Automatic Creation of Researcher’s Competence Profiles Based on Semantic Integration of Heterogeneous Data sources

Vinaya Khadgi (8309081357)
Tianyi Wang (8609095339)

MASTER THESIS 2012
INFORMATICS
Automatic Creation of Researcher’s Competence Profiles Based on Semantic Integration of Heterogeneous Data sources

Vinaya Khadgi
Tianyi Wang

Detta examensarbete är utfört vid Tekniska Högskolan i Jönköping inom ämnet informatik. Arbetet ärr ett led i masterutbildningen med inriktning informationsteknik och management. Författarnas varsararsjälvförfrågade, slutsats och resultat.

Handledare: Dr. Vladimir Tarasov
Examinator: Dr. Ulf Seigerroth
Omfattning: 30 hp (D-nivå)

Datum: 14/12/2012

Arkiveringsnummer:
Abstract

The research journals and publications are great source of knowledge produced by the virtue of hard work done by researchers. Several digital libraries have been maintaining the records of such research publications in order for general people and other researchers to find and study the previous work done in the research field they are interested in. In order to make the search criteria effective and easier, all of these digital libraries keep a record/database to store the meta-data of the publications. These meta-data records are generally well design to keep the vital records of the publications/articles, which has the potential to give information about the researcher, their research activities, and hence the competence profile.

This thesis work is a study and search of method for building the competence profile of researchers’ base on the records of their publications in the well-known digital libraries. The publications of researchers publish in different publication houses, so, in order to make a complete profile, the data from several of these heterogeneous digital libraries sources have to be integrated semantically. Several of the semantic technologies were studied in order to investigate the challenges of integration of the heterogeneous sources and modeling the researchers’ competence profile. An approach of on-demand profile creation was chosen where user of system could enter some basic name detail of the researcher whose profile is to be created. In this thesis work, Design Science Research methodology was used as the method for research work and to complement this research method with a working artifact, scrum- an agile software development methodology was used to develop a competence profile system as proof of concept.
Sammanfattning

.(Abstract in Swedish)
Acknowledgements

I am very thankful to our teacher/supervisor Dr. Vladimir Tarasov for providing ideas, feedback and encouragement during this thesis work. I am also thankful to the entire team of CLICK project, the project from where the whole concept of this thesis work originates. These people were always accessible for discussion, and help us throughout this thesis period with some crucial vision.

Sincerest thanks to my family in Nepal, for their encouragement since the first moment I came in Sweden. Their immense support and motivation made the whole work a lot easier than it actual was.

At the end, I am also very thankful to my friends for their help and time for discussion.

Vinaya Khadgi

---

I am very thankful to Dr. Vladimir Tarasov, our thesis supervisor, for providing ideas, feedback and encouragement. He was always accessible for discussion, and guided us throughout this thesis period.

Sincerest thanks to my family in China, for their encouragement and support. At the end, I am also very grateful to all my friends for their help and time for discussion.

Tianyi Wang
Key words

Competence Model, Researchers Competency Profile, Heterogeneous Data-source Integration, Bibliographic Data sources, Competency Profile creation system
# Contents

## 1 Introduction

1.1 Background ........................................................................................................... 1

1.2 Purpose/Objectives ............................................................................................... 2

1.3 Limitations ........................................................................................................... 3

1.4 Thesis Outline ..................................................................................................... 3

## 2 Theoretical Background

2.1 Existing Researcher’s Competence Profile Creation Systems ......................... 5

  2.1.1 Profiles Research Networking Software (PRNS) ......................................... 5

  2.1.2 VIVO ............................................................................................................ 6

2.2 Researcher Competence Profile and Competence Modeling ......................... 7

2.3 Bibliographic/Publication Data Sources ............................................................. 7

  2.3.1 Three Examples of Bibliographic Data Sources ........................................... 8

  2.3.2 Citation of Publication and keywords ......................................................... 9

2.4 Data Integration Approaches ............................................................................. 10

2.5 Challenges in Bibliographic Data Source Integration for Researchers’ Profile .... 10

  2.5.1 Syntactic and Schematic Heterogeneity of the Data Sources .................... 11

  2.5.2 Semantic Heterogeneity of the Data Sources ............................................. 11

  2.5.3 Author Name Disambiguation ..................................................................... 11

  2.5.4 Same Entity in Different Sources ............................................................... 12

  2.5.5 Maintenance of Integrated data and Profile updating ............................... 12

2.6 Author Name Disambiguation in Bibliographic Digital Library ..................... 12

  2.6.1 Different Approaches ................................................................................. 13

  2.6.2 Open Challenges in author name disambiguation ................................... 16

2.7 Semantic Technology: Terms and Tools ............................................................. 17

  2.7.1 Ontology .................................................................................................... 17

  2.7.2 XML ........................................................................................................... 18

  2.7.3 RDF and RDFs ......................................................................................... 18

  2.7.4 Web Ontology Language (OWL) ............................................................... 19

  2.7.5 JENA- Ontology Framework ................................................................. 19

  2.7.6 SPARQL-Query Language ...................................................................... 19

  2.7.7 Inference Engine/ Reasoners .................................................................... 20

  2.7.8 XSL and XSLT ......................................................................................... 20

2.8 The Role of Ontology in Data Source Integration .............................................. 20

  2.8.1 Ontology base data integration approaches ............................................. 20

2.9 Summary of Literature Review .......................................................................... 22

## 3 Research Methods

3.1 Research Methods in Information System ......................................................... 23
3.1.1 Action Research (AR) .................................................. 23
3.1.2 Constructive Research Process ....................................... 24
3.1.3 Design Science Research (DSR) ..................................... 24
3.1.4 Selection of Research Method ....................................... 25
3.2 IMPLEMENTATION OF DSR .......................................... 26
3.2.1 DSR Guidelines .................................................. 26
3.2.2 Design Science Research Methodology and Activities .......... 29
3.3 DESIGN OPTIONS FOR SYSTEM DEVELOPMENT .................. 32
3.3.1 SCRUM-Agile Software Development Methodology .......... 33
3.3.2 Methontology methodology for building Ontology ............ 34
3.4 RESEARCH FRAMEWORK ........................................... 37

4 Realization .................................................................. 39

4.1 STUDY OF EXISTING SYSTEM APPROACHES ....................... 39
4.2 SPECIAL FEATURES AND DELIMITATION OF EXISTING PROFILE CREATION APPROACHES .......... 40
4.2.1 Profiles Research Networking Software ............................. 40
4.2.2 VIVO ................................................................ 41
4.3 PROBLEMS AND SOLUTIONS ......................................... 41
4.3.1 How can we resolve the syntactic and schematic heterogeneity of data in bibliographic data sources? ................................ 41
4.3.2 How can we resolve the semantic heterogeneity problem in meta-data representation of data sources? ............................ 42
4.3.3 How can we solve the author name ambiguity problem? ........... 42
4.3.4 How can we identify the identity of a document existing in multiple sources? .............. 42
4.3.5 How can we maintain/update the profile of researchers in our knowledge base to make it more up-to-date? ....................... 43
4.3.6 How can we give relevance weightage to competence level of researchers based on their publication only? .......................... 43
4.4 METHOD FOR COMPETENCY PROFILE SYSTEM DEVELOPMENT .................. 44
4.4.1 Study of the Data sources ...................................... 45
4.4.2 Development of competence profile Ontology for researcher ...... 45
4.4.3 Development of Researcher registration system .................. 45
4.4.4 Development of Data access module ............................. 46
4.4.5 Parsing/Cleansing the data for all the heterogeneity issues including name disambiguation ........................................ 46
4.4.6 Transformation of data into common data model (xml/xml-rdf) ......................................................... 46
4.4.7 Ontology Population ............................................ 47
4.4.8 Showing result to Researcher ..................................... 47
4.5 LAYERED-ARCHITECTURE OF PROFILE CREATION SOFTWARE .................. 47
4.6 ONTOLOGY BASED RESEARCHERS COMPETENCE MODELING .................. 48
4.7 REUSE OF WELL KNOWN VOCABULARY AND ONTOLOGIES FOR INTEROPERABILITY .................. 49
4.8 Refined research objectives ........................................ 49
Contents

5 Results .................................................................................................................. 50
  5.1 Result from the Realization................................................................................. 50
    5.1.1 System Component Architecture................................................................. 50
    5.1.2 Conceptual Modeling for Competence Profile ............................................. 53
    5.1.3 UML Modeling.......................................................................................... 53
  5.2 Prototype Development Using Agile (Scrum) Software Development Methodology 57
    5.2.1 Scrum Roles ............................................................................................... 58
    5.2.2 Scrum Artifacts .......................................................................................... 58
  5.3 Implementation Details and Screenshots ......................................................... 59
  5.4 System Analysis and Evaluation .................................................................... 69
    5.4.1 System Analysis ......................................................................................... 69
    5.4.2 Evaluation .................................................................................................. 72

6 Conclusion and Reflection ................................................................................... 75
  6.1 Conclusion ......................................................................................................... 75
  6.2 Reflection .......................................................................................................... 76

7 Recommendation and Future work ................................................................. 78
  7.1 Recommendations ........................................................................................... 78
  7.2 Future Work ..................................................................................................... 79

8 References ........................................................................................................... 80

9 Appendix ................................................................................................................ 85
## List of Figures

<table>
<thead>
<tr>
<th>Figures number</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 2.1</td>
<td>Profiles Research Networking System [61]</td>
</tr>
<tr>
<td>Figure 2.2</td>
<td>VIVO homepage [60]</td>
</tr>
<tr>
<td>Figure 2.3</td>
<td>PubMed publication metadata [2]</td>
</tr>
<tr>
<td>Figure 2.4</td>
<td>Author Name Disambiguation Taxonomy [65]</td>
</tr>
<tr>
<td>Figure 2.5</td>
<td>ORCiD [70]</td>
</tr>
<tr>
<td>Figure 2.6</td>
<td>Researcher ID homepage [58]</td>
</tr>
<tr>
<td>Figure 2.7</td>
<td>RDF triplet describing Joe Smith-“Joe has homepage identified by URI <a href="http://www.example.org/~joe%E2%80%9D%5B12">http://www.example.org/~joe”[12</a>]</td>
</tr>
<tr>
<td>Figure 2.8</td>
<td>Global Ontology Approach [35]</td>
</tr>
<tr>
<td>Figure 2.9</td>
<td>Multiple Ontology Approach [35]</td>
</tr>
<tr>
<td>Figure 2.10</td>
<td>Hybrid Ontology Approach [35]</td>
</tr>
<tr>
<td>Figure 3.1</td>
<td>Research Method</td>
</tr>
<tr>
<td>Figure 3.2</td>
<td>Design Science Research Cycles [72]</td>
</tr>
<tr>
<td>Figure 3.3</td>
<td>Scrum Methodology [71]</td>
</tr>
<tr>
<td>Figure 3.4</td>
<td>Ontology life cycle in METHONTOLOGY [14]</td>
</tr>
<tr>
<td>Figure 3.5</td>
<td>Ontology integration process [54]</td>
</tr>
<tr>
<td>Figure 3.6</td>
<td>Research Framework</td>
</tr>
<tr>
<td>Figure 4.1</td>
<td>ViVO attributes</td>
</tr>
<tr>
<td>Figure 4.2</td>
<td>Method outline</td>
</tr>
<tr>
<td>Figure 4.3</td>
<td>Layered Architecture of Profile Creation Software</td>
</tr>
<tr>
<td>Figure 4.4</td>
<td>Composition of Researcher Competence Profile</td>
</tr>
<tr>
<td>Figure 5.1</td>
<td>System Architecture for competence profile creator system.</td>
</tr>
<tr>
<td>Figure 5.2</td>
<td>Conceptual model of Researcher’s competence</td>
</tr>
<tr>
<td>Figure 5.3</td>
<td>Use case diagram for Competence Profile System</td>
</tr>
<tr>
<td>Figure 5.4</td>
<td>Sequence diagram</td>
</tr>
<tr>
<td>-----------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>Figure 5.5</td>
<td>Class diagram for Competence profile system</td>
</tr>
<tr>
<td>Figure 5.6</td>
<td>IEEE data schema analysis</td>
</tr>
<tr>
<td>Figure 5.7</td>
<td>Example of reuse and restrictions</td>
</tr>
<tr>
<td>Figure 5.8</td>
<td>Researcher’s Competence Profile Ontology Schema</td>
</tr>
<tr>
<td>Figure 5.9</td>
<td>Code excerpt for xml data retrieval from IEEE</td>
</tr>
<tr>
<td>Figure 5.10</td>
<td>Code excerpt for data mapping/cleansing/parsing</td>
</tr>
<tr>
<td>Figure 5.11</td>
<td>Code excerpt for loading data to ontology</td>
</tr>
<tr>
<td>Figure 5.12</td>
<td>Database model</td>
</tr>
<tr>
<td>Figure 5.13</td>
<td>Login web-form</td>
</tr>
<tr>
<td>Figure 5.14</td>
<td>Part of registration form</td>
</tr>
<tr>
<td>Figure 5.15</td>
<td>Code excerpt of SPARQL query</td>
</tr>
<tr>
<td>Figure 5.16</td>
<td>Researcher’s Profile Visualization</td>
</tr>
<tr>
<td>Figure 5.17</td>
<td>Data flow diagram of the system (POC)</td>
</tr>
<tr>
<td>Figure 5.18</td>
<td>Graphical view of result of SPARQL</td>
</tr>
<tr>
<td>Figure 5.19</td>
<td>Data extraction from IEEE only</td>
</tr>
<tr>
<td>Figure 5.20</td>
<td>Data extraction from IEEE and DiVA</td>
</tr>
<tr>
<td>Figure 5.21</td>
<td>Competence profile extracted from two sources</td>
</tr>
<tr>
<td>Figure 5.22</td>
<td>Reuse of VIVO and FOAF</td>
</tr>
</tbody>
</table>
## List of Tables

<table>
<thead>
<tr>
<th>Tables number</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 3.1</td>
<td>Seven DSR guidelines [18]</td>
</tr>
<tr>
<td>Table 3.2</td>
<td>DSRM activities. [43]</td>
</tr>
<tr>
<td>Table 3.3</td>
<td>DSR Evaluation Methods [18]</td>
</tr>
<tr>
<td>Table 3.4</td>
<td>Classification of METHONTOLOGY activities [54]</td>
</tr>
<tr>
<td>Table 5.1</td>
<td>Product Backlog [Appendix]</td>
</tr>
<tr>
<td>Table 5.2</td>
<td>Sprint Backlog [Appendix]</td>
</tr>
</tbody>
</table>
### List of Abbreviations

<table>
<thead>
<tr>
<th>S.N</th>
<th>Abbreviations</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Owl</td>
<td>Web Ontology Language</td>
</tr>
<tr>
<td>2.</td>
<td>RDF</td>
<td>Resource Description Framework</td>
</tr>
<tr>
<td>3.</td>
<td>RDFs</td>
<td>RDF schema</td>
</tr>
<tr>
<td>4.</td>
<td>XML</td>
<td>Extensive Markup Language</td>
</tr>
<tr>
<td>5.</td>
<td>XSL</td>
<td>EXtensible- Stylesheet Language</td>
</tr>
<tr>
<td>6.</td>
<td>XSLT</td>
<td>XSL-Transformation</td>
</tr>
<tr>
<td>7.</td>
<td>URL</td>
<td>Uniform Resource Locator</td>
</tr>
<tr>
<td>8.</td>
<td>URI</td>
<td>Uniform Resource Indicator</td>
</tr>
<tr>
<td>9.</td>
<td>POC</td>
<td>Proof of Concept</td>
</tr>
<tr>
<td>10.</td>
<td>IEEE</td>
<td>Institute of Electrical and Electronics Engineers</td>
</tr>
<tr>
<td>11.</td>
<td>AR</td>
<td>Action Research</td>
</tr>
<tr>
<td>12.</td>
<td>DSR</td>
<td>Design Science Research</td>
</tr>
<tr>
<td>13.</td>
<td>FOAF</td>
<td>Friend Of A Friend</td>
</tr>
<tr>
<td>14.</td>
<td>DOAP</td>
<td>Description Of A Project</td>
</tr>
<tr>
<td>15.</td>
<td>RSS</td>
<td>Really Simple Syndication</td>
</tr>
<tr>
<td>16.</td>
<td>WWW</td>
<td>World Wide Web</td>
</tr>
</tbody>
</table>
## 1 Introduction

The World Wide Web has brought revolutionary changes in the way people communicate today. The information thrive in the web is enormous and the society today is changing more as knowledge society. The use of computer has changed from being simply computing machine to information processing, database management, text processing, and critical analysis system of all the information. Computers are now the entry point highways of information [1].

The continuous research work done by different researchers in various field generate enormous amount of information and knowledge in the form of research publication, journals, articles, books etc. The importance of this information has made different universities, research institutes, social and government organizations from different countries store the records of all the research work done, articles and journals published on different practice and experimental works as the bibliography data repository. This information helps other researchers and students to see the latest of work in their interests and use citation counts to measure the impact of any article in the research field. The immense information in those bibliography data repository about those articles, journals and other publish material can be used for different knowledge acquisition process. The contribution of any individual to such information repository in terms of research papers articles, books and journals could actually define the competence of the individual/researcher. Competence profiles generated from such sources are reliable information about the researchers as they have been produced from the actual work done by them in different research field. Hence, researchers looking for collaborating with other researchers, consulting organization seeking for competent researchers, students and researchers searching for information for their researches could refer to the profile generated from these sources and know the researchers competent in the field they are looking for. Ravikarnet. Al [38], proposed the approach of using skill classification ontology and data mining techniques to determine the expertise of the researcher based on their research publications.

In this thesis work, bibliographic data repositories of different biomedical, educational, and other scientific journals/publications such as PubMed, DiVA, IEEE and Science Direct have been studied for data integration in order to generate the Competence Profile of the contributing researchers in automatic way. While creating the researcher’s competence profile, we focus on the general information available in papers/metadata repositories about researcher, the research papers and the keywords used in them to indicate the research interest field or their active field of research. Different semantic technology and tools have been used in order to implement the prototype project to ground the concepts presented.

This very first chapter introduces the knowledge gap and research problem of thesis work carried out. Details of how the research work was done, have been described in this chapter. It also summarizes the limitation and scope of the research work along with brief introduction to the rest of the report structure.

### 1.1 Background

**Knowledge Gap:** The competence profile of any researcher would help to summarize and present the skills and expertise gain by the researcher as a result of extensive research work done during the course of time. Such profiles can be beneficial for dif-
different individuals and purposes including the researcher himself/herself. Creating comprehensive profiles of researchers based on their own input and effort has some major challenges. The profile created by the researchers themselves might be incorrect, inaccurate or not comprehensive. Either being unaware of certain of the qualities/skill, researcher may not include the complete data in competence profile or some may exaggerate the certain skill description in self-made profiles. Apart from that, the competence of the researcher is dynamic in nature; over time the researchers’ knowledge and experience change which the researchers might not update in the profiles [5]. Tim et.al [39] empirically evaluated the semi-automatic profile generation where users were involved in choosing the textual documents to generate their profile for expert recommendation systems. In their study, they felt the need of some sort of expertise weighing algorithm to specify the expertise level and the semi-automatic profile generation process are mostly not comprehensive.

Hence there is a need of a method for creating the competence profile of researchers in automatic way. This could open up many beneficial possibilities as helping research institutions in finding the correct researcher for work, can help a researcher to locate the most suitable/competent researcher for collaboration work etc. The above-mentioned scopes and shortcoming gave us motivation for our research work.

Among the many probable alternatives to create researcher’s competence profile, we focus on creating the profile based on data integration of different research journals/articles and other documents published in different bibliographic data repositories using different semantic tool kits.

**Research Questions:** The main research questions that we aim to answer in this thesis work are as follow:

1. **What are the main challenges of bibliographic data sources integration for creating the researcher’s competence profile?**

2. **How can data from heterogeneous bibliographic data sources be converted into the common data model and common knowledge model for extracting the competence profile?**

3. **What are the benefits of creating researcher’s competence profile using semantic data source integration method?**

As a use-case to support our thesis work, we use CLICK (Computer Supported Collaborative Work through Dynamic Social Networking) project carried out by at Jonkoping Academy for improvement of Heath and Welfare. The subject area of this thesis is to figure out an approach of integrating heterogeneous research publication data sources available in the Web semantically such that the competence profile of the publishers/researchers can be drawn out of the integrated information from the data repositories.

**1.2 Purpose/Objectives**

The main purpose of this thesis work is to come up with a method and architecture for building the competence profile of researchers based on their contribution in the different academic and other research fields. There are many ways of creating the competency profiles but the approach chosen here in this thesis work is the automatic creation of profile by semantically integrating the different data gather from the bibliographic data repositories. To ground the concepts of the research work done, we also
Introduction

aim to build a simple prototype as a proof of concept (POC). In order to fulfill the above-mentioned objectives, the thesis work is carried out in the following steps:

- Study of different publication database available in the web to know about the different methods/approaches of collecting/extracting the metadata of the all the citation published in them.
- Study of competence model for the researcher’s in-order to map their competence in ontological knowledge repository.
- Obtain an interoperable knowledge base of researcher competence.
- Evaluation of the researcher’s competency model and profile created based on the semantic integration of different data-sources.

1.3 Limitations

Though the studies of different bibliographic data-sources available in the web was done, integration of all of the available publication data-repositories was not practically possible due the time constrain we had for our thesis work. We used some of the existing ontologies and models by different researchers in order to save some time otherwise necessary to build them ourselves, which actually increases the interoperability of the ontology.

Conversion of the data sources in any format (comma/tab separated files, sql-database etc.) into RDF is possible using corresponding tools available. But for implementing our POC, we have chosen the XML only, which is the one of the main form of data exchange in web. Our future work will be focused on implementing an architecture that would be able to integrate the data sources in any other above-mentioned formats that will give a new approach of data integration.

Even though we used Scrum agile software development methodology for implementation of the proof of concept, we had to adjust our work and could not fully justify the methodology. And lastly, our evaluation of the final system is based on verification of system performing according to the research objectives set in the course of this work. Evaluation of competence profile based on actual end user’s feedback (getting feedback from use case project’s user) and recommendation would definitely help this work for further enhancement, which we aim to do in future.

1.4 Thesis outline

This thesis report is divided in following way. In chapter two, we have presented the theoretical overview of the research problem and terminologies associated with it. We first defined the competence profile and model and how researcher’s competence profile should be modeled. We also present some of the approaches that are being used in data source integration and problems related to them. Then we describe the role of ontologies in data source integration.

Chapter three is related with the research methodologies, which we used during the research. We have presented the brief description of the research methodologies that are commonly used in research related to information systems and computer science. Then we described in detail about Design science research, the actual research method that we chose for our research and how we used it in our research to solve the research questions. Since the research method that we used needs designing the solution by constructing an artifact, we also described the software development methods that we
used for developing the artifacts. Combining the understandings from both the research method and the software development methodology, we designed our own research framework as a roadmap for this research.

In chapter four, we discuss some of realizations based on the literature review by presenting the high-level layered architecture diagram of the overall system. We also discuss the general concept behind the design of our architectural diagram as well as design and reuse of different well-known vocabulary for interoperability of the knowledge base we create. We refined our thesis objective and focus towards limited goals.

In chapter five, we have presented the results of our research; the system design for automatic creation of researcher’s competence profile and conceptual models of the competence profile, which we developed as the artifacts using the approach of design science research methodology. Implementation and process we followed for it have been discussed in detail in chapter five. We have given some technical explanations of how scrum software development methodology was implemented to build the proof of concept and also presented the code excerpts to make the reader understand the procedure. Brief analysis and evaluation of the system based on our achievements and goals have also been presented in chapter five. In chapter six, we concluded our report by giving a summary of our study and depicting answers to our research questions. We have also presented our reflection on the process that we followed and mentioned our main learning from this work.

In chapter seven, we have written some recommendations that the future researchers in the same domain can use as guidelines for making the competence profile system and also discuss for future work that we would like to work on.
Theoretical Background

2 Theoretical Background

Before the implementation of the core ideas and views, literature review of the subject matter and relevant works previously done by other researchers are very important to build the foundation. Hence, literature review was done to explore the relevant research work done by other researchers, which helped us to learn insight details of the concepts that had been implemented so far as well as gives new ideas that could be used in the implementation of this research topic’s vision. In this chapter, detail description of literature review done for the research topic has been presented.

2.1 Existing Researcher’s competence profile creation systems

There are plenty of approaches and tools that could create researchers competence profile based on publication data manually, semi-automatically or automatically. In this section we are going to introduce two typical systems, each of them has more or less problem regarding to our research. Among those problems, we will discuss and choose some of them as our main focus then define objectives of our solution.

2.1.1 Profiles Research Networking Software (PRNS)

Profiles Research Networking Software [61], is an open source system developed by Harvard University for the purpose of finding collaborators in certain area based on researchers’ publication or research area.

Profiles Research Networking Software is funded by National Institute of Health (NIH), which helps in speeding up the process of finding researchers with specific
areas of expertise for collaboration and professional networking. Profiles RNS imports and analyzes "white pages" information, publications, and other data sources to create and maintain a complete searchable library of web-based electronic CV's. Built-in network analysis and data visualization tools allow administrators to generate research portfolios of their institution, discover connections between parts of their organization, and understand what factors influence collaboration [61].

2.1.2 VIVO

![VIVO homepage](image)

**Figure 2.2: VIVO homepage [60]**

VIVO [60] is an open source semantic web application originally developed and implemented at Cornell University. When installed and populated with researcher interests, activities, and accomplishments, VIVO enables the discovery of research and scholarship across disciplines at that institution. VIVO supports browsing and a search function for returning faceted results for rapid retrieval of desired information. The original purpose of VIVO is to create ontology to store the information about the researchers. Besides, it also provides the application that uses this ontology.

Content in any local VIVO installation may be maintained manually, brought into VIVO in automated ways from local systems of record, such as HR, grants, course, and faculty activity databases, or from database providers such as publication aggregators and funding agencies.

VIVO is a Java web application that runs over the Tomcat servlet container. It uses numerous open source libraries including HP's Jena semantic web framework. It is currently available under the terms of the Open Source Initiative BSD License [60].


2.2 Researcher competence profile and competence modeling

To build the researcher’s competence profile and model, we first have to understand the concept of competence, which is actually a complex term [16] to understand. Definition given by several of the scholars and organization about competence is actually not enough to be accepted by whole community member of different research field. The most formidable definition of competence so far can be found in [17]. According to Coi [17], competence consists of three underlying dimensions:

- Competency – represents skills
- Context – domain for using skills
- Proficiency level – expertise level of skill for performing task.

So, while making a competence profile in any domain, these three dimensions should be considered. The term competence profile is very often used in the field of human resource management (HRM) to describe the well structured documents that consist of different sets of competence and skills either defining the ability of the employee or that the job under consideration needs. Level of each competence is quantifiable [3]. HR-XML, the technical consortium working to standardize the HR-XML define competency as “a specific, identifiable, definable and measurable knowledge, skill, ability and/or other employment-related characteristic (e.g., attitude, behavior, physical ability) which a human resource may possess and which is necessary for, or material to, the performance of an activity within a specific business context”. Based on above definition, competency can be seen as a smallest set of capability or the combination of resources in a specific context for accomplishing an object or mission effectively and efficiently.

Competence modeling thus can be seen as process of understanding and capturing the information about different competence shown by individual or organization in more intuitive way. Vladimir et. Al[16], used the enterprise modeling techniques for capturing the existing competence of enterprises and individual so that they can be evaluated systematically. While modeling the competence, it is very important to see the relation of competency with the work situation and also that the proper evaluation or weighing of the competence is needed.

So, in the context of Researchers, the competence profile not only should be able to show the general competency of the researcher but also should be able to give detail about the research work area or the research field of interest, the information about previous works accomplished and the information about any publication the researcher is associated with. The competence profile having more relevant information about the researcher would be considered better/thicker profile in terms of information.

2.3 Bibliographic /Publication Data Sources

The core idea of the thesis is to generate researcher’s profile from the different accomplished research work done and their corresponding papers published on the work within different bibliographic data repositories. The method gives a high degree of reliability in so produced profiles of researchers. Hence the first work was to do a throughout study of such available repositories to have needed access to them. Below are the details of some of the data sources we studied.
2.3.1  Three Examples of Bibliographic Data Sources

A. Diva

DiVA is an online data archive and research publication finding tool of academic research publication by researchers/teachers/students and student thesis written at 30 different universities and higher education colleges (see appendix). DIVA was initiated by the EPC at Uppsala University Library in the year 2000 A.D to preserve the academic and research work for long term. Bibliographic registrations of documents from 1995 can be found in DiVA with some document being older too. Participating universities and publicly financed research institutions are from all across Sweden and abroad too. The technical development of DiVA is being carried out by EPC in collaboration with all the participating institutions.

All the participating institutions have their own local interface of DIVA portal and research publications. Student thesis can be published and registered locally at the university or college of origin with the bibliographic information for every title, abstract and a link to full text [10].

DiVA contains more than 11,000 publications in full text, most of them being doctoral thesis, research papers from different academic researchers, student thesis, reports, articles and publications of different types. In addition to the publications, the DiVA repository also holds records of more than 130,000 references to the publications produced by the different researchers and employees from different Universities. Moreover, DiVA is a freely accessible full text archive for everyone to read, download and print out. The authors/publishers of the documents retain the full copyrights of their work if any republication or other use of the document is needed [11]. The DiVA repository helps researchers and students to have easy access to numerous publications and also preserves them securely for long time.

B. PubMed/Medline

U.S National Library of Medicine (NLM) has indexed the biological literature since 1879 in an effort to facilitate health professionals to have a convenient access to information on different research work, experiments and other relevant information from all across the world. This information will help the health professional in their own research work, health care that they being providing to their patient and also enhance their knowledge and education. The printed indexed database is known as the MEDLINE database, which contains journal citations, titles and author names including the abstract etc. for biomedical literature from all around the world. MEDLINE has been publicly available to general users as well from 1996 as free access and U.S National Library of Medicine National Institute of Health (NIH) provide web interface to the repository search tool known as PUBMED: http://www.ncbi.nlm.nih.gov/pubmed/. Most of the PubMed data come from MEDLINE citations [2], which is actually the largest component of PubMed. The MEDLINE records are indexed according the controlled vocabulary developed by NLM known as the MeSH (Medical Subject Heading) vocabulary.

PubMed web interface can be used to search the information from various other data repository apart from the PubMed and MEDLINE data-repositories. Using PubMed, user can have access to over 21 millions records from different biological literature.
C. IEEE

IEEE (Institute of Electrical and Electronics Engineers) is the world’s largest professional association dedicated to the advancing technological innovation and excellence for the benefit of humanity. The history of the IEEE is rooted to 1884 and earlier, when development of electricity and other communication system was growing fast all over the world. The main aim of the association was to help professionals in their nascent field and to aid them in their endeavor to apply innovation for the betterment of humanity [7].

IEEE now has more than 400,000 members including the professionals and students in over 160 countries of the world with the publications documents exceeding three millions. IEEE publishes about one third of world’s total technical literature in electrical engineering, computer science and electronics. The IEEE documents can be explored from the IEEE Xplore Digital Library; statistics shows that the documents are being downloaded over eight millions time each month from its repository. Apart from maintaining the staggering collection of technical documents, journals, magazines, conference papers etc., the main contribution of IEEE in developing the international standards that governs the major of the telecommunications, information technology and other services [7][8]. Apart from that, IEEE uses INSPEC Thesaurus to assign most appropriate terms or preferred terms to represent/index the source documents. Such terms are assigned by subject specialists and termed as “controlled terms” in XML output generated by IEEE Xplore gateway [53].

2.3.2 Citation of Publication and keywords

Publication repositories (PubMed, IEEE etc.) as mentioned above, receive the publication from different publishers/researchers mostly in electronic forms using scanning and Optical Character Recognition (OCR)]2[7][53]. These publication organizations have different policies to publish such publication. For instance, PubMed has definite status of the published document according to the case of the publication. When a publication is first published as it was send by publisher, the status of such document is clearly mentioned as “as supplied by publisher” so that it is clear to readers.

As in figure above, the publication has clear PubmedID (PMID) but the status is clear to reader that this document has not gone through the quality control analysis to improve its bibliographic and bibliometric meta-data improvisation. When these sort of documents are processed by quality control team (a specialist team for analysis meta-data provided for the publication), they access the document for its bibliographic accuracy [2] [7].

One other very important process involve in the quality control process is the improvement in the keywords used in the publication. Subject specialist analyzes the keywords provided by researcher who published the publication so that they could be indexed with most appropriate keywords that the general users could use for searching purpose too [2][7][53]. For instance, in case of PubMed, the quality control team uses
Theoretical Background

MeSH database while IEEE uses INSPEC Thesaurus. Hence keywords to the documents are chosen in such specific way that they represent the core content and subject research field of the paper.

2.4 Data integration approaches

The bibliographic data repositories primarily stores the information or meta-data about the research articles. For example, title of publication, publication year, name of publisher/s, keywords mentioned by authors, author’s details mentioned in the publication etc. Even though the primary focus of such data source is the information about the research publication, but the collection of meta-data is also equally informative for the data about researchers. The contributions of researchers in any research reflect their competence. But the publication of research materials is not the work of single body of authority. There are many universities, research institutes, consortiums and organization of researchers, which does the work of publication. Hence, most of the publications are either published by some of the existing publication house or might even be publish in multiple repositories. So, in order to create a complete profile of any researcher, we need to integrate meta-data information from as different publication repositories. As the number of integrated data sources increased, the created profiles of researcher will be better or thicker.

According to [36], generally two kinds of data integration approaches are used for heterogeneous data source integration: Materialized view approach and Virtual View Approach.

Materialized View Approach: In this approach, the copy of data from all the distributed heterogeneous sources are physically extracted from the sources, transformed into the desired models and formatting, and finally loaded into the global or central data source for use. This approach of Extraction, Transformation and Loading (ETL) is used in all the traditional data-warehouses. This approach has serious pitfalls of data redundancy, data synchronization problem with the sources and global view and also the system has to do a lot of re-configuration and rebuilding in case of addition of new data source [36]. This approach is mainly used while building artifacts such as Business Intelligence and Data mining tools for the organizations where it is important for organization to analyze all the existing organizational data for getting information from them.

Virtual View Approach: In this approach, user’s queries are used to describe the local views of the distributed data sources and all these local views are integrated to form a global view of the entire data sources. The main difference of virtual view approach with materialized view approach is that data are always in the primary distributed sources; physical transformation of all data is not done. Virtual View approach is recognized by the use of Wrapped/Mediator pattern, which processes the user queries and integrate the local views from sources as the global view for the user [36].

2.5 Challenges in bibliographic data source integration for researchers’ profile

The different data sources chosen to integrate for this thesis work were created by different organizations/individuals and for different context. Even though the domain that these different data sources represent is same i.e. publication metadata, but the information that these sources contain depend on the level of granularity the data
sources has been design to capture. The nomenclatures used in these sources were also different according to the creator of the source. Apart from the meta-data disambiguates, there were several challenges while integrating the actual information content of these sources that needed to be handled in the process. They are being mentioned as below:

2.5.1 Syntactic and Schematic Heterogeneity of the Data Sources

Syntactic heterogeneity is caused by the use of different syntactic languages or models (entity-relationship, object oriented etc.) to construct the data-source whereas schematic or structural heterogeneity is caused by different structural schema between the sources [35].

2.5.2 Semantic Heterogeneity of the Data Sources

Semantic heterogeneity is caused due to the different meaning of the data in different context/sources where they were used. This kind of heterogeneity is hard to correct and map properly. Semantic interoperability in heterogeneous information sources can be achieved only if the actual meaning of the information interchanged across the system is understood [35].

2.5.3 Author Name Disambiguation

Author name disambiguates have long been a serious problem for all the bibliographic data-repositories. Literature reviews show that Name Disambiguation problem in bibliographic record keeping was first noticed and discussed by Eugene Garfield [49] in 1969 in his letter to Institute of Scientific Information, USA. It is very hard to find which particular author wrote the document and also hard to figure out all the documents written by a particular author. It requires extensive metadata analysis or even the throughout reading of the full text of the publication for a reader to make an educated guess regarding the identity of the publication’s author [9]. The task of author name disambiguation becomes even harder for computer programs. Not only in the field of bibliometric records, but name disambiguation is prevalent problem in many fields where individuality or identity is a concern. But still the name disambiguation problem has remained the biggest problem in bibliometric analyses at individual level and scientist have struggled to come up with robust methods to solve the problem [50].

According to [9], there are four main distinct challenges in author name disambiguation process:

a) Single author publishing under different names/format, with different spelling or changing name after marriage or religion conversion, gender re-assignment etc.

b) Different authors having the same name.

c) Inadequate meta-data collection for bibliographic database regarding author name like author’s full name, author’s institution, email address etc. There is no consistent rules/standard followed by bibliographic repositories for their indexes.

d) Multi-authored articles creates disambiguates in proper credential to authors.

It might be possible to identify the particular authors manually in case of one small library maintaining indexes of its own collection, but manual process for name disambiguates in case of data integration of indexes from multiple massive repositories is
near to being impossible. The name disambiguation problem has become even more profound with the increase of large number of Chinese authors and their contribution to scholar publications [51]. For example, in Medline database, the name “Wang, Y” appears more than in 7000 articles [52].

Author name disambiguation is probably the biggest of the hurdle to overcome in our research topic and we need in-detail study of the subject area to get ideas for the solutions for the proof of concept, which our research method demands. Hence, we have dedicated a separate section on Author Name disambiguation and work done on finding a better solution to the problem. Our own research on this topic is presented in section 2.6.

2.5.4 Same Entity in Different Sources

It is highly possible that a particular article/research publication is published in different publication and is being indexed in various repositories. For instance, there are cases of same article’s record in PubMed and IEEE. In such a case, it becomes difficult to identify the individual article, as they will be indexed as per convention being followed by the indexing institution. Different processing methods for managing documents are needed to avoid redundancy. Repositories often maintain a unique id to every article; for example every document published in PubMed has unique PubMed ID [2]. But the reference ids maintained by different repositories are not consistent/interoperable to each other and has no regards in between them. So, those kind of unique id maintain in a particular repository has its meaning in its own repository only.

2.5.5 Maintenance of Integrated data and Profile updating

Once the competence profile of the researcher is created and instantiated in the competence profile ontology, the biggest of the challenge is updating the profile over time. Competence of the researcher is dynamic in nature [5]; over time the researchers’ knowledge and experience change which results in new publications being publish and printed in any of the bibliographic data source. For making the competence profile up-to-date, the competence profile in our knowledge base ontology should also reflect to added publication in the bibliographic source.

It is hard to track the addition of new publications of particular researcher in any publication database because that information is mostly not available.

2.6 Author Name Disambiguation in Bibliographic Digital Library

There are numerous digital publication libraries such as PubMed, IEEE, CiteSeer, DiVA etc. which list millions of bibliographic citation records or the metadata/attributes about the publications (for example: author and co-authors names, work and publication venue titles, keywords of the publication content material etc.). These have become a very important source of information for the scholars and other academic research worker as they provide many features, functionalities and search services that helps to discover relevant publications in centralized manner. Besides these metadata are great source of information for analyzing different scopes, quality and impact of research publication, collaboration pattern in researchers community, topic coverage etc. which could be used for many purposes such as determining the fund for
Theoretical Background

Fund agencies, searching for most cited publications etc. [62][63]. Author name disambiguation is very important element for all the digital libraries as it improves the quality of information from the digital libraries by locating all the name variants of the author and consolidates their citations into a single definitive name [68].

Author name disambiguation is a very important task in digital library, which helps in giving proper credits for authors for their publication, as well as helps in bibliometric analysis [67]. Disambiguation of author names open up immense scope of bibliographic and bibliometric data analysis. If every individual author is identified properly, then from the bibliographic meta-data analysis of author publications, the impact on the research field, author’s collaboration network as well as suggestions for collaborating with certain other researchers etc. can be determined.

Managing the bibliographic data repository is very complex task as there is no standardization of the process yet. Lee et al [63] point out the challenges of having high quality contents in bibliographic repositories data mainly arises from sources like human errors while data entry (spellings and other typos), lack of standards and common practice of using citations format in papers, ambiguous author names, use of abbreviations etc.; among these author name disambiguation being the hardest one to resolve. Many of the bibliographic repositories also use automatic metadata harvesting techniques like CiteSeerX [64] that are more prone to above-mentioned problems.

2.6.1 Different Approaches

Author name disambiguation has been addressed in many researches [62][63][65][67][68] as the one of the most challenging problem in dealing with bibliographic data. Numerous independent researches by researcher community all over world have generated a myriad of disambiguation methods. Author name disambiguation can be either accomplish automatically with some computer programs or by assigning task to human assistant to do the task manually. Below, we have summarized both ways in brief.

2.6.1.1 Automatic Methods for Author Name Disambiguation

![Author Name Disambiguation Taxonomy](image)

*Figure 2.4: Author Name Disambiguation Taxonomy [65]*
Anderson A. Ferreira et. Al [65] proposed the taxonomy for automatic process of author name disambiguation on the base of the survey study conducted on the subject. Their classification of taxonomy is based on the approach/method used and evidence explored to help in the process of disambiguation. Below we have summarized the taxonomy as proposed in the paper [65].

1) Types of Approaches

The approaches used in different researches for author name disambiguation can be divided highly into two main categories; Author grouping and Author assignment.

a) Author Grouping: In author grouping approach, the references of the same author are grouped according to some kind of reference attributes by applying a similarity function, which decides whether to group corresponding references using the clustering technique such as partition, archival agglomerative clustering, and density based clustering etc. Such similarity function may be either any predefined function calculated based on existing attributes, or may be learned function using some supervised machine learning algorithms, or may be a function extracted from relationship mapping graph among the authors and co-authors. Clustering methods using such similarity functions are then used to partition the author’s references into groups having maximum intra group similarities.

b) Author Assignment: In author assignment approach, the references to the given author are tried to directly assign to their respective author by constructing a model that represents the author either using a supervised classification technique or a mathematical model-based clustering technique.

2) Explored Evidence

For assisting the disambiguation process, any method needs the evidence data from which it can judge whether the publication belongs to certain researcher/author or not. Below we describe the main kinds of evidence which are explored by most commonly used disambiguation methods for author names -

a) Citation Information: Most of the methods in practice for name disambiguation use citation information such as author/co-author names, work title, publication venue, author affiliation etc. available from the bibliographic repositories. The problem however is that these information are not always sufficient for disambiguation as the data repositories do not have any standardized criteria of attribute list to maintain the records. Attributes such as emails, address of authors, paper headers are not always present in all the data sources.

b) Web Information: Often information about researchers/authors can be retrieved from their personal websites and blogs or from the website of the institute which the author is affiliated to. Such information is used as addition source of information in disambiguation process to calculate the similarities among the references. The drawback here is the additional cost for retrieving the information from web sources.

c) Implicit Evidence: On analyzing the visible elements of attributes from the citations of particular references, some implicit evidence can be inferred. For instance, if different publication titles of same authors are compared, we will find much similarity in the words used in them. So, given the citations, such techniques can be developed which will give the probability of each citation being of same author.
2.6.1.2 Manual Methods for Author Name Disambiguation

Several of the digital libraries have attempted to use manual procedure to disambiguate the publication by individual authors by assigning the task to librarians [69] or collaborative efforts. Though this is highly accurate procedure but the efficiency depends on human resource and is not feasible solution for massive disambiguation task for the gigantic digital libraries such as PubMed and IEEE. Other attempts of assigning unique digital identification for authors have also shown promising potential for name ambiguity problem. Among such systems, we have given brief summary of two of such systems.

a) ORCiD

![ORCiD](image)

**Figure 2.5: ORCiD [70]**

ORCiD stands for Open Researcher and Contributor Identification and its main mission is to provide unique digital identification to each author and connect researchers to their researches. ORCiD is an open, non-profit, community based effort to create and maintain a central registry of researchers with unique identification such that it facilitates researchers to link to their research publication to account their credits solving the name ambiguity problem. Researcher can then use the unique identifiers provided by ORCiD to publish any publications; and it also provides API for system-to-system communication and authentication, which is also compatible with other similar attempts like ResearcherID. The identifiers, and the relationships among them, can be linked to the researcher’s output to enhance the scientific discovery process and to improve the efficiency of research funding and the collaboration within the research community [70]. Besides, it also helps readers of the literature to distinguish between authors with resembling names.

Registering into ORCiD system is simple and easy which gives researcher the unique digital identification. Such identification can be used by researchers in their publication and get their work recognized explicitly.
Theoretical Background

b) RESEARCHERID

Figure 2.6: Researcher ID homepage [58]

ResearcherID is a project initiated by Thomson Reuters [58], one of the world’s leading information sources for businesses and professionals, to identify researchers with unique identification numbers and help in author name disambiguation process within the scholarly research community. The ResearcherID project adopts the manual way of creating the researcher’s profile where researcher is given a unique identification code during the registration process and then by, the researchers will be able to visit their profile pages online and upload their publications and other information to make their profile. In addition, ResearcherID information integrates with the Web of Knowledge and is ORCID compliant, allowing researchers to claim and showcase their publications from a single one account. Search the registry to find collaborators, review publication lists and explore how research is used around the world [59].

2.6.2 Open Challenges in author name disambiguation

With the advancement in computer science, there are so many sophisticated algorithms and tools to solve a lot of problems complimented by highly reliable, powerful and robust computational machine. But still author name disambiguation in bibliographic repositories is a one of the hardest problem to solve.

“Why is author name disambiguation a difficult problem to solve?”

In [65][68], some of the open challenges in that need to be overcome for making the reliable solutions for author name disambiguation have been discussed. They are mainly as follows:

Not sufficient citation data: Only a minimum set of attributes (in most cases only author names, publication title, and venue) are available to work with author name disambiguating problem [65]. And to add to it, all the repositories follow different policies for data entry resulting in different schemas and attributes in the dataset. There is no standardized rules and set of attributes that the data repositories must follow while maintaining the records.

Ambiguous nature of problem cases: Since the probability of authors name being exactly same with most of the other attributes also matching to each other can’t be neglected, it is highly challenging task to develop a solution that is 100% reliable. As
in most cases of problems of probability, even the slight margin of probability can turn the table. Author name disambiguation is also a case of probability. For instance, in coauthor-based heuristics, the primary hypothesis is that the author name can be disambiguated with analysis of the co-authors they have published journal with. This may be successful in many cases but there is a probability that author with same names might have work with same group of coauthors or even there might be case where even coauthors names are ambiguous [65].

**Citations with Errors:** The process of data fetching for storing citation is complex. Be it manual or automated, errors during data fetching during publication database maintenance or the mistakes by the original authors of the publication while writing citation information can’t be unconsidered [65].

**Efficiency:** Every publication repository is huge in their content size and the contents keep on growing rapidly over the years. So, the method for disambiguation needs to be highly efficient and scalable being accurate at the same time [65][68].

**Changes in author profile:** Over the course of time, author’s focused field of research may change, either due to change in personal interest or due to changes in professional work circumstances or natural evolution of a research field. In either of the cases, it is obvious that they will affect the publication profile of the researcher and it adds more difficulties to the methods for name disambiguation [65].

**New Author:** When a completely new author with ambiguous name has to be identified, due to lack of any data from previous publication, the method for disambiguation have to address this issue in different way. Disambiguating a new author with ambiguous name is a hard task for any method [65].

### 2.7 Semantic technology: Terms and Tools

In this section, we have presented the overview of some of the fundamental terms, data models, tools used in the frameworks of semantic web. These data models and frameworks are the backbone part of building any semantic application.

#### 2.7.1 Ontology

Tom Gruber [27] originally proposed the definition of ontology as “explicit specification of a conceptualization”. The word “ontology” has a lot of difference in its use in the field of philosophy (where it means the subject of existence) and knowledge sharing (where it means specification of a conceptualization). So, ontology can be thought as a knowledge base on a conceptualization of a specific domain of interest and which stores the concepts/objects, other entities and the relationships that exist between them. Ontology has brought revolutionaries changes in the way the knowledge engineering field has flourish in the recent past. Ontology is key tool in processing data to knowledge. Berners-Lee recognized ontologies as an important component in building the new form of web, the Semantic web [55].

Ontologies help in facilitating the interoperability among the information system, intelligent processing by sharing and re-using the knowledge among different system agents. Ontology presents a shared and common understanding of domain knowledge by capturing different concepts/entities in them with the specification of their meaning in the context and makes it possible to be communicated between the different people, heterogeneous applications that might be spread around [54].
2.7.2 XML

In the history of computing, XML is one of the most vital innovations in document syntax. XML stands for Extensible Markup Language, which is a W3C-endorsed standard for document markup. In XML, all the data are included as strings of text only and the markup can be well designed to describe the document’s semantics which could be also easily understood by human. XML is a meta-markup language, which does not have fixed sets of tags as other markup languages. The extensible nature of XML allows it to be extended and adapted to different needs making it as one of the most flexible markup language [6].

The significant benefit of XML however is the enormous possibility it gives for cross platform communication being a long-term data formats. It is incredibly simple structure and all details about the structure of the documents are always explicitly found in the document itself. Hence, working with XML is fairly straightforward [6].

2.7.3 RDF and RDFs

![RDF triplet describing Joe Smith](http://www.example.org/~joe/contact.rdf#joesmith)

![RDF triplet describing Joe Smith](http://xmlns.com/foaf/0.1/homepage)

![RDF triplet describing Joe Smith](http://www.example.org/~joe)

*Figure 2.7: RDF triplet describing Joe Smith- “Joe has homepage identified by URI [http://www.example.org/~joe](http://www.example.org/~joe)” [12]*

RDF is the basic building block for the supporting the Semantic web [12]. RDF stands for Resource Description Framework and is the primarily intended for representing the metadata about the WWW (World Wide Web) resources using URI (Uniform Resource Identifier). In RDF the information is represented in a graph form using triples subject-predicate-object where all the resources are defined by their URI.

RDF is a XML based language recommended by W3C (WWW Consortium), which is capable of describing any fact/resource about any domain. RDF is structured and machine-understandable which makes it possible for computers programs to do useful operations with the represented knowledge [12].

Even though RDF can be used to describe any domain knowledge with the help of subject-predicate-object triplet/statement, but RDF lack other vocabularies to define classes, sub-classes, class member variables and also the relation between the classes. Hence, W3C then recommended RDFs (RDF schema) as a language that can define such vocabularies and which adds more semantics to RDF predicates and resources [12]. RDFs also extend some of definition of RDF elements, for example- it sets the domain and range of properties and relate the RDF classes and properties into the
taxonomies of RDFs vocabulary. RDFs thus can be thought as both syntactic and semantic extension of RDF. The root class of everything in RDFs is the rdfs:resource with following URI:

http://www.w3.org/2001/01/rdf-schema#resource

2.7.4 Web Ontology Language (OWL)

Owl is the latest recommendation of W3C and is arguably the most popular ontologies development language. OWL is the semantic markup language for sharing and publishing the ontologies in web. The simplest of the mathematical definition for OWL given in [12] is as follow:

OWL = RDF schema + new constructs for expressiveness

So, while developing ontologies with OWL, all the classes and properties provided by RDFs can be used. In fact, OWL gives much more complex and richer expressiveness in detailing the relationship between classes and other attributes. The root of all classes in a ontology developed using OWL is a owl:thing with URI as follows:

http://www.w3.org/2002/07/owl#Thing

The added constructs in OWL (e.g., cardinality constraints, someValuesFrom, allValuesFrom, hasValue construct etc.) make its inferring power more sophisticated and advance than before.

2.7.5 JENA- Ontology Framework

Jena is java language based framework for working with semantic web applications. It provides a collection of tools and libraries to build semantic web application and linked-data application, different semantic tools and server. Jena is the top-level open source Apache project, which was originally developed by researchers at HP-Labs, UK, 2000 [28].

From developer’s prospect, Jena provides an extensive java libraries/API for developing codes to handle RDF, RDFS, RDFa, OWL, SPARQL and varieties of storage strategies to store RDF triplets. It also includes a rule based inference engine, which can perform reasoning based on OWL and RDFS ontologies [28]. Apart from that, Jena has a comprehensive documentation for developers with tutorials and examples in its official website [29].

2.7.6 SPARQL-Query Language

SPARQL is a data-oriented query language and protocol to query RDF data models. It is W3C recommendation query language for Semantic web. It can pull values from structured as well as semi-structured data and explore the unknown relationship in data. It can be used to pull the data from RDF data vocabulary and transform it into another RDF vocabulary [30].

A SPARQL query can consist of following [30]:

- Prefix declarations – used to abbreviate URIs
- Dataset definition – giving information about the RDF graph being queried
- Result clause – pointing out what information the query will return
- Query pattern – specifying the query for the underlying dataset
Theoretical Background

- Query modifier – ordering, slicing, grouping and other result formatting.

2.7.7 Inference Engine/Reasoners

The significant benefit of ontology and semantic web is the power of inferring new facts based on the set of asserted facts or axioms. Such a software program, which does the task of inferring logical consequences and generating new facts, are called Inference engine or reasoners. They are rule based inference engine and works mostly on OWL description logics, for example, FACT++ (an open source, C++ based reasoner), Pallet (open source java based OWL DL).

2.7.8 XSL and XSLT

XSL stands for Extensible Style-sheet Language (XSL) and consists of two parts: the XSL Transformation (XSLT) and XSL Formatting Object (XSL-FO). As XML is only a markup language, if formatted output or any conversation of data from XML format to other is needed, a XML transformation application can be used. Such a transformer is known as XSLT (XSL Transformation) and consist of the XSLT processor which compares the elements and other nodes in an XML input document to the template-rule patterns in a Stylesheet, and further serialize the output into the prescribe format. XSL can be used to transform XML to plain text, html, RDF etc. In order to match the nodes, XSLT uses XPath, which is a non-XML language for identifying particular parts of XML documents [6].

2.8 The role of Ontology in data source integration

According to [4], there are five distinct use of ontology in the area of data integration. They are listed as below:

Metadata Representation: The metadata (data about the data) of the source schemas to be integrated can be represented in a local ontology using any formable single language like owl or RDF.

Global Conceptualization: The single top-level global ontology can be prepared to give the conceptual view of all the schematically heterogeneous source schemas of the data being integrated.

Support for High Level Queries: The global conceptualization of the all the data sources with a single ontology helps in formulating the queries without the particular knowledge of any of the source schemas.

Declarative Mediation: in the hybrid peer-to-peer systems, query processing uses the global ontology as a declarative mediator for query reformulating between the peers.

Mapping Support: While integrating the data-sources, the ontology can be used as thesaurus or collection of vocabularies for the automation of mapping process.

2.8.1 Ontology base data integration approaches

Base on the [35], mainly following three approaches are used to integrate data sources using ontology.
2.8.1.1 Global Ontology Approach

In this approach, a single global ontology is used to describe the data in all the sources to be integrated and query to the system is based on this global ontology. The process of designing the global ontology is very difficult, as it needs the developer/designer to have all the information about the semantics of the data structure used in various sources to be integrated. Domain experts need precise designing of the global ontology and it is always hard for system to adapt to the changes of merging data from new data source. In many cases, specialized ontologies from different data source domain are combined to make the global ontology [35].

2.8.1.2 Multiple Ontology Approach

In multiple ontology approach, individual data sources are first represented by corresponding local ontology separately. These local ontologies representing the distributed data sources are then mapped for interoperability and queries to the system are always executed through local ontology by mapping them to each of local ontology. Data integration in global level is not needed and changes to the local ontology or data source does not affect the whole system, which is a definite advantage over global ontology model [35].
2.8.1.3 Hybrid Ontology Approach

As the name suggests, in this approach the advantages of both the approaches are combined with the intention to overcome the drawbacks the previous approaches have. As in multiple ontology approach, each of the source ontologies are firstly represented by the corresponding local ontology and a share vocabulary is built around the local ontologies to make them interoperable. Queries to the system are always mapped and executed from the shared vocabulary [35].

2.9 Summary of Literature Review

The literature review that we did for the research gave us the foundation knowledge about the problem and relevant work done previously by other researchers. We utilized this knowledge to generate new ideas to solve the research problem in innovative way applying research methods. The literature review is also very important because it helps us not to invest time into doing something that has already been done, rather build-up upon the present available materials on the subject domain.

In chapter 4 and 5 we have utilized our learning from the literature review to come up with the method, model and system architecture of competence profile making system. Rather than trying to build such a system from scratch, we re-used ideas from various systems we studied, chose the best of alternative for data-integration for our purpose and design ontologies based on well knowledge vocabularies etc. We added our own finding into the outcomes of literature reviews to finally build a prototype which works as per the theories we generate and helps us to practically demonstrate the concepts we depicted in theoretical models.
3 Research Methods

The aim of this thesis work is to produce new scientific knowledge in the domain of creating competence profile for researcher. For this, we need to apply accepted scientific research methods while conducting all the research work. The application of research methods help in managing the research in scientific way, and also make the outcome result more reliable. Research methods also help in getting the best solution to the given problem in hand.

Selection of appropriate research method is very hard in the field of Information System (IS) because of lack of proper guidance in choosing best alternative amongst all methods. The researcher should always choose the research method according to the research problem. However, all the research methods in the IS has different known flaws or weakness because of which they only provide limited, qualified evidence about the research problem being analyzed. So, research strategy of most of the researchers is to use multiple methods in such a way that weaknesses of one method is addressed or complemented by other method [13].

3.1 Research methods in Information System

To understand the applicability, strengths and weaknesses of different alternative research methods, we studied the most used research methods in the field of software engineering. The research methods that are used for doing the research process and the methods that are used to produce the artifacts are both studied. We studied Action Research (AR), Constructive Research Process and Design Science Research (DSR) to know their applicability in our research problem and we chose DSR for implementing to our research problem. Below we have given brief introduction to both AR and Constructive research process and then given detail about DSR and its implementation.

3.1.1 Action Research (AR)

Action Research has been widely discussed in the IS literatures [24][25] as one of the research methods applicable to research in software engineering and information systems. According to Rapoport [23], “Action research aims to contribute both to the practical concerns of people in an immediate problematic situation and to the goals of social science by joint collaboration within a mutually acceptable ethical framework”.

This definition of AR gives the dual goal of the method contributing to both practice and research concurrently. Action researchers may start their work with theory free action learning however theories are needed to guide them and make them focus on their learning. AR involves high degree of client involvement that makes it highly context dependent while attempting to address the specific client’s research concern. Such an involvement of client for the research is not always possible in general. Again much of AR is done to understand the existing reality, such as complex working en-
vironment of a big organization and complex human behaviors. Constructing the new and innovative ways of solving problem or class of problem producing a new reality or artifact is not the mandatory purpose of AR [34].

3.1.2 Constructive Research Process

Constructive research is the process of construction, based on the existing knowledge with novel approaches and possibility to add the missing links. The construction procedure starts through the designing work that helps in projecting the vision of future solution to the problem (artifacts, theory), and fills the conceptual/theoretical or other kind of knowledge gap. The creation of knowledge in constructive research process is mainly through interaction between the observer and the observed or real world constructs [31]. In constructive research work, research is done being limited to constrain of the law of world even though the interaction and new conceptualizations often changes the relationships that exist between the observer and the world.

The constructive research approach focus on solving problems by constructing model, diagram, plans, etc. in both the practical or theoretical level to solve a domain specific problem in order to figure out the new knowledge of solving the problem. The constructive research not only focuses on generating the new artifact, but also solves several other problems related to the knowledge gap concerning feasibility, improvement and novelty. In doing so, the main emphasis is always on theoretical relevance of the problem and construct, the elements of solution that are central to the benefits and the way to present the solution in the most condensed form [31][37].

Research in engineering field is more concerned on how to do things and how to create new objects. The main knowledge production is usually occurs in construction phase of the research process. Similarly, in software engineering research as well, the fundamental concern is always constructive. The software engineering research groups often work predominantly in ‘constructive mode’, inventing new tools and models. It is observed that less effort is given in understanding the actual mechanisms of knowledge production in constructive research by using cognitive artifacts. To start the constructive research in software engineering, it is firstly important to have well understood knowledge about the research problem and domain field. So, mostly empirical investigations using either quantitative or qualitative methods are needed to do constructive research in software engineering field [31].

3.1.3 Design Science Research (DSR)

Among the several of the well-known definition of DSR, we picked some of the important ones. Hevner et al. [18] described design science as a paradigm that seeks to extend the boundaries of human and organizational capabilities by creating new and innovative artifacts, including constructs, models, methods and instantiations. In the design science paradigm, knowledge and understanding of problem domain and its solution are achieved in the building and application of the designed artifacts.

Vaishnavi and Kuechler [73] characterize Design Research as “yet another lens or set of analytical techniques and perspectives (complementing the Positivist and Interpretive perspectives) for performing research in IS. Design research involves the analysis of the use and performance of designed artifacts to understand, explain and very frequently to improve on the behavior of aspects of Information Systems.”
Venable [20] identified “solution technology invention” as the core of DSR. They defined DSR as a research activity that invents or builds new, innovative artifacts for solving problems or achieving improvements, i.e. DSR creates new means for achieving some general (un-situated) goal, as its major research contributions. Such new and innovative artifacts create new reality, rather than explaining existing reality or helping to make sense of it. Simon [19] described that the design science paradigm as fundamentally a problem solving paradigm and has its roots in the engineering and the sciences of the artificial.

Design science research aimed at developing executive information systems and systems to support emerging knowledge processes, respectively, within the context of IS design theories. Such theories prescribe effective development practices/methods and a type of system solutions or instantiations for the particular case of user requirements/models [21][22].

Research process in software engineering requires both the process for conducting the set of research activities and a product artifact as a proof of concept. The DSR supports the design process, which is a sequence of expert activities that produces the innovative products/artifacts. The so produced artifacts are evaluated giving the feedback information for the better understanding of the problem that subsequently improves the quality of the product and the design process itself. In DSR, the build-and-evaluate process loop is iterated a number of times till the final design is built making the researchers learn better of both the design process and design artifact [22].

3.1.4 Selection of Research Method

Hevner et al. [18] categorized research in information systems into two scientific paradigms including behavioral science, which is to develop and verify theories, that explain or predict human or organizational behavior and design science which extends the boundaries of human and organizational capabilities by creating new and innovative artifacts, including constructs, models, methods, and instantiations. However, when we do research work in IS research field based on this category, there are two main views around this research’s core focus: a narrow view focusing on the IT artifact as the core subject matter of IS research, and a broad view that focuses on the interplay between social and technical aspects of IT that is embedded into a dynamic evolving context [18]. DSR creates new means for achieving some general (un-situated) goal, as its major research contributions creating a new reality, rather than explaining existing reality or helping to make sense of it.

According to Rapoport [24], AR aims to contribute both to practical concerns of people in an immediate problematic situation and to the goals of social science by joint collaboration within a mutually acceptable ethical framework. This definition assumes that there is a concrete client involvement and is highly context dependent, attempting to address specific concern/problem of the specific client. So, when compared with DSR, an essential difference is that DSR assumes neither any specific client nor joint collaboration between researchers and the client and the solutions given by DSR are generic and useful to other relevant domains as well [78].

On the other hand, it is important to have prior well-understood knowledge about the research problem and the domain field to start a constructive research. Mostly, empirical investigations are done using either qualitative or quantitative methods to understand the domain more clearly [31]. Again, if we compare Constructive research
method process and outcome with DSR method’s process and outcomes, we can easily observe that DSR covers all the things as that of constructive research.

For our case, this research is regarding to information systems, and its detailed purpose is to develop a method and architectures to solve a domain problem with a proof of concept. After both field and theoretical background study, we learned that the related domain contains many research questions and problems. This leads us to create an IS artifact with the purpose of building innovative utility solutions which could answer our research questions. Therefore the research method must be suitable for generating new theory and developing innovative artifacts. Beside, as we are doing a academic research work, we neither have any particular client nor we want to limit our research work to any particular context of user group (i.e we are not limited to making profile of any particular domain/field researcher’s competency profile only).

Hence, after studying the above mentioned research methods and after consulting with our super-visor, we decided that DSR is the best suit for our research purpose and we implemented it for our thesis research.

3.2 Implementation of DSR

3.2.1 DSR Guidelines

After both field and theoretical background study, we learned that the related domain contains many research questions and problems. This leads us to create an IS artifact with the purpose of building innovative utility solutions which could answer our research questions. That’s why we have chosen DSR as our research methodology.

Hevner [18] has indicated that the fundamental questions for design science research method are:

“What utility does the new artifact provide?”

“What demonstrates that utility?”

Regarding to those fundamental questions for DSR, there are seven guidelines (table 3.1) provided by Hevner’s research that guide the research process [18]. We are following some of them later on in our own methodology as well as when we design and develop the artifact.

<table>
<thead>
<tr>
<th>Guidelines</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guideline 1: Design as an Artifact</td>
<td>Design-science research must produce a viable artifact in the form of a construct, a model, a method, or an instantiation.</td>
</tr>
<tr>
<td>Guideline 2: Problem Relevance</td>
<td>The objective of design-science research is to develop technology-based solutions to important and relevant business problems.</td>
</tr>
<tr>
<td>Guideline 3: Design Evaluation</td>
<td>The utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods.</td>
</tr>
</tbody>
</table>
Guideline 4: Research Contributions  
Effective design-science research must provide clear and verifiable contributions in the areas of the design artifact, design foundations, and/or design methodologies.

Guideline 5: Research Rigor  
Design-science research relies upon the application of rigorous methods in both the construction and evaluation of the design artifact.

Guideline 6: Design as a Search  
The search for an effective artifact requires utilizing available means to reach desired ends while satisfying laws in the problem environment.

Guideline 7: Communication of Research  
Design-science research must be presented effectively both to technology-oriented as well as management-oriented audiences.

<table>
<thead>
<tr>
<th>Table 3.1: Seven DSR guidelines [18]</th>
</tr>
</thead>
</table>

In the following section we will show how we used these guidelines that are closely related to our research process and explain how we followed them and how does guideline helped us to improve our research quality.

3.2.1.1 Implementation of DSR guidelines

For properly applying the Design Science Research Methodology into our research, it is very important to follow DSR guidelines. Hevner quoted that the guidelines that he suggested from his research are of necessity, can be adaptive and are process oriented [18]. In the following section, we will choose some guidelines that are closely related to our research process and explain how we followed them and how does guideline helped us to improve our research quality.

**Design as an Artifact**

Hevner [18] indicated that Artifacts are innovations that define the ideas, practices, technical capabilities, and products through which the analysis, design, implementation, and use of information systems can be effectively and efficiently accomplished.

As we introduced in the very beginning, our research objective is to generate a method, architecture of system for automatically creating competence profile and develop a prototype to prove this method and architecture works. To develop the method and system architecture of such a system, we studied some of the existing system to know their model and how they work so that we could work on the shortcoming they have. Since this method, architecture and proof of concept are going to solve particular problems and also answer our research questions, we can consider our objective to develop an artifact according Hevner’s definition of artifact above.

**Problem Relevance**

In the theoretical background and introduction chapter we have cleared the field background, existing systems, actual needs and problems present in the domain. In order to fully follow the guideline: Problem relevance, we also analyzed the problem
and requirements step by step. In the chapter realization, we improved our research objectives and specific requirements after all the analysis of problem relevance to make sure that our work is going to solve exact problem. In chapter 5, we also demonstrated a technology-based implementation of simple prototype of the system according to method and system architecture we built for the system which also addresses the refined research objectives.

**Design Evaluation**

The goodness and efficacy of an artifact can be rigorously demonstrated via well-selected evaluation methods [18]. The evaluation of designed artifacts typically uses methodologies available in the knowledge base. These are some of evaluation method we used in our work, like observational evaluation, analytical evaluation, and so on. The selection of evaluation methods must match appropriately with the designed artifact and the selected evaluation metrics. And we followed all those evaluation methods for both method and artifact during application of the DSR activities in chapter 5 (5.4 System Analysis and Evaluation).

**Research Contributions**

Criteria for assessing contribution focus on representational fidelity and its applicability. Artifacts must accurately represent the business and technology environments used in the research, information systems themselves being models of the business. These artifacts must be implementable, hence the importance of instantiating design science artifacts. Beyond these, however, the research must demonstrate a clear contribution to the business environment, solving an important, previously unsolved problem. After checking the present field situation and existing systems, in the chapter realization and result we ensured that our work is innovative and will bring new knowledge to this filed. So, we can say that our main contributions to this domain of research are:

1. Our unique model, methodology and architecture that addresses the research objectives.
2. The prototype/proof of concept, which demonstrate the ideas we presented do work.

**Research Rigor**

First, our research draws from a clearly defined and tested base of modeling literature and knowledge in theoretical background. Then we have used rigorous research approaches in both the development and evaluation phases. For e.g we developed our knowledge-based ontology using a well-defined and developed ontology development methodology, also we evaluated our artifact to test its accuracy according to our objectives.

**Design as a Search**

Problem solving can be viewed as utilizing available means to reach desired ends while satisfying laws existing in the environment [19]. The DSR cycle guided us to revisit and search the knowledge base (theoretical background where we can find scientific theories and methods) till we were determine that the solution is achieved. We studied the relevant literatures and other scientific publication documents to find best solution (to our capabilities) for the research problems we were trying to answer.
Communication of Research

Throughout our research, we documented the process we followed to develop the artifacts which addresses the research objectives. In chapter 4, we presented our entire theoretical findings and realizations for the broad range of audience in most simple non-technical form whereas in chapter 6, we have given the technical details of the system we developed specifically for the audiences who have technical background.

3.2.2 Design Science Research Methodology and Activities

In order to contribute to the domain we are doing the research in, and to create an artifact that solves the field problems, we are using DSR methodology. The two basic objectives of DSR Methodology is to improve the production, presentation, and evaluation of design science research while being consistent with the principles and guidelines of design science research established in previous research studies as mentioned in [18][40][41].

Peffers [42] created the design science research methodology or DSRM with three objectives in mind:

1. Provide a nominal process for the conduct of DS research;
2. Build upon prior literature about DS in IS and reference disciplines
3. Provide researchers with a mental model or template for a structure for research outputs.

For accomplishing objectives mentioned above, we used six activities that make up the DSRM as a nominal sequence, which was slightly different from Peffers’s research [42]. This table shows each activity in detail: “what to do” and links the knowledge base with the different activities: “how the activities are executed” [43].
<table>
<thead>
<tr>
<th>DSRM activities</th>
<th>Activity description</th>
<th>Knowledge base</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem identification and motivation</td>
<td><em>What is the problem?</em> Define the research problem and justify the value of a solution</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Understand the problem’s relevance and its current solutions and their weaknesses</td>
<td></td>
</tr>
<tr>
<td>Define the objectives of a solution</td>
<td><em>How should the problem be solved?</em></td>
<td>Knowledge of what is possible and what is feasible. Methods, technologies and theoretical background</td>
</tr>
<tr>
<td></td>
<td>In addition to general objectives such as feasibility and performance, what are the specific criteria that a solution for the problem defined in step one should meet?</td>
<td></td>
</tr>
<tr>
<td>Design and development</td>
<td><em>Create an artifact that solves the problem.</em></td>
<td>Application of methods, technologies and theory to create an artifact that solves problem.</td>
</tr>
<tr>
<td></td>
<td>Create constructs, models, methods, or instantiations in which a research contribution is embedded.</td>
<td></td>
</tr>
<tr>
<td>Demonstration</td>
<td><em>Demonstrate the use of the artifact.</em></td>
<td>Knowledge of how to use the artifact to solve the problem.</td>
</tr>
<tr>
<td></td>
<td>Prove that the artifact works by solving one or more instances of the problem.</td>
<td></td>
</tr>
<tr>
<td>Evaluation</td>
<td><em>How well does the artifact work?</em></td>
<td>Knowledge of relevant metrics and evaluation techniques.</td>
</tr>
<tr>
<td></td>
<td>Observe and measure how well the artifact supports a solution to the problem by comparing the objectives with observed results.</td>
<td></td>
</tr>
<tr>
<td>Communication</td>
<td>Communicate the problem, its solution, and the utility, novelty, and effectiveness of the solution to researchers and other relevant audiences.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Knowledge of the disciplinary culture.</td>
<td></td>
</tr>
</tbody>
</table>
3.2.2.1 DSRM activities

In this section, we have given brief explanation of how the activities of DSRM applied in our research work.

**Problem identification and motivation**

We have discussed about the problem and challenges in section 2.5 and 2.6.

**Define the objective of a solution**

The main objective of this thesis work is to develop a methodology and design an IS artifact that is able to automatically create researcher's competency profile based on semantic integration of different data source and figure out the main challenges involved in it.

**Design and development**

In the course of doing this thesis work, we designed an IS artifact with a method for automatic creating competence profile of researcher by integrating information from heterogeneous bibliographic data sources, an system architecture based on the method and the working prototype including an ontology scheme for researcher’s competence profile.

**Demonstration**

The method for automatic creation of competency profile and layered architecture both have been presented in chapter 4, and the detail of implementation of the artifact is demonstrated in chapter 5.

**Evaluation**

We take Click Project as the demand of this artifact and result of this research. Therefore the functional performance of our result that generated by survey of participants in Click could be the evaluation result of this research.

Beside, as an artifact, our research result can be evaluated in terms of functionality, completeness, consistency, accuracy, performance, reliability, usability, fit with the organization, and other relevant quality attributes [18]. Hence, there are some DSR evaluation methods that are also available for this research such as the ones shown in table below:

<table>
<thead>
<tr>
<th>Design Evaluation Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Observational</strong></td>
</tr>
<tr>
<td><strong>Case Study:</strong> Study artifact in depth in business environment</td>
</tr>
<tr>
<td><strong>Field Study:</strong> Monitor use of artifact in multiple projects</td>
</tr>
<tr>
<td><strong>2. Analytical</strong></td>
</tr>
<tr>
<td><strong>Static Analysis</strong> (e.g., complexity); <strong>Architecture Analysis</strong> (e.g., performance); <strong>Dynamic Analysis</strong> (e.g., performance);</td>
</tr>
<tr>
<td><strong>3. Experimental</strong></td>
</tr>
<tr>
<td><strong>Controlled Experiment</strong> (e.g., usability); <strong>Simulation</strong> (e.g., performance); <strong>Execute artifact with artificial data</strong></td>
</tr>
<tr>
<td><strong>4. Testing</strong></td>
</tr>
<tr>
<td><strong>Functional</strong> (Black Box); <strong>Structural</strong> (White Box)</td>
</tr>
</tbody>
</table>
5. Descriptive

| Informed Argument (e.g., relevant research); |
| Scenarios: Construct detailed scenarios around the artifact to demonstrate its utility. |

Table 3.3 DSR Evaluation Methods [18]

For our final chosen evaluation method for this research, we will describe and discuss later in chapter 6.

Communication

As this paper is going to be uploaded in DiVA as a research article, other researcher who wants to make contribution in the relevant field could review it and get some information.

3.2.2.2 Design Science Research Cycle

![Design Science Research Cycles](image)

Figure 3.2: Design Science Research Cycles [72]

The DSR activities are framed by the Design Science Research Cycles [72]. As we followed the DSR our research involved two main steps:

1. Requirements and evaluation criteria analysis.
   
   In order to contribute an innovative artifact as DSR requires, and fulfill our purpose of this thesis work, we needed to define the requirements for our work and the acceptance criteria for evaluation.

2. System design.
   
   Based on the requirements specifications from step 1, we designed and implemented our prototype of this Automatic profile creation system.

3.3 Design options for system development

To develop the artifacts as the proof of concept for our thesis work, we chose following software development methodologies.
3.3.1 SCRUM-Agile Software Development Methodology

Among the software developers, scrum is one of the most popular forms of methodology in agile framework beside XP in recent time. “Empirical process control” is the feature that makes scrum unique among the other agile methodologies. While implemented in any project, scrum uses the real-world progress of a project and not just a best guess or any forecast to plan and schedule the releases. Scrum is an innovative way of getting the work done so not only in software development; it works well in other work scope as well [32][33].

Iterative and incremental development of software in phases is the main feature of any agile methodology. In scrum such small iterative phases or the basic unit of development are called sprints. A sprint in a scrum generally last from one to four weeks period only. A meeting is organized at the end of each such sprint to assess the progress of project and also to plan for works to be done in up-coming sprints. Knowing the state of ongoing works in the project gives advantage to project manager to flexibly decide and adjust the next steps. Though the method gives project manager and the developer’s flexibility to adjust and re-orient the direction of project, but the set of roles of individuals, their responsibilities and the schedule meetings are not changed in general [32][33].
We will describe in detail about the different process iterations/activities done for software development as an artifact for our thesis in detail in the chapter 5 (Section 5.2 and 5.3).

### 3.3.2 Methontology methodology for building Ontology.

As our research work also involves the construction of ontology for researcher’s competence profile, we also studied Methontology methodology used in ontology engineering.

Ontology Engineering has received much of attention during past years to enhance and develop a robust and reliable ontology building process, namely its life cycle and methodologies to guide the construction of ontology [54]. There had been a series of approaches reported for developing ontologies; SENSUS, On-to-Knowledge, Methontology methodology to name few. Among them, Methontology methodology is among the most popular one to build ontology. It was created by the Artificial Intelligence Lab from the Technical University of Madrid (UPM), for building ontologies either from the scratch level, reusing the other ontologies in their original form, or re-engineering them for adapting to own need [14]. Methontology framework prescribes building ontology from knowledge level. This methodology guides inexperience ontology builder by defining intermediate representations in clear and easily understandable form, which in turn helps in having a focused acquisition and conceptualization activity to easy formalization [54].

![Figure 3.4: Ontology life cycle in METHONTOLOGY [14]](image_url)

The METHONTOLOGY framework [14] consist of identification of the ontology development process, a life cycle based on evolving prototypes, and particular techniques to carry out each activity.
Ontologies can either be built from scratch or by reusing other available ontologies. Different ontology building process is associated with different types (Representational, general or upper-level ontologies, Domain ontologies) of ontologies. Methontology supports both the process of building ontology.

### 3.3.2.1 Ontology Building from Scratch

The roots of this methodology for building ontologies are mainly from the IEEE software development process and other knowledge engineering methodologies [15][54], therefore, the terminology related to development and support in Methontology is analogous to that of software development methodology. The table 3.4 below shows the general comparison of the activities involve in ontology development from scratch using Methontology methodology and that of software development life cycle. The process starts with different knowledge acquisition techniques to gather the requirement specification followed by conceptualization of those knowledge objects according to the domain. Evaluation of the developed ontology finishes the cycle of development, which is to be continued by another cycle. As Methontology is second generation of ontology development methodology, key point in its development cycle is the maintenance phase that was not present in the first generation methodologies [54].

<table>
<thead>
<tr>
<th>Methontology activity</th>
<th>Corresponds to</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requirement specification</td>
<td>Specification</td>
</tr>
<tr>
<td>Conceptualization of domain knowledge</td>
<td>Conceptualization</td>
</tr>
<tr>
<td>Formalization of conceptual model</td>
<td>Formalization</td>
</tr>
<tr>
<td>Implementation of formal model</td>
<td>Implementation</td>
</tr>
<tr>
<td>Maintenance</td>
<td>Maintenance</td>
</tr>
<tr>
<td>Knowledge acquisition</td>
<td>Knowledge acquisition</td>
</tr>
<tr>
<td>Documentation</td>
<td>Documentation</td>
</tr>
<tr>
<td>Evaluation</td>
<td>Evaluation</td>
</tr>
</tbody>
</table>

*Table 3.4: Classification of METHONTOLOGY activities [54]*

### 3.3.2.2 Ontology Building with reuse (fusion and composition)

Ontology reuse is an activity in the development phase of ontology, which generally lies in between formalization and implementation stages [54]. Ontology reuse is the process of building new ontology either by getting ideas and concepts from existing ontology in the same domain or completely reusing and specializing the existing one. According to [56][57], there are two ways of reusing ontologies:

i) **Fusion/merging:** In fusion, the knowledge from different source ontologies from the same domain/subject is unified to make the new ontology. As the new ontology is just build from the knowledge of the source ontology, the new ontology can’t be considered as the simple revisions, improvements or variations of the source ontology. It
is generally very hard to identify the regions that were taken from the source ontologies. In order words, the knowledge from the source ontologies is homogenized. The source ontologies normally may have much of overlapping in the knowledge of which only the distinct concepts are chosen.

**ii) Composition/integration:** In case of composition, two or more source ontologies from different subject/domain are integrated in order to build the new ontology, which satisfies the context. The source ontology goes through the process of extension, specialization or adaptations to form the new ontology. After composition, the regions in the new ontology can still be identified for their sources. As source ontologies on different subject/domains are chosen to make new ontology, the overlapping of the concepts in this form of ontology building is small.

![Ontology integration process](image)

Figure 3.5: Ontology integration process [54]

Well known ontology and vocabularies like People.Owl, FOAF etc. can be used for making Researcher Ontology that actually increases the interoperability of the built ontology. For building the competence model for researchers using METHONTOL-OGY methodology, the recommendations (see appendix) given by [16] can be uti-
lized. The implementation process we followed for building ontology for this thesis work will be explained in detail in chapter 6 (Implementation).

3.4 Research Framework

After studying in detail about the problem domain, literature review and previous work done in the same field, research method that is applicable to our research domain and the software development methodologies that we need to develop the proof of concept which our research method demands, we developed a research framework as a roadmap to conduct this thesis work. This framework helped us to understand the development procedure that we needed to follow in order to accomplish the research work in a scientific way.

DSR addresses important unsolved problem in unique/innovative way or get more efficient and effective ways of solving the problems which makes the research work different from professional design and development work. The design science research methodology prescribes to conduct the research activities to problem identification and motivation for doing the research. In the literature review phase, not only we identified the problems in the domain, but also cleared our objectives of building the competence profile system. The analysis done on problem area, study of similar artifacts helped us build a method for developing a competence system in automatic way by integrating data sources and the system architecture, which we then used for developing various models using UML. So developed UML models helped us in specifying the requirements for the system development to demonstrate that our ideas and finding should work. We chose scrum software development methodology to develop the software artifact and as we design the system to use ontology as core knowledge base, we followed Methontology methodology to build the system’s core ontology. The successful implementation of prototype we developed demonstrated that the researcher’s profile could be generated from the bibliographic meta-data from different source; by integration and semantically processing them to have a clear visualization of the competence model. The pictorial depiction of research framework (figure 3.6) we followed for this research work shows the process involved in each step.

Our research process started with the theoretical literature reviews and studying/observing the existing systems, which resembles our thoughts of the artifact capable of making researcher’s competence profile, and generating a theory that such system is practically possible; which is a typical inductive approach of research. Again, we then used the theory and logically confirmed that the idea actually works by implementing a system possessing the features that we wanted to focus; which is a typical way of accepting or rejecting hypothesis in deductive research approach. Thus, to our knowledge, we believe that we have used both the inductive and deductive orientation of research in our thesis work.
Figure 3.6: Research Framework
4 Realization

Two important characteristics of design science artifacts are relevance and novelty. First, an artifact must solve an important problem: i.e., being relevant. Second, to differentiate design science research from routine design [43]. Hevner [18] suggests that design science research should address either an unsolved problem in a unique and innovative way or a solved problem in a more effective or efficient way. Because of this, and what we have introduced above, the guidelines 1 and 2 indicated a requirement for our research to present an innovative, purposeful artifact for a specified problem domain. Regarding to our DSRM activities, there is a need for problem identification and motivation. Therefore, it is necessary to run a field background check and try to locate out specified problem domain and be more particular about our purpose of a new solution for specific problem that haven’t been solved yet.

In this section, firstly we describe the reasons of why we have chosen those existing systems to study in Chapter 2 and how actually we benefited from them. Then we give the description of analysis of the existing profile creation systems developed by some other research institute and Universities. As the guidelines of DSR suggest us to study the problem domain and what has already been done in the field, it is very important for our thesis to see some of the previous relevant projects and analyze their work and figure out some specific problem that we can focus on. Then in the subsequent section we proposed a layered architecture for the Profile Creation Software system and the designing of the competence profile for research in ontology language.

4.1 Study of Existing System approaches

As a part of knowledge base, study of existing approaches are important to define the knowledge of what is possible and what is feasible, and to bring the knowledge of methods, technologies, and theories that can help with defining the research objectives [43]. As we mentioned in the previous sections, there are three main kinds of approaches for competence profile creation:

1. **Manually profile creation**
   
   ![Manually profile creation](Diagram)
   
   Manually profile creation approach usually offers a form for the user to fill, and then generate a profile based on the data that users provide.

2. **Semi-Automatic creation**

   ![Semi-Automatic creation](Diagram)
   
   Semi-Automatic profile creation approach uses user feed data and integrate it will some other data from different sources with manual intervention to create the user profile.

3. **Automatic creation**

   ![Automatic creation](Diagram)
   
   Lastly, in automatic profile creation systems, users profiles are created based on data from some predefine source of data repositories. Mostly the procedure is automatic, system only need basic information regarding name details of the person for whom it should generate profile. Among many different automatic creation approaches, PRNS
and VIVO are the most typical ones, which create profiles by extracting data from pre-defined sources. These two systems can be taken as representatives of the automatic profile creation systems. Therefore we analyze them in detailed and described below.

In general, there is no doubt that these three kinds of approaches have their unique features for better profile creation, and because of those distinct features they are wildly used in different information systems for different needs. However, regarding to our research topic, we are focused to create better competence profile for researcher based on researcher’s publication data from heterogeneous publication data sources. Therefore we need to analyze the approaches/features of all the kinds of profile creation systems to be able to realize the gap where we could contribute an innovative solution. So, even though the manually profile creation should not be our concern, there are still some interesting parts of some manually creation approaches that we could study and use.

Obviously the biggest difference between three existing systems is the basic model of getting data: user feed data; user feed data plus data source and maintain staff; and lastly just use data direct from data source. Each model has its own strength and weakness, regarding to the reliability and comprehensiveness. The study of all the existing approaches helped us to complement/address some of the shortcoming of presently available automatic systems. Beside, the study of manual and semi-automatic approaches also helped us learn how we could build a model/method for automatic profile creation where human intervention for profile creation is minimal.

4.2 Special features and delimitation of existing profile creation approaches

4.2.1 Profiles Research Networking Software

Data source: PRNS uses PubMed and Harvard’s internal Database as the source.

Type of Profile Creation: PRNS offers both automatically and manually profile creation, which it called “passive” and “active”.

Profile competency: It used the publication’s keywords as base for researcher’s competency profile.

Special features: The main features of PRNS are time line, visualization, and concepts cloud.

Other features: PRNS has built in Author Disambiguation Engine.

- Algorithms analyze publication data to define a researcher's professional interests with a set of prioritized keywords.
- Offer a "Disambiguation Engine" system to self-populates the individual researcher overviews in Profiles RNS, and identifies the specific keywords that characterize each researcher.
- For anyone that tries to research this domain, the algorithms of disambiguation are most important thing. And there is no 100 percent guaranteed algorithm that resolves ambiguities.
Limitations:

- Cannot add new data source
- Update must be done manually
- Local database
- Cannot generate researcher’s contact information from web

4.2.2 VIVO

For VIVO, it used closed database that come from the participate universities. The interesting part of VIVO is that it offers an RDF-based standard for sharing information about researchers.

![VIVO attributes](image)

**Figure 4.1: ViVO attributes**

**Limitation:** The ontology it provides has missed some parts that we think could be important to make the researcher profile rich in information. As shown in the figure above, ViVO focuses mainly on author’s collaboration and hence does not include some of the personal attributes.

4.3 Problems and Solutions

In this section, we will describe the problems and the potential solutions that we figure out after studying the problem relevant literature and previous works in the field.

4.3.1 How can we resolve the syntactic and schematic heterogeneity of data in bibliographic data sources?

These are the basic problem that all heterogeneous data sources’ schemas have. In our system design, we first need to transform the data output from these sources to a common data model, which we have chosen as RDF for general model (XML for the proof of concept). So, while doing the data model transformation to RDF, we will deal with both the syntactic and schematic heterogeneity problem by mapping them to our core OWL ontology knowledge base’s schema, syntax and semantic of each of the element.
4.3.2 How can we resolve the semantic heterogeneity problem in metadata representation of data sources?

As mentioned previously, semantic heterogeneity is one of the hardest problems to solve and needs more manual study of the data source schema as well as understanding of context meaning of the data. Then only, while transformation to RDF model, we will be able to map the data semantically to our Knowledge base to solve the problem. Proper data analysis of each of the source is the key to this problem; hence immense study of available documentation of data sources, whitepapers, API documentations etc. is needed.

4.3.3 How can we solve the author name ambiguity problem?

We studied several of the approach that has been researched for resolving the name ambiguity problem of researcher. Many of these techniques have produced good results in closed domain but for multiple data sources, the problem still remains unsolved to a larger extend. Below we list down some of the solution, which we think could be applied for resolving this problem:

- Metadata in most of the data sources also contain email address of the paper publisher, which are unique identity. But for the older papers might not contain email address or some of the researcher might not have included their email id in the research paper.
- The different affiliation of researcher can give us some idea about the identity of the researcher.
- To identity that the different papers have been written by same researcher, the topic correlation (same author name and related topics) and web correlations [26] can be one way to resolve the problem.
- Researchers have tendency to cite/refer to their previous work, so this bibliometric feature can be exploited.
- Authors and co-authors correlations in the papers, as discussed in [77].
- Using ontology for name disambiguation, as discussed in [76].
- Several of machine learning algorithms have shown enormous potential in the task of automatic name disambiguation. So, both the supervised and non-supervised machine learning algorithms could be the potential problem solver.

These solutions might be used in combine to obtain better results.

4.3.4 How can we identify the identity of a document existing in multiple sources?

There is a high probability that the same publication has been published in multiple bibliographic publications and while extracting data about the researchers from multiple sources, there will be redundancy of some information. The title, keywords and authors name could be compare and then the documents could be uniquely identified to figure out their existence in the multiple data sources. Different text analysis for similarity comparisons rather than exact matching is needed in order to solve this problem, as we can’t neglect the human error and inconsistency in data entry phase in bibliographic record maintenance process.
4.3.5 How can we maintain/update the profile of researchers in our knowledge base to make it more up-to-date?

The competency of researcher is dynamic in nature. The researcher might be involved in different research area and in course of time, they will surely publish papers of their new accomplishment, which might include the new research area too. Again, the profile creation software that we aim to build now is incorporating data from only few data sources available in web. There are many more bibliographic data sources maintained by different universities, research institute and other organizations which might contain some different papers in new research field previously not contained in the data sources integrated. So, when we aim to incorporate a new data source into our system, we have to reflect the profile for any possible addition of competence information available from new system.

For resolving this problem, we have decided an innovative solution of creating a database to store the basic information of researchers, which can also be used to generate all the unique user queries to the bibliographic repositories in order to get publication records as needed. So, instead of updating the user profile, we can drop the initial knowledge base (instances only), and then recreate all the profiles executing the profile creation query for each entity in the database. Generally, the resources in terms of time and other processing required for creating new profile is very much close to updating the profile.

4.3.6 How can we give relevance weightage to competence level of researchers based on their publication only?

It is very important for the quality and reliability of competence profile of researcher that only the relevant, main research area and expertise of researcher is depicted in the profile. It is often found that researchers are partly involved in many of the research area, and publications related to such research fields may also be found associated with them. Again the research done in multi-disciplinary field may list the researcher in the research field that he/she is not the expert of. The most of the bibliographic data sources lack the information about the contribution of individual work when the publication has multiple authors. Because of the above mention problem scenario, it is a difficult task to figure out the genuine competence profile retrieving and integrating information from bibliographic data source.

One possible solution to this problem is categorization of research papers based on the related research area. Their contribution to the field in terms of number of publications gives the general idea about relevance of publication for making researcher’s competence profile. This is also very complex problem and needs a throughout analysis. Some of the factors that could be considered while giving weightage to the competence level of researchers are listed below:

- Number of papers published in same research area (categorization and keywords count).
- Number of years the researcher has been active in certain research area field
- Number of times the researcher has published any paper with another co-authors having competence in certain field (relevant weightage).
4.4 Method for Competency Profile System Development

The theoretical background studies, analysis of the existing system and the problem they have, helped us to have a clear understanding of how a competence profile system should be built. We applied the knowledge to develop a method that can be applied general to all the similar systems.

The fundamental basis for building competence profile of researcher begins with the assumption that research papers reflects the knowledge and skill sets of the researchers which they gain from their work experience and studies. The bibliographic repositories use well specified keywords for every publication in order for them to be retrieved based on those keywords. These keywords in most of the cases reflect the main contribution fields of the research paper. From there, we draw an assumption that such keywords also reflect the research fields that the authors are competent in. The developed methodology helped us to come up with the design decisions for the proof of concepts.

In the context of integrating bibliographic data sources, either due to technical or organizational rules/laws of exposing entire data, materialize view approach (cf. section 2.4 Data integration approaches) is not suitable. The huge collection of metadata of bibliographic publication is the sole property of organization, and organizational laws determine exposing data to outside world and licensing rules. So, to make competence profile of the individual researcher, queries to get data about specific data information about researcher can be used as in virtual view approach or the on-demand profile creation approach. So, rather than integrating the entire data of all the data sources only the certain views of data (specific to particular researcher) from each source is taken and integrated upon the request/registration of researcher into the system.

The whole process we followed are listed below -

- Study of data-sources
- Development of competence profile Ontology for researcher
4.4.1 Study of the Data sources

The whole purpose of studying data sources is to get the knowledge of information granularity the targeted data repositories have regarding the research publications. It helped us to determine whether or not those meta-data are good enough to draw competence profile for the researchers. Based on what sort of information we get from these sources, we will be able to design the ontology schema for the system. According to the level of data/information granularity available from the sources, the conceptual model and the schema of the ontology should be designed.

Again for data integration purpose, it is vital to understand the core meaning of each of the data attributes used by different heterogeneous system for them to be mapped with each other. Normally, the bibliography data repository also provides the meaning of the data attribute they have been using.

The main questions that the data source study will answer are as follow-

- What kind of data access privilege is available?
- What type/format of data is given back by the services?
- What are the content of the data we get from each repository?
- Which of the data elements can be used for building competence profile?

4.4.2 Development of competence profile Ontology for researcher

The ontology acts as the knowledge base where integrated information is to be stored. Adding to that, ontology also helps to get new insights from that information by the virtue of inference. Again, class and concepts of the ontology depends/reflects on what sort of information we can extract from the targeted data sources. The final ontology built is supposed to have all information necessary to depict the researchers’ competence profile in detail form.

The other important aspect while building the ontology is its semantic interoperability with some other well-known vocabularies and ontology. Extending and re-using such well-known vocabularies ensures the interoperability as well as helps in maintaining the structural patterns. As we are using the Methontology, it supports both building ontology from scratch or reusing and fusion with existing ones.

4.4.3 Development of Researcher registration system

The aim of the automatic profile creation software is to get least of the information from researcher and built the competence profile for them. For doing so, we need a process of registering the researcher into our system where researcher are prompt to give some of the basic information about themselves. The information provided by
researcher plays a vital role in name disambiguation issue. For an automatic system, it is very important that the system is fed with adequate information for it to be able to analyze and generate all information, but at the same time it should not be a daunting task for researcher. Hence only certain vital information such as name details and profession should be taken as input to the system or made mandatory while registering user. Based on that information, the system will make queries for individual data sources to fetch data and generate the competence profile.

Just for storing the used fed data to system, a relational database system can be developed and used to store every researcher’s general detail permanently.

4.4.4 Development of Data access module

The different publication houses have different policies and rules for exposing the stored-data to outer world. The owner publication house determines the amount of data access privilege to the outside users/system. The publication repositories may expose their data in different forms/format and provide different access-mechanism; for example, the possible format of data from different sources might be xml, csv, sql tables etc. and the possible data access mechanism may be the web-services, RSS Feed, direct access to data source via FTP. Hence, the development of any system, which depends on data from such kind of publication houses need to determine which kind of available data format and access-mechanism is sufficient enough for their requirements.

Based on the information provided by the researcher and the access level available to the data repositories, the competence profile system sends the queries (web queries using the provided access APIs) to system servers. The queries are prepared as per the specifications given by the service APIs of corresponding repositories.

4.4.5 Parsing/Cleansing the data for all the heterogeneity issues including name disambiguation

The data retrieved from different bibliography repositories have different structures, naming convention, and different granularity level of information even though the main purpose of all the repositories are to make the search criteria easier. The level of information granularity is different. For our purpose, we have chosen only the XML data output format from the different sources, which makes our work little easier.

After getting the data, the first thing we have to look at is similarity in schema of each of the output data. Data elements from each of the sources having similar meaning and information have to be mapped. Matching and extracting only the elements of interest from different data sources to each other primarily makes the data in common data model. Besides the different heterogeneity issues, it is very important at this point that ambiguous publication information to be filtered out from the list of so accumulated publication information. Different machine learning, text similarity matching have to be implemented in order to avoid ambiguous or wrong information into the list of publication.

4.4.6 Transformation of data into common data model (xml/xml-rdf)

This additional step of data transformation is needed only if the original data provided from the data source is not in the XML format. As mentioned above, data repositories may have different access method and may provide data in different format such as
csv, database tables etc. So, if the data output is not in the XML format, we need to convert data into firstly into one common format as XML. XSL transformation can be used to convert different sources into XML.

### 4.4.7 Ontology Population

The common data model that has been parsed, cleansed and transformed (if required) needed to be integrated and populated into the ontology model build for storing the researcher’s competency. In case of XML data format, the data elements from XML can be directly populated into the source ontology with the help of Jena API.

### 4.4.8 Showing result to Researcher

Showing the generated profile to the researcher is the final step of the whole process. The amount of information and the way of data visualization both become equally important for making profile comprehensive and attractive. Different data visualization methods can be adapted such as graph-base data visualization, cloud tags etc to show the competence profile. In our research, we have given primary focus on the method of extraction of data and creation of profile than for visualization.

### 4.5 Layered-architecture of Profile Creation Software

![Layered Architecture of Profile Creation Software](image)

*Figure 4.3: Layered Architecture of Profile Creation Software*

As an intermediate level between the theoretical method that we developed and the practical system architecture specific to our proof of concept (detail given is section 5.1.1), we developed a layered architecture that helped us to visualize the whole system in more intuitive way. In the very bottom of the architecture, we have kept the OWL ontology as our knowledge base. This OWL ontology will have information about the researchers in the form of instances and will be linked to their competence and other information as properties. On top of it, the next two layers will be that of Semantic technology. The Jena Ontology Framework will handle the RDF data model to populate instances generated after merging the data-sources. SPARQL engine will be used to query the underlying OWL file for generating profile while OWL inference engine will be used to infer some new knowledge from the facts present in OWL.
On the top of the layered architecture, we have kept the Data Source Integration module, which will deal with all the challenge in the integration and conversion of data into common model (RDF). This layered architecture helped us build the system design diagram (cf. section 5.1) which shows in detail about the about the system with some technical detail.

### 4.6 Ontology based researchers competence modeling

While modeling the competence profile for researcher in the ontology, we extend the concept of competence given by [17] in order to make it more applicable for our purpose of application usability. We added some more dimension of looking at the competence profile in order to have more information about researcher apart from the bibliometric information.

The competence profile of researcher should in general have all the detail about the researcher’s publication information, scholar activities and the achievements. The profile should have information about the researchers’ affiliation to the various educational or other research related institution. The list of published papers by the researcher and projects the researcher was involved will give the general idea about the work field and accomplishment of the researcher. The general information about the researcher should include information such as name details, contact/correspondence detail. The skill sets of the researcher will be extracted from the different publication produced by researchers.

![Figure 4.4: Composition of Researcher Competence Profile](image)

In our scope for the thesis work, we have described that we aim to produce the competence profile for researchers with thicker information, so that the generated profile could be used for the purposes as researchers’ collaborations and finding researchers in near geographical proximity. Hence, we also include the geographical location component in the structural design of the competence profile of the researchers. In figure above, we show the exception of the conceptual model of the competence profile model (block diagram) for the thesis purpose where the components of competence of researcher are shown as linked classes.
4.7 Reuse of well known vocabulary and ontologies for interoperability

Interoperability of Ontologies is one of the most important objectives for building the Semantic web. Ontology interoperability is mainly concerned with allowing the knowledge exchange between ontologies either specified in the same language or in different semantic language. Ontology interoperability is closely related to Ontology integration and Ontology merging [44].

Combining existing RDF vocabularies and reusing existing OWL ontologies is common practice to save effort and achieve semantic interoperability [47][48]. While building the competence profile ontology using Methontology methodology, we realized that reusing the existing well-known vocabulary and ontologies relevant to our domain would give us immense benefit for interoperability and scaling the output artifact to Web level. Building ontology from scratch and figuring out each and every classes and relation is hard task and it will make the knowledge-base closed one regarding the accessibility. It would also require a domain expert who knows the taxonomy and their relationship with each other while designing the ontology. We studied some of the well-known vocabulary and ontology like FOAF (Friend Of A Friend) [45] vocabulary to describe a researcher’s general/personal information, DOAP (Description Of A Project)[46], VIVO 1.4 etc. to merge and utilized their taxonomy and properties. We have chosen VIVO and FOAF for reusing, and the detail of this procedure is given in section 5.3.

4.8 Refined research objectives

After researching those existing systems, their features and delimitations, we have generated several problems that need innovative or creative solution. Since we have chosen a quite big research domain, and there are plenty of problems in this domain, we have to focus our research purpose to solve specific problems. On the other hand, it is obviously over our capability to try to solve all the problems in the given constraint of time. Thus, in order to present contribution to the field, we must limit our focus and refine our research objectives.

According to the analysis of those two systems above, we learned that there is some common delimitation, which we are aiming to solve.

- Add/Remove data source
- Updated profile
- Ontology refinement
5 Results

In this chapter, firstly we present the results we obtained from our realization of the literature reviews, study of the existing system etc. We draw the system component architecture for the realization of the layered model we initially drew. Then we present the different models and explain about their importance and use for the whole system. Then we described detail of how system development methodology (Scrum) and Methontology was used for the implementation of the proof of concept. We also did brief evaluation of the developed system.

5.1 Result from the realization

5.1.1 System Component Architecture

The above figure gives the high-level system architectural diagram of the competence profile creation system that depicts the specific instant of competence profile system that we built based on the generic layered-architecture presented in section 4.4.

Figure 5.1: System Architecture for competence profile creator system.
Among the two approaches for integration, we chose virtual view approach for data source integration where the users queries will be proceed to generate the local views of data from the individual data sources. Our objective of this thesis work is to generate profile of the researchers as needed or on-demand. So, materialized view approach was not suited for our need. Again, the data amount in well known bibliographic data sources are really huge because of which materialized view approach is not feasible for the scalability of the system.

We purposed a J2EE web application running on top of Tomcat application server. As of this moment, the system integrates data from 3 different data sources as depicted in figure above. Below we describe each component of the system in detail.

5.1.1.1 User Interface
The user interface component of the system basically comprises of two parts; the user registration/login component and the profile visualization part. The process of profile creation starts with user giving basic information (name: first name, last name, middle name) of the researcher whose profile is to be created. A basic User interface is provided for users to enter the query. The so created competence profile of research is also displayed to user in basic web interface.

5.1.1.2 Query Constructor
Based on the information provided by user in the form of query, the “query constructor” module will create individual queries for each of the data sources. Since the most of the bibliographic data sources expose their querying services in the web via web-service, as feed or RSS, APIs etc., the query constructor will be responsible to make the appropriate query module according to service provided by sources. Then the individual queries are sent to corresponding sources and output/extracted data is receive back by the system for further processing.

5.1.1.3 Data Integration
The process of data integration comprise of several other process like parsing and cleansing data, applying algorithms to resolve author name disambiguation problem, populating the data into ontology etc. Apart from that we have also included a profile analyzer module that analyzes profile for giving relevant weightage to competence level while extracting specific profile for researcher.

i) Data Conversion (Cleansing/Parsing): The data received after executing queries in data-sources may be in different format (XML, Comma-separated files, JSON data format etc.) in general. If the sources return back data in different data format, then conversion of data into the same data model is necessary for the interoperability of the data output from different sources and merging them together. In this case, XML data model is the conversion format that is we needed for semantic interoperability of data.

If the data model we have chosen was XML-RDF, then the data received from different heterogeneous sources should be converted into RDF by using different data conversion techniques (for example, XSL is used to transform XML data to RDF). While converting the data received into XML-RDF data model, further processing of data can also be done by transforming the data into conceptually equivalent schema of the ontology that we are supposed to use for competence profile. As all the data conver-
sion from different sources refer to same base ontology schema model, the transform
data model will have alignment in their schema automatically.

The very last process in this module is to semantically integrate such processed data
using different semantic technology for alignment.

   ii) Author Name disambiguation Engine: The integrated publication list must
be processed for any of the ambiguous element that does not belong to the researcher
of interest. A robust author name disambiguation must be implemented for reliable
profile.

   iii) Ontology builder: All the functionality of inserting data into core ontology is
included into this module. Extracted list of publication is inserted into core ontology
after they have been processed by data conversion module.

   iv) Profile analyzer: The last model for system is the profile creation model,
which uses semantic technology like Jena, SPAQRDL and power of inference engine to
produce a competence profile of the researcher from the underlying knowledge base
OWL file. All the task of giving weightage to individual competence level, listing
other attributes is done in this model. The generated profile of research is then sending
to User Interface for user to visualize.

5.1.1.4 Competence Profile OWL Ontology

The knowledge base of the system will be the competence profile OWL file that will
be populated with data instances from the integrated RDF (or XML) data model. The
schema of the OWL file is modeled according to the requirement of producing the
competence profile of researcher with better information availability.

5.1.1.5 Security

The system uses the form-based authentication that verifies the user credential with
the underlying database. This data source security module can be integrated with
J2EE security module to achieve the fine-grained data permissions and user privilege.

Though the system architecture that we build can be generally used for any form of
data type from data sources (with some data-type specific changes), we will focus our
implementation on XML specific data source and hence the later specification given
will be targeted to XML sources only. And also, since our implementation is based on
java platform, so some of our descriptions are java platform specific.
5.1.2 Conceptual Modeling for Competence Profile

Before implementing the Ontological model of researcher’s competence model using ontology development tools like Topbraid and protégé, we first drew the conceptual model of the researcher’s competence profile. Based on the block diagram model we presented on section 4.3, we first realized all the concepts necessary to model the competence profile of researcher in granular level. Then we related the concepts according to the physical relations they have in real world. This model helped us to visualize the semantics and relations between the various concepts that the researcher’s competence profile should comprise of. The figure above is the part of model we drew for conceptual modeling of researcher competence profile. In figure, we can see that the main concept “Researcher_Competence_Profile” had been sliced up into various concepts that would make up the detail profile of the researcher.

5.1.3 UML Modeling

UML modeling techniques are very handy and helpful tool to specify, visualize the model of an artifact especially in object oriented software development procedure. UML modeling helps to visualize the system’s architectural blueprints and different
elements involved in the entire systems such as activities, actors, business process, programming language etc. UML diagrams not only help in building the model but also help in clearing the requirements of the system (analysis phase), expand the analysis (design phase) and to confirm whether or not the development work has been accomplish as per specifications (implementation phase). Hence, based on the system architecture we depicted in section 5.1.1, we used UML models to realize the system and its constitutional components and generate requirements initially. Below we have explained the use case, sequence diagram and class diagram drawn in the process.

5.1.3.1 Use Case for the Proof of Concept

![Use case diagram for Competence Profile System.](image)

After having the sufficient overview of the tasks that needed to be accomplished and the feature that the system should provide, a general use case of the competence profile system were drawn. This use case diagram helped us to understand more about the general features that we should have in such systems and helped us to visualize the core modules that the system should have.

The above diagram demonstrates the use case model for the registration and login part of the system. The general visitor should have the privilege to know about the system and its terms of condition for the registration. If the researcher wants to register into system, the system should provide basic registration facility and once the researcher
registered into the system, then researcher should be able to login and logout. For the administrative part, a special privilege user “admin” should have some supreme user privileges like updating and deleting profiles.

Several of other use cases were done to get the graphical understanding of the user’s need and capabilities that the system should have.

5.1.3.2 Sequence Diagram

![Sequence Diagram]

Figure 5.4: Sequence diagram

The sequence diagram we built for the system under development helped us to have a logical view of all the interactions between the objects of the system in the time sequence. The above diagram depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects in order to carry out the process of making the competence profile for user/researcher who registers into the system.

The process of making competence profile starts as the researcher register into the system and feed the system with some generic information. Only the researcher’s full name is used to generate the user and data source specific URL queries and such queries are subsequently executed to get the researcher data from the sources if available. Then all such data is analyze/process for different measures and fed into the core ontology, which maps the researcher’s competence profile. For showing the so prepared profile to researcher, firstly a SPAQRL query is executed against the core competence ontology and other user specified generic data from local database is merging to show the full profile back to researcher.
5.1.3.3 Class Diagram

The class diagram is a static form of UML modeling which depicts the structure of the system by showing the system classes, their attributes, operations they do and mainly the relationship among the classes. Based on the system architecture that we have shown in section 5.1.1, we sketch the classes diagram for the system. In the diagram above, we have shown the part of class diagram that we used for understanding the system in class level. In the query construction section, we realized that there should be a interface class that should have a method to take name description of the author’s whose profile is to be made and return the data source specific URL which could later be used by URL_Fetcher class to get the data from the web URL. Similar, there should be definitive interface for XML handling process that should then be implemented later specific to the data-source scheme. Class diagram above also shows the

Figure 5.5: Class diagram for Competence profile system
Plain Old Java Object (POJO) type of classes (Author and Publication) which we used for parsing the XML output and storing the publication specific information.

5.2 Prototype development using agile (scrum) software development methodology

The DSR research method consolidates the general finding of the research with an artifact that fulfills the objective of research. So, in order to ground the knowledge we gain in our research for automatic generation of researcher’s competence profile from the bibliographic publications of the researcher, we need to build an artifact as a proof of concept. The development of artifact needs to be done with some design guidelines and methods which follow the finding of our research on this topic.

As we are building a software artifact, a dynamic web project as our proof of concept, we followed agile development methodology/principals as our software prototype development methodology. Among the different agile methodologies, we chose SCRUM as the best fit for our case as we have to come up with working prototype in fast pace and also due to the fact that the implementation is comparatively small project. As artifact that we aim to build has distinct semantic web part, which consists of the core ontology modules to map the researcher’s competence profile, we also have to use ontology development methodology for building the ontologies. We chose Methontology methodology to develop the ontology that we needed to build our POC application.

Agile software development is one the most used software development methodology. Our aim in this thesis project is to build an artifact as the proof of concept to our research understanding and findings; hence we developed a prototype that demonstrates the major concepts and realization of study. We chose scrum software development methodology that is an iterative and incremental form of agile software development methodology. Again, as this is a proof of concept of the student thesis work, we have to adjust some of the principal and rules different than what the real industry world would use while implementing the agile (scrum) methodology due to the limitations we had with human resources, time etc. We followed the scrum manual/handbook prescribed by Scrum Training Institute Press [71] to develop our prototype. Before we proceeded with the actual implementation, we also had a meeting with the choice of tools and technology for development of proof of concepts that we have listed below:

a) Java programming language
b) MySQL& MySQL workbench tool
c) Apache Tomcat webserver
d) Topbraid/Protoge
e) Tomcat application server
f) Prime faces UI components
g) Jena Ontology framework
h) SPARQL query engine
i) Eclipse/Netbeans
5.2.1 Scrum Roles

During this thesis work, much of requirement gathering and problem areas were collected from the meetings and brainstorming sessions organized for the CLICK project. The core roles that we can formulate are as follow:

**Product Owner:** The entire team of CLICK project, which is our use case study project, played the major role of product owner. The five meetings and one brainstorming workshop that we attended for CLICK project were very helpful for us to gather the requirements, user stories and problem areas of the research field. Again, our supervisor also played substantial role of product owner by guiding us for the major requirement for the prototype that we are building and ensure that we follow the proper methodology fulfilling the requirements we picked. The UML models that we build according to the system architecture helped to identify the requirements in fine granular level that we formulated into product backlog.

**Development Team:** The development team for the proof of concept was composed of two students doing this thesis. Both the students were responsible for all the research and development work. The work of analysis, design, development, test and document writing were done iteratively by organizing the equal work distribution.

**Scrum Master:** The role of scrum master is very crucial in the scrum as he/she facilitate the whole procedure by acting as the buffer between the development team and any negative influencing factor. Within this project, both the students partially had to play the role of scrum master depending on the situation and demand.

5.2.2 Scrum Artifacts

5.2.2.1 Product Backlog

The initial study of the theoretical backgrounds, problems and challenges involved and the observation of some cases (PNRS and ViVO), meetings, brainstorming session attended for CLICK project and the discussion we did with our supervisor helped us to articulate the whole product vision. This vision on product was then listed in general and prioritized based on their criticality and importance to the system, which formed the product backlog. The product backlog was used as the roadmap for the development work as it could give the single definitive view of everything that could be done by development team in order of priority. However, in scrum methodology, this product backlog is an ever-evolving artifact, with inception of new ideas and requirements, changes and prioritization was made (product backlog refinement workshops) to it during the life cycle of the product. The relative estimates were made for each item in terms of “points”. Detail on each individual task and their implementation are discussed in sprint section below Section 5.3. The product backlog table we draw during the implementation of system has been presented in appendix. (See appendix)

5.2.2.2 Sprint Planning and Sprint Backlog

The purpose of sprint planning is to break down the product backlog items into specific sprints such that they could be prioritized and assigned to development teams based on importance and requirement to the product. Normally, sprint planning is done in two sprint planning meetings but in this thesis project we organized several small meetings (mostly online Skype meetings) to manage the sprint planning. As product backlog was prepared with excessive planning and theoretical background
study, the prioritization adapted in product backlog was followed and items were break down into sprint. (See Appendix for spring backlog table.)

5.3 Implementation details and screenshots

The product backlog items were carefully divided into relevant sprints. Each of the sprints was modularize such that none of them take more than 4 weeks of time for the development team of two students. Below we have presented the implementation detail of each of the sprint.

i) Study the IEEE xml schema and extraction procedure

http://ieeexplore.ieee.org/gateway/ipsSearch.jsp?au=Vladimir Tarasov&hc=200&rs=1

The one of the data source we studied in detail was IEEE. The official website and other web-materials in Internet had little information about getting the data of particular researcher in xml format. We email the IEEE support team for getting more information on the access way to the IEEE data repositories and got the document with detail about IEEE gateway API to access data.

The example web query given above retrieves the publication of author “Vladimir Tarasov” from IEEE’s repository in xml format. The xml output data were studied in detail to figure out the main attributes that can be found in IEEE data source schema. The analysis of schema was also necessary to build the xml parsing modules for the implementation of proof of concept. Below we have presented an excerpt of the data analysis being done on IEEE data output.
ii) Study the Diva xml schema and extraction procedure

http://www.diva-portal.org/dice/mods?query=+authorName:(Vladimir Tarasov)&start=0&rows=500&sort=author_sortasc

Study was done to figure out the way to access DiVA repositories and get xml data output. We mailed Uppsala University library to know more about access methods to DiVA. The DiVA data source provides links to feed bibliographic data in different data format as RSS, CSV, HTML, Mods etc. As we were to use only the xml data format for our poc, we chose Mods data format, which is actually a variant of xml. The web query given as example above retrieves all the publication data in mods (xml) format of “Dr. Vladimir Tarasov” as author from DiVA data source.

The xml output data were studied in detail to figure out the main attributes that can be found in DiVA data source schema. The study of these sources helped us to determine the concepts that we should have in our final ontology model.

iii) Study the PubMed xml schema and extraction procedure

Among the three data sources we used, it was most difficult to figure out the access method to PubMed data. PubMed data source does not provide a direct way to access
the repository and retrieve the xml output of publication. Every document stored in PubMed is given a unique identification number called as “PMID”. So, for getting specific publication data about the researcher, first a query should be executed to get the “PMID” of documents that were publish by researcher with specific name and then based on such “PMID” second query should be executed to get the actual xml output of publication data.

Similar to IEEE and DiVA, the PubMed data output schema was also analyzed to figure out the common attributes and concepts that the final ontology schema should have to store data from these sources.

**iv) Finalize dimensions to be included in Competence Profile Ontology**

After the careful study of xml scheme of all the three data repositories that we were to use for the proof of concept, we had to sideline the “project” dimension of the competence profile ontology that we modeled in section 4.5. The projects attribute is not present in the xml output in the data sources that we are using. So, retrieving information on project accomplished by certain researchers by extracting bibliographic data sources is not feasible. Apart of “projects” concepts, all other concepts were used and are visible in the final profile of researcher.

**v) Study of well-known ontologies for interoperability**

In section 4.4, we mentioned about some of the well know ontology that are being used by many of the semantic web projects. These ontologies are well designed and are being continuously enhanced. Reusing these ontology will benefit the system as they will ensure the ontology to be interoperable with other systems using such ontology. The well-known ontologies like FOAF, BIBO, People.owl were studied from core technical prospect to see figure out which concepts and relations could be reused in the ontology for competence profile system.

While designing our owl schema, we reused some well-known ontology as we mentioned in previous chapter. For example, in our schema the class “Researcher” is a subclass of “Person” which already been defined in VIVO or FOAF. With the restriction we added for the class “Researcher”, we can easily get all “Person” relations defined by FOAF towards other things by this reuse.

![Class Form Diagram]

*Figure 5.7: Example of reuse and restrictions*
vi) Creating Ontology using Methontology methodology and Topbraid tool

Before the implementation of the actual ontology into the machine-readable form, we first study the domain of competence modeling and figure out the main concepts that would make the profile rich in information. The study of data source also helped us in finalizing the concepts that we implement in the final design, as there is no point in just making the classes and concept for the information that we can’t get from data sources in general for example, we remove the “projects” dimension from our final ontology as we figure out that project information is not available in all the data sources we were working with. As recommended in [16], we did the throughout designs and iterate it until we were convinced that the model we have is comprehensive.

We used the integration approach of Methontology to build the ontology. The author class in our ontology was derived as a subclass of person class from FOAF ontology, which makes our ontology interoperable with FOAF. Similarly, location class is derived from the ViVO. Other main classes that were chosen for the profile are publication, keywords the publication has, the source the papers were found from, the affiliation of researcher, and general author’s name details.
vii) Create xml extraction module for each data-source

```java
public String[] readIEEExml(String ieeeurl, String attribute) throws Exception{
    DocumentBuilderFactory dbf = DocumentBuilderFactory.newInstance();
    DocumentBuilder db = dbf.newDocumentBuilder();
    Document doc = db.parse(ieeeurl);
    NodeList pubmed = doc.getElementsByTagName("document");
    String[] string = new String[pubmed.getLength()];
    stringize = (pubmed.getLength());
    for (int i = 0; i < pubmed.getLength(); i++) {
        Node node = pubmed.item(i);
        if (node.getNodeType() == Node.ELEMENT_NODE) {
            NodeList list = node.getElementsByTagName(attribute);
            if (list.item(1) != null) {
                Node element = (Element) list.item(0);
                Element element = (Element) node;
                if (element.toJSON().getElementsByTagName(attribute)) {
                    list = element.getElementsByTagName(attribute);
                    if (list.item(1) != null) {
                        NodeList text = element.getElementsByTagName(attribute);
                        if (text.item(1) != null) {
                            string[i] = StringFilter(((Node) text.item(1)).getTextContent());
                        }
                    }
                }
            } else {
                string[i] = StringFilter(((Node) text.item(1)).getTextContent());
            }
        }
    }
    return string;
}
```

*Figure 5.9: Code excerpt for xml data retrieval from IEEE*

After figuring out the web request queries for each of the data sources, the next step was to retrieve the xml output that those sources generate. Each of the data sources was queried and the outputs were programmatically analyzed. The so generated output first needs to be check for its valid xml structure and then read with specifically designed program. As each of the xml documents we used had organization specific xml document tags to represent the publication information, we had to tailor design xml reader programs for each of the source. The above code snippet is the function for reading the IEEE xml document. IEEE uses “document” tag as the root node for representing each of the publication.

viii) Create mapping/cleansing/parsing module for each data source

Data heterogeneity is another problem that is common while integrating data from different sources. Here, we had to firstly map each of the xml document node with each other after knowing the semantics of them. For example, the publication keywords are denoted by term “Keyword” in IEEE and PubMed whereas term “ControlledTerm” is used in DiVA for the same. So, for specific data source, we first did the data analysis to know the semantics of each of the node in the xml, and then we mapped the relevant terms so that while loading into the final ontology schema, we could load the terms with similar semantics into same concept/class of ontology. For example, in case of parsing keywords for each document in that we got from 3 different sources, we parse node “keywords” from IEEE and PubMed and node “ControlledTerms” from DiVA and populate them to the class “keyword” of our competence profile ontology.
ix) Develop a module to load data to ontology

```java
public void loadModel(String owlfilepath) {
    try {
        FileInputStream file = new FileInputStream(owlfilepath);
        InputStreamReader in = new InputStreamReader(file, "UTF-8");
        OntModel model = OntModelFactory.createOntologyModel(modelFactory);
        model.read(model, in, null);
        file.close();
    } catch (FileNotFoundException e) {
        e.printStackTrace();
    } catch (IOException e) {
        e.printStackTrace();
    }
}
```

Figure 5.10: Code excerpt for data mapping/cleansing/parsing

x) Study on different potential solutions and implementation of a solution for Name disambiguation problem (especially Random Forest)

Different potential solutions for the problem of disambiguating the author name were studied. Methods such as using co-authors name to disambiguate author, taking affiliations features from bibliographic data, considering word similarity in keywords and title used etc. were considered. Even though feature like email address of an author can actually disambiguate the author from others, but from study of xml dtd (me-
ta-data) from all sources we know that email of authors are not always present in the sources. Hence, we can’t use email for most of the cases.

Since the problem is related to probability of author being the same person of interest or not, it is very hard to rely on any specific method for 100% guaranteed result. However, just to make the simple working prototype for our research, we used author’s most used affiliation as our main feature to identify author uniquely even though we know from our own research that affiliations alone can not give a satisfactory result for all the cases.

Our research on author name disambiguation is still on run and we are determined to use Random forest machine learning algorithm which to our knowledge and research study has provided best result among all the algorithms used so far by different researchers and in different studies. We have given a brief introduction to the Random forest algorithm in the appendix for readers to have basic concept of the methods.

xi) Create a database

A database schema was design for storing basic information about the research. The schema is designed for implementing a login/registration system for proof of concept. The researchers need to provide some basic information to get registered into the system. The purpose of getting some basic information from researcher is mainly for the author name disambiguation problem as mentioned in section 2.6. Also, from our study of bibliographic data sources, we know that all the data that we need to make

![Database Model Diagram]

Figure 5.12: Database model

A database schema was designed for storing basic information about the research. The schema is designed for implementing a login/registration system for proof of concept. The researchers need to provide some basic information to get registered into the system. The purpose of getting some basic information from researchers is mainly for the author name disambiguation problem as mentioned in section 2.6. Also, from our study of bibliographic data sources, we know that all the data that we need to make
comprehensive profile of researcher are not always present in the sources; we take some of the researcher own fed personal information like contact details from this local database in our final profile.

xii) Create login webpage (UI)

For researcher to login into the competence profile system, a simple login webpage was created using the primefaces UI component. The webpage basically contains two text input fields for researcher to input the username and password, which is then validated against the stored credential in database by pressing the login button. The new researchers can navigate to registration page by pressing the “New User” button.

![Login Form](image)

*Figure 5.13: Login web-form*

xiii) Create registration webpage (UI)

![Part of registration form](image)

*Figure 5.14: Part of registration form*
The registration web page allows the researchers to feed some of the basic information about themselves. During design, we made the registration process as simple as possible and made only specific parameters mandatory for researchers to fill that we could later use for name disambiguation problem. Apart from that, there is also a provision for taking user consent for making the competence profile in the registration process.

xiv) Develop security module for login system

A database-based security model was developed where user fed credential (username and password) are checked with underlying database to authorize user to login. However, the design of the system is flexible and any other security model like J2EE security model can be easily adapted into the system.

xv) Creating SPAQRL for data extraction from ontology.

Since we used ontology to store the extracted data/information from the repositories, we need SPARQL queries to retrieve back the information from ontology. Several SPARQL queries were formulated to retrieve data from the ontology that we build for competence profile modeling. The above example retrieves the details about researchers with specified keywords and language. The system also comprise of SPARQL queries that could get researchers competent in certain research field.

xvi) Create webpage for profile visualization (UI)

The end user/researcher needs to see the built profile. All the dimensions like general information of researcher, work field, list of publication etc. that were considered for the rich profile were included in the build profile. To show the competence level of the researcher in a particular field, we simply used a ranking function on the all keywords that we could get from the list of publication from all sources we integrated. So, the repetition of any particular keywords, which actually represents the specific research field, gave us relevant weightage of competency level the researcher has in that field. Even though visualization is very important aspect of the end product, in our thesis work and the proof of concept, we mainly focused on the method and system architecture (theoretic level) than in presenting the premium visualization.
Results

Vladimir Tarasov
vladimir.tarasov@jth.hj.se
Doctor, Science
1966
Swedish
Jönköping University, School of Engineering

Figure 5.16: Researcher’s Profile Visualization

After the complete implementation of these entire sprint tasks, we had our first working prototype, which also proofed that our idea of generating a rich competence profile in a automatic way just by integrating the meta-data that is available from different bibliographic repositories is practical. The next step we needed to do was to analyze and evaluate the developed artifact for its comprehensiveness.
5.4 System Analysis and Evaluation

In this section we present the analysis done on the built system and the system evaluation we performed.

5.4.1 System Analysis

DSR give emphasis on system evaluation in terms of utility, quality and efficacy of the design, so we need to analyze the built artifact rigorously to demonstrate how well the system works. We performed the data flow analysis of the system, followed by profile retrieval from system with SPARQL and profile update analysis.

5.4.1.1 Data Flow Diagram

The above diagram is the graphical explanation of how the system built as the proof of concept works. The process of creation of profile start as the researcher firstly registered into the system giving the basic information to the system. Such user details are firstly registered into the system and at the same time data source specific web queries are prepared using the full name (first name and last name) of the researcher. Such queries are then send to the data-repositories as the HTTP request and the data sources return the xml file (xml feed, xml data output as a web service) to the program. The individual xml data from each of the sources then is mapped against the
competence profile ontology to figure out the attribute/elements of interest in them. Only the specific elements like publication title, authors, keywords, affiliation etc. are extracted from the individual sources, processed (name disambiguation, multiple record removal etc.) and then inserted into the competence ontology with the help on Jena API (see appendix).

When it comes to showing the built profile to the researcher, the user specific SPARQL queries are executed against the data in the competence profile ontology. Using the inference, some extra information is also deduced from the store data in ontology (for example, author is competent in “English” language if he has written paper in English).

5.4.1.2 Retrieving Researcher’s Profile

If a researcher is needed with competence modeling experience and who speaks English, then we can easily find qualified people from our ontology by running this SPARQL query:

**SPARQL query:**

```sparql
PREFIX profileDEMO: <http://example.org/profileTEST3#>
SELECT ?Researcher
WHERE {
  ?Researcher profileDEMO:has_competency_keyword profileDEMO:Competence_Modeling.
  ?Researcher profileDEMO:speaks_language "eng" .
}
```

Let’s make an example of the query above and assume that we found two qualified persons, and the diagram below will be the outlook.

![Graphical view of result of SPARQL](image)

*Figure 5.18: Graphical view of result of SPARQL*
From this example diagram, we can easily get the qualified person’s name, and also publication list, affiliation or the data source in which he/she published the paper. According to our system design, if the researcher has registered into the system providing basic information, then we can even use it to generate some personal information such as email address and phone number in the profile.

5.4.1.3 Updating the Researchers Profile

As stated above, researcher’s competence profile is dynamic in nature. As the researchers get involved in new researches, they will most certainly publish their new accomplishments, which might also include new research field that has not yet been included in the researcher’s competence profile previously created. Again, even if the new accomplishment is in the very same field as the previous works, we still need to include the newly published articles to researcher’s credit which then will help to evaluate the researcher’s competence level in those particular fields. The amount of work done by any researcher in any particular field is directly related to the experience and knowledge of researcher in the field, which again gives the value to competence level of the researcher by citing the number of published papers in the field.

In case of our research topic, there are potentially two cases when the researcher’s competence profile needs to be updated.

Case A: When researcher new paper/article get published in the data repositories that have been integrated.

When new paper of any researcher in the system’s knowledge base is published in the data sources that the system is already using to create the competence profile, then the competence profile of that particular researcher needs to be updated. But most of the bibliographic data sources might not have notification services like RSS for new addition to their publication. So, it becomes difficult to know which particular researcher’s profile need to be updated. Certain repositories have option of searching articles/paper of particular researcher that was published recently (for example last 15 days or last 1 month). So, querying data source periodically for publications of certain researcher within certain time frame could be one of the most suitable options for updating profile based on new addition of researcher publication.

Case B: When a new data source is to be integrated into the existing system of competence profile software.

As mentioned in section., in most of the cases it is observed that most of the researcher’s publications are published by different scientific publication/ organizations. The scientific publication houses have certain criteria that the papers need to fulfill in order for them to be recognized by those bibliographic publication houses and publish the papers. It also might be the case that the researchers only send their paper to certain publication house to be published. In any of the case, it becomes important for the competence profile creation software to have data from as many bibliographic publication houses as possible to make the profile most comprehensive. So, whenever the system is to integrate a new data source, the existing competence profile of all the researchers in the knowledge base need to be updated in order to reflect any new addition to the competence from the new source.

Profile Updating Strategy

As we are following the on-demand competence profile creation of researchers, it is possible to store every request of profile creation to a data-source other than the
knowledge base (competence profile Ontology). Such data source could be common
database tables where all unique requests from researchers are maintain. So, when
there is a need for updating the competence profile, firstly whole of the instances of
researchers can be dropped from the competence profile ontology just keeping the
schema of ontology and then whole of the profiles of researchers can be rebuilt with
the help of the secondary data source which has the name detail of the researchers
who has previously built their profile. Which means, in the case of update, the query
constructor module will get the request parameter from database instead of user feed
input.

5.4.2 Evaluation

One of the very conclusive and important steps of DSRM activity is the evaluation of
the implemented artifact. It is very important for the DSR project to verify and validate
that the system achieves the milestones set as the objectives of the research. After
studying the similar artifacts (PNRS and ViVO) and the shortcomings that we observed in those systems, we refined our research objectives (c.f. to section 4.5) into 3 focused areas. In this section, we evaluate our proof of concept against the above mentioned refined research objectives.

5.4.2.1 Adding new data source

One of the major limitations of the existing systems that we studied were that they
could not integrate data from new sources. We refined our research to focus on this
problem and we empirically proved that the system that we built is capable of adding
new data sources. Below with the help of 3 figures, we explain how the process of
adding new data source works in our system.

Figure 5.19: Data extraction from IEEE only

In the above diagram, the system is using only one data source (IEEE) to generate the
profile. In this particular case, there was no research publication from the author, and
hence the URL query we used did not produce any result.
In the second case, we modified our system to query two data sources (IEEE and DiVA). Here we found few numbers of papers for the same author and the system extracted the related metadata in order to generate the researcher’s competence profile. The figure below shows the extraction of researcher basic information and key competence terms from the meta-data of the papers found in the DiVA data source.

![Figure 5.20: Data extraction from IEEE and DiVA](image)

**Figure 5.20: Data extraction from IEEE and DiVA**

In the second case, we modified our system to query two data sources (IEEE and DiVA). Here we found few numbers of papers for the same author and the system extracted the related metadata in order to generate the researcher’s competence profile. The figure below shows the extraction of researcher basic information and key competence terms from the meta-data of the papers found in the DiVA data source.

![Figure 5.21: Competence profile extracted from two sources](image)

**Figure 5.21: Competence profile extracted from two sources**

### 5.4.2.2 Updating researcher’s profile

We used the on-demand profile creation approach for building the competence profile where the system will query the data sources each time the user request to view the profile. Because of the approach, the prepared profile is up-to-date with the information that is present in the data sources.

Again, in the case addition of new data source, the system creates the profile based on the meta-data at the current state of all the data sources and hence the profile is always updated. So, the system design and the methodology we used for the system helped us to deal with the profile update system to a great extend. Nevertheless, we should also consider the limitations of this approach. The organizational policy of different publication houses that maintain the bibliographic records might not allow more than certain number of web queries from a definite IP Address (Internet protocol Address). For example, PubMed only allows 3 requests per second.
5.4.2.3 Ontology refinement

The final OWL ontology that we used for mapping the competence profile reuses VIVO and FOAF ontologies. By reusing these ontologies, we achieved semantic interoperability between the competence ontology that we developed and the VIVO and FOAF. Reusing these well known ontology not only helped in getting semantic interoperability, but as these ontologies are refined and developed by other researchers, they already included a lot of concepts that we could reused in our ontology. For instance, the location class in VIVO was reused in the competence ontology to map the location of the researcher.

The other study that we did for refining the competency ontology was add different dimensions that researcher’s competence profile should have. As mentioned in section 4.3, we included all the dimensions like general information (name, date of birth, email, phone etc.), Publication list, Research area (shown by competence terms) Affiliations of the researcher and geographical location. But due to the fact that the data sources we used had no information about the projects that researchers were involved, we could not show this information in our final profile.

---

Figure 5.22: Reuse of VIVO and FOAF
6 Conclusion and Reflection

6.1 Conclusion

Research in information system should respond to dual missions: to make valuable theoretical contributions and assist in solving the current and anticipated problems of the practitioners [74]. In our thesis work, we work on both the aspects to identify the potential problems areas in the domain and how such problems can be tackle, which helped us to draw three research questions.

1. What are the main challenges of bibliographic data sources integration for creating the researcher’s competence profile?

In order to answer the first research question, we did an extensive literature reviews and studied the similar work done by other researchers. We figured out the main potential problem areas for integrating heterogeneous bibliographic data sources in order to generate the competence model for the researchers with respect to their contribution in the research domain. There were many challenges like semantic, schematic, syntactic heterogeneity of data sources but author name disambiguation stood as the main challenge in accomplishing the integration task.

2. How can data from heterogeneous bibliographic data sources be converted into the common data model and common knowledge model for extracting the competence profile?

Initially during our literature review, we thought of using RDF as our common data model and studied different technologies like XSLT to convert any of the data format to intermediate common model of RDF. However, when we implemented an artifact to support the problems of the practitioners in same domain, we chose XML as our common data model and our implementation of proof of concept did not require any intermediate conversion of data format as we specifically chose XML data from all three sources we used.

We used ontology built with OWL as the common knowledge model for the artifact that we built but as the data we extracted from different sources do not have some information that we needed in the final profile for researcher, we also had to use a local database to store and extract some user specific generic information as name details.

3. What are the benefits of creating researcher’s competence profile using semantic data source integration method?

As we reused well-known vocabularies for building the core knowledge model for the system, we achieved semantic interoperability with them. This is the basic advantage of using semantic technology in data source integration and specifically ontology as the knowledge base.

Hence, we conclude that we have been able to answer all our research questions to a good extent and we also provided some contribution to both the theoretical domain (method, model and system architecture) and practical domain (working prototype to verify that the concept we presented works practically). The ideas that were generated from this research work can be implication for similar systems as well.
6.2 Reflection

Our research work started with the hypothesis/assumption that the publications papers published by different researchers reflect to their knowledge and skill sets and has potential to extract their competency profile. Rather than analyzing the content of whole research papers, we focused on the bibliographic meta-data that the different repositories use for easier information retrieval process. While doing so, we came across different data and data integration related issues. Access to different bibliographic repositories, the meta-data heterogeneity, and author name disambiguation problems were some of the major issues while complete-ness of so created profile was also in question.

The literature review and background check was done to learn from the similar work done in past. To accomplish the work in more systematic and scientific way, DSR method was applied to the research process that not only helped us to focus on research but also helped us to achieve the results we aimed for. A general methodology for automatic profile creation system for researcher was drawn based on literature review and it's learning. A layer-architecture based on the methodology was sketched which helped us to visualize how the particular instantiation of such system would work.

Finally, to implement the proof of concept, careful analysis of ways to access data from the different bibliographic sources it became clear that XML form of data is available from most of the sources and we decided to use virtual view approach of data integration as its most appropriate when access to whole of the data source is not available/possible. Web services provided by different bibliographic repositories were used to send specific researcher inquiry related web-queries to get publication meta-data in XML format. Among the 3 sources that we used for POC, access to XML bibliometric data of IEEE and DiVA were straightforward whereas some additional steps were needed for PubMed.

For integration, mapping of each of the attributes from different XML data was done and corresponding attributes whose semantic matches were mapped to the core-competency ontology that was build as the knowledge base for the POC. Disambiguating the authors just based on registration information provided by authors and the information from the publication meta-data was the hardest of the issue in bibliographic data source integration. As it is the case of probability, none of the algorithms have shown 100% guaranteed results till date. Even though we studied different machine learning algorithms like Random forest and support vectors, we limited our work of disambiguation by just comparing the registration data with publication XML data from sources.

As far as the complete-ness of the so extracted competence profile was concerned, we studied different research done on competency modeling topic and summarized the core-competency component that would be comprehensive to show researchers knowledge and skill sets. As our information base/source for making competence profile is the meta-data that the bibliographic repositories use to index for information retrieval, the comprehensive-ness and complete-ness of competency profile depends thoroughly on what sort of information have been indexed by those repositories. Apart from general information, we also included the different publications the researcher has contributed into the profile, making it information richer.
Conclusion and discussion

The initial results we had were encouraging and ground the assumptions that a rich competency profile of researchers can be built out of the meta-data of publications. So, rather than being just used as indexes for faster search and information retrieval, the bibliographic meta-data posses vital source of information regarding the authors/researchers of those publication if analyzed well. Competency profile of researchers is one of the potential outcome of such repositories and we truly belief that other vital information can be proceed from such repositories.

So, during this research work, we felt need of standardization of indexing process of publication by defining a standard set of attributes that must be used as we already seen that they could be a potential source of enormous information. Secondly, there should also be a standard body or consortium like “ResearherID” to take care of unique identity of researchers so that the problem of disambiguating the researcher’s name and work becomes easier. The above-mentioned two issues were among the most substantial issues we observed while working for our thesis research and need further work in future.
7 Recommendation and Future work

7.1 Recommendations

The research work we did on the subject helped us to have a better understanding of the problem domain. What we learnt during the course of this research work could provide some valuable insight for the other researchers doing their research work in the similar domain. We hope that the recommendations presented below will help the researchers to obtain better results by implying them into the practice.

**Recommendation 1: Identify the main purpose of this created profile; define the use case, scope of related project.**

We suggest defining the motivation of doing this kind of research; and more important is to have a clear purpose of making this profile. Design use case for profile usage, define the end user and indicate the scope of related project or system.

**Recommendation 2: Make sure the ontology is build with acceptable quality, suitability, and extensibility, and follow the design guidelines.**

The ontology works as the knowledge base in this system, so the schema for the competence profile ontology, its quality, suitability and extensibility will directly effect to the final profile. Therefore we recommend carefully designing the ontology structures, classes, properties and reuse necessary component according to the requirement and the research purpose.

**Recommendation 3: Carefully choose and analyze the data sources and their data structure the system is going to use.**

As the system is using the publicly available data sources, we recommend to chose the data sources by thinking of the main focus area, creditability, availability and openness. As data sources will have organization specific schema/structures, we also recommend giving more emphasis on analyzing the schemas and picking up only the relevant attributes for building the competence model.

**Recommendation 4: Design and apply at least one name disambiguation method.**

In our study, we figure out that author name disambiguation is the main challenge for the realization of this sort of system. Hence, we suggest designing at least one robust algorithm addressing this issue.

**Recommendation 5: Design and apply at least one method/algorithm to determine researcher’s competence level.**

Competence profile with the indication of the level of competence will enrich the information level for the user of the system. Hence, its very important that while integrating information from various sources, it is important to apply different algorithms to give weightage to the competency level of the researcher. Number of publications in certain research field, number of years the researcher have been involved in the research area and other bibliometric data could be analyzed to get these information.

**Recommendation 6: Make sure that there is open window for external sources to help creating better profile.**

The automatic profile creation here does not mean it has to be a sealed system or the data sources used should only be the bibliographic digital libraries. To create better
profile in the end, it is recommended and necessary to get some other data from the outside systems, organizations like researcher’s personal blog, social media sites or even ask people for the first hand data.

**Recommendation 7: Consider of the privacy fact of sensitive information or protected data.**

As the final point for researcher in this field, it is necessary to consider the privacy issues when automatically deal with large numbers of data from different sources. Researchers should be aware of data usage terms and users consent for creating profile should always be taken.

### 7.2 Future Work

Research is a never-ending process or cycle. The research we did for this thesis topic has helped us to understand some of the major challenges in the subject domain, we were able to solve some of them and we came across some of the problems, which we still have to work more rigorously. We also realized that there are more scopes on this domain to enhance the results that we achieved.

In future, we aim to work on a more robust author name disambiguation algorithm and imply some machine learning approaches like Random forest algorithm to address the issues. Again, we also like to enrich the profile by doing some bibliometric data analysis and giving weightage to competency level in more organized way. We also see a high scope of using researcher’s personal blogs, LinkedIn profiles and other social media as potential data sources to enhance the information in the competency profile.
8 References


[3] Christiaens, S., Kerremans, K., Meersman, R., & Tang, Y., “Profile Compiler: Ontology-Based Community grounded, multilingual Online services to support collaborative decision-making”.


[14] Oscar Corcho, Mariano F. Lopez, Asuncion G. Perez, “Methodologies, tools and languages for building ontologies. Where is their meeting point?”


80
References


[31] Gordana DODIG CRNKOVIC, “Constructive Research and In-Form-Computational Knowledge Generation”, Malardalen University, Sweden.


[35] Michel Gagnon, “Ontology-Based Integration of Data Sources” Quebec, Canada.

References


[38] Ravikarn Punnarut, Gridaphat Sriharee, (2010), “A researcher Expertise Search System using Ontology-Based Data Mining”, Department of Computer and Information Science, King Mongkut’s University of Technology North Bangkok.


82
References


[56] Pinto HS (1999), “Towards ontology reuse”.

[57] Pinto HS (2000), “Ontology integration: characterization of the process and a methodology to per it”.


[60] About VIVO, Retrieved on June 20, 2012 from: http://vivoweb.org/about


[64] About CiteseerX, Retrieve on October 1, 2012 from: http://citeseerx.ist.psu.edu/about/site


[70] Open Researcher and Contributor Identification (ORCID), Retrieved on October 10, 2012 from: www.orcid.org

[71] Jeff Sutherland, (2010)” Scrum Handbook- Everything that you need to know to start Scrum project in your organization”, Scrum Training Institute Press.
References


[74] Rosemann, M., and Vessey, I. (2008),” Toward Improving the Relevance of Information System Research to Practice: The Role of Applicability Checks”.


# Appendix

Table 9.1: List of Universities and Institutes associated with DiVA (Uppsala University, 2012)

<table>
<thead>
<tr>
<th>Code</th>
<th>University Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>DU</td>
<td>Dalarna University</td>
</tr>
<tr>
<td>ESH</td>
<td>ErstaSköndal University College</td>
</tr>
<tr>
<td>FHS</td>
<td>Swedish National Defence College</td>
</tr>
<tr>
<td>GIH</td>
<td>The Swedish School of Sport and Health Sciences</td>
</tr>
<tr>
<td>HGO</td>
<td>Gotland University</td>
</tr>
<tr>
<td>HH</td>
<td>Halmstad University</td>
</tr>
<tr>
<td>HHS</td>
<td>Stockholm School of Economics</td>
</tr>
<tr>
<td>HIG</td>
<td>University of Gävle</td>
</tr>
<tr>
<td>HIS</td>
<td>University of Skövde</td>
</tr>
<tr>
<td>HJ</td>
<td>Jönköping University</td>
</tr>
<tr>
<td>HKR</td>
<td>Kristianstad University</td>
</tr>
<tr>
<td>HV</td>
<td>University West</td>
</tr>
<tr>
<td>KAU</td>
<td>Karlstad University</td>
</tr>
<tr>
<td>KMH</td>
<td>Royal College of Music</td>
</tr>
<tr>
<td>Konstfack</td>
<td>University College of Arts, Crafts and Design</td>
</tr>
<tr>
<td>KTH</td>
<td>Royal Institute of Technology</td>
</tr>
<tr>
<td>LIU</td>
<td>Linköping University</td>
</tr>
<tr>
<td>LNU</td>
<td>Linnaeus University</td>
</tr>
<tr>
<td>MDH</td>
<td>Mälardalen University</td>
</tr>
<tr>
<td>MIUN</td>
<td>Mid Sweden University</td>
</tr>
<tr>
<td>NAI</td>
<td>The Nordic Africa Institute</td>
</tr>
<tr>
<td>NTNU</td>
<td>The Norwegian University of Science and Technology</td>
</tr>
</tbody>
</table>
## Appendix

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORU</td>
<td>Örebro University</td>
</tr>
<tr>
<td>Polar</td>
<td>Swedish Polar Research Secretariat</td>
</tr>
<tr>
<td>RKH</td>
<td>Red Cross University College</td>
</tr>
<tr>
<td>SH</td>
<td>Södertörn University</td>
</tr>
<tr>
<td>SHH</td>
<td>Sophiahemmet University College</td>
</tr>
<tr>
<td>SU</td>
<td>Stockholm University</td>
</tr>
<tr>
<td>UMU</td>
<td>Umeå University</td>
</tr>
<tr>
<td>UU</td>
<td>Upsala University</td>
</tr>
</tbody>
</table>
Summary of Recommendations for Competence Profile Modeling, [16]

**Recommendation 1:** Identify the core area of the competence modeling as early in the modeling process as possible: Activity, product knowledge or individual competence focus?

As a rule of thumb, we recommend the following. If organizational competences are the core interest of the project: chose an activity based perspective first and keep an open mind for individual competences. If the purpose of the project is strongly related to products or their lifecycle, select a product knowledge focus first and try to understand the activities to be performed. If the human resources are of specific interest for the project, start with individual focus of the project and observe the importance of product knowledge in case of manufacturing enterprises.

**Recommendation 2:** Investigate what ways of structuring and describing competences exist in the scope of the competence-modeling project

We recommend involving the human resource department or the responsible manager for this issue. Questions to investigate are ways of describing and structuring competences (local scope, global definition, integration of standards, etc.), existing task and responsibility descriptions for roles, and competence development plans and their basis.

**Recommendation 3:** Check quality and suitability of competence catalogues or classifications before deciding to apply them

As a simple initial check, we recommend to express the competences for one selected role or activity with the standard under consideration. If such standards shall be applied, the motivation for this requirement has to be analyzed. A requirement regarding compatibility between the competence model to be developed and the existing standard can turn out as severe constraint for the projects, where the effects have to be analyzed. Standard as starting point, without compatibility constraints, can be beneficial for the modeling project.

**Recommendation 4:** If assessment of individual abilities/skills is necessary, find out reliable methods for doing this beforehand

Assessment of abilities/skills through self-evaluation and interviews is error-prone and inaccurate. More reliable measures are needed that can be created with the help of professional psychologists and human resource departments.

**Recommendation 5:** Include both domain experts and modeling experts in the modeling process

The domain expert has knowledge and experiences from the domain that are needed for competence modeling. The modeling expert can through this expand their situational knowledge in relation to the domain in focus.

**Recommendation 6:** Include in a competence model the task, competence, organization, and context.

As soon as a competence model is modular, other parts of the model are optional. The first three mentioned modules are necessary to express what competencies are necessary to fulfill the task. A context can represent a number of situational aspects that will affect the competence demand even if the formal task(s) is the same.
**Recommendation 7:** Conduct modeling in an incremental and iterative way

The modeling should include several steps addressing the process with activities, competences that are needed to perform these activities, relation between roles and competencies, and recourses that are needed to accomplish a task.

**Recommendation 8:** When a competence model needs to be implemented in machine-readable form, it is advisable to follow a distinct design phase, choose an implementation language that has good tool support, and make sure that technical specialists skilled in the chosen technology exist on place.

The design documentation is needed for revisions of the competence model. If a competence model is supposed to be a basis for a software application, extensive tool support of the implementation language simplifies application development. Lack of relevant technical skills will be an obstacle for application deployment.
**Random Forest Algorithm**

**RANDOM FOREST (RF):** is a machine learning technique or statistical learning techniques, which is considered as the one of the blackbox tool for doing predictions, but there are many scenarios where it can be used as variable selection. RF is a classification and regression method based on the aggregation of a large number of decision trees.

“Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest” (Breiman, 2001).

In general, RF algorithm can be viewed in following steps-

- Construct random trees and grow them
- Trees grown using B independent bootstrap samples
- Different random subsets of “mtry” variables are used at each tree node.
- Grow deep trees with a fixed node-size.

![Random Forest Diagram](Image)

Details on Random Forest and its implementation can be found on following link:

[http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm](http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm)
## Appendix

### New Estimates of Effort Remaining as of Sprint..

<table>
<thead>
<tr>
<th>Item</th>
<th>Detail (wiki URL)</th>
<th>Priority</th>
<th>Estimate of Value “points”</th>
<th>Initial Estimate of Effort</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study data source to investigate the data access methods for each data source.</td>
<td>…</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Develop and enhance competence profile ontology</td>
<td>…</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Develop ETL module for each data source</td>
<td>…</td>
<td>3</td>
<td>10</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Investigate on the solution for Author Name Disambiguation problem</td>
<td>…</td>
<td>4</td>
<td>15</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The user should be able to register/login into system.</td>
<td>…</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The registered user should be able to view profile</td>
<td>…</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Investigate on benefits of inference power of Ontology</td>
<td>…</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Table 5.1: The initial Product backlog*

<table>
<thead>
<tr>
<th>Product Backlog Item</th>
<th>Sprint Task</th>
<th>Volunteer</th>
<th>Initial Estimate of Effort</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study the data source to investigate the data access methods for each data source.</td>
<td>Study the Diva xml schema and extraction procedure</td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

|                              | Study the IEEE xml schema and extraction procedure. |            |                             |   |   |   |   |   |   |
|                              | Study the PubMed xml schema and extraction procedure. |            |                             |   |   |   |   |   |   |

90
<table>
<thead>
<tr>
<th>Task</th>
<th>Notes</th>
<th>Milestone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Develop and enhance competence profile ontology</td>
<td>Finalize dimensions to be included in Competence Profile Ontology</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Study of well known ontologies for interoperability</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Creating Ontology using methontology methodology and Topbraid tool</td>
<td></td>
</tr>
<tr>
<td>Develop ETL module for each data source</td>
<td>Create xml extraction module for each data-source</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Create mapping/cleansing/parsing module for each data source.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Develop a module to load data to ontology.</td>
<td></td>
</tr>
<tr>
<td>Investigate on the solution for Author Name Disambiguation problem</td>
<td>Study on Different potential solutions specially Random Forest</td>
<td>12</td>
</tr>
<tr>
<td>The user should be able to register/login into system.</td>
<td>Create a database</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Create login webpage (UI)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Create registration webpage (UI)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Develop security module for login system</td>
<td></td>
</tr>
<tr>
<td>The registered user should be able to view profile</td>
<td>Write SPAQRL for data extraction from ontology.</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Create webpage for profile visualization (UI)</td>
<td></td>
</tr>
<tr>
<td>Investigate on benefits of inference power of Ontology</td>
<td></td>
<td>5</td>
</tr>
</tbody>
</table>

Table 5.2: Spring backlog
Initial visualization of data flow for the Proof of concept
IEEE XML data feedback screenshot

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```xml
<root>
  <totalfound>0</totalfound>
  <totalsearched>0</totalsearched>
  <rank>1</rank>
  <title><![CDATA[
An ultra-dispersive optically controlled atomic prism
]]></title>
  <author><![CDATA[
Li, Robin; Sautenkov, Vladimir A.; Postovtsev, Yuri V.; Scully, Harlan O.
]]></author>
  <affiliations><![CDATA[
Department of Physics, Texas A&M University, College Station, 77843-4242, USA
]]></affiliations>
  <thesaurus terms><![CDATA[
Atom optics
Electromagnetic refraction
Frequency
Laser beams
Magnetic field measurement
Nonlinear optics
]]></thesaurus terms>
</root>
```
Appendix

DiVA XML data feedback screenshot

This XML file does not appear to have any style information associated with it. The document tree is shown below.
Identification of the Components of a Glycolytic Enzyme Metabolon on the Human Red Blood Cell Membrane

Glycolytic enzymes (GEs) have been shown to exist in multi-enzyme complexes on the inner surface of the h protein other than band 3 has been found to interact with GEs, and because several GEs do not bind band 3, membrane proteins that serve as docking sites for GE on the membrane. For this purpose, a method known as
Appendix

Profile creation process screenshots

Retrieve process

Insert process

Result