Protocol of A Systematic Mapping Study of Requirements Engineering Approaches for Big Data Project Development

Belachew Regane[1], Sarah Beecham[2], Dagmawi Lemma[1], Norah Power[2]

[1]Department of Computer Science, Addis Ababa University, Addis Ababa, Ethiopia

[2]Department of Computer Science, Limerick University, Ireland

Preamble

The objective of developing this protocol is to have a guideline for conducting a systematic mapping study on requirements engineering approaches in big data applications development. In designing and developing software, choosing the right approach is an important activity that has been performed to have a good product. This systematic mapping protocol will help to summarize the requirements engineering approaches that have been used in existing works in the domain of big data to address the research questions.

Based on the protocol, relevant data from journals, conferences, and workshops in software engineering and big data will be collected. Then, the papers will be reviewed to come up with a piece of empirical results.

1. Introduction

Big data technology benefits the user of data by providing insight for decision-making. However, big data application development has become a challenging task. The frequently changing requirements due to its Vs characteristics are among the factors that make the task challenging. Requirements engineering (RE) impacts positively or negatively the success of software development [1], [2]. To have good requirements, a RE approach plays an important role in enhancing the software development process. Yet, the traditional requirements engineering approach could not handle the frequently changing requirements of big data [3]; and it is important to understand the challenges in RE in the case of developing big data technology. Thus, conducting a systematic mapping study, to get evidence on the challenges and potential opportunities by investigating previous works and by considering the state-of-the-art RE of big data applications development process is vital.

The main goal of systematic mapping studies is to get a comprehensive overview of a research area and identify the quantity and type of research and results available within the study area. Furthermore, it explained that the analysis of the results focuses on presenting the frequencies of publications for each category and this makes it possible to see which categories have been emphasized in past research and thus to identify gaps and possibilities for future research [4]. Similarly, it is discussed in [5] that the rationale for conducting a systematic mapping is (i) to summarize the evidence for existing works in software engineering (ii) to identify gaps in existing research works (iii) to provide background for new research studies. Therefore, the rationale for this systematic mapping is to get empirical evidence of the challenges and benefits of RE approaches that have been used in big data project development. According to Peterson et

al. [4] findings, a systematic mapping study is different from a systematic review in terms of their goals, breadth, and depth. But they can complement each other. Therefore, “a systematic mapping can be conducted first, to get an overview of the topic area and then the state of evidence in specific topics can be investigated using a systematic review”.

The result of this systematic mapping will benefit stakeholders in the domain area by providing evidence that justifies how to choose the right RE approach. Therefore, it could help software developers, and requirements engineers to understand the advantage and disadvantages of those RE approaches. This enables them to choose the right approach to identify and analyze requirements for designing software. Thus, conducting a systematic mapping of RE approaches for big data project development gives evidence on the challenges and potential benefits by investigating previous works looking at the state-of-the-art in the area.

2.Method

In this section, the method and procedures that will be followed while conducting this systematic mapping study are presented. To conduct this systematic mapping, the researchers adopt Petersen et al., [4] i.e Systematic Mapping Studies in Software Engineering.

2.1 Research Questions

To achieve the objective of the systematic mapping the following research questions are formulated. These research questions are:

1. How are requirements engineering (RE) activities performed to address the needs of stakeholders in the context of big data?
2. How are users' perspectives addressed in the RE activities in a big data context?

3. What are the requirements engineering approaches that have been proposed for big data project development?

As indicated in the above questions, the researchers want to start by investigating the first two research questions to have a comprehensive understanding of the area and to provide empirical evidence for the research gap. Then, continued with a wider view of software development approaches to get an insight into the general methodologies used. And then, the discussion will be narrowed down to RE approaches and problems related to it to get empirical evidence that helps for proceeding the study forward.

2.2 Search Strategy

The strategy that will be used for identifying search terms from Beecham et al. [6] will be followed as a search strategy. Accordingly, major terms are derived from the questions by identifying the population, intervention, and outcome; alternative spellings and synonyms for major terms will be identified; keywords in any relevant papers will be checked; when the database allows, the Boolean OR will be used to incorporate alternative spellings and synonyms; and the Boolean AND will be used to link the major terms from the population, intervention, and outcome. Following these steps, the following search keys will be identified.

2.3 Constructing Search Terms

According to Beecham et al. [6], it is better to look at three different views while constructing search terms. These views are intervention, population, and outcome. In this systematic mapping, the intervention is the requirements engineering approach, the population is stakeholders (users, software developers, and requirements engineers) and the outcome is requirements that satisfy

the stakeholders' need to enhance software development of big data applications. This will help us as a basis to construct search terms while preparing a systematic mapping study protocol.

2.4 Search Terms and Synonyms

To identify search terms and synonyms, tasks will be performed according to the following two steps.

Step 1: Key terms will be identified from research questions.

Step 2: Synonyms and search strings will be identified.

2.5 Search Strings

An identical search string shown below will be used for both IEEE Xplore and ACM Digital Library. The first search string will be used for research questions one and two. The second search string will be used for research question three.

i. ((“Big data”) AND (“Requirements Engineering” OR "Requirements Analysis") AND (Stakeholder* OR User*))
ii. ((“Big data”) AND ("Software Development" OR "Application Development" OR "Project Development") AND (“Requirements Engineering” OR "Requirements Analysis") AND (Approach* OR Method OR Process))

2.6 Resources searched

To collect relevant data while searching from electronic databases, the search engine of IEEE Xplore and ACM Digital Library will be used to get journals, conferences, and workshop papers of requirements engineering in the context of big data. In addition, a snowball search will be

used. A snowball search helps to collect sufficient relevant documents for the study. Kitchenham and Charters [5] recommends searching manually for relevant documents from "reference lists of relevant primary studies and review articles, grey literature (i.e. technical reports, work in progress)".

2.7 Document Selection Criteria

2.7.1 Inclusion and Exclusion Criteria

This section lists the inclusion and exclusion criteria that will be used for considering or rejecting a published work as a form of evidence for addressing the research questions.

To be included, the study needs to comply with one or more of the following criteria:

- Subject matter criteria
  - A paper may directly answer any of the research questions;
  - All papers must be focused on big data – no exception.
  - Paper that focuses on the challenges in the context of requirements engineering.
  - Solutions/approaches in the context of requirements engineering
- Publication criteria
  - Published as a journal article, conference/workshop proceedings.
  - To catch the views of a wider range of stakeholders (that may not publish in peer-reviewed venues) the grey literature is also included. Such articles published online as work in progress, reports [annual, research, technical, project, etc.], working papers, government documents, white papers, and evaluations are included;

- Published from 2011 to 2021 (research has progressed over the past decade, so the focus will be on more recent work – since the literature review process included snowballing, key works published before this cut off date of 2010 will be identified in the next phase);
- Reports primary research (empirical studies);
- Literature surveys that summarise evidence in the field.
- Published in the English language

Exclusion criteria

Studies that fall into the following categories excluded (E):

- Studies on the following topics are excluded:
  - Studies on design;
  - Studies on architecture;
  - Studies on computer hardware and infrastructure, e.g. networks.
  - Studies that do not explicitly discuss big data;
  - Studies that do not focus on software engineering or requirements in any form (analysis, engineering).
- Types of publication excluded, include:
  - Posters, opinion pieces papers, viewpoints, power-point presentations, introduction to conferences;
  - Books, book chapters, thesis, and dissertation;
  - Keynote speech;
  - Short papers (less than four pages);

- Summaries of conferences/editorials or guidelines/templates for conducting mapping studies;
- Studies not accessible in full-text.

2.7.2 Selection process

The selection process will be started by looking at carefully the title of the journal articles, conferences, or workshop papers to check as it is in the needed domain area. Following this, the abstracts part of the documents read to see their relevance. This helps to determine to review the document. Then, if it is relevant, the entire document will be reviewed. Then, the supervising team [members in the team] will validate each step of the process.

How the final set of papers selected was recorded based on 1. The initial number of papers found (from running searches - search terms in DBs), 2. After removing duplicates, 3. After checking Titles, and Abstracts, 4. After reading the full paper and applying inclusion-exclusion criteria.

2.8 Data Synthesis

The data synthesis is descriptive/narrative using tables to show the link between intervention, population, context, sample sizes, outcomes, and study quality with the study questions. This helps to show how the data will answer the research questions. The reason is the nature of the software engineering survey is qualitative – descriptive [5], [7].

To do data synthesis, the first data extraction will be performed. **T**he data extraction protocol will be used while surveying literature from the studies. Data extraction is a way in which information the researchers require is obtained from primary studies [5]. In data extraction procedures, the data will be extracted and then, the extracted data will be sent to the supervising team/team members to look at, discuss and finally, validate it. Categorization of the data will be

done based on their similarities/commonalities. The data will be categorized into some classes to make the analysis and discussion easier. For example: Based on their contributions, RE activity types, RE approaches, etc.

2.8.1 Data Extraction Form

The contents of the form that will be used to extract the data include authors, year of publication, source, reference, title, topic category of study, research question, keynotes that will answer the systematic mapping questions.

Table 1: Data extraction form

S_No	Paper ID	Date_of_search	Year_of_publication	Title	Authors	Full reference	Source	Topic_category_of_the_study	Research_question	Approach_used_to_answer_RQ	Requirements_types	contribution	Key_notes_that_answer_RQ

2.9 Search Process Documentation

For bibliography management, the Mendeley reference manager application will be used and to document data extracted from primary studies, Microsoft excel will be used.

2.9.1 Documenting the Search

The process will be documented based on the Mendeley reference manager application as follows.

Table 2: Search process documentation

Data source	Documentation
Journals	Title Authors Journal Year Volume Issue Page Date of search
Conferences	Title Authors Proceeding title Year Page Date of search
Workshop papers	Title Authors Year Page Date of search
Digital library	Name of database Search strategy for database Title Authors Year Date of search
Grey literature	Title Authors Year Page Date of search

3.Evaluating the Protocol

To evaluate the protocol, the evaluation checklists recommended by Kitchenham and Charters [5] will be followed and implemented before conducting the study. Besides, supervising team will validate it. These checklists are:

- What are the review's objectives?

- What sources were searched to identify primary studies? Were there any restrictions?
- What were the inclusion/exclusion criteria and how were they applied?
- What criteria were used to assess the quality of primary studies?
- How were quality criteria applied?
- How were the data extracted from the primary studies?
- How were the data synthesized?
- How were differences between studies investigated?
- How was the data combined?
- Was it reasonable to combine the studies?
- Do the conclusions flow from the evidence?"

4.Study Quality Assessment

For quality assessment, a checklist for qualitative study from Kitchenham and Charters [5] will be followed to validate the design, conduct, analysis, and conclusion of the study.

Table 3: Quality assessment checklist

Number	Question
1	How credible are the findings?
1.1	If credible, are they important?
2	How has knowledge or understanding been extended by the research?
3	How well does the evaluation address its original aims and purpose?
4	How well is the scope for drawing wider inference explained?
5	How clear is the basis of evaluative appraisal?
6	How defensible is the research design?
7	How well defined are the sample design/target selection of cases/documents?
8	How well is the eventual sample composition and coverage described?
9	How well was data collection carried out?
10	How well has the approach to, and formulation of, analysis been conveyed?
11	How well are the contexts and data sources retained and portrayed?
12	How well has diversity of perspective and context been explored?
13	How well have detail, depth, and complexity (i.e. richness) of the data been conveyed?
14	How clear are the links between data, interpretation and conclusions – i.e. how well can the route to any conclusions be seen?
15	How clear and coherent is the reporting?
16	How clear are the assumptions/theoretical perspectives/values that have shaped the form and output of the evaluation?
17	What evidence is there of attention to ethical issues?
18	How adequately has the research process been documented?

5.Dissemination strategy

To communicate the results, the report will be sent to academic journals for publication. The structure and contents of the report might include the following Kitchenham and Charters [5].

i. Title

ii. Authors

iii. Structured Abstract

a. Context
b. Objectives
c. Methods
d. Results
e. Conclusions

iv. Background
v. Review questions
vi. Review Methods
    a. Data sources and search strategy
    b. Study selection
    c. Study quality assessment
    d. Data extraction
    e. Data synthesis
vii. Included and excluded studies
viii. Results
    a. Findings
    b. Sensitivity analysis
ix. Discussion
    a. Principal findings
    b. Strengths and weaknesses
    c. Meaning of finding
x. Conclusions
xi. Acknowledgments

xii. Conflict of Interest

xiii. References and Appendices

6.Activities timetable

Table 4: Schedule of activities for conducting a systematic review on approaches for big data application development.

Activities	**Start Date**	**People Involved**	**Completion Date**
Planning and Preparation			
Discussion to conduct SMS	2 Mar 2021	All team members	30 Mar 2021
Construct research questions	31 Mar 2021	Belachew	2 Apr 2021
Developing protocol V1	3 Apr 2021	Belachew	6 Apr 2021
Protocol v2 will be circulated for comment (will be revised accordingly)	7 Apr 2021	All team members	13 Apr 2021
Produce final v3 of protocol (incorporate feedback from all)	15 Apr 2021	Belachew	20 Apr 2021
Conduct Review			
Stage 1: Download references based on face value papers	22 Apr 2021	Belachew	24 Apr 2021
Stage 2: Check Inclusion/Exclusion Criteria	25 Apr 2021	Belachew	27 Arl 2021
Stage 3: Quality assessment and result form	28 Apr 2021	Belachew	3 May 2021
Stage 4: (Secondary studies) checking result form	5 May 2021	Belachew	10 May 2021
Validate 1: Review process (accepted and rejected papers)	11 May 2021	Belachew	12 May 2021
Validate 2: Results	15 May 2021	Belachew	30 May 2021
Synthesis data	1 Jun 2021	Belachew	29 June 2021
Report the result			
Reporting the review and Discussion on the report	30 Jun 2021	All team members	30 Jun 2021
Report the finding		Belachew et al.	

References

[1] F. Chen, N. Power, J. J. Collins, and F. Ishikawa, “Contemporary requirements challenges and issues: An empirical study in 11 organizations,” *Proc. ACM Symp. Appl. Comput.*, vol. Part F1477, pp. 1592–1599, 2019, doi: 10.1145/3297280.3297435.

[2] A. Eberlein and J. C. Sampaio do Prado Leite, “Agile Requirements Definition : A View from Requirements Engineering Julio Cesar Sampaio do Prado Leite,” *Int. Work. Time-Constrained Requir. Eng.*, no. July 2014, 2002.

[3] S. R. Kourla, E. Putti, and M. Maleki, “REBD: A Conceptual Framework for Big Data Requirements Engineering,” *Comput. Sci. Inf. Technol. (CS IT)*, pp. 79–87, 2020, doi: 10.5121/csit.2020.100608.

[4] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, “Systematic Mapping Studies in Software,” *12th Int. Conf. Eval. Assess. Softw. Eng.*, vol. 12, pp. 1–12, 2008, [Online]. Available: http://content.ebscohost.com/ContentServer.asp?T=P&P=AN&K=22674743&S=R&D=bth&EbscoContent=dGJyMNHX8kSeqK44zdnyOLCmr0qeprZSr6e4SrCWxWXS&ContentCustomer=dGJyMPGosk+xq65QuePfgeyx44Dt6fIA%5Cnhttp://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=2447601.

[5] Barbara Kitchenham and S. Charters, “Guidelines for performing Systematic Literature Reviews in Software Engineering,” 2007. doi: 10.1145/1134285.1134500.

[6] S. Beecham, N. Baddoo, T. Hall, and H. Robinson, “Protocol for a Systematic Literature Review of Motivation in Software Engineering,” 2006.

[7] P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, and M. Khalil, “Lessons from applying the systematic literature review process within the software engineering domain,” *J. Syst. Softw.*, vol. 80, no. 4, pp. 571–583, 2007, doi: 10.1016/j.jss.2006.07.009.