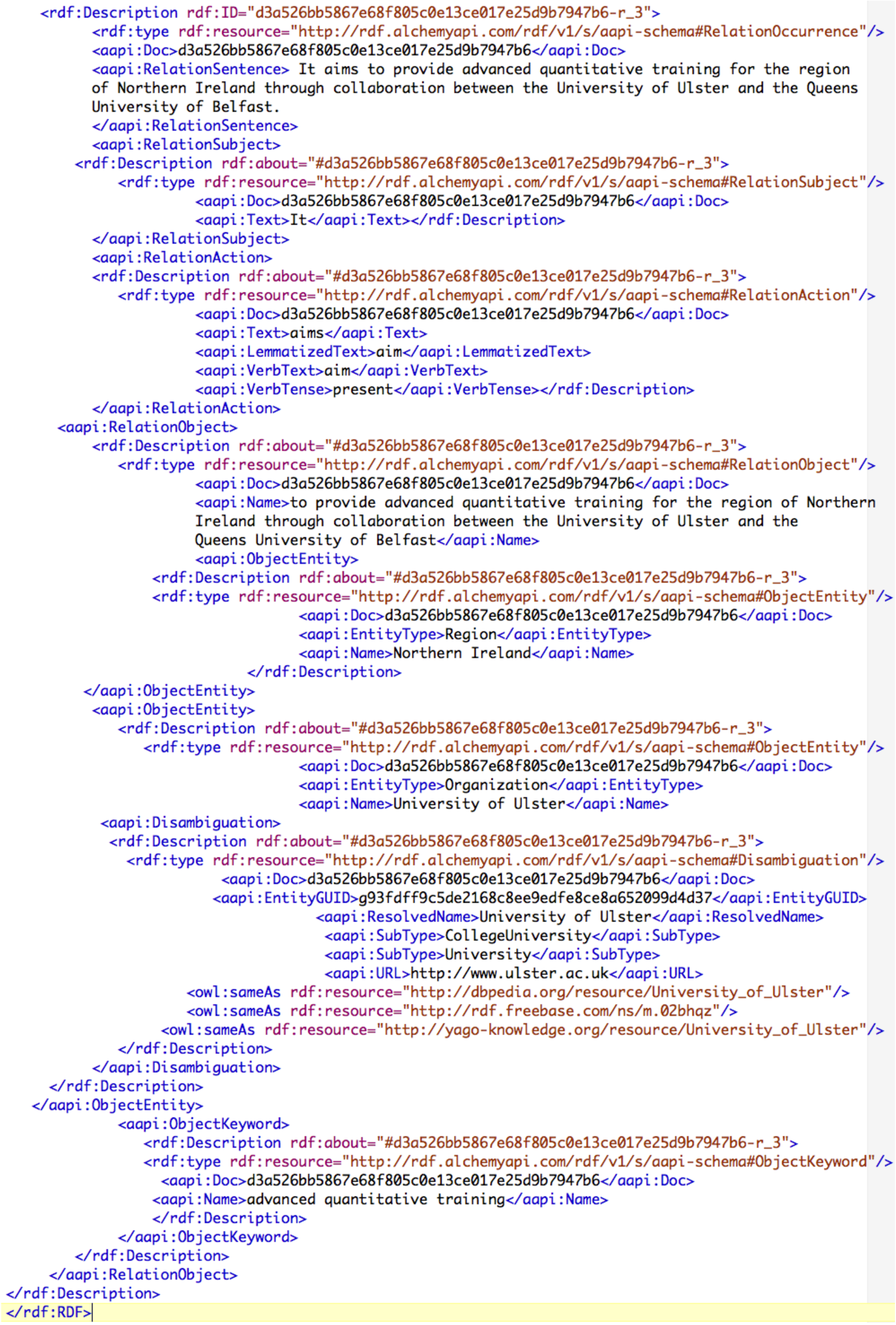
```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:aapi="http://rdf.alchemyapi.com/rdf/v1/s/aapi-schema#"
        xmlns:owl="http://www.w3.org/2002/07/owl#"
        xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#"
        xml:base="http://www.restore.ac.uk/rdf/v1/r/response.rdf">
    <rdf:Description rdf:ID="d58a112a3c63f723549f042831c82ad2a18d915f7">
          <rdf:type rdf:resource="http://rdf.alchemyapi.com/rdf/v1/s/aapi-schema#DocInfo"/>
          <aapi:ResultStatus>OK</aapi:ResultStatus>
          <aapi:URL>http://www.restore.ac.uk/Using%20Quantitative%20Research/</aapi:URL>
          <aapi:Language>english</aapi:Language>
    </rdf:Description>
    <rdf:Description rdf:ID="d58a112a3c63f723549f042831c82ad2a18d915f7-gc_0">
          <rdf:type rdf:resource="http://rdf.alchemyapi.com/rdf/v1/s/aapi-schema#ConceptOccurrence"/>
          <aapi:Relevance>0.949016</aapi:Relevance>
          <aapi:Name>Quantitative research</aapi:Name>
              <owl:sameAs rdf:resource="http://dbpedia.org/resource/Quantitative_research"/>
              <owl:sameAs rdf:resource="http://rdf.freebase.com/ns/m.022hfn"/>
              <owl:sameAs rdf:resource="http://yago-knowledge.org/resource/Quantitative_research"/>
    </rdf:Description>
    <rdf:Description rdf:ID="d58a112a3c63f723549f042831c82ad2a18d915f7-gc_5">
          <rdf:type rdf:resource="http://rdf.alchemyapi.com/rdf/v1/s/aapi-schema#ConceptOccurrence"/>
          <aapi:Relevance>0.661622</aapi:Relevance>
          <aapi:Name>Research methods</aapi:Name>
              <owl:sameAs rdf:resource="http://dbpedia.org/resource/Research_methods"/>
    </rdf:Description>
    <rdf:Description rdf:ID="d58a112a3c63f723549f042831c82ad2a18d915f7-gc_7">
          <rdf:type rdf:resource="http://rdf.alchemyapi.com/rdf/v1/s/aapi-schema#ConceptOccurrence"/>
          <aapi:Relevance>0.574858</aapi:Relevance>
          <aapi:Name>Qualitative research</aapi:Name>
              <owl:sameAs rdf:resource="http://dbpedia.org/resource/Qualitative_research"/>
              <owl:sameAs rdf:resource="http://rdf.freebase.com/ns/m.020gjv"/>
              <owl:sameAs rdf:resource="http://yago-knowledge.org/resource/Qualitative_research"/>
    </rdf:Description>
  </rdf:RDF>
```

Figure D.1: A screenshot of RDF file showing concepts-based triples along with LoD-based resources along with individual relevance scores



```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:aapi="http://rdf.alchemyapi.com/rdf/v1/s/aapi-schema#"
         xmlns:owl="http://www.w3.org/2002/07/owl#"
         xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#"
         xml:base="http://rdf.alchemyapi.com/rdf/v1/r/response.rdf">
    <rdf:Description rdf:ID="d3a526bb5867e68f805c0e13ce017e25d9b7947b6">
          <rdf:type rdf:resource="http://rdf.alchemyapi.com/rdf/v1/s/aapi-schema#DocInfo"/>
          <aapi:ResultStatus>OK</aapi:ResultStatus>
          <aapi:URL>http://www.restore.ac.uk/resources/latentvariable.html</aapi:URL>
          <aapi:Language>english</aapi:Language>
    </rdf:Description>
    <rdf:Description rdf:ID="d3a526bb5867e68f805c0e13ce017e25d9b7947b6-r_1">
          <rdf:type rdf:resource="http://rdf.alchemyapi.com/rdf/v1/s/aapi-schema#RelationOccurrence"/>
          <aapi:Doc>d3a526bb5867e68f805c0e13ce017e25d9b7947b6</aapi:Doc>
         <aapi:RelationSentence> This web resource contains materials on training in latent variable
         modelling based on a summer school and master class workshops for research students and
         researchers in Northern Ireland.</aapi:RelationSentence>
          <aapi:RelationSubject>
            <rdf:Description rdf:about="#d3a526bb5867e68f805c0e13ce017e25d9b7947b6-r_1">
             <rdf:type rdf:resource="http://rdf.alchemyapi.com/rdf/v1/s/aapi-schema#RelationSubject"/>
                      <aapi:Doc>d3a526bb5867e68f805c0e13ce017e25d9b7947b6</aapi:Doc>
                      <aapi:Text>This web resource</aapi:Text>
                <rdf:Description>
          </aapi:RelationSubject>
          <aapi:RelationAction>
            <rdf:Description rdf:about="#d3a526bb5867e68f805c0e13ce017e25d9b7947b6-r_1">
             <rdf:type rdf:resource="http://rdf.alchemyapi.com/rdf/v1/s/aapi-schema#RelationAction"/>
                      <aapi:Doc>d3a526bb5867e68f805c0e13ce017e25d9b7947b6</aapi:Doc>
                      <aapi:Text>contains</aapi:Text>
                      <aapi:LemmatizedText>contain</aapi:LemmatizedText>
                      <aapi:VerbText>contain</aapi:VerbText>
                      <aapi:VerbTense>present</aapi:VerbTense>
                </rdf:Description>
          </aapi:RelationAction>
          <aapi:RelationObject>
            <rdf:Description rdf:about="#d3a526bb5867e68f805c0e13ce017e25d9b7947b6-r_1">
             <rdf:type rdf:resource="http://rdf.alchemyapi.com/rdf/v1/s/aapi-schema#RelationObject"/>
                      <aapi:Doc>d3a526bb5867e68f805c0e13ce017e25d9b7947b6</aapi:Doc>
                <aapi:Name>materials on training in latent variable modelling based on a summer school
                      and master class workshops for research students and researchers in
                      Northern Ireland
                      </aapi:Name>
                      <aapi:ObjectEntity>
            <rdf:Description rdf:about="#d3a526bb5867e68f805c0e13ce017e25d9b7947b6-r_1">
              <rdf:type rdf:resource="http://rdf.alchemyapi.com/rdf/v1/s/aapi-schema#ObjectEntity"/>
                                  <aapi:Doc>d3a526bb5867e68f805c0e13ce017e25d9b7947b6</aapi:Doc>
                                  <aapi:EntityType>Region</aapi:EntityType>
                                  <aapi:Name>Northern Ireland</aapi:Name>
                            </rdf:Description>
                      </aapi:ObjectEntity>
                      <aapi:ObjectKeyword>
          <rdf:Description rdf:about="#d3a526bb5867e68f805c0e13ce017e25d9b7947b6-r_1">
             <rdf:type rdf:resource="http://rdf.alchemyapi.com/rdf/v1/s/aapi-schema#ObjectKeyword"/>
                                  <aapi:Doc>d3a526bb5867e68f805c0e13ce017e25d9b7947b6</aapi:Doc>
                                  <aapi:Name>latent variable modelling</aapi:Name>
                            </rdf:Description>
                      </aapi:ObjectKeyword>
                </rdf:Description>
          </aapi:RelationObject>
    </rdf:Description>
```

Figure D.2: A prototype of RDF file showing concepts-based triples along with LoD-based resources along with individual relevance scores (1/2)



```
    <rdf:Description rdf:ID="d3a526bb5867e68f805c0e13ce017e25d9b7947b6-r_3">
          <rdf:type rdf:resource="http://rdf.alchemyapi.com/rdf/v1/s/aapi-schema#RelationOccurrence"/>
          <aapi:Doc>d3a526bb5867e68f805c0e13ce017e25d9b7947b6</aapi:Doc>
          <aapi:RelationSentence> It aims to provide advanced quantitative training for the region
          of Northern Ireland through collaboration between the University of Ulster and the Queens
          University of Belfast.
          </aapi:RelationSentence>
          <aapi:RelationSubject>
        <rdf:Description rdf:about="#d3a526bb5867e68f805c0e13ce017e25d9b7947b6-r_3">
             <rdf:type rdf:resource="http://rdf.alchemyapi.com/rdf/v1/s/aapi-schema#RelationSubject"/>
                      <aapi:Doc>d3a526bb5867e68f805c0e13ce017e25d9b7947b6</aapi:Doc>
                      <aapi:Text>It</aapi:Text></rdf:Description>
          </aapi:RelationSubject>
          <aapi:RelationAction>
          <rdf:Description rdf:about="#d3a526bb5867e68f805c0e13ce017e25d9b7947b6-r_3">
             <rdf:type rdf:resource="http://rdf.alchemyapi.com/rdf/v1/s/aapi-schema#RelationAction"/>
                      <aapi:Doc>d3a526bb5867e68f805c0e13ce017e25d9b7947b6</aapi:Doc>
                      <aapi:Text>aims</aapi:Text>
                      <aapi:LemmatizedText>aim</aapi:LemmatizedText>
                      <aapi:VerbText>aim</aapi:VerbText>
                      <aapi:VerbTense>present</aapi:VerbTense></rdf:Description>
          </aapi:RelationAction>
      <aapi:RelationObject>
          <rdf:Description rdf:about="#d3a526bb5867e68f805c0e13ce017e25d9b7947b6-r_3">
             <rdf:type rdf:resource="http://rdf.alchemyapi.com/rdf/v1/s/aapi-schema#RelationObject"/>
                      <aapi:Doc>d3a526bb5867e68f805c0e13ce017e25d9b7947b6</aapi:Doc>
                      <aapi:Name>to provide advanced quantitative training for the region of Northern
                      Ireland through collaboration between the University of Ulster and the
                      Queens University of Belfast</aapi:Name>
                      <aapi:ObjectEntity>
                 <rdf:Description rdf:about="#d3a526bb5867e68f805c0e13ce017e25d9b7947b6-r_3">
                 <rdf:type rdf:resource="http://rdf.alchemyapi.com/rdf/v1/s/aapi-schema#ObjectEntity"/>
                                  <aapi:Doc>d3a526bb5867e68f805c0e13ce017e25d9b7947b6</aapi:Doc>
                                  <aapi:EntityType>Region</aapi:EntityType>
                                  <aapi:Name>Northern Ireland</aapi:Name>
                            </rdf:Description>
         </aapi:ObjectEntity>
          <aapi:ObjectEntity>
             <rdf:Description rdf:about="#d3a526bb5867e68f805c0e13ce017e25d9b7947b6-r_3">
                <rdf:type rdf:resource="http://rdf.alchemyapi.com/rdf/v1/s/aapi-schema#ObjectEntity"/>
                                  <aapi:Doc>d3a526bb5867e68f805c0e13ce017e25d9b7947b6</aapi:Doc>
                                  <aapi:EntityType>Organization</aapi:EntityType>
                                  <aapi:Name>University of Ulster</aapi:Name>
           <aapi:Disambiguation>
            <rdf:Description rdf:about="#d3a526bb5867e68f805c0e13ce017e25d9b7947b6-r_3">
              <rdf:type rdf:resource="http://rdf.alchemyapi.com/rdf/v1/s/aapi-schema#Disambiguation"/>
                         <aapi:Doc>d3a526bb5867e68f805c0e13ce017e25d9b7947b6</aapi:Doc>
                        <aapi:EntityGUID>g93fdff9c5de2168c8ee9edfe8ce8a652099d4d37</aapi:EntityGUID>
                                    <aapi:ResolvedName>University of Ulster</aapi:ResolvedName>
                                     <aapi:SubType>CollegeUniversity</aapi:SubType>
                                     <aapi:SubType>University</aapi:SubType>
                                     <aapi:URL>http://www.ulster.ac.uk</aapi:URL>
                     <owl:sameAs rdf:resource="http://dbpedia.org/resource/University_of_Ulster"/>
                     <owl:sameAs rdf:resource="http://rdf.freebase.com/ns/m.02bhqz"/>
                  <owl:sameAs rdf:resource="http://yago-knowledge.org/resource/University_of_Ulster"/>
             </rdf:Description>
          </aapi:Disambiguation>
     </rdf:Description>
   </aapi:ObjectEntity>
             <aapi:ObjectKeyword>
                <rdf:Description rdf:about="#d3a526bb5867e68f805c0e13ce017e25d9b7947b6-r_3">
                <rdf:type rdf:resource="http://rdf.alchemyapi.com/rdf/v1/s/aapi-schema#ObjectKeyword"/>
                  <aapi:Doc>d3a526bb5867e68f805c0e13ce017e25d9b7947b6</aapi:Doc>
                 <aapi:Name>advanced quantitative training</aapi:Name>
                 </rdf:Description>
             </aapi:ObjectKeyword>
        </rdf:Description>
     </aapi:RelationObject>
</rdf:Description>
</rdf:RDF>
```

Figure D.3: A prototype of RDF file showing concepts-based triples ... (2/2)

References

Alba, A., Coden, A., Gentile, A. L., Gruhl, D., Ristoski, P., and Welch, S. (2017). Multi-lingual concept extraction with linked data and human-in-the-loop. In *Proceedings of the Knowledge Capture Conference*, page 24. ACM.

Alobaidi, M., Malik, K. M., and Sabra, S. (2018). Linked open data-based framework for automatic biomedical ontology generation. *BMC Bioinformatics*, 19(1):319.

Alves, H. and Santanchè, A. (2013). Folksonomized ontology and the 3e steps technique to support ontology evolvement. *Web Semantics: Science, Services and Agents on the World Wide Web*, 18(1):19 – 30. Special Section on the Semantic and Social Web.

Andersen, E. (2012). Making enterprise search work: From simple search box to big data navigation. *Center for Information Systems Research (CISR) Massachusetts Institute of Technology (MIT) Sloan School Management*, 12(11).

Azouaou, F., Chen, W., and Desmoulins, C. (2004). Semantic annotation tools for learning material. In *International Workshop on Applications of Semantic Web Technologies for E-Learning (SW-EL*. Citeseer.

Bailey, K. D. (1994a). *Typologies and taxonomies: an introduction to classification techniques*, volume 102. Sage.

Bailey, K. D. (1994b). *Typologies and taxonomies: an introduction to classification techniques*, volume 102. Sage.

Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., and Su, Z. (2007). Optimizing web search using social annotations. In *Proceedings of the 16th international conference on World Wide Web*, pages 501–510. ACM.

Bast, H., Buchhold, B., and Haussmann, E. (2016). Semantic search on text and knowledge bases. *Foundations and Trends® in Information Retrieval*, 10(2-3):119–271.

Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online review*, 13(5):407–424.

Bedi, P., Thukral, A., and Banati, H. (2013). Focused crawling of tagged web resources using ontology. *Computers & Electrical Engineering*, 39(2):613–628.

Beissel-Durrant, G. (2004). A typology of research methods within the social sciences.

Benjamins, R., Contreras, J., Corcho, O., and Gomez-Perez, A. (2002). The six challenges of the semantic web.

Berners-Lee, T., Harmelen, F. v., and Giannandrea, J. (2012). Big graph data panel.

Bi, B., Ma, H., Hsu, B.-J. P., Chu, W., Wang, K., and Cho, J. (2015). Learning to recommend related entities to search users. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 139–148. ACM.

Bizer, C., Heath, T., and Berners-Lee, T. (2009a). Linked data-the story so far. *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, pages 205–227.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009b). Dbpedia - a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154 – 165. The Web of Data.

Boiński, T. and Ambrożewicz, A. (2017). Dbpedia and yago as knowledge base for natural language based question answering—the evaluation. In *International Conference on Man–Machine Interactions*, pages 251–260. Springer.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM.

Bontcheva, K., Tablan, V., and Cunningham, H. (2014). *Semantic Search over Documents and Ontologies*, volume 8173 of *Lecture Notes in Computer Science*, chapter 2, pages 31–53. Springer Berlin Heidelberg.

Buckland, M., Jiang, H., Kim, Y., and Petras, V. (2001). Domain-based indexes: Indexing for communities of users. *3e Congrès du Chapitre français de l'ISKO, 5–6 juillet 2001. Filtrage et résumé informatique de l'information sur les réseaux*, pages 181–185.

Buckland, M. K. (2012). Obsolescence in subject description. *Journal of documentation*, 68(2):154–161.

Butt, A. S., Haller, A., and Xie, L. (2015). A taxonomy of semantic web data retrieval techniques. In *Proceedings of the 8th international conference on knowledge capture*, page 9. ACM.

Cameron, D., Sheth, A. P., Jaykumar, N., Thirunarayan, K., Anand, G., and Smith, G. A. (2014). A hybrid approach to finding relevant social media content for complex domain specific information needs. *Web Semantics: Science, Services and Agents on the World Wide Web*, 29:39 – 52. Life Science and e-Science.

Castells, P., Fernandez, M., and Vallet, D. (2007). An adaptation of the vector-space model for ontology-based information retrieval. *IEEE transactions on knowledge and data engineering*, 19(2).

Chauhan, R., Goudar, R., Sharma, R., and Chauhan, A. (2013). Domain ontology based semantic search for efficient information retrieval through automatic query expansion. In *Intelligent Systems and Signal Processing (ISSP), 2013 International Conference on*, pages 397–402. IEEE.

Chi, E. H. and Bernstein, M. S. (2012). Leveraging online populations for crowdsourcing. *IEEE Internet Computing*, 16(5):10–12.

Clark, S. (2013). *Our ambiguous world of words.*

Cleverley, P. H. and Burnett, S. M. (2015). The best of both worlds: highlighting the synergies of combining manual and automatic knowledge organization methods to improve information search and discovery.

Dale, R. (2018). Text analytics apis, part 1: The bigger players. *Natural Language Engineering*, 24(2):317–324.

Damova, M., Kiryakov, A., Simov, K., and Petrov, S. (2010). Mapping the central lod ontologies to proton upper-level ontology. In *Proceedings of the 5th International Conference on Ontology Matching-Volume 689*, pages 61–72. CEUR-WS. org.

d'Aquin, M. and Jay, N. (2013). Interpreting data mining results with linked data for learning analytics: motivation, case study and directions. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pages 155–164. ACM.

Davies, J. F., Grobelnik, M., and Mladenic, D. (2008). *Semantic knowledge management: Integrating ontology management, knowledge discovery, and human language technologies.* Springer Science & Business Media.

Davis, R. C. (2016). Annotate the web: Four ways to mark up web content. *Behavioral & Social Sciences Librarian*, 35(1):46–49.

De Virgilio, R. (2011). Rdfa based annotation of web pages through keyphrases extraction. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 644–661. Springer.

Derntl, M., Hampel, T., Motschnig-Pitrik, R., and Pitner, T. (2011). Inclusive social tagging and its support in web 2.0 services. *Computers in Human Behavior*, 27(4):1460–1466.

Doty, D. H. and Glick, W. H. (1994). Typologies as a unique form of theory building: Toward improved understanding and modeling. *Academy of management review*, 19(2):230–251.

Du Toit, J. L. and Mouton, J. (2013). A typology of designs for social research in the built environment. *International Journal of Social Research Methodology*, 16(2):125–139.

Dumitrache, A., Aroyo, L., and Welty, C. (2015). Achieving expert-level annotation quality with crowdtruth: The case of medical relation extraction. In *BDM2I@ ISWC*.

Estellés-Arolas, E. and González-Ladrón-de Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information Science*, 38(2):189–200.

Fairthorne, R. A. (1971). Temporal structure in bibliographical classification. *Chan LM, Richmond PA, E. Svenonius (Eds.), Theory of subject analysis: a sourcebook*, pages 359–366.

Fatima, A., Luca, C., and Wilson, G. (2014). New framework for semantic search engine. In *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*, pages 446–451.

Fensel, D., van Harmelen, F., Andersson, B., Brennan, P., Cunningham, H., Della Valle, E., Fischer, F., Huang, Z., Kiryakov, A., Lee, T. K.-i., et al. (2008). Towards larkc: a platform for web-scale reasoning. In *Semantic Computing, 2008 IEEE International Conference on*, pages 524–529. IEEE.

Fernández, M., Cantador, I., López, V., Vallet, D., Castells, P., and Motta, E. (2011). Semantically enhanced information retrieval: An ontology-based approach. *Web semantics: Science, services and agents on the world wide web*, 9(4):434–452.

Fernandez, M., Lopez, V., Sabou, M., Uren, V., Vallet, D., Motta, E., and Castells, P. (2009). Using trec for cross-comparison between classic ir and ontology-based search models at a web scale.

Fernandez-Lopez, M., Gomez-Perez, A., and Suarez-Figueroa, M. C. (2013). Methodological guidelines for reusing general ontologies. *Data & Knowledge Engineering*, 86:242–275.

Fernández, M., Cantador, I., López, V., Vallet, D., Castells, P., and Motta, E. (2011). Semantically enhanced information retrieval: An ontology-based approach. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(4):434 – 452. JWS special issue on Semantic Search.

Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., and Dredze, M. (2010). Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 80–88. Association for Computational Linguistics.

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJcAI*, volume 7, pages 1606–1611.

Gagnon, M., Zouaq, A., and Jean-Louis, L. (2013). Can we use linked data semantic annotators for the extraction of domain-relevant expressions? In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1239–1246. ACM.

Gal, A., Modica, G., and Jamil, H. (2003). Improving web search with automatic ontology matching. *Submitted for publication. Available upon request from avigal@ ie. technion. ac. il.*

Gangemi, A. (2013a). A comparison of knowledge extraction tools for the semantic web. In *Extended Semantic Web Conference*, pages 351–366. Springer.

Gangemi, A. (2013b). A comparison of knowledge extraction tools for the semantic web. In *Extended Semantic Web Conference*, pages 351–366. Springer.

Gangemi, A., Presutti, V., Reforgiato Recupero, D., Nuzzolese, A. G., Draicchio, F., and Mongiovì, M. (2017). Semantic web machine reading with fred. *Semantic Web*, 8(6):873–893.

Georgiev, G., Popov, B., Osenova, P., and Dimitrov, M. (2013). Adaptive semantic publishing. In *Proceedings of the 2013th International Conference on Semantic Web Enterprise Adoption and Best Practice-Volume 1106*, pages 35–44. CEUR-WS. org.

Gerolimos, M. (2013). Tagging for libraries: A review of the effectiveness of tagging systems for library catalogs. *Journal of Library Metadata*, 13(1):36–58.

Given, L. M. (2008). *The Sage encyclopedia of qualitative research methods.* Sage Publications.

Gonzalo, J., Li, H., Moschitti, A., and Xu, J. (2014). Sigir 2014 workshop on semantic matching 4uj7axdk786 information retrieval. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 1296–1296. ACM.

Grady, C. and Lease, M. (2010). Crowdsourcing document relevance assessment with mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's mechanical turk*, pages 172–179. Association for Computational Linguistics.

Grassi, M., Morbidoni, C., Nucci, M., Fonda, S., and Piazza, F. (2013). Pundit: augmenting web contents with semantics. *Literary and linguistic computing*, 28(4):640–659.

Gregor, S. (2006). The nature of theory in information systems. *MIS quarterly*, pages 611–642.

Grix, J. (2002). Introducing students to the generic terminology of social research. *Politics*, 22(3):175–186.

Halpin, H. (2013). *Social Semantics: The Search for Meaning on the Web*, volume 13. Springer US, USA.

Haslhofer, B., Momeni, E., Gay, M., and Simon, R. (2010). Augmenting europeana content with linked data resources. In *Proceedings of the 6th International Conference on Semantic Systems*, page 40. ACM.

Heese, R., Luczak-Rösch, M., Paschke, A., Oldakowski, R., and Streibel, O. (2010). One click annotation. In *SFSW*.

Hiemstra, D. (2009). Information retrieval models. *Information Retrieval: searching in the 21st Century*, pages 1–17.

Hinze, A., Heese, R., Luczak-Rösch, M., and Paschke, A. (2012). Semantic enrichment by non-experts: usability of manual annotation tools. In *International Semantic Web Conference*, pages 165–181. Springer.

Howe, J. (2008). *Crowdsourcing: How the power of the crowd is driving the future of business.* Random House.

Jay, N. and d'Aquin, M. (2013). Linked data and online classifications to organise mined patterns in patient data. In *AMIA Annual Symposium Proceedings*, volume 2013, page 681. American Medical Informatics Association.

Jellouli, I. and El Mohajir, M. (2009). Towards automatic semantic annotation of data rich web pages. In *2009 Third International Conference on Research Challenges in Information Science*, pages 139–142. IEEE.

Jiang, J., Hassan Awadallah, A., Shi, X., and White, R. W. (2015). Understanding and predicting graded search satisfaction. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 57–66. ACM.

Kamnardsiri, T., Choosri, N., Sureephong, P., and Lbath, A. (2013). A study of domain centric tourist recommendation system using ontology search. In *7th International Conference on Software, Knowledge, Information Management and Applications (SKIMA 2013).*

Karpf, D. (2012). Social science research methods in internet time. *Information, Communication & Society*, 15(5):639–661.

Khan, A., Martin, D. J., and Tiropanis, T. (2013). Using semantic indexing to improve searching performance in web archives. *ThinkMind*, page 101 to 104.

Khan, A., Tiropanis, T., and Martin, D. (2015). Exploiting semantic annotation of content with linked open data (lod) to improve searching performance in web repositories of multi-disciplinary research data. In *9th Russian Summer School, RuSSIR 2015, Saint Petersburg, Russia, August 24-28, 2015*, volume 573, pages 130–145. Springer International Publishing.

Khan, A., Tiropanis, T., and Martin, D. (2017). Crowd-annotation and lod-based semantic indexing of content in multi-disciplinary web repositories to improve search results. In *Proceedings of the Australasian Computer Science Week Multiconference*, page 53. ACM.

Kiryakov, A., Popov, B., Terziev, I., Manov, D., and Ognyanoff, D. (2004). Semantic annotation, indexing, and retrieval. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2(1):49–79.

Kondreddi, S. K., Triantafillou, P., and Weikum, G. (2014). Combining information extraction and human computing for crowdsourced knowledge acquisition. In *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*, pages 988–999. IEEE.

Lashkari, F., Ensan, F., Bagheri, E., and Ghorbani, A. A. (2017). Efficient indexing for semantic search. *Expert Systems With Applications*, 73:92–114.

Lawson, K. G. (2009). Mining social tagging data for enhanced subject access for readers and researchers. *The Journal of Academic Librarianship*, 35(6):574–582.

Lease, M. and Yilmaz, E. (2012). Crowdsourcing for information retrieval. In *ACM SIGIR Forum*, volume 45, pages 66–75. ACM.

Lee, J., Cho, H., Park, J.-W., Cha, Y.-r., Hwang, S.-w., Nie, Z., and Wen, J.-R. (2013). Hybrid entity clustering using crowds and data. *The VLDB Journal*, 22(5):711–726.

Lei Zeng, M. (2008). Knowledge organization systems (kos). *Knowledge organization*, 35(2-3):160–182.

Levene, M. (2011). *An introduction to search engines and web navigation*. John Wiley & Sons.

Lu, K. and Kipp, M. E. (2014). Understanding the retrieval effectiveness of collaborative tags and author keywords in different retrieval environments: An experimental study on medical collections. *Journal of the Association for Information Science and Technology*, 65(3):483–500.

Luff, R., Byatt, D., and Martin, D. (2015a). Review of the typology of research methods within the social sciences.

Luff, R., Byatt, D., and Martin, D. (2015b). Review of the typology of research methods within the social sciences.

Marshall, J. K. (1977). *On equal terms: a thesaurus for nonsexist indexing and catalogin*. Neal Schuman Publishers.

McCloskey, J. C. and Bulechek, G. M. (1994). Standardizing the language for nursing treatments: an overview of the issues. *Nursing Outlook*, 42(2):56–63.

Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. (2011). Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*, pages 1–8. ACM.

Mestrovic, A. and Calì, A. (2017). *An Ontology-Based Approach to Information Retrieval*, pages 150–156. Springer International Publishing, Cham.

Milne, D. and Witten, I. H. (2008). Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM.

Mirizzia, R., Di Noiaa, A. R. T., and Di Sciascioa, E. (2010). Lookup, explore, discover: how dbpedia can improve your web search.

Mukherjee, S., Bhayani, J. V., Chand, J., and Raj, R. N. (2014). Keyword recommendation for internet search engines. US Patent 8,676,830.

Müller, B., Hagelstein, A., and Gübitz, T. (2016). Life science ontologies in literature retrieval: A comparison of linked data sets for use in semantic search on a heterogeneous corpus. In *European Knowledge Acquisition Workshop*, pages 158–161. Springer.

Navas-Delgado, I., Moreno-Vergara, N., Gomez-Lora, A. C., del Mar Roldan-Garcia, M., Ruiz-Mostazo, I., and Aldana-Montes, J. F. (2004). Embedding semantic annotations into dynamic web contents. In *Database and Expert Systems Applications, 2004. Proceedings. 15th International Workshop on*, pages 231–235. IEEE.

Nebot, V. and Berlanga, R. (2014). Exploiting semantic annotations for open information extraction: an experience in the biomedical domain. *Knowledge and information Systems*, 38(2):365–389.

Nickerson, R. C., Varshney, U., and Muntermann, J. (2013). A method for taxonomy development and its application in information systems. *European Journal of Information Systems*, 22(3):336–359.

Noy, N. F., Mortensen, J., Musen, M. A., and Alexander, P. R. (2013). Mechanical turk as an ontology engineer?: using microtasks as a component of an ontology-engineering workflow. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 262–271. ACM.

Oliveira, P. and Rocha, J. (2013). Semantic annotation tools survey. In *Computational Intelligence and Data Mining (CIDM), 2013 IEEE Symposium on*, pages 301–307. IEEE.

Olson, H. A. and Wolfram, D. (2006). Indexing consistency and its implications for information architecture: A pilot study. *IA Summit*.

Ou, S., Pekar, V., Orasan, C., Spurk, C., and Negri, M. (2008). Development and alignment of a domain-specific ontology for question answering. In *LREC*.

Packer, H. S. (2011). Evolving ontologies with online learning and forgetting algorithms.

Paulheim, H. (2012a). Generating possible interpretations for statistics from linked open data. In *Extended Semantic Web Conference*, pages 560–574. Springer.

Paulheim, H. (2012b). Nobody wants to live in a cold city where no music has been recorded. In *Extended Semantic Web Conference*, pages 387–391. Springer.

Pech, F., Martinez, A., Estrada, H., and Hernandez, Y. (2017). Semantic annotation of unstructured documents using concepts similarity. *Scientific Programming*, 2017.

Petras, V. (2006). *Translating dialects in search: Mapping between specialized languages of discourse and documentary languages.* PhD thesis.

Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D., and Goranov, M. (2003). Kim-semantic annotation platform. In *International Semantic Web Conference*, pages 834–849. Springer.

Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D., and Kirilov, A. (2004). Kim-a semantic platform for information extraction and retrieval. *Natural language engineering*, 10(3-4):375–392.

Raimond, Y., Ferne, T., Smethurst, M., and Adams, G. (2014). The bbc world service archive prototype. *Web Semantics: Science, Services and Agents on the World Wide Web*, 27:2–9.

Raimond, Y., Smethurst, M., McParland, A., and Lowis, C. (2013). Using the past to explain the present: interlinking current affairs with archives via the semantic web. In *International Semantic Web Conference*, pages 146–161. Springer.

Ranwez, S., Duthil, B., Sy, M. F., Montmain, J., Augereau, P., and Ranwez, V. (2013). How ontology based information retrieval systems may benefit from lexical text analysis. In *New Trends of Research in Ontologies and Lexical Resources*, pages 209–231. Springer.

Repko, A. F. (2008). *Interdisciplinary research: Process and theory.* Sage.

Riggs, F. W. (1981). *Interconcept report: a new paradigm for solving the terminology problems of the social sciences*, volume 44. Unesco.

Ristoski, P. and Paulheim, H. (2013). Analyzing statistics with background knowledge from linked open data. In *Workshop on Semantic Statistics.*

Ristoski, P. and Paulheim, H. (2015). Visual analysis of statistical data on maps using linked open data. In *International Semantic Web Conference*, pages 138–143. Springer.

Ristoski, P. and Paulheim, H. (2016). Semantic web in data mining and knowledge discovery: A comprehensive survey. *Web Semantics: Science, Services and Agents on the World Wide Web*, 36:1 – 22.

Rizzo, G., Troncy, R., Hellmann, S., and Bruemmer, M. (2012a). Nerd meets nif: Lifting nlp extraction results to the linked data cloud. *LDOW*, 937.

Rizzo, G., Troncy, R., Hellmann, S., and Bruemmer, M. (2012b). Nerd meets nif: Lifting nlp extraction results to the linked data cloud. *LDOW*, 937.

Rizzo, G., van Erp, M., and Troncy, R. (2014). Benchmarking the extraction and disambiguation of named entities on the semantic web. In *LREC*, pages 4593–4600.

Royo, J. A., Mena, E., Bernad, J., and Illarramendi, A. (2005). Searching the web: From keywords to semantic queries. In *Third International Conference on Information Technology and Applications (ICITA'05)*, volume 1, pages 244–249. IEEE.

Rujiang, B. and Xiaoyue, W. (2010). A semantic information retrieval system based on kim. In *E-Health Networking, Digital Ecosystems and Technologies (EDT), 2010 International Conference on*, volume 2, pages 392–395. IEEE.

Rusu, D., Fortuna, B., and Mladenic, D. (2011). Automatically annotating text with linked open data. *LDOW*, 813.

Sánchez, D., Isern, D., and Millan, M. (2011). Content annotation for the semantic web: an automatic web-based approach. *Knowledge and Information Systems*, 27(3):393–418.

Sanderson, M., Manning, Raghavan, Schütze, C. D., Prabhakar, and Hinrich (2010). Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103. Copyright - Copyright © Cambridge University Press 2010; Last updated - 2015-05-30.

Shabanzadeh, M., Nematbakhsh, M. A., and Nematbakhsh, N. (2010). A semantic based query expansion to search. In *Intelligent Control and Information Processing (ICICIP), 2010 International Conference on*, pages 523–528. IEEE.

Shirkey, C. (2013). Power laws, weblogs, and inequality. pages 65–72.

Sicilia, M.-A., Yang, C., Yang, K.-C., and Yuan, H.-C. (2007). Improving the search process through ontology-based adaptive semantic search. *The Electronic Library*, 25(2):234–248.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast-but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics.

Souza, R. R., Tudhope, D., and Almeida, M. B. (2012). Towards a taxonomy of kos: Dimensions for classifying knowledge organization systems. *Knowledge organization*, 39(3):179–192.

Subhashree, S., Irny, R., and Kumar, P. S. (2018). Review of approaches for linked data ontology enrichment. In *International Conference on Distributed Computing and Internet Technology*, pages 27–49. Springer.

Suchanek, F. M., Vojnovic, M., and Gunawardena, D. (2008). Social tags: meaning and suggestions. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 223–232. ACM.

Teddlie, C. and Tashakkori, A. (2006). A general typology of research designs featuring mixed methods. *Research in the Schools*, 13(1):12–28.

Tiddi, I. (2013). Explaining data patterns using background knowledge from linked data.

Tiddi, I., d'Aquin, M., and Motta, E. (2013). Explaining clusters with inductive logic programming and linked data.

Tiddi, I., d'Aquin, M., and Motta, E. (2014a). Dedalo: Looking for clusters explanations in a labyrinth of linked data. In *European Semantic Web Conference*, pages 333–348. Springer.

Tiddi, I., d'Aquin, M., and Motta, E. (2014b). Using neural networks to aggregate linked data rules. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 547–562. Springer.

Vargas, S., Blanco, R., and Mika, P. (2016). Term-by-term query auto-completion for mobile search. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 143–152. ACM.

Vavpetič, A., Novak, P. K., Grčar, M., Mozetič, I., and Lavrač, N. (2013). Semantic data mining of financial news articles. In *International Conference on Discovery Science*, pages 294–307. Springer.

Verma, M., Yilmaz, E., and Craswell, N. (2016). On obtaining effort based judgements for information retrieval. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 277–286. ACM.

Wang, K., Wang, Z., Topor, R., Pan, J., and Antoniou, G. (2009). Concept and role forgetting in ALC ontologies. *The Semantic Web-ISWC 2009*, pages 666–681.

Wang, T. D., Parsia, B., and Hendler, J. (2006). A survey of the web ontology landscape. In *International Semantic Web Conference*, pages 682–694. Springer.

Weber, M. (1949). *Max Weber on the methodology of the social sciences.* Free Press.

Weiss, S. M., Indurkhya, N., and Zhang, T. (2010). *Fundamentals of predictive text mining*, volume 41. Springer.

Wilson, T. D. (2000). Human information behavior. *Informing science*, 3(2):49–56.

Wu, J., Killian, J., Yang, H., Williams, K., Choudhury, S. R., Tuarob, S., Caragea, C., and Giles, C. L. (2015). Pdfmef: A multi-entity knowledge extraction framework for scholarly documents and semantic search. In *Proceedings of the 8th International Conference on Knowledge Capture-K-CAP*, page 13. ACM.

Wu, P. H., Heok, A. K., and Tamsir, I. P. (2006a). Annotating the web archives-an exploration of web archives cataloging and semantic web. In *International Conference on Asian Digital Libraries*, pages 12–21. Springer.

Wu, P. H., Heok, A. K., and Tamsir, I. P. (2007). Annotating web archives—structure, provenance, and context through archival cataloguing. *New Review of Hypermedia and Multimedia*, 13(1):55–75.

Wu, X., Zhang, L., and Yu, Y. (2006b). Exploring social annotations for the semantic web. In *Proceedings of the 15th international conference on World Wide Web*, pages 417–426. ACM.

Xiong, C., Power, R., and Callan, J. (2017). Explicit semantic ranking for academic search via knowledge graph embedding. In *Proceedings of the 26th international conference on world wide web*, pages 1271–1279. International World Wide Web Conferences Steering Committee.

Zervanou, K., Korkontzelos, I., Van Den Bosch, A., and Ananiadou, S. (2011). Enrichment and structuring of archival description metadata. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 44–53. Association for Computational Linguistics.

Zhai, H., Lingren, T., Deleger, L., Li, Q., Kaiser, M., Stoutenborough, L., and Solti, I. (2013). Web 2.0-based crowdsourcing for high-quality gold standard development in clinical natural language processing. *Journal of medical Internet research*, 15(4):e73.