A Specialized Global Software Engineering Taxonomy for Effort Estimation

Ricardo Britto
Department of Software Engineering
Blekinge Institute of Technology
Sweden
ricardo.britto@bth.se

Emilia Mendes
Department of Computer Science
and Engineering
Blekinge Institute of Technology
Sweden
emilia.mendes@bth.se

Claes Wohlin
Department of Software Engineering
Blekinge Institute of Technology
Sweden
claes.wohlin@bth.se

***Abstract*—To facilitate the sharing and combination of knowledge by Global Software Engineering (GSE) researchers and practitioners, the need for a common terminology and knowledge classification scheme has been identified, and as a consequence, a taxonomy and an extension were proposed. In addition, one systematic literature review and a survey on respectively the state of the art and practice of effort estimation in GSE were conducted, showing that despite its importance in practice, the GSE effort estimation literature is rare and reported in an ad-hoc way. Therefore, this paper proposes a specialized GSE taxonomy for effort estimation, which was built on the recently proposed general GSE taxonomy (including the extension) and was also based on the findings from two empirical studies and expert knowledge. The specialized taxonomy was validated using data from eight finished GSE projects. Our effort estimation taxonomy for GSE can help both researchers and practitioners by supporting the reporting of new GSE effort estimation studies, i.e. new studies are to be easier to identify, compare, aggregate and synthesize. Further, it can also help practitioners by providing them with an initial set of factors that can be considered when estimating effort for GSE projects.**

I. Introduction

For many years, several organizations worldwide have developed software in a globally distributed manner (Global Software Engineering - GSE), aiming at achieving benefits such as reduced time-to-market and access to skillful people all over the world [1]–[4].

Despite all the argued benefits, there are several challenges associated with GSE, which often impact the productivity [5] and effectiveness [6] of distributed software development, leading to delayed projects [7], [8]. The large number of delayed projects reported provides evidence that practitioners have fallen short of obtaining accurate and reliable effort estimates in both collocated and globally distributed projects.

Although many of the challenges that hinder software development are present in both collocated and distributed contexts, they are amplified by geographical, temporal and sociocultural distances, which are present in distributed projects [9]. To better understand how these challenges impact effort estimation processes in globally distributed projects, Britto *et al.* conducted two empirical studies: i) a systematic literature review (SLR) [10], focusing on the state of the art, and ii) a survey, focusing on the state of the practice [11].

Despite the relevance of this research topic, Britto *et al.* identified just a few studies investigating it empirically. In addition, the authors also found out that the related studies are described in an ad-hoc manner, i.e. no common terminology or classification scheme was used to report effort-related studies in the context of GSE projects. The absence of a common terminology and lack of structured knowledge can be detrimental to understand a study's context, making it harder to analyze, compare and aggregating its results to those from other similar studies. Thus, it can hinder the advances in the field and the transfer of research results to industry.

A classification scheme can mitigate the aforementioned problems, in addition to fostering the dissemination of a studys research results to academia and industry [12].

To date, some taxonomies [1] have been designed for organizing the existing knowledge in the GSE field [18]–[20]. A classification scheme, such as a taxonomy, can be beneficial for both researchers and practitioners in four different ways:

1) It can ease the sharing of knowledge [12], [21], [22].
2) It can help to identify gaps in a particular knowledge area [12], [21], [22].
3) It can provide a better understanding of the interrelationships between the factors associated to a particular knowledge area [12].
4) It can support decision making processes [12].

Taxonomies are meant to classify knowledge, which can help describe a researchs context. Specifically in software engineering, the description of an investigations context is mandatory to enable researchers and practitioners to understand the extent to which the reported findings are relevant in their own contexts.

Smite *et al.* proposed a GSE taxonomy [20], but their proposal does not include an exhaustive list of GSE factors; however a taxonomy is a classification scheme that is expected to evolve over time [12]. Considering the findings of Britto *et al.* [10], [11], we identified a need for incorporating additional factors into Smite *et al.*'s taxonomy, enabling this taxonomy

[1] According to the Oxford English Dictionary [13], a taxonomy is "a scheme of classification". This concept was initially devised to classify organisms [14], although it has been applied in many different domains, e.g. Education [15], Psychology [16] and Computer Science [17].

for classifying GSE scenarios from an effort estimation perspective. Furthermore, we identified that there is a distinction between the identified factors; some of them are only relevant for classifying GSE scenarios from an effort estimation perspective, while the others can be used for classifying GSE scenarios from a more general perspective.

Thus, in a previous study we proposed an extension of Smite *et al.*'s taxonomy [23], which includes factors of general use. In this paper we specialize Smite *et al.*'s taxonomy, including factors that are relevant for classifying GSE scenarios from an effort estimation perspective, and not for GSE in general. We refer to this as being a specialization based on a specific perspective (here effort estimation, but other specializations may also be possible). Therefore, the contribution of this paper is two-fold:

- A specialized GSE taxonomy based on evidence from an SLR and a survey, and expert knowledge.
- A validation of the specialized GSE taxonomy via the classification of eight completed GSE industrial projects.

The remainder of this paper is organized as follows: The related work is presented in Section II. Section III details the applied research design and methodology. Section IV presents the specialized taxonomy, followed by its validation in Section V. Section VI provides a discussion of the academic/industrial implications and limitations of this work. Finally, in Section VII we draw our conclusions and present directions for future work.

II. Related work

We identified five taxonomies [18]–[20], [23], [24] and two ontologies [25], [26] in the GSE context[2].

The identified knowledge organization schemes can be categorized as follows:

- **Description approach**: Three studies [19], [25], [26] proposed graphical-based approaches that are more adequate to describe rather than to classify GSE projects. It comes from the fact that none of these approaches has dimensions with clear classification criteria associated to them; rather, they provide a set of "variables" that should be instantiated.
- **Classification approach**: The other two studies [18], [20] are more adequate to classify GSE projects, because they are organized in dimensions that have categories with associated classification criteria. Britto *et al.* [23] and Techio *et al.* [24] have extended Smite *et al.*'s taxonomy

A. Smite et al.'s GSE taxonomy

Smite *et al.* [20] conducted a Delphi-inspired study with GSE researchers to develop an empirically based glossary and taxonomy, focused on the sourcing strategy aspect of GSE projects. The taxonomy was developed for classifying the relationship between pairs of sites, although it is equally possible to describe more complex GSE projects, with more than two sites. The taxonomy is presented in Figure 1 and detailed in Table I. In Figure 1, the rounded boxes represent the dimensions (factors) of the taxonomy used to classify GSE projects, while the squared boxes represent the categories (possible classification values) of each dimension[3].

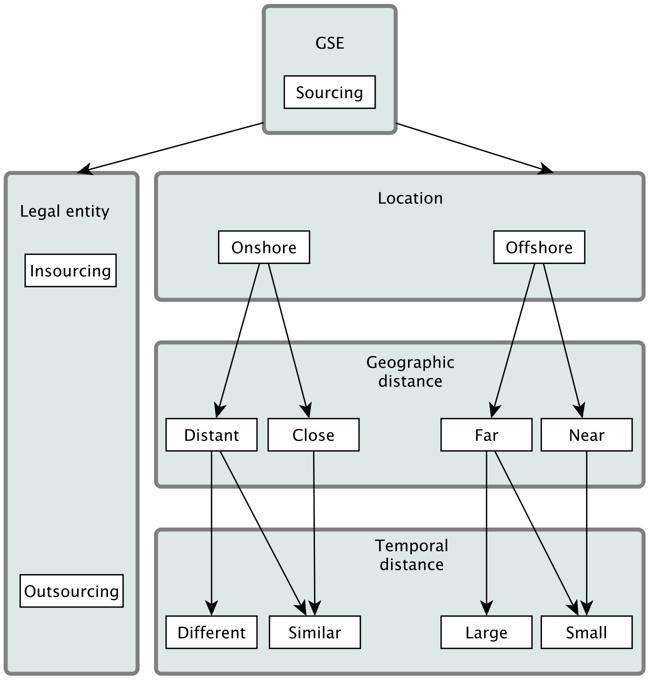


Figure 1. The original GSE taxonomy (Adapted from Smite *et al.* [20]).

B. Britto et al.'s extended GSE taxonomy

Smite *et al.*'s taxonomy (base GSE taxonomy) was extended by Britto *et al.* [23], who based their extension on the results from two empirical studies [10], [11] and expert knowledge. Britto *et al.* added seven new dimensions to the original taxonomy. The extended taxonomy is presented in Figure 2 and the extended dimensions are detailed in Table II. Note that in Figure 2 the black rounded boxes represent the original dimensions from the base GSE taxonomy (see Figure 1).

C. Additional related work

Gumm [18] developed a taxonomy to classify GSE projects in terms of distribution dimensions. Its goal was to provide a foundation to discuss the challenges related to GSE projects and was based on an earlier literature study performed by the same author. The proposed taxonomy uses four different dimensions (physical distribution, organizational distribution, temporal distribution and distribution between stakeholder groups) to classify the ways in which people and artifacts can be distributed in GSE projects.

Laurent *et al.* [19] proposed a taxonomy and a visual notation to address the requirements engineering aspect of GSE projects. The main goal of the authors was to design

[2] Other taxonomies were also identified, but either they were synthesized by Smite et al. [27]–[29] or they have no supporting empirical evidence [30].

[3] Note that the same idea applies to Figures 2 and 4.

Table I
DIMENSIONS OF SMITE *et al.*'S TAXONOMY.

Dimension	Categories	Description
GSE	Sourcing	This dimension contains the root of the taxonomy, called sourcing. In this context, sourcing means some form of external software development.
Location	Onshore, Offshore	A sourcing can be delegated to a site in the same country, i.e. onshore, or to a site in another country, i.e. offshore.
Legal entity	Insourcing, Outsourcing	Independently from the location, a sourcing can be transferred to a different branch (site) of the company, i.e. insourcing, or subcontracted to a different legal entity (company), i.e. outsourcing.
Geographical distance	Close, Distant, Near, Far	In onshore projects, the geographical distance is considered: **close** when it is possible to have relatively frequent face-to-face meetings, since no flights are required to go from one site to the other; **distant** when at least one flight is required to have face-to-face meetings, which yields time and cost increases. In offshore projects, the geographical distance is considered: **near** when the required flying time is less than two hours; **far** when the flying time is longer than two hours and staying overnight is usually required.
Temporal distance	Similar, Different, Small, Large	In onshore projects, the temporal distance is considered: **similar** when there is a time difference of one hour or less; **different** when the time difference between two sites is longer than one hour. In offshore projects, the temporal distance is considered: **small** when there is a time distance between sites of four hours or less; **large** when there is a time distance between two sites of more than four hours.

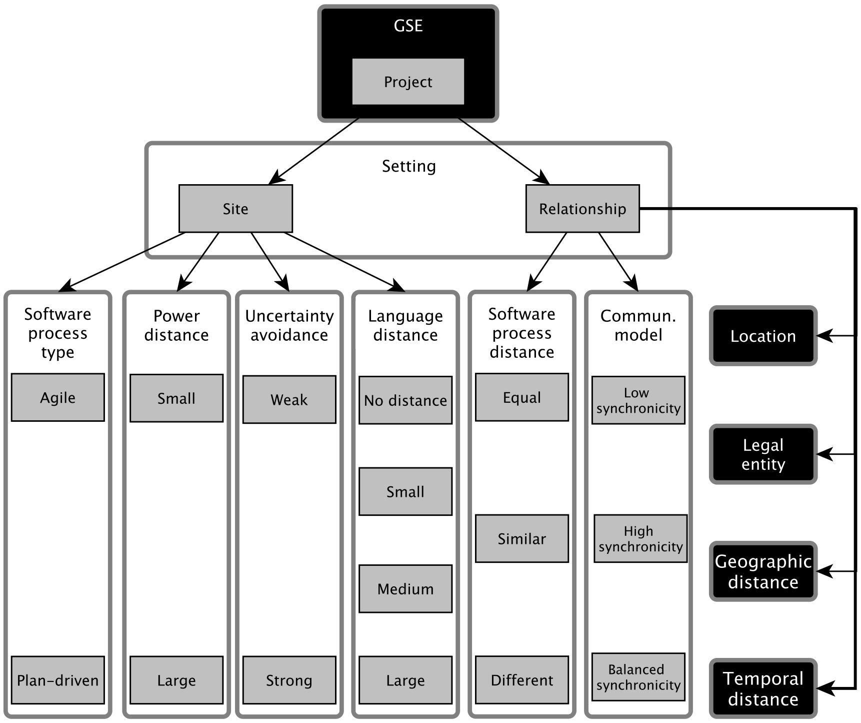


Figure 2. The extended GSE taxonomy (Adapted from Britto *et al.* [23]).

a common language for modeling the requirements of GSE projects and to allow project managers to manage distributed requirements in a better way. The proposal was derived from the findings of a broad study performed with industrial partners (seven different projects). Interviews were performed with the team leaders responsible for eliciting and gathering the requirements in each project.

Vizcaino *et al.* [25] developed an ontology, called O-GSD, which was aimed at easing the communication and avoiding misunderstanding in GSE projects. This ontology was iteratively developed in the context of a project that involved five companies and two universities in Spain. The authors used the REFSENO (representation formalism for software engineering) [31] to create the ontology.

Marques *et al.* [26] introduced an ontology for team task allocation in GSE projects. This ontology was developed based on the findings from a systematic mapping study performed by the authors and aimed at clarifying the concepts related to team task allocation in distributed projects. The authors used UML class diagrams to represent the ontology, and the main concepts they addressed were artifact, competence and constraints.

Techio *et al.* [24] extended Smite *et al.* taxonomy aiming at classifying empirical evidence in GSE. To do so, the authors performed an expert opinion survey with GSE researchers and practitioners. They added seven new dimensions to Smite *et*

Table II
EXTENDED DIMENSIONS OF BRITTO *et al.*'S EXTENDED TAXONOMY.

Dimension	Categories	Description
Setting	Site, Relationship	A GSE project can be classified both at the **site** level or **relationship-between-pair-of-sites** level.
Software process type	Agile, Plan-driven	A site is to be classified as **agile** if its software process is mainly based on agile practices. Otherwise, it is to be classified as **plan-driven**.
Power distance	Small, Large	A site has **large** power distance when its power distance index is greater than 50; otherwise, it is considered **small**.
Uncertainty avoidance	Weak, Strong	A site has **strong** uncertainty avoidance when its uncertainty avoidance index is greater than 63; otherwise, it is considered **weak**.
Language distance	No distance, Small, Medium, large	The language distance is measured in terms of the distance between a site's mother tongue and English (language distance index), which is a number that varies from 0 to 1. There is **no distance** when a site's mother tongue is English or there is no need for a *lingua franca*. The distance is **small** when the a site's language distance index is smaller or equal to 0.4. It is considered **medium** when a site's language distance index is greater than 0.4 and smaller than 0.57. Finally, it is considered **large** when it is greater than 0.57 and smaller or equal to 1.
Software process distance	Equal, Similar, Different	The software processes of two sites are considered **equal** when they use the same workflows, roles and practices to develop software. They have **similar** processes when they have the same software process type, but the workflows, roles and practices are not exactly the same. The processes are considered different when there are no commonalities.
Communication model	Low synchronicity, high synchronicity, balanced synchronicity	It is **low** when the communication is mainly based in asynchronous media (e.g. email). It is **high** when the communication is mainly based in synchronous media (e.g. instant messaging tools). It is **balanced** when the communication model has both synchronous and asynchronous media and each media type is used for its most adequate purpose.

al.'s taxonomy and validated their proposal by classifying 26 papers. According to the authors, their extended taxonomy was able to classify most of the studies.

D. Research gaps

Only the taxonomies proposed by Gumm [18] and the base GSE taxonomy [20] (and its extensions) are considered as knowledge classification approaches. The base GSE taxonomy is more comprehensive, providing a wider range of relevant dimensions and clear criteria to classify GSE projects. In addition, it was also developed with the input from several GSE experts, which adds to such a taxonomy more strength and credibility.

The base GSE taxonomy (and its extensions) was designed to enable the classification of GSE research in general, accounting for many different factors. However, GSE research with a specialized scope (e.g. effort estimation or testing in the GSE context) may demand additional aspects that are relevant only within the specific scope. In this paper, we complement into the base GSE taxonomy dimensions that are relevant only within the effort estimation scope. Therefore, we included into the specialization presented herein the following aspects, as identified by Britto *et al.* [10], [11]:

- The effort estimation process itself can be carried out in different ways, e.g. only one site is responsible for the entire estimation or the estimation is distributed among different sites.
- The sites might have different roles within an effort estimation process, like providing data or managing the estimation process.
- Effort estimates can be obtained in different stages of a project, like before or after the requirements elicitation and analysis.

Britto *et al.*'s extension (presented in our previous work [23]) and the specialization presented herein were developed in parallel, i.e. they have the same starting point (two empirical studies and Smite *et al.*'s taxonomy). However, the two works deal with different types of factors; the aforementioned aspects are relevant only from an effort estimation perspective, while the extension dealt with general factors.

The core of the GSE base taxonomy is composed by its root and the dimension "setting" (added in the extension), i.e. the extension also influenced the design of the specialized dimensions. Figure 3 shows how the original taxonomy, the extension and the specialization relate to each other.

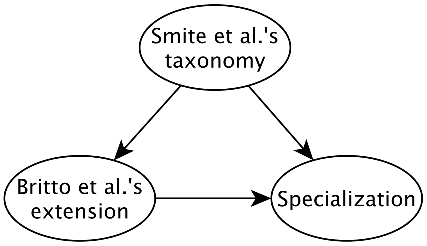


Figure 3. The relationship between the specialization, Smite *et al.*'s taxonomy and the extension.

III. Research design and methodology

This section presents the research design and methodology used herein. The following research questions drove the work reported in this paper:

- **RQ1**: What dimensions are needed to enable Smite *et al.*'s taxonomy to classify GSE scenarios from an effort estimation perspective?
- **RQ2**: What is the utility of the specialized taxonomy?

To answer RQ1 and RQ2, **first**, we identified the factors to be incorporated into the specialized taxonomy. We used the knowledge of the authors about the topic, the results of a systematic literature review [10] and the result of a survey [11] as input. We identified three new factors ("estimation stage", "estimation process role" and "estimation process architectural model") that we judged as essential to capture in the effort estimation specialized taxonomy.

Second, we identified categories for each dimension. To do so, we used relevant literature related to each dimension (see Section IV) and our own knowledge for identifying meaningful categories with clear classification criteria. Clear classification criteria facilitate the usage of the taxonomy and help in making correct categorizations of the subject matter [32].

Third, we combined the new dimensions with the dimensions of Smite *et al.*'s taxonomy and Britto *et al.*'s extension; no inconsistency was identified.

Fourth, we validated our specialized taxonomy. A taxonomy can be validated in three ways [20]:

- **Orthogonality demonstration** - The orthogonality of the taxonomy dimensions and categories should be demonstrated.
- **Benchmarking** - The taxonomy should be compared with other similar classification schemes.
- **Utility demonstration** - The utility of the taxonomy should be demonstrated through the classification of existing knowledge.

The orthogonality of the new dimensions was ensured by defining categories with clear classification criteria (see Section IV). It was not possible to perform benchmarking, because there was no taxonomy with a focus on effort estimation. Finally, to demonstrate the utility of our extended taxonomy, we illustrate its usage by classifying eight industrial completed GSE projects (see Section V).

IV. The specialized GSE taxonomy

Figure 4 displays the specialized GSE taxonomy proposed herein[4]. Three new dimensions were incorporated: "estimation stage", "estimation process role" and "estimation process architectural model".

The relationships between the specialized dimensions, the extended dimensions and the original dimensions of Smite *et al.'s* GSE taxonomy are as follows:

- The dimension "GSE" is parent of all other dimensions.
- The classifications by means of the dimensions "estimation stage" and "estimation process role" are related to the category "site" of dimension "setting".
- The classification by means of the dimension "estimation process architectural model" is related to category "relationship" of the dimension "setting".

Note that the subject to be classified is a *project*, which is herein conceptualized as "a temporary endeavor undertaken to create a unique product, service, or result" [33]. We consider that a project has sites and relationship-between-pair-of-sites. A "site" is defined as a unit composed of human resources that interact with other sites (nodes). We define a "relationship" as the relationship between two sites interacting in a project (edge).

Although we focus on the specialized dimensions herein, the specialized taxonomy is composed by all the dimensions related to the GSE taxonomy proposed by Smite *et al.* (original, extended and specialized), i.e. one should use all the dimensions to classify projects in effort estimation-related research.

The three new dimension are further detailed in Sections IV-A, IV-B and IV-C respectively.

A. Estimation process role

Considering the activities that are part of the effort estimation process, a site can play different roles. We define the following possible categories:

- **Estimator** - A site plays this role when it is responsible for applying an effort estimation technique to obtain effort estimates.
- **Provider** - A provider site is the one that is not able or allowed to obtain effort estimates by itself. Thus, it should supply historical data or expert's knowledge to an estimator site. Eventually, a site that plays this kind of role could give feedback on the obtained effort estimates.
- **Estimator & Provider** - A site of this type calculates its own effort estimates, considering its own historical data and/or expert's knowledge. However, to understand the overall effort related to a given project, the effort estimates obtained by sites of this kind have to be combined.

B. Estimation stage

During the lifetime of a particular software project, the effort estimates can be calculated or refined many times. Considering the defined roles that a site can play into an effort estimation process, it is also important to describe when historical data from past projects or knowledge from experts are collected and when the effort estimates are obtained. The difficulty to perform these activities can vary depending on the moment in time they are carried out.

Thus, we have defined three possible stages upon when data/knowledge collection and effort estimates' attainment can be carried out by a site, which led to the following categories:

[4] The details of the original (black rounded boxes) and extended (dark gray rounded boxes) dimensions are not shown here to facilitate the presentation of the specialized dimensions. See Section II for more details about the original and extended dimensions.

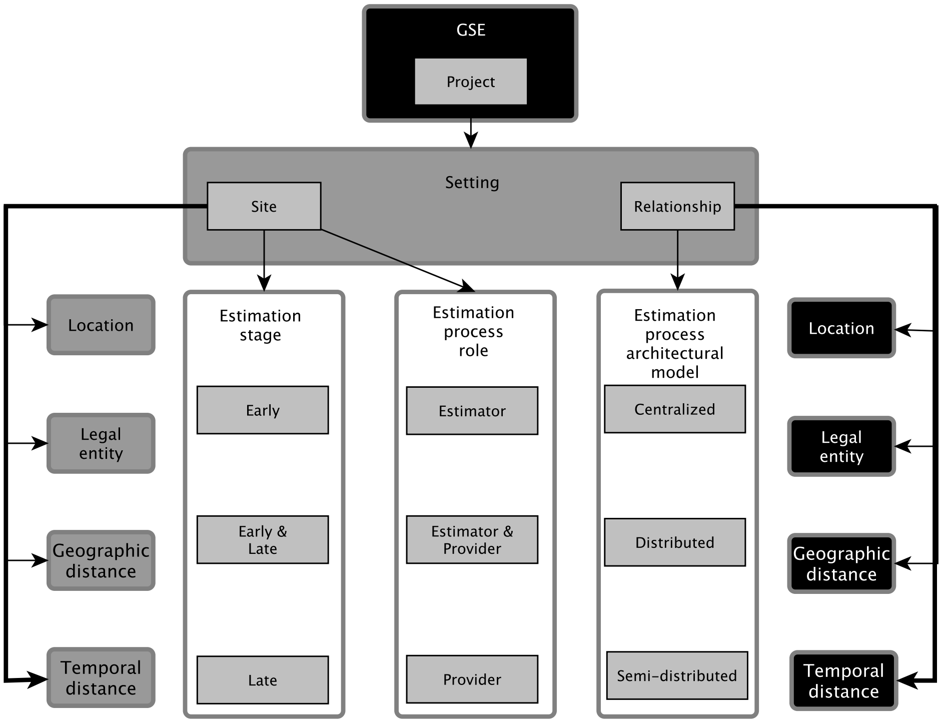


Figure 4. The specialized GSE taxonomy.

- **Early** - When the effort estimation process is performed just after requirements' elicitation, we call it early effort estimation process.
- **Late** - When the effort estimation process is performed after requirements' analysis and design, we call it late effort estimation process. In general, this type of estimation is expected to be performed to refine early effort estimates.
- **Early & Late** - A combination of both the abovementioned types.

C. Estimation process architectural model

The purpose of estimating effort as part of managing a project is to predict the amount of person-hours required to accomplish the set of tasks needed as part of a project's life cycle, based on a set of inputs such as the knowledge/data of previous similar projects and other application and project characteristics that are believed to be related to effort [34].

A generic effort estimation process encompasses four different activities, as follows [34]:

- **Activity 1** - Collect data and/or knowledge from similar finished projects.
- **Activity 2** - Build an estimation model based on the collect data/knowledge.
- **Activity 3** - Estimate the size and determine the cost drivers' values of the new software project.
- **Activity 4** - Estimate the effort using the estimated size and determined cost drivers' values as inputs.

Effort estimation techniques can be classified as expert-based, algorithmic or artificial intelligence [35]. The sequence in which the aforementioned activities are executed into the effort estimation process depends on the type of the effort estimation technique. In addition, activity 2 is not explicitly executed by expert-based effort estimation techniques, since the estimates are subjectively obtained by this kind of technique [34].

On one hand, there is no major concern about the way in which the activities of the effort estimation process are coordinated and performed in collocated projects, since all the steps are performed in the same place. On the other hand, the distributed nature of GSE projects allows for different interaction manners between sites, which can affect the entire effort estimation process.

To the best of our knowledge, there is no existing classification of possible collaboration patterns when estimating effort in GSE projects. Therefore, we combined effort estimation and the distributed systems literature to design this dimension of our taxonomy.

According to Coulouris *et al.* [36], a distributed system is defined as "a system in which hardware or software components located at networked computers communicate and coordinate their actions only by passing messages". The computers of a distributed system can be positioned in the same room, in the same building, in different countries or even in different continents. If we replace the word "computer" by "teams", the definition of distributed systems becomes very similar to the definition of GSE.

In distributed systems, an architectural model describes the structure of a system in terms of its components and their relationships. There are two main architectural models: Client-server and peer-to-peer.

In the **client-server** architectural model, client computers consume the resources managed by server computers. Each computer in a given distributed system is called an element. The roles of both client and server elements are well-determined, despite that in some cases a server can be a client from the point of view of another server. In general, when a distributed system is built based on the client-server model, the server elements have better hardware resources than the client ones.

Client-server systems are reasonably easy to construct and manage. However, it is difficult to scale systems based on this model, since the services (or "expensive" tasks) are supplied by just a few system elements.

In the **peer-to-peer** architectural model, all elements (computers) of a distributed system play similar roles, working in a cooperative way to fulfill tasks. There is no such distinction between client and server elements, like in the client-server model. Although peer-to-peer systems tend to be harder to construct and manage, they are highly scalable, since each element can carry out "expensive" tasks.

The effort estimation process architectural model dimension of our taxonomy was inspired by the above-described concepts from the distributed systems knowledge area. Thus, we defined three different architectural models, which led to three different dimensions in this dimension, respectively called centralized, distributed and semi-distributed.

The **centralized** model (Figure 5) is based on the client-server distributed system model. Considering a given relationship between pairs of sites, one of the sites plays a role as a "server" (site x), centralizing all the effort estimation process, in a top-down fashion. The other site of the pair can be seen as a "client" (site y), being able of just using the effort estimates calculated by the "server" site. Eventually, the client site can provide complementary data to support the effort estimation.

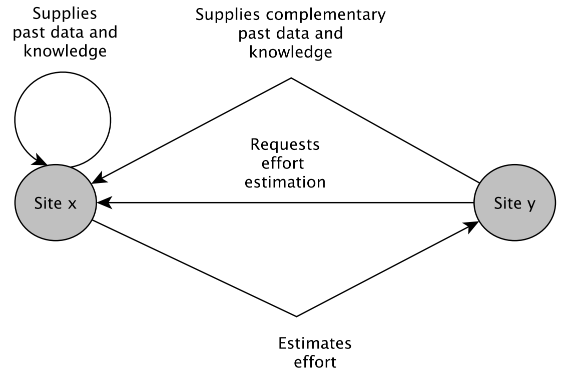

Supplies past data and knowledge
Supplies complementary past data and knowledge
Requests effort estimation
Site x
Site y
Estimates effort

Figure 5. Centralized model.

The **distributed** model (Figure 6) is based on the peer-to-peer distributed system model. Considering a given relationship between two sites, both sites (sites x and y) are responsible for calculating their own effort estimates based on the local data or knowledge, in a bottom-up fashion. The effort estimates calculated in each site are later on combined and shared by the "integrator" site (site x), to allow both sites to understand the overall effort inherent to the project.

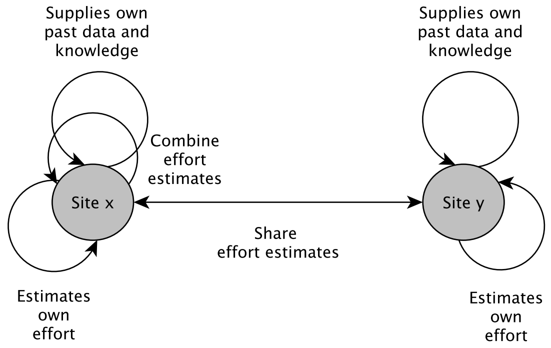


Figure 6. Distributed model.

The **semi-distributed** model (Figure 7) is based on both distributed system models. In this model, a "server" site (site x) is responsible for calculating all the effort estimates regarding both sites of the pair. However, differently of the centralized model, the client site (site Y) can provide feedback about the calculated effort estimates, so that the "server" site could perform further refinements. Likewise the centralized model, the client site can eventually supply complementary data to support the effort estimation.

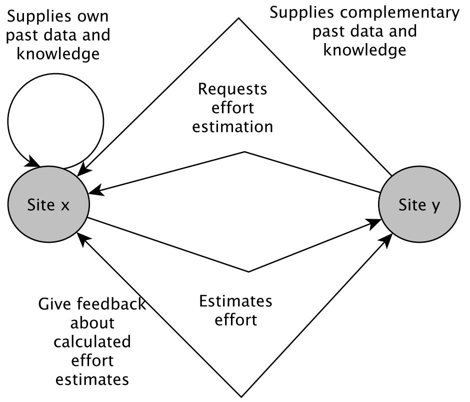

Supplies own past data and knowledge
Supplies complementary past data and knowledge
Requests effort estimation
Site x
Site y
Give feedback about calculated effort estimates
Estimates effort

Figure 7. Semi-distributed model.

V. Validation

As discussed in Section III, a taxonomy can be validated through orthogonality demonstration, benchmarking and utility demonstration. The dimensions' orthogonality was ensured by defining categories with clear classification criteria (see Section IV). We were not able to perform benchmarking, since there is no other taxonomy that is capable of classifying GSE projects from an effort estimation perspective. To demonstrate the utility of our extended taxonomy, we classified eight finished GSE projects.

The projects were obtained from Ramasubbu *et al.*[5] [37]. Since the paper did not report all the data needed to perform

[5] We sent emails to the authors of all primary studies included in Britto *et al.* [10], but only Ramasubbu *et al.* responded, which is the reason why we only used their data to validate our proposal.

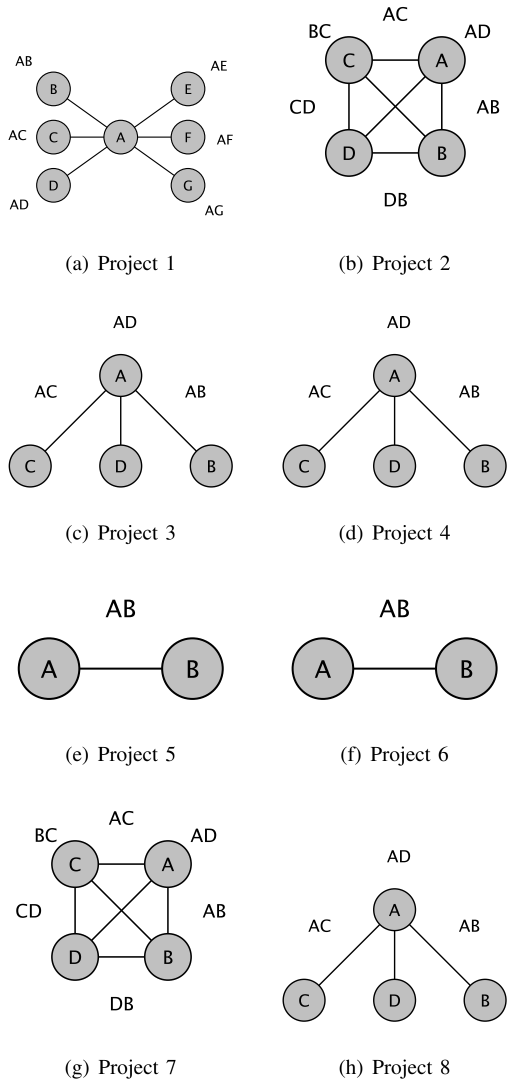
the classification, we contacted the authors to obtain the required data by filling out an excel spreadsheet containing eight tabs (one per project). Each tab had the following fields: project's ID (PID), company's name, software domain, site's ID (SID), site's country, site's city, site's main language, estimation process role, estimation stage, estimation process architectural model, relationship ID (RID) and relationship's sites. Clarifications related to the spreadsheet was provided whenever required.

Due to a non-disclosure agreement, the name of the companies and the domain of the software applications developed/maintained were not provided. The cities wherein the sites were placed were also not explicitly stated due to the same confidentiality issues. Rather, they presented the sites' locations using countries instead. In the case of sites placed in the USA, the region of each site was also provided.

We used the provided data for conducting the classification of the eight projects. Figure 8 shows the setup of each project, i.e. the connections between the involved sites[6], while Table III shows the classification results. The projects classified by means of the proposed specialized taxonomy provided a variety of different setups:

- **Project 1** had seven sites and project management was concentrated in site A (USA), i.e. there was no interaction between sites B, C, D, E, F and G; all the sites only interact with A.
- In **Project 2**, four onshore sites (USA) were involved and all the sites directly interact with each other, although site A had the biggest responsibility regarding project management.
- **Project 3** had four sites and project management was concentrated in site A (India). Despite the fact that sites B, C and D were all located in the USA, they did not interact with each other, only with site A.
- In **Project 4**, four sites were involved and project management was concentrated in site A (India). The other sites only interact with site A.
- **Project 5** and **project 6** had two sites involved each and project management was mainly the responsibility of site A for both project 5 and project 6 (India and Germany respectively).
- In **Project 7**, four onshore sites (India) were involved and all the sites directly interact with each other, although site A had the main responsibility regarding project management.
- **Project 8** had four sites and project management was concentrated in site A (USA). Despite the fact that sites B, C and D were all located in India, they did not interact with each other, only with site A.

Note that the specialized taxonomy was capable of classifying projects that have different configurations, number of sites, locations and distinct estimation processes. It is also important to highlight that the specialized taxonomy is consistent with the base taxonomy [20] and its extension [23].



(a) Project 1

(b) Project 2

(c) Project 3

(d) Project 4

(e) Project 5

(f) Project 6

(g) Project 7

(h) Project 8

Figure 8. Setup of the classified projects.

VI. Discussion

To lead the research community to "speak the same language", it is important to ensure that there is a consensual terminology and that this consensual terminology will not fragment over time. Our previous work [23] illustrates that it is possible to keep evolving a taxonomy of general use without fragmenting its content. Herein we showed that it is possible to go beyond and specializing such a taxonomy accounting for factors that are relevant only for specific perspectives (effort estimation in our case). As in our previous work, the specialized taxonomy remains consistent with the original taxonomy.

The specialized taxonomy presented herein can help GSE researchers to report the context of new GSE effort estimation research in a more systematic, clear and comparable way. Therefore, it can facilitate the analysis, comparison and aggregation of results from new studies, fostering the advancement

[6] The nodes represent the sites and the edges represent the relationships between sites.

Table III
CLASSIFICATION RESULTS.

PID	SID	Country	Estimation process role	Estimation stage	Architectural model
1	A	USA	Estimator & Provider	Early & Late	Semi-distributed
	B	USA	Provider	Early & Late	
	C	India	Provider	Early & Late	
	D	India	Provider	Early & Late	
	E	Singapore	Provider	Early & Late	
	F	Germany	Provider	Early & Late	
	G	Australia	Provider	Early & Late	
2	A	USA	Estimator & Provider	Early	Distributed
	B	USA	Estimator & Provider	Early	
	C	USA	Estimator & Provider	Early	
	D	USA	Estimator & Provider	Early	
3	A	India	Estimator & Provider	Early & Late	Semi-distributed
	B	USA	Provider	Early & Late	
	C	USA	Provider	Early & Late	
	D	USA	Provider	Early & Late	
4	A	India	Estimator & Provider	Early & Late	Semi-distributed
	B	Germany	Provider	Early & Late	
	C	Spain	Provider	Early & Late	
	D	England	Provider	Early & Late	
5	A	India	Estimator & Provider	Early & Late	Distributed
	B	Japan	Estimator & Provider	Early & Late	
6	A	Germany	Estimator & Provider	Early	Distributed
	B	India	Estimator & Provider	Early	
7	A	India	Estimator & Provider	Early & Late	Distributed
	B	India	Estimator & Provider	Early & Late	
	C	India	Estimator & Provider	Early & Late	
	D	India	Estimator & Provider	Early & Late	
8	A	USA	Estimator & Provider	Early & Late	Distributed
	B	India	Estimator & Provider	Early	
	C	India	Estimator & Provider	Early	
	D	India	Estimator & Provider	Early	

of this research field. Further, it can also help researchers to identify literature of interest.

With regard to the contribution that our specialized taxonomy can make to practice, it can help practitioners to identify useful literature related to different contexts and consequently also help them to address different problems related to effort estimation in GSE.

The specialized taxonomy, in addition to the extended version of Smite *et al.*'s taxonomy [23], can also be used for designing a checklist, which can: i) help practitioners when estimating the effort of new projects; ii) provide an initial set of factors to be considered by the practitioners when estimating their GSE projects; and iii) be used for recording the factors used when estimating the effort for projects, which can be later on used as a knowledge base for subsequent project estimation.

As with most studies, the research reported herein comes with some limitations, as follows:

- The dimensions proposed in this paper do not represent an exhaustive list. However, the specialized taxonomy can evolve over time as long as new factors are identified.
- Only the authors of this paper provided expert knowledge during the design of the specialized dimensions, thus the participation of other experts would increase the reliability of the specialized taxonomy, and perhaps lead to a wider set of specialized factors.
- We validated the specialized taxonomy by classifying only eight projects, thus it is important to classify more GSE projects to strengthen the taxonomy's external validity, specifically because not all the possible scenarios were available in the sample of projects that we used to validate our proposal.
- The specialized taxonomy has not yet been used by practitioners in ongoing projects, i.e. there is no empirical evidence ensuring the utility of our proposal in this kind of situation. So, we also encourage GSE researchers to apply our proposed specialized taxonomy in industrial settings so to validate further its usefulness in practice.
- Herein we just focused on specializing Smite *et al.*'s taxonomy from the effort estimation perspective. To show that it is possible to specialize such a taxonomy accounting for other specific factors, we encourage other GSE researchers to look at GSE projects from other perspectives (e.g. software quality and software testing).

VII. CONCLUSIONS AND FUTURE WORK

This paper presents a specialized taxonomy for classifying GSE projects from an effort estimation perspective, which is based on a previous taxonomy [20], two empirical studies [10], [11] and on expert knowledge.

We addressed the first research question of this study (RQ1) by incorporating three new dimensions into Smite *et al.*'s taxonomy, named "estimation process role", "estimation stage" and "estimation process architectural model".

To validate the specialized taxonomy and demonstrate its utility (RQ2), we illustrated its usage by classifying eight finished GSE projects.

The results show that the specialized taxonomy can help both researchers and practitioners by facilitating the reporting and understanding of GSE effort estimation related research. Jointly with the extended dimensions, it can also help practitioners to identify factors that can be incorporated in their effort estimation processes.

To further help practitioners, we intend to design a checklist, which will combine the specialized dimensions presented herein, and the original and extended dimensions of Smite *et al.*'s taxonomy. We also intend to investigate other perspectives that can be used to specialize Smite *et al.*'s taxonomy, such as software testing and software quality.

VIII. Acknowledgments

We would like to thank CNPq, UFPI and INES for partially supporting this work, and Dr. Narayan Ramasubbu for helping us with the validation of the proposal put forward in this paper.

References

[1] N. Ramasubbu, M. Cataldo, R. K. Balan, and J. D. Herbsleb, "Configuring global software teams: A multi-company analysis of project productivity, quality, and profits," in *Proceedings of the 33rd International Conference on Software Engineering - ICSE'11*, 2011, pp. 261–270.

[2] E. Conchúir, P. J. Ågerfalk, H. Holmström, and B. Fitzgerald, "Global software development: Where are the benefits?" *Communications of the ACM*, vol. 52, no. 8, pp. 127–131, aug 2009.

[3] A. B. Bondi and J. P. Ros, "Experience with training a remotely located performance test team in a quasi-agile global environment," in *Proceedings of the 2009 Fourth IEEE International Conference on Global Software Engineering - ICGSE'09*. Washington, DC, USA: IEEE Computer Society, 2009, pp. 254–261.

[4] J. Herbsleb and D. Moitra, "Global software development," *IEEE Software*, vol. 18, no. 2, pp. 16–20, 2001.

[5] J. D. Herbsleb, A. Mockus, T. A. Finholt, and R. E. Grinter, "Distance, dependencies, and delay in a global collaboration," in *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work - CSCW'00*. New York, NY, USA: ACM, 2000, pp. 319–328.

[6] J. A. Espinosa, N. Nan, and E. Carmel, "Do gradations of time zone separation make a difference in performance? a first laboratory study," in *Second IEEE International Conference on Global Software Engineering - ICGSE'07.*, Aug 2007, pp. 12–22.

[7] J. D. Herbsleb and A. Mockus, "An empirical study of speed and communication in globally distributed software development," *IEEE Transactions on Software Engineering*, vol. 29, no. 6, pp. 481–494, June 2003.

[8] D. Šmite, "Global software development projects in one of the biggest companies in latvia: Is geographical distribution a problem?" *Software Process: Improvement and Practice*, vol. 11, no. 1, 2006.

[9] J. Ågerfalk and B. Fitzgerald, "Flexible and distributed software processes: old petunias in new bowls," *Communications of the ACM*, vol. 49, pp. 27–34, 2006.

[10] R. Britto, V. Freitas, E. Mendes, and M. Usman, "Effort estimation in global software development: A systematic literature review," in *Global Software Engineering (ICGSE), 2014 IEEE 9th International Conference on*, Aug 2014, pp. 135–144.

[11] R. Britto, E. Mendes, and J. Börstler, "An empirical investigation on effort estimation in agile global software development," in *Proceedings of the IEEE International Conference on Global Software Engineering - ICGSE'15*, 2015.

[12] S. Vegas, N. Juristo, and V. Basili, "Maturing software engineering knowledge through classifications: A case study on unit testing techniques," *Software Engineering, IEEE Transactions on*, vol. 35, no. 4, pp. 551–565, July 2009.

[13] O. Dictionaries, *Oxford Dictionary of English*. OUP Oxford, 2010.

[14] C. Linnaeus, *System of nature through the three kingdoms of nature, according to classes, orders, genera and species, with characters, differences, synonyms, places (in Latin)*, 10th ed. Laurentius Salvius, 1758.

[15] B. S. Bloom, *Taxonomy of educational objectives. Vol. 1: Cognitive domain*. McKay, 1956.

[16] T. E. Moffitt, "Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy." *Psychological Review*, vol. 100, no. 4, pp. 674–701, Oct 1993.

[17] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1-3, pp. 7–42, 2002.

[18] D. C. Gumm, "Distribution Dimensions in Software Development Projects: A Taxonomy," *IEEE Software*, vol. 23, no. 5, pp. 45–51, 2006.

[19] P. Laurent, P. Mader, J. Cleland-Huang, and A. Steele, "A Taxonomy and Visual Notation for Modeling Globally Distributed Requirements Engineering Projects," in *Proceedings of 5th IEEE International Conference on Global Software Engineering - ICGSE'10*, Princeton, USA, 2010, pp. 35–44.

[20] D. Šmite, C. Wohlin, Z. Galvina, and R. Prikladnicki, "An empirically based terminology and taxonomy for global software engineering," *Empirical Software Engineering*, vol. 19, no. 1, pp. 105–153, 2014.

[21] I. Vessey, V. Ramesh, and R. L. Glass, "A unified classification system for research in the computing disciplines," *Information and Software Technology*, vol. 47, no. 4, pp. 245–255, Mar. 2005.

[22] C. Wohlin, "Writing for synthesis of evidence in empirical software engineering," in *Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement - ESEM '14*. New York, NY, USA: ACM, 2014, pp. 46:1–46:4.

[23] R. Britto, C. Wohlin, and E. Mendes, "An extended global software engineering taxonomy," 2015.

[24] A. Techio, R. Prikladniki, and S. Marczak, "Reporting empirical evidence in distributed software development: An extended taxonomy," in *Proceedings of the IEEE International Conference on Global Software Engineering - ICGSE'15*, 2015.

[25] A. Vizcaíno, F. García, I. Caballero, J. C. Villar, and M. Piattini, "Towards an ontology for global software development," *IET Software*, vol. 6, no. 3, p. 214, 2012.

[26] A. B. Marques, J. R. Carvalho, R. Rodrigues, T. Conte, R. Prikladnicki, and S. Marczak, "An Ontology for Task Allocation to Teams in Distributed Software Development," in *Proceedings of 8th International Conference on Global Software Engineering - ICGSE'13*, Bari, Italy, 2013, pp. 21–30.

[27] P. J. Ågerfalk and B. Fitzgerald, "Outsourcing to an unknown workforce: Exploring opensourcing as a global sourcing strategy," *MIS Q.*, vol. 32, no. 2, pp. 385–409, Jun. 2008.

[28] M. Robinson and R. Kalakota, *Offshore Outsourcing: Business Models, ROI and Best Practices*, 2nd ed. Mivar Press, 2010.

[29] G. Hofner, V. S. Mani, R. Nambiar, and M. Apte, "Fostering a High-Performance Culture in Offshore Software Engineering Teams Using Balanced Scorecards and Project Scorecards," in *Proceedings of IEEE 6th International Conference on Global Software Engineering - ICGSE'11*. Helsinki, Finland: Ieee, Aug. 2011, pp. 35–39.

[30] M. Narasipuram, "Towards a Taxonomy for Globally Distributed Work," in *Proceedings of the Twelfth Americas Conference on Information Systems - AMCIS'06*, Acapulco, Mexico, 2006, pp. 867–871.

[31] C. Tautz and C. G. Von Wangenheim, "REFSENO: a representation formalism for software engineering ontologies," Fraunhofer IESE, Tech. Rep., 1998.

[32] G. R. Wheaton, "Development of a taxonomy of human performance: A review of classificatory systems relating to tasks and performance," American Institute for Research, Washington DC, Tech. Rep., 1968.

[33] P. M. Institute, *A Guide to the Project Management Body of Knowledge (PMBOK® Guide)*, ser. PMI Standard. Project Management Institute, Incorporated, 2013.

[34] E. Mendes, *Cost Estimation Techniques for Web Projects*. IGI Publishing, 2007.

[35] M. J. Shepperd and G. Kadoda, "Using simulation to evaluate prediction techniques," in *Proceedings of the IEEE 7th International Software Metrics Symposium*, 2001, pp. 349–358.

[36] G. Coulouris, J. Dollimore, T. Kindberg, and G. Blair, *Distributed Systems: Concepts and Design*, 5th ed. Addison Wesley, 2011.

[37] N. Ramasubbu and R. K. Balan, "Overcoming the challenges in cost estimation for distributed software projects," in *Proceedings of 34th International Conference on Software Engineering - ICSE'12*, Zurich, Switzerland, 2012, pp. 91–101.