Collaborative Exchange of Systematic Literature Review Results: The Case of Empirical Software Engineering

Fajar J. Ekaputra[1] Marta Sabou[1] Estefanía Serral[2] Stefan Biffl[1]

[1)] Vienna University of Technology, Christian Doppler Laboratory CDL-Flex
Favoritenstrasse 9-11/188, AT 1040 Vienna, Austria
{firstname.lastname}@tuwien.ac.at

[2)] KU Leuven, Dept. of Decision Sciences and Information Management
Naamsestraat 69, 3000 Leuven, Belgium.
estefania.serralasensio@kuleuven.be

ABSTRACT

Complementary to managing bibliographic information as done by digital libraries, the management of concrete research objects (e.g., experimental workflows, design patterns) is a pre-requisite to foster collaboration and re-use of research results. In this paper we describe the case of the Empirical Software Engineering domain, where researchers use systematic literature reviews (SLRs) to conduct and report on literature studies. Given their structured nature, the outputs of such SLR processes are a special and complex type of research object. Since performing SLRs is a time consuming process, it is highly desirable to enable sharing and reuse of the complex knowledge structures produced through SLRs. This would enable, for example, conducting new studies that build on the findings of previous studies. To support collaborative features necessary for multiple research groups to share and re-use each other's work, we hereby propose a solution approach that is inspired by software engineering best-practices and is implemented using Semantic Web technologies.

Categories and Subject Descriptors

H.2.8 [**Database Applications**] Scientific databases; H.5.3 [**Group & Organization Interfaces**] Collaborative computing

Keywords

Collaboration, Research Publication, SLR, EMSE

1. Introduction

Scholarly data has diverse facets. For this paper we distinguish between 1) bibliographic data and 2) content data (including research objects, e.g., experimental workflows). Digital libraries such as Google Scholar or Microsoft Academic Search focus on bibliographic data such as authors, venues, co-authorship relations as well as broad topic domains. The representation and management of content data, such as experimental setups and results, is another important aspect and has received much attention especially in the life-sciences domain. For example, MyExperiment [6] allows sharing and reusing scientific workflows, while the ontologydesignpatterns.org [4] platform supports a community based effort for sharing, validating and reusing ontology design patterns. A key motivation behind representing, storing and providing improved access to research objects is to foster their re-use by other researchers. It is therefore natural that collaborative features are important for systems storing research objects. To that end, MyExperiment and ontologydesignpattern.org build around communities of users and allow their interaction and collaboration through Web2.0 elements. In the Empirical Software Engineering Domain (EMSE), the complexity of the research objects, however, requires a more complex collaboration infrastructure.

In this paper we focus on the issue of *collaboratively managing complex research objects*. In particular, we exemplify such issues in the area of EMSE, where the extraction of such complex research objects are typically performed according to Systematic Literature Study (SLR) approach [5]. We describe the current issues of the EMSE domain in Section 2. In Section 3 we propose the Collaborative Exchange of SLR results (CESLR) approach, a technology agnostic guideline for sharing and publishing EMSE research publication data through integration of results from SLR studies. We discuss the strengths and weaknesses of Semantic Web technologies to implement our solution approach in Section 4, where we also conclude and provide insights into future work.

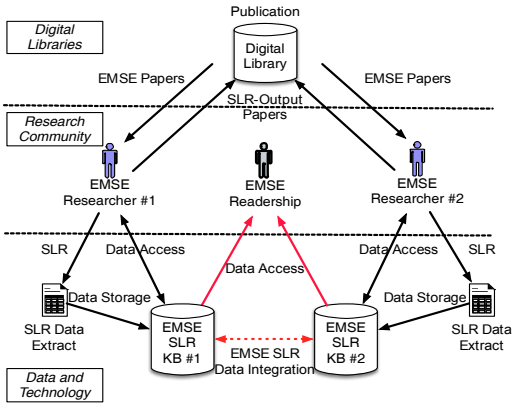

Figure 1 EMSE Stakeholders and Challenges

2. Publication Data Analysis in EMSE

Figure 1 illustrates the current problem setting in EMSE. Researchers inspect previously published papers drawn from digital libraries and apply the SLR process to extract detailed and structured data about the EMSE experiments reported in those publications. Currently the main result of a SLR process is, in general, a specific research synthesis report (i.e., a conference paper, a journal article) made public through digital libraries [2]. Meanwhile the accumulated knowledge in the SLR working material (SLR Data Extract in Figure 1) is typically not made available to other researchers or the general readership.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742027

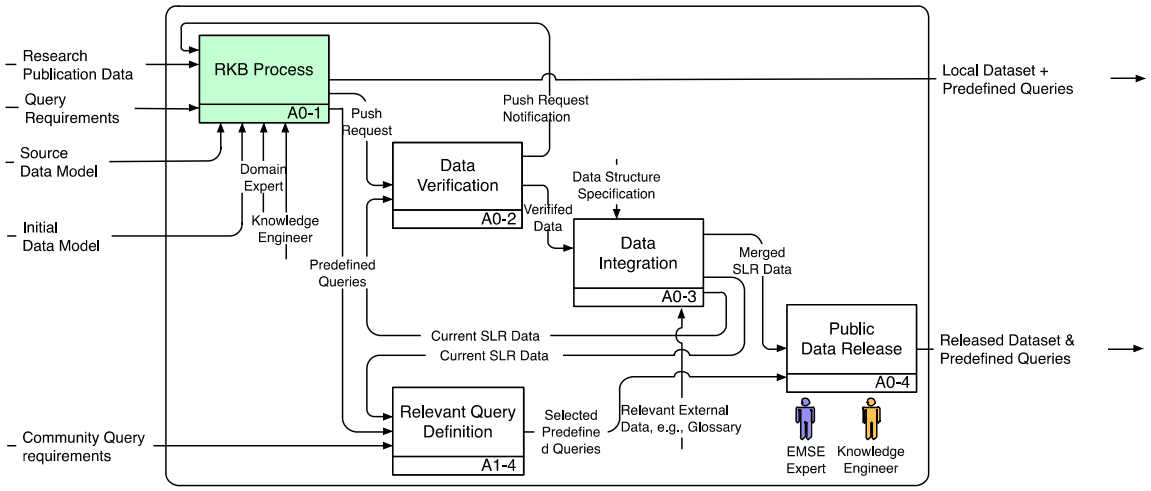

Figure 2 IDEF-0 diagram of CESLR approach.

The lack of sharing the actual SLR output has several drawbacks: (1) the general EMSE readership is deprived from accessing and querying accurate and well-curated data that are the basis for the SLR-based scientific publications; (2) other EMSE researchers cannot explore the underlying extracted information to answer questions related to their specific research goals; (3) there is no possibility for complex, meta-analyses tasks that would rely on combining data from more than one SLR study.

3. CESLR Approach

The Collaborative Exchange of Systematic Literature Review results (CESLR) approach is built on top of the Research Knowledge Base (RKB) [3], an approach that focuses on creating Knowledge Base for a single SLR result. CESLR aims to address drawbacks from traditional SLR and RKB process to manage the research data by providing means for collaborative exchange of SLR study results and provide users with a consistent view of integrated SLR data and reproducible query results. The CESLR approach consists of four main phases as discussed next and shown in Figure 2.

Data Verification. In this process, the pull request generated by one of the RKB processes will be checked and analyzed by both the EMSE domain experts and knowledge engineers. The goal of the verification is to check the relevance and correctness of the data against the content in the integrated repository.

Data Integration. The integration phase decide whether to merge the new data or to create a separate repository for it. It is also possible to add and/or integrate external data into CESLR repository, e.g., integration of glossary data, which contains sets of synonym of EMSE concepts to enhance the term-based search feature.

Relevant Query Definition. Along with the pull request, a list of predefined queries could be sent into CESLR repository. Selected queries coming from the RKB process and additional queries from the domain experts will provide a set of queries for the users of the released data.

Public Data Release. The integration data and the query definitions is aggregated into a release of data. The goal of this phase is to provide a persistent set of data releases that could be accessed and queried consistently.

4. Conclusion and Further Work

We have introduced the CESLR approach, a technology agnostic guideline for integrating EMSE research publication data through integration of results from SLR studies. We are working to build a working prototype of CESLR based on Semantic Web and we have identified the following advantages of using Semantic Web to represent scholarly data in EMSE domain:

- **Flexible data model for data integration.** Semantic web technologies provide a flexible way to handle different data models and to integrate them into a common repository.
- **Concept-based search.** Semantic web technologies provide the capability of concept-based search and improves search results [1].

Furthermore, we have identified several important points for our further work as follows:

- **Extraction of content data**. To support for a (semi-) automatic extraction of content data of research publications.
- **Change management of scholarly data**. To provide a robust ontology and data change support for complex research publication content.
- **Automatic concept classification detection.** To provide an automatic classification of SLR concepts hierarchy from heterogeneous sources.

References

[1] Biffl, S., Kalinowski, M., Ekaputra, F.J., Serral, E. and Winkler, D. 2014. Building Empirical Software Engineering Bodies of Knowledge with Systematic Knowledge Engineering. *Proceeding of 26th SEKE Conference* (2014), 552–559.

[2] Cruzes, D.S. and Dybå, T. 2011. Research synthesis in software engineering: A tertiary study. *Information and Software Technology*. 53, 5 (May 2011), 440–455.

[3] Ekaputra, F.J., Serral, E. and Biffl, S. 2014. Building an Empirical Software Engineering Research Knowledge Base from Heterogeneous Data Sources. *14th I-KNOW Conference* (2014).

[4] Gangemi, A., Gómez-Pérez, A., Presutti, V. and Suárez-Figueroa, M.C. 2007. *Towards a Catalog of OWL-based Ontology Design Patterns*.

[5] Kitchenham, B. and Charters, S. 2007. Guidelines for performing Systematic Literature Review in Software Engineering. *Keele University Technical Report - EBSE-2007-01* (2007).

[6] Roure, D. De, Goble, C. and Stevens, R. 2009. The design and realisation of the Virtual Research Environment for social sharing of workflows. *Future Generation Computer Systems*. (2009).