ISSN 2277-3061

An Analysis of the PROMISE and ISBSG Software Engineering Data Repositories

Laila Cheikhi, Alain Abran
Ecole Nationale Supérieure d'Informatique et d'Analyse des Systèmes, Université Mohammed V – Souissi
BP: 713, Rabat-Morocco
cheikhi@ensias.ma
École de Technologie Supérieure, Université du Québec, 1100 Notre-Dame Ouest, Montréal- Canada
alain.abran@etsmtl.ca

ABSTRACT

There exist two ongoing large repositories of software projects in the software engineering community: the repository of the International Software Benchmarking Standards Group (ISBSG) and the Repository referred to as PROMISE (PRedictOr Models In Software Engineering). Researchers interested in using the datasets have to conduct their own analysis of the datasets within these repositories that figure out what contents are suitable for their purposes. Repositories designed without users (researchers and Industrial) needs in mind greatly are more challenging to use. This paper present an analysis of both repositories and provide users with additional information on these datasets by identifying the topics addressed, highlighting their descriptiveness, and their availability, and by indicating whether or not further details are available in order to enhance their reusability in further empirical studies. Recommendations to both PROMISE managers and datasets owners are also suggested to improve the usefulness of the data provided.

Indexing terms/Keywords

Software engineering data repositories, ISBSG, PROMISE, software datasets, availability, descriptiveness, reusability, classification framework.

Academic Discipline And Sub-Disciplines

Software engineering: Software engineering repositories; Information Technology: empirical studies.

SUBJECT CLASSIFICATION

Software development, Software project data, software defects, software maintainability.

TYPE (METHOD/APPROACH)

Survey

Council for Innovative Research

Peer Review Research Publishing System

Journal: INTERNATION JOURNAL OF COMPUTERS AND TECHNOLOGY

Vol. 13, No. 5

editorijctonline@gmail.com

www.cirworld.org/journals

ISSN 2277-3061

1. INTRODUCTION

A software engineering data repository is defined as a set of well-defined, useful, and pertinent real-world data related to software projects, called datasets, which include quantitative and descriptive information about resources, products, processes, techniques, management, etc. Such data are being collected for various purposes by recognized organizations, as well as by individual software organizations and researchers. In most scientific and engineering disciplines, these data are useful for conducting benchmarking, experimental, and empirical studies. While highly varied and widely available in mature disciplines, data repositories are much less frequently found in emerging disciplines, including software engineering key process area [1].

Collecting reliable data is a critical step in any engineering study where the validity of the results and findings is highly dependent on the data used. Data collection in software engineering is not an easy task, as explained by Fenton and Pfleeger [2]: "Data collection is easier said than done, especially when data must be collected across a diverse set of projects," and by Kan [3]: "Gathering software engineering data can be expensive, especially if it is done as part of a research program." Obtaining useful and reliable data is not only an expensive endeavor, it is also quite time-consuming. Furthermore, as Jones [4] stresses: "Quality data must be measured, and it must be accumulated for quite some time before enough of it is available to be useful."

To overcome some of the burden of collecting software project data, a few group initiatives have emerged over the past decade for creating public data repositories [5] to serve the needs of the software engineering community. This paper analyzes the two largest of the small number of software engineering repositories publicly available: the ISBSG Repository [6] which contains datasets covering a considerable number of fields, and the PROMISE [7] repository with its large number of different datasets. Because of their size, it is difficult for researchers to quickly find in them what is relevant to their work. Moreover, there is no structured documentation for these repositories to help with this search. The novelty and the benefit of this paper is that it presents a structured overview of the datasets within these repositories, which will allow researchers and practitioners to find more quickly the information they need. In particular, this paper presents a survey of the datasets within these repositories to document which topics they address and which ones provide, or fail to provide, a reliable description of their content and attributes, with a view to enhancing their reusability in other empirical studies. Recommendations are also presented to improve the usefulness of these repositories.

The paper is structured as follows. Section 2 identifies the various datasets within the ISBSG and PROMISE repositories and presents the classification framework we have adopted. Section 3 surveys and sorts the datasets according to this classification framework. Section 4 presents the results of the survey and discusses the findings. Section 5 provides some recommendations to improve the quality of the repositories.

2. IDENTIFICATION OF THE DATASETS WITHIN THE REPOSITORIES

This section presents an overview of the software engineering datasets (see Table I) within the ISBSG and PROMISE repositories.

2.1 The ISBSG Repository and its Datasets

The International Software Benchmarking Standards Group (ISBSG) is a non-profit organization established in the late of 1990s, and its members are the software measurement associations of 13 countries. The ISBSG manages a repository of software-related data with the aim of improving software practices by encouraging benchmarking studies and the development of estimation and prediction models [6].

The ISBSG provides two software datasets for researchers and practitioners:

- The projects dataset for software development and enhancement, the purpose of which is to "improve Information Technology (IT) performance through estimation, benchmarking, project planning and management, and IT infrastructure planning" [6].
- The applications dataset for software maintenance and support, which provides organizations with "an opportunity to see how their maintenance and support compare with the industry, and to contribute to a body of software engineering knowledge that is open, available and used for the betterment of the IT profession" [8].

These ISBSG datasets are updated on a continual basis with new real project data submitted by software organizations from around the world. The latest releases of these datasets are R12 – 2013 for the projects dataset [9], and R6 – 2012 for the applications dataset [10]. These are provided in the form of an Excel spreadsheet. We refer to them as the ISBSG projects dataset and the ISBSG applications dataset respectively.

The ISBSG projects dataset for development and enhancement can be used for purposes such as[1] :

- Estimating project size, effort, duration, and cost,
- Checking the completeness of project requirements,
- Reducing project risk,

[1] http://www.isbsg.org/ An introduction to the ISBSG

- Managing project process,
- Negotiating/controlling software development,
- Acquiring custom-built software on the basis of price per functional unit,
- Planning development infrastructure,
- Benchmarking performance.

The ISBSG applications dataset for software maintenance and support can be used for purposes such as:

- Planning and managing resource allocation,
- Maintaining awareness of likely effort and cost requirements,
- Benchmarking against similar organizations,
- Better management of a software portfolio,
- Planning the maintenance and support infrastructure,
- Negotiating service level agreements.

Table I: Repositories and Datasets

Repositories and their Datasets	Datasets Surveyed
ISBSG Repository	
• Projects for software development and enhancement (ISBSG projects dataset)	1
• Applications for software maintenance and support (ISBSG applications dataset)	1
Total	2
PROMISE Repository and Categories	
• Defect prediction	56
• Effort prediction	12
• Text mining	8
• Model-based software engineering	3
• General	8
Total	87
Overall Number of Datasets	**89**

2.2 The PROMISE Repository and its Datasets

The PRedictOr Models In Software Engineering (PROMISE) repository was begun in December, 2004, by two researchers, Shirabad and Menzies, to encourage the development of predictive models for software engineering [7]. This repository is a collection of datasets from various sources (research, open source projects, etc.) provided to the public free of charge by the software engineering community to serve researchers and the software industry (available at www.promisedata.googlecode.com; last accessed in March, 2014). The first version of this repository (Repo v1.0) was created from NASA data and hosted at the University of Ottawa (Canada). In 2006, the PROMISE repository contained 23 datasets, and has since expanded to include 145 datasets as of March, 2012. It was subsequently reduced to 87 datasets to eliminate sets that did not provide detailed data. In the 2014 version of the PROMISE repository (Repo v4.0), the 87 datasets are grouped into 5 categories, based on the topic addressed (see Table I):

1. Defect prediction, with 56 datasets,
2. Effort prediction, with 12 datasets,
3. Text mining, with 8 datasets,
4. Model-based software engineering, with 3 datasets,
5. General, with 8 datasets.

Each of these 89 datasets (ISBSG and PROMISE) is explored in the next section, using the following classification framework criteria:

- **Dataset name:** the name of the dataset as it appears on the website. For example, one dataset name can include many data files (distinct files or modified versions of the same file). In this case, only the name of the dataset is listed in this table, and not all the files that it contains.
- **Year the dataset was originally donated**: the year the dataset was originally made available on the PROMISE website. Over the years, the datasets may have been updated, and it is possible that more than one file will be found under a single dataset name (in this case, only the year in which it was originally donated to PROMISE is reported here).
- **Dataset source**: the source of the dataset, or the person who made it available (donor).
- **Availability of the descriptions of the attributes**: indicates whether or not a description of the attributes of the dataset is available. It could be available with the data or in a separate file (readme), or in an available reference paper, in which case that paper must be clearly identified and made available through a valid link.
- **Availability of the data file**: indicates whether or not the data file is available. An available dataset provides data in an appropriate format, such as Arrf[2], CSV[3], Txt, or Excel. (These formats can be used in the Weka[4] data mining software, provided by the University of Waikato.) This software has gained popularity among researchers in a number of domains, and can be downloaded from their website.
- **Type of software project**: indicates the kind of projects used by the datasets: academic or Industrial.
- **Number of attributes**: number of data fields for a project within the dataset.
- **Number of instances**: indicates the number of instances of projects; that is the size of the datasets.

3. SURVEY OF THE DATASETS

In the following subsections, we present a summary of the inventory of each dataset based on the 8 classification criteria. In Tables II, III, IV, V, VI, and VII, the 1st column presents the dataset names, the 2nd column the year in which it was originally donated, the 3rd column the source/donor of the dataset, the 4th column the availability of the dataset, the 5th column the descriptiveness of the dataset, the 6th column the type of software project used by the dataset, the 7th column the number of attributes, and the 8th column the number of instances (the size of the dataset) :

- The availability of the dataset refers the availability of the data file criterion, and is defined in this context as "*the degree to which a dataset is available through the website when required for use*."
- The descriptiveness of the dataset refers to the availability of the description of the attributes criterion, and is defined in this context as "*the degree to which a dataset has a clear and complete description of [its] attributes and the context of its usage to achieve its specified goal in its specified context of use*."

Our purpose is to identify which datasets are readily available and which require further detailed analysis in order to make them usable in empirical studies. The details of the individual surveys, using additional criteria, are not presented in this article, but are available from the authors in the form of an Excel spreadsheet upon request.

3.1 ISBSG Datasets

The ISBSG, through its two datasets (See Figure1) provides "software development practitioners with industry output standards against which they may compare their aggregated or individual projects, and real data of international software development that can be analyzed to help improve the management of IT resources by both business and government" [6]. To achieve these goals, the ISBSG makes two kinds of questionnaires[5] available to the public with a view to collecting data on projects (for software development and enhancement) and on applications (for software maintenance and support), including mandatory software functional size measured with any of the measurement standards recognized by the ISO[6] (e.g. COSMIC[7] – ISO 19761, IFPUG[8] ISO 20926, and so on). The collected data are then assembled, evaluated, and stored in a database in Australia. A standardized extract of a number of data fields is provided, for a fee, in ISBSG Releases, while extracts of additional data fields are available upon request for research purposes in the form of an MS-Excel file.

[2] Attribute-Relation File Format (ARRF)
[3] Comma-separated values (CSV)
[4] http://www.cs.waikato.ac.nz/ml/weka (Waikato environment for knowledge analysis)

[5] www.isbsg.org
[6] International Organization for Standardization (ISO)
[7] COmmon Software Measurement International Consortium (COSMIC)

[8] International Function Point Users Group (IFPUG)

ISSN 2277-3061

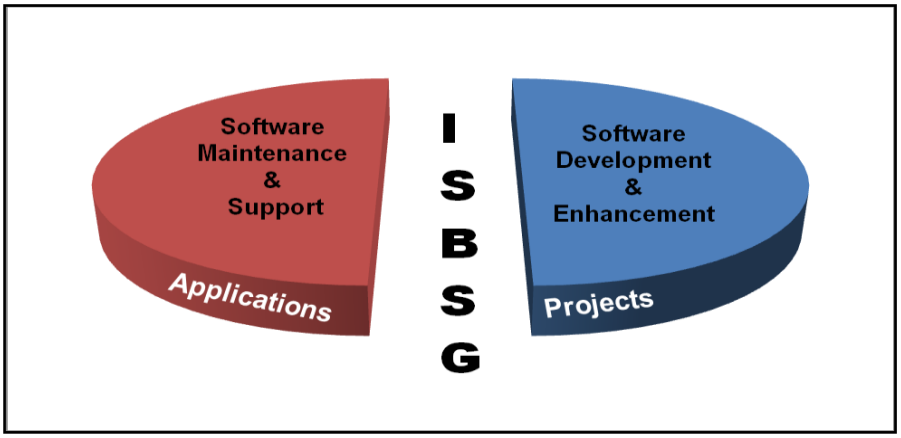

Figure 1: ISBSG Datasets

The *ISBSG projects* dataset is multi-organizational, multi-application, and multi-environment dataset with more than 100 data fields on more than 6,000 projects, the majority of which were collected after 2001. These +100 data fields include information about project staffing, effort by phase, development methods and techniques, team size, project type, organization type, software process along with the various life cycle phases, technology and tools used for developing and carrying out the project, software product characteristics, functional size attributes, defects reported and so on.

The *ISBSG applications* dataset is also a multi-organizational, multi-application, and multi-environment dataset, and provides more than 50 data fields on almost 1,200 applications, the majority of the benchmarking periods falling into the 2000 to 2010 time frame [11]. The data fields include information on the organization's domain of application, benchmarking period, type of application, functional size, maintenance and support hours, types of platform, types of programming language, defects reported, size of end user base, documents available for maintenance, number of change requests, system availability, and so on.

Table II: ISBSG Datasets

Dataset Name	Year Originally Donated	Dataset Source/ Donor	Dataset Availability	Dataset Descriptiveness	Type of Software Project	Number of Attributes	Number of Instances
ISBSG projects dataset (for software development and enhancement)	1994 to the present	ISBSG	Upon request	Yes	IT Industry software projects	More than 100	More than 6,000
ISBSG applications dataset (for software maintenance and support)	1994 to the present	ISBSG	Upon request	Yes	IT Industry software applications	More than 50	1,200

The two ISBSG datasets (see Table II) are collected by the ISBSG organization, and include real data from IT industry software projects and applications. Each is related to a specific process area of the software engineering discipline. They can be used for multiple topics (section 2), depending on the purpose of the research, and the cause and effect relationships studied[9], such as software defect prediction, software productivity, software quality, software effort prediction, etc. For example, the ISBSG provides data related, among other things, to:

- Defect prediction: topics like number of defects recorded during the various software life cycle phases, effort, size in Function Points and LOC(Lines Of Code), number of requests for specification changes during the software life cycle, type of application, etc.
- Effort prediction: topics like effort by phases, summary work effort, normalized work effort, etc.

With regard to the availability of the data fields within these datasets (4th column), ISBSG does not provide its full dataset, both for reasons of confidentiality and of cost: ISBSG has indicated that a key factor considered by the ISBSG board for the inclusion of a data field in its MS-Excel data extract made available to the public is preservation of the anonymity of the

[9] http://www.isbsg.org/isbsgnew.nsf/webpages/~GBL~Academic

data sources; in practice, this is done by analyzing whether there are few enough data points to ensure that no one data point can be traced back to a single organization; this also includes a cross-analysis with other variables, such as organization types, and so on. ISBSG makes available a standardized extract of its dataset, at a moderate price, to both industry and researchers. In addition, researchers can ask for more data details by submitting a research plan through the ISBSG Research Application Pack (available at http://www.isbsg.org/isbsgnew.nsf/webpages/~GBL~Academic; last accessed in April, 2014).

To minimize ambiguity when filling in the data collection questionnaires, the ISBSG has documented with great care a glossary of terms [12,13]. Specifically, the description of the attributes (5th column – descriptiveness of the datasets) is standardized across all the users of each kind of questionnaire and its corresponding ISBSG data repository.

3.2 PROMISE Datasets

This section provides a detailed investigation of PROMISE datasets. But in general, looking at PROMISE (87 datasets), what we can see is the diversity of the sources of the datasets, such as Published software engineering research work, Open source projects, NASA software, Industry software projects, etc. and each dataset was collected for one specific purpose and addresses a single topic. Moreover, from Figure 2, we can see that 64% of the datasets in the PROMISE repository concern defect prediction, the other 36% being related to other topics: 14% for effort prediction, 9% for text mining, 9% for general purpose, and 4% for model-based software engineering.

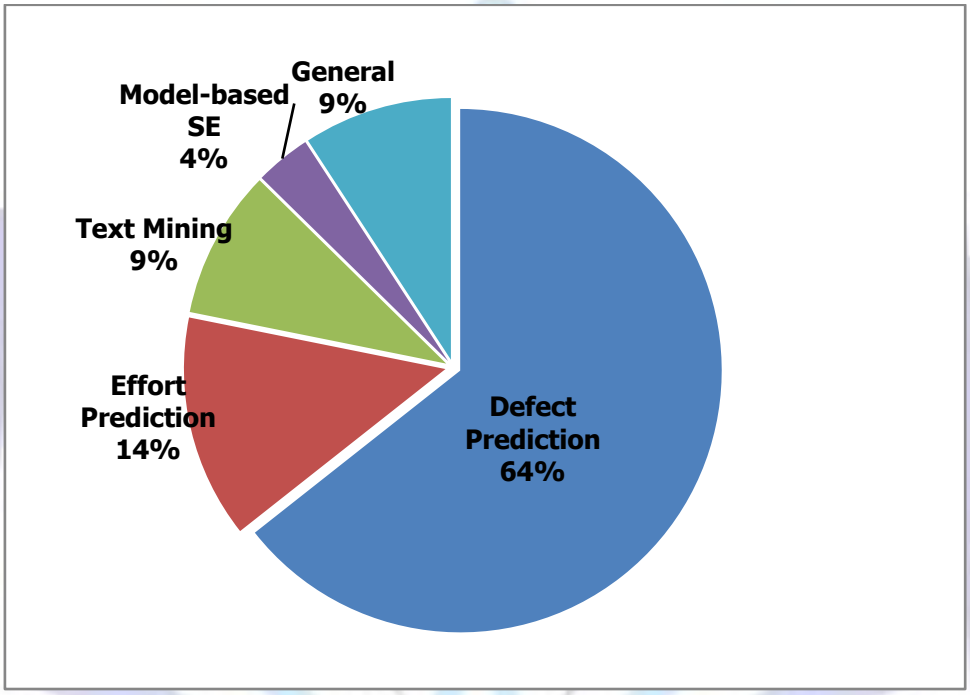

Figure 2: PROMISE – Distribution of Datasets by Topic

3.2.1 Defect Prediction

Defect prediction is the topic with the largest number of datasets from various sources in the PROMISE repository (i.e. 56 datasets out of 87); for example, datasets *PC5 to 1, MC1 to 2, KC2 to 3, JM1,* and *CM1* were provided by Menzies (NASA[10]) in 2004, datasets *Ar1 to 6* were provided by Softlab[11] in 2009, and datasets *Ant* to *Zuzel* were provided by Jureczko in 2010 (more details- upon request).

[10] National Aeronautics and Space Administration (NASA)
[11] Software Research Laboratory (Softlab) of Istanbul

Table III: Defects Prediction Datasets

Dataset Name	Year Originally Donated	Dataset Source/ Donor	Dataset Availability	Dataset Descriptiveness	Type of Software Project	Number of Attributes	Number of Instances
CM1	2004	PatCallis	Yes	Yes	NASA Projects	38	327
JM1	2004	Tim Menzies	Yes	Yes	NASA Projects	22	7782
KC2 to KC3 (2 datasets)	2004	Tim Menzies	Yes	Yes	NASA Projects	22 (KC2)	522 (KC2)
MC1 to MC2 (2 datasets)	2004	Tim Menzies	Yes	Yes	NASA Projects	39 (MC1)	1988 (MC1)
PC1 to PC5 (5 datasets)	2004	Tim Menzies	Yes	Yes	NASA Projects	39 (PC5)	17186 (PC5)
Datatrieve	2005	Guenther Ruhe	Yes	Yes	Project carried out at Digital Engineering, Italy	9	130
Coc81-dem	2006	Tim Menzies-Data presented in Boehm [16]	Yes	Yes	NASA Projects	27	63
Nasa93-dem	2006	Tim Menzies	Yes	Yes	NASA Projects	27	93
Mozilla4	2007	A. Gunes Koru	Yes	Yes	Open Source Mozilla product	6	15545
MW1	2007	Tim Menzies	Yes	Yes	NASA Projects	38	253
MB2	2008	A. Gunes Koru	Yes	Yes	NASA Projects	-	-
AR1 to AR6	2009	Softlab	Yes	Yes	Turkish white goods manufacturer	30 (AR6)	101 (AR6)
Bugreport	2010	Martin Pinzger, Emanuel Giger	Yes	Yes	Open source: Eclipse, Gnome, Mozilla bug database	122	18592
Ant to Zuzel (33 datasets)[12]	2010	Marian Jureczko	Yes	Yes	Open source and academic projects	24 (Ant)	745(Ant)
Am1	2012	Audris Mockus	Yes	Yes	Open source software systems	18	3299
Spe	2012	Ning Chen	**No**	Yes	Industrial software projects	-	-
Mtdjedit	2013	Josée Tassé	Yes	**No**	Open source software	13	1379
ABACUS2013	2013	André Riboira and Rui Abreu	**No**	**No**	Open source software	-	-
- Total = 56 datasets **- Availability (4th column) = 54 of 56** **- Descriptiveness (5th column) = 53 of 54**							

Each dataset has been built and collected for one specific purpose with locally based definitions of the attributes collected. For example:

- The *AR1 to AR6* datasets concern embedded software in a white-goods product, implemented in C, with 30 attributes each (such as McCabe, Halstead, LOC, and defects) and different numbers of instances.

[12] For the groups of datasets, the number of attributes and instances reported are given for one dataset as an example; the other datasets of the same group have the same number of attributes but differs in number of instances.

- The *Bugreport* dataset includes bug report data on 6 systems of 3 open source projects: Eclipse, Mozilla, and Gnome. Although the data are available (a different number of instances for each project), the descriptions of the 122 attributes are not available directly, but can be extracted from [14]. This dataset was initially used to build prediction models to recommend whether or not a new bug should, or will, be fixed quickly, or require more time to be resolved [14].
- The *Ant* to *Zuzel* group of 33 datasets includes 24 attributes, including the Chidamber and Kemerer Object Oriented metrics, and a different number of instances for each. These datasets were initially used to "perform clustering on software projects in order to identify groups of software projects with similar characteristics from the defect prediction point of view" [15].

With regard to the availability of the datasets (4th column), 54 of the 56 datasets related to defects prediction (96%) are available. Of these, 53 (98%) provide descriptions of their attributes (descriptiveness – 5th column), and 1 (2%) does not, which makes using those tow datasets with no descriptiveness or availability quite challenging.

3.2.2 Effort Prediction

The effort prediction topic includes 12 datasets (out of 87) (see Table IV), again provided by different sources, with different purposes and using different attributes, different definitions and different numbers of instances.

Table IV: Effort Prediction Datasets

Dataset Name	Year Originally Donated	Dataset Source/ Donor	Dataset Availability	Dataset Descriptiveness	Type of Software Project	Number of Attributes	Number of Instances
Nasa93	2006	Tim Menzies	Yes	Yes	NASA Projects	24	93
Coc81, Coc81-inh	2006	Tim Menzies-Data presented in Boehm [16]	Yes	Yes	NASA Projects	19	63
Boetticher	2008	Gary Boetticher	Yes	**No**	-	61	172
Maxwell	2009	Li Yanfu	Yes	Yes	Finish banking data	27	62
Cocomo_sdr	2009	Ekrem Kocaguneli	Yes	Yes	Turkish software industry	25	12
Miyazaki94	2010	Sousuke Amasaki	Yes	Yes	Software projects	9	48
Kemerer	2010	Jacky W. Keung	Yes	**No**	-	8	15
China	2010	Hon Yun Fang	Yes	Yes	-	19	499
Albrecht	2010	Li Yanfu, Jacky W. Keung	Yes	**No**	-	8	24
Kitchenham	2011	Masateru Tsunoda	Yes	Yes	Maintenance and development projects	10	145
COSMIC	2012	ISBSG.org	Yes	Yes	ISBSG-IT Industry software projects	86	42
ISBSG 10	2012	ISBSG.org	Yes	Yes	ISBSG-IT Industry software projects	94	37
- Total = 12 datasets **- Availability (4th column) = 12 of 12** **- Descriptiveness (5th column) = 8 of 12**							

Some of these software effort prediction are the following:

- *Coc81, Cocomo_sdr,* and *Nasa93*: These datasets use the COCOMO[13] I [16] or COCOMO II [17] attributes, among others. The *Nasa93* data were collected from various centers of NASA projects from 1971 to 1987.
- *COSMIC*, *Maxwell*, and *Kitchenham*: These are the only datasets that collect data on the characteristics of software projects, and not on the software product itself. For example, the *Kitchenham* dataset was initially used to "analyze data from 145 enhancements and development projects managed by a single outsourcing company, including estimates of effort and duration, actual values of effort and duration, and function point counts" [18]. These datasets include different numbers of instances and different numbers of attributes, some of which are common to all of them.
- *Kemerer* and *Albrecht*: These datasets do not provide a description of the attributes or the context of their collection. They require further analysis of their data fields for usability for research purposes.

With regard to the availability of the datasets, the data files of all 12 of the datasets containing effort data are available, of which 8 (67%) include a description of their attributes (descriptiveness), while the other 4 datasets (33%) do not.

3.2.3 Text Mining

The text mining topic includes 8 datasets (see Table V). Most of these are from NASA, and were provided between 2005 and 2008: the MODIS dataset concerns requirements and their traceability, the NFR dataset concerns non functional requirements, since the "early detection of NFRs is useful because it enables system level constraints to be considered and incorporated into early architectural designs, as opposed to being refactored in at a later time" [19]. The Pits (Project and Issue Tracking System) data have been collected for more than 10 years, and include issues on robotic satellite missions and human-rated systems captured by NASA's IV&V Program for software testing [20].

Table V: Text Mining Datasets

Dataset Name	Year Originally Donated	Dataset Source/ Donor	Dataset Availability	Dataset Descriptiveness	Type of Software Project	Number of Attributes	Number of Instances
Modis	2005	Jane Huffman Hayes	**No**	Yes	NASA Software projects	3	49 low level requirements 19 high level requirements
NFR	2007	Jane Cleland-Huang	Yes	Yes	-	3	625
Pits A,B, C,D,E,F (6 datasets)	2008	Tim Menzies	Yes	**No**	NASA Software projects	5 (PitsA)	966 (PitsA)
- Total = 8 datasets **- Availability (4th column) = 7 of 8** **- Descriptiveness (5th column) = 1 of 7**							

With respect to the availability of the datasets, 7 of the 8 related to text mining provide a data file. Of these, 1 (14%) provides a description of its attributes (descriptiveness) , while the other 6 (86%) do not, making these datasets quite challenging to use.

3.2.4 Model-Based Software Engineering

The model-based software engineering topic includes 3 datasets (see Table VI). These datasets have been provided since 2009 and were collected for different purposes. For example, the goal of the Bike dataset is "to show that, in this real-world industrial setting, treatment learning offers a faster, higher-quality identification of the factors likely to cause a failure in a complex system than traditional optimization techniques" [21]. The CM1-bn dataset describes the node probability tables (NPT) for every node in the Bayesian network and contains data on the quality measures of the 7 attributes collected (change effort, state, probability, comments_ratio, average_cyclomatic_complexity, average_module_size, fix effort) [22].

[13] Constructive Cost Model - COCOMO

Table VI: Model-based SE Datasets

Dataset Name	Year Originally Donated	Dataset Source/ Donor	Dataset Availability	Dataset Descriptiveness	Type of Software Project	Number of Attributes	Number of Instances
CM1-fix/ CM1-bn	2009	Stefan Wagner	Yes	Yes	NASA spacecraft instrument written in 'C'	1 / 7	25/ 15
Bike	2009	Misty Davies	Yes	Yes	NASA Industrial projects + Real-world data from a bicycle ride	11	4435
Quantitative	2012	Emmanuel Leiter	**No**	**No**	Quantitative goal model of an ambulance service system	**-**	**-**
-Total = 3 datasets **- Availability (4th column) = 2 of 3** **- Descriptiveness (5th column) = 2 of 2**							

With respect to the availability of the datasets, 2 of the 3 related to model-based software engineering are available (67%) and include a description of the attributes (descriptiveness), while 1 (33%) does not.

3.2.5 General

The general topic includes 8 datasets (see Table VII) that address purposes different from those cited in the previous sections. Some of these datasets are the following:

- *Reuse*: This dataset consists of a set of candidate reusability factors [23]. It contains a set of 29 software project management, process, and product attributes on 24 projects.
- *Nickle*, *Xfree86*, and *Xorg*: These datasets provide data generated from CVS archive files of the Nickel, Xorg, and Xfree86 open source projects [24].
- *Qosdata*: This dataset provides data on the Quality of Service (QoS) behaviors. It proposes "a non functional testing process in which time, space, and quality attributes can be collected through the process and used to predict the QoS behaviors of individual components or a collection of integrated components" [25] with 6 attributes and 272 instances.
- *CMMI*: the dataset provides results from two recent surveys conducted by the Software Engineering Institute (SEI) on CMMI high maturity practices. The survey includes close to 175 questions and sub questions: 156 organizations provided responses to the questionnaire in 2008 and in 2009 to which as set of 84 organizations has provided responses to a variant of the questionnaire. The collected data is about the measurement and analysis activities within organizations.

Table VII: General Datasets

Dataset Name	Year Originally Donated	Dataset Source/ Donor	Dataset Availability	Dataset Descriptiveness	Type of Software Project	Number of Attributes	Number of Instances
Reuse	2004	Morisio, Ezran, Tully-Data presented in [23]	Yes	Yes	Industrial organisations	29	24
Nickle	2005	Bart Massey	Yes	Yes	Public CVS archives of NICKEL programming language)	10	2972
Xfree	2005	Bart Massey	Yes	Yes	Public CVS archives of X.free86 project)	10	176658
Xorg	2005	Bart Massey	Yes	Yes	Public CVS archives of X.org)	10	136435
CM1-Maintain	2006	Hany H. Ammar, K. Goseva-Popstojanova, W. Abdelmoez	Yes	Yes	NASA spacecraft instrument written in 'C'	4	12
Qosdata	2006	Zhou, Cooper, Yen	Yes	Yes	Open source software (Windows, Unix, Linux)	6	272
Generics	2013	Chris Parin	**No**	**No**	-	-	-
CMMI	2013	High maturity research data from SEI- Dave Zubrow	Yes	Yes	Industrial organisation	Around 175	156 /2008 84/ 2009
- Total = 8 datasets - Availability (4th column) = 7 of 8 - Descriptiveness (5th column) = 7 of 7							

With regard to the availability of the datasets and their descriptiveness (i.e. description of the attributes), 7 datasets are available and 1 is not.

4. DISCUSSION ON THE AVAILABILITY OF THE DATASETS AND REUSABILITY

From the survey results of the two repositories conducted in the previous section, we can see in general, that the ISBSG practitioners have made their own data publicly available since 1994, while the software engineering research community only began to share their data later, in 2004, although these data were available before that time (see Figure 3). The main goal is to solve the problem related to the non availability of useful data from software projects that could be used as a reference to allow researchers to compare their results. Since 2004, the number of datasets submitted to PROMISE has been increasing, as this community now recognizes the importance of the data for conducting studies and gaining a better understanding of ways to successfully achieve their objectives, such as increasing productivity, improving quality, etc.

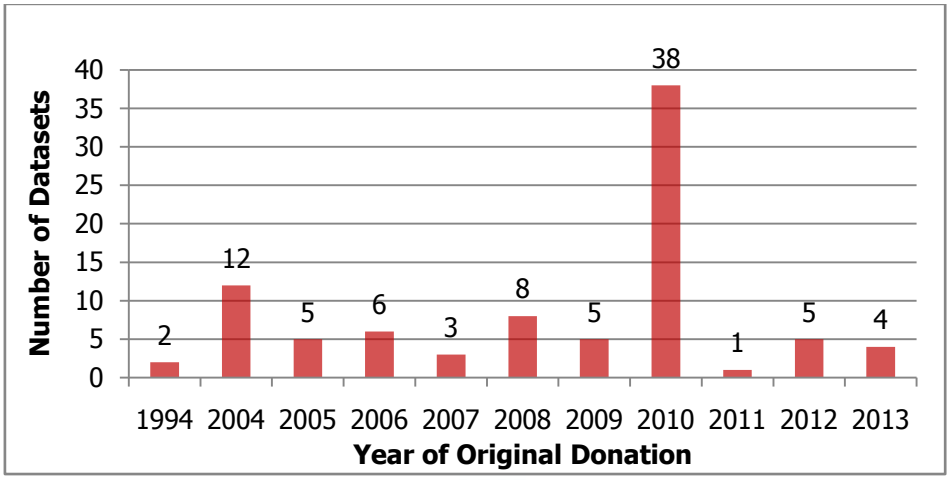

Figure 3: Distribution of datasets by the year in which they were originally donated (PROMISE & ISBSG)

Hereafter, a discussion of the survey results on the datasets proceeds from two perspectives:

- their categories of reusability, and
- their usefulness for benchmarking.

4.1 Datasets Reusability

From the survey results, we highlight two main issues related to datasets in particular – their availability and their descriptiveness – in order to identify the most readily reusable of the 89 datasets within the PROMISE and ISBSG repositories. Table VIII provides a summary of the descriptiveness and availability of the 87 PROMISE datasets.

Table VIII: Availability and Descriptiveness of the PROMISE Datasets

		Dataset Availability		
Dataset Descriptiveness		Yes	No	Total
	Yes	72	2	74
	No	10	3	13
	Total	82	5	87

In general, reusability refers to the ability of an item to be used in another context or environment. The reusability of a software product is recognized by ISO 25000 experts as an attribute of its quality models defined as “the degree to which an asset can be used in more than one system, or in building other assets” [26]. In the context of datasets, we define reusability as “the degree to which a dataset can be used in more than one study or in building other predictive models. Therefore, the reusability of the dataset is expressed in terms of its availability and descriptiveness.”

In Table VIII, the datasets with:

- a *Yes* for both availability and descriptiveness are considered to have good reusability (they do not require additional information);
- a *Yes* for availability and a *No* for descriptiveness are considered to have moderate reusability (they require additional information from the dataset owners); and
- a *No* for availability and either a *Yes* or a *No* for descriptiveness are considered to have poor reusability (and dataset owners should update their links on the website).

In summary, the reusability of 83% of the 87 datasets is good, that of 11% of them is moderate, and that of 6% of them is poor (see Figure 4). In contrast, both the ISBSG datasets are available upon request and both have good descriptiveness. Therefore, their reusability is good.

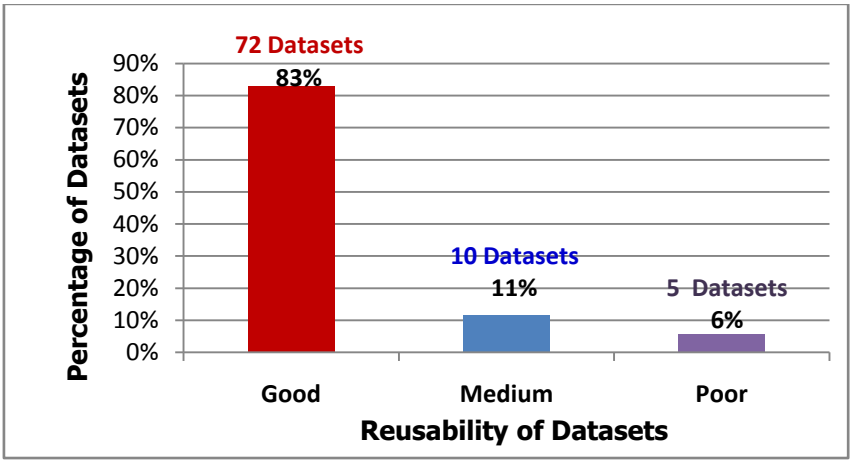

Figure 4: Distribution of poor, moderate, and good PROMISE dataset reusability

Moreover, we have identified the datasets that are readily reusable based on their reusability score, which is made up of availability (data file availability) and descriptiveness (attribute description). Table IX summarizes the principal finding of the survey carried out in section 3, which is that 74 of the 89 datasets provided by the two repositories are readily reusable.

Table IX: Distribution of Reusable Datasets by Topic (PROMISE & ISBSG)

Repositories and their Datasets	Datasets Surveyed	Datasets Readily Reusable	Datasets Name
ISBSG Repository			
• Projects for software development and enhancement	1	1	ISBSG Projects
• Applications for software maintenance and support	1	1	ISBSG Applications
Total	2	2	2
PROMISE Repository and Categories			
• Defect Prediction	56	53	Ant to Zuzel(33), CM1, Am1, AR1 to AR6, MW1, Mozilla4, PC1 to PC5 (5), MC2, MC1, KC3, KC2, M1, MB2, Datatrieve, Bugreport, Nasa93, COC81/inh
• Effort Prediction	12	9	Nasa93-dem, Coc81-dem Kitchenham, Miyazaki94 Cocomo_sdr, COSMIC ISBSG, China, Maxwell,
• Text Mining	8	1	NFR
• Model-based Software Engineering	3	2	CM1-fix/bn, Bike
• General	8	7	CM1-Maintain, Reuse, Nickle, Xfree, Xorg, Qosdata, CMMI
Total	87	72	72
Overall number of datasets	89	74	74

4.2 Benchmarking Usefulness

According to the PROMISE community, the past usage of these datasets in research work should be available in order to "encourage repeatable, verifiable, refutable, and/or improvable predictive models of software engineering" [7] and to conduct benchmarking studies using these datasets. By *past usage*, we mean that *the authors mention having used their own datasets in the past, and they include whatever references or links exist to published papers in which these datasets were used, if any*. The availability of information on the past usage of a dataset is very important for further research work,

since this information can attract the interest of researchers and practitioners, and instill confidence in them in terms of using the dataset.

From our survey of the 87 PROMISE datasets, only 15 (17%) report directly past usage and the remaining 72 datasets (83%) do not. In particular (see Figure 5):

- Of the 72 datasets with good reusability, past usage is only recorded for 13 of them (18%).
- Of the 10 datasets of medium reusability, past usage is recorded for only 1 dataset (10%).

There are several possible reasons for this lack of reference to past usage of the PROMISE datasets. The data sources may not have provided this information; the only past usage might be the reference paper given (if so, it should be reported in the section on past usage); or perhaps the datasets had not been used before (this should also be mentioned in the section on past usage). For example, the authors of the Nasa93 dataset have stated that there is no past usage with their specific dataset, and refer to some old published research conducted using data similar to theirs.

With regard to the ISBSG datasets, the past usage is available and up to date, and documented on the website in two ways: research papers that have used ISBSG datasets or referred to them, and research projects that have used the ISBSG datasets.

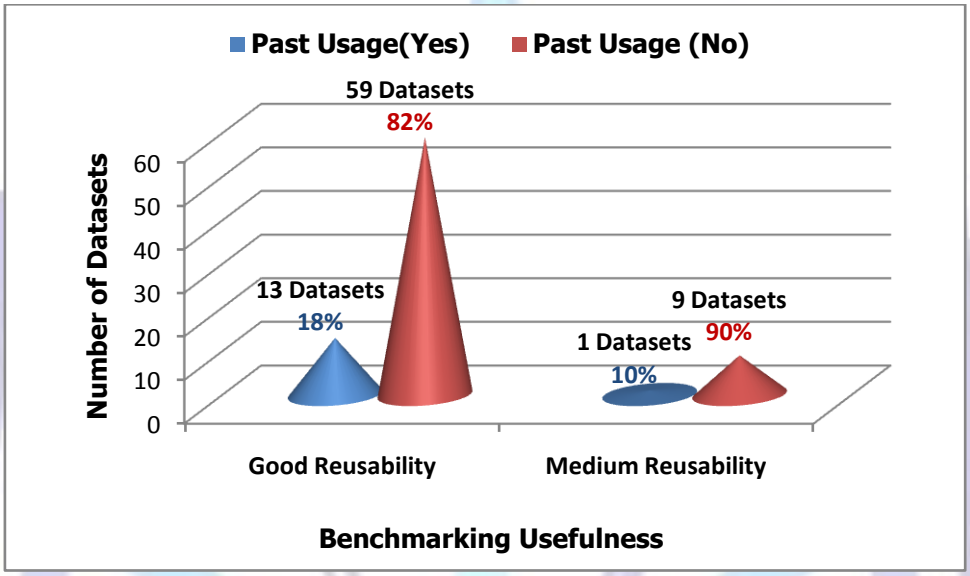

Figure 5: Distribution of moderate and good PROMISE dataset reusability for benchmarking usefulness

Moreover, we have identified among the datasets with good reusability score – see Table IX, those that can be useful for benchmarking studies. Table X summarizes the principal finding: only 15 (13 for PROMISE and 2 for ISBSG) of the 74 readily reusable datasets provided by the two repositories are useful for benchmarking.

ISSN 2277-3061

Table X : Distribution of benchmarking usefulness datasets by Topic (PROMISE & ISBSG)

Repositories and their Datasets	Datasets Surveyed	Datasets Readily Reusable	Datasets Benchmarking Usefulness	Datasets Names
ISBSG Repository				
• Projects for software development and enhancement	1	1	1	ISBSG projects
• Applications for software maintenance and support	1	1	1	ISBSG applications
Total	2	2	2	2
PROMISE Repository and Categories				
• Defect Prediction	56	53	8	Mozilla4, MW1, MC2, MC1, KC2, JM1, Datatrieve, Nasa93
• Effort Prediction	12	9	1	Nasa93-dem
• Text Mining	8	1	0	-
• Model-based Software Engineering	3	2	1	CM1-fix/bn
• General	8	7	3	Reuse, Qosdata, CMMI
Total	87	72	13	13
Overall number of datasets	**89**	**74**	**15**	**15**

5. CONCLUSION

This paper has provided an overview of the two major software engineering data repositories made available by the software engineering community to serve the needs of this community, namely the ISBSG Repository and PROMISE.

The ISBSG Repository, which is a multi-organizational, multi-application, and multi-environment data repository, offers, at a cost, the largest publicly available datasets of software project, unlike the repositories owned by commercial entities, access to which is restricted. The goal of the ISBSG is to provide researchers and the software industry with industry data to conduct empirical and benchmarking studies (see examples of use in [27, 28, 29]). This repository is continuously updated with new industry data, and it has been used for multiple purposes, among them productivity studies, software effort estimation, and defects density analysis.

The PROMISE repository offers free access to 87 datasets originating either from research work or from open source software systems. The goal of PROMISE is to provide researchers and experts in this field with the opportunity to test their assumptions and so improve their practices (see examples of its use in [30, 31, 32]). This set of datasets has been used for various purposes, among them software defect prediction and software effort estimation.

This survey of these datasets looked into the topics addressed, the source of the datasets, the year in which the dataset was originally donated, the availability, the descriptiveness of the dataset, the type of software project used by the dataset, the number of attributes, and the number of instances (the size of the datasets). This survey has led us to three principal conclusions:

- Of the 87 datasets available in PROMISE Repository, 72 provide information on their availability and descriptiveness, and therefore can be used directly without contacting the dataset owners. However, the lack of the context of the studies, the software product used, the size, the development type, domain of application, etc. can make the comparative studies very difficult in terms of the analysis of the results obtained through the use of datasets for two different contexts.

- The purpose of the PROMISE repository is to provide software engineering communities with reliable and real data that could be used in replication studies. However, to conduct this kind of study, information on the past usage of these data, that is, context of use and results, is required. This information is only available in published papers that have used these data before. Unfortunately, past usage information is readily available for only 13 of the 87 PROMISE datasets.
- In spite of the diversity of topics addressed by the PROMISE and ISBSG datasets, more in-depth analyses with their data are required to identify which software product quality attributes (such as those proposed by ISO 25000) are addressed with these datasets and in which phase of the software product life cycle, as well as the kind of data or measurements collected.

In summary, improving the quality of the PROMISE repository, in terms of usefulness, accessibility, descriptiveness, availability, and reusability, will require the joint effort of the following three groups of participants:

- The managers of the PROMISE repository should check the availability of the following information, before accepting the dataset: not only the name of the dataset, the year it was made available on the PROMISE website (the year it was originally donated), the source of the dataset or the donors' names, the reference for the paper in which it was used, the number of attributes, and the number of instances, but also the past usage of the dataset, if any (that is, published papers), a description of the attributes and useful information about the dataset (such as the year of collection of the data and a link to the reference paper that provides this information), the year the dataset was made available on the PROMISE website, the source of the dataset or the donors' names, the reference for the paper in which it was used, the number of attributes and their number of instances, the past usage of the dataset, if any, a description of the attributes and useful information about the dataset.
- The owners of the datasets should regularly check the availability of the data files through the links provided in the PROMISE repository, and update them whenever necessary.
- The users of the available datasets (both researchers and practitioners) should provide references for the published papers (using the datasets), both to update the past usage of this dataset and to "encourage repeatable, verifiable, refutable, and/or improvable predictive models of software engineering" [7].

Suggestions for further improvements to the PROMISE repository include:

- Descriptiveness of the dataset: when no data description is directly available from the PROMISE repository, such information may be available indirectly from the papers that initially reported on a dataset. However, many such papers are neither identified and easily accessible: if the authors were to provide such information directly into the PROMISE repository, they would save considerable research effort for all subsequent researchers interested in reusing such datasets.
- Although some datasets are self-descriptive (such as for Albrecht and Kemerer datasets where the name of the attributes are based on Function Point terminology or on well-known measures such as LCOM, CBO, DIT, LOC, etc.) it is not sufficient to say that the dataset descriptiveness is good. For instance: many datasets have the attribute name: "LOC" which is related to a size measure applied to source code in terms of lines of code, but without providing information on the programming language and measurement procedures for data collection: such lack of details makes comparatives studies with two LOC attributes from two different datasets not reliable unless they have been defined and measured in the same way. Therefore, the attributes names are not sufficient for the descriptiveness of the dataset and the PROMISE managers should address this issue.
- For benchmarking usefulness: we have noticed a lack of information on the past usage of these datasets and we have first want to search for the published papers (using PROMISE datasets) and add these references to our survey, but it requires a considerable effort and it could be very useful for the SE community. So we suggest donors of the datasets to add also this information to the PROMISE Repository or to be done by the repository managers themselves, since they need to improve the usefulness of their datasets.
- Regarding the quality of the datasets: For PROMISE Repository, some datasets owners report if their datasets are changed (add new instances) or corrected (invalid data) such as: CM1, JM1, MW1, MC1, MC2, KC2, KC3, Coc81, etc. For example, the CM1 dataset was donated in 2004 and corrected in 2011; the Coc81dataset was donated in 2006 and corrected in 2009. On the other hand, the ISBSG rates each submitted project with a code of A, B, C or D which is applied to the project data by the ISBSG data administrator using the following ratings and criteria [12]:
    - A = the data submitted was assessed as being sound with nothing being identified that might affect its integrity.
    - B = the submission appears fundamentally sound but there are some factors which could affect the integrity of the submitted data.
    - C = due to significant data not being provided, it was not possible to assess the integrity of the submitted data.
    - D = due to one factor or a combination of factors, little credibility should be given to the submitted data.

The quality of the data provided within a dataset should also be taken into account by the PROMISE mangers to provide a way to rank the quality of the data within the datasets. This is of course challenging since the datasets have been designed individually with their own locally-based definitions and data collection procedures.

Research in progress includes working on a more detailed survey, based on additional criteria to explore in greater depth the content of these datasets, in particular those that focus on object oriented metrics. Our objective is to identify datasets that allow studies to be conducted: (1) to investigate the relationships between processes, attributes, and product quality; (2) to establish software product predictive quality models; and (3) to identify ISO quality attributes that are not addressed by these software engineering datasets, in order to encourage researchers to focus on these quality attributes in their future research work.

REFERENCES

[1] IEEE. 2004. "Guide to the Software Engineering Body Of knowledge- SWEBOK." Los Alamitos, California: IEEE Computer Society, 204 p. (last accessed on 30/01/2013). http://www.computer.org/portal/web/swebok

[2] Fenton, Norman E. and Shari Lawrence Pfleeger. 1997. "Software Metrics: A Rigorous and Practical Approach," 2nd ed. Boston, MA, USA: PWS Publishing Company, 656 p.

[3] Kan, Stephen H. 2003. "Metrics and models in software quality engineering," 2nd ed. Massachusetts: Boston: Addison-Wesley, 528 p.

[4] Jones, Capers. 1996. "Applied Software Measurement – Assuring Productivity and Quality," 2nd ed. United States: New York, N.Y.: McGraw-Hill, 618 p.

[5] Cukic, Bojan. 2005. "Guest Editor's Introduction: The Promise of Public Software Engineering Data Repositories," IEEE Software, vol. 22, no. 6, pp. 20-22.

[6] ISBSG, 2012. "Data collection questionnaire new development, redevelopment or enhancement sized using COSMIC function points," version 5.16. Australia: International Software Benchmarking Standards Group.

[7] T. Menzies, B. Caglayan, E. Kocaguneli, J. Krall, F. Peters, and B. Turhan.2012. "The PROMISE Repository of empirical software engineering data," http://promisedata.googlecode.com, West Virginia University, Department of Computer Science (last accessed March, 2014).

[8] ISBSG, 2010. "Data collection questionnaire application software maintenance and support," version 2.3. Australia: International Software Benchmarking Standards Group.

[9] ISBSG, 2013. "Software project repository for software development and enhancement – Release 12". Australia: International Software Benchmarking Standards Group.

[10] ISBSG. 2012. "Software application repository for maintenance and support – Release 6". Australia: International Software Benchmarking Standards Group.

[11] ISBSG, 2010. "Software application repository for maintenance and support – Data Demographics Release 4. Australia: International Software Benchmarking Standards Group.

[12] ISBSG. 2012. "Glossary of terms for software project development and enhancement," v 5.16. Australia: International Software Benchmarking Standards Group.

[13] ISBSG. 2010. "Glossary of terms for application software maintenance and support," v 2.5. Australia: International Software Benchmarking Standards Group.

[14] Giger, Emanuel and Pinzger, Martin and Gall, and Harald. 2010. "Predicting the fix time of bugs," Proceedings of the 2nd International Workshop on Recommendation Systems for Software Engineering. Cape Town, South Africa. 52-56. ACM. New York, NY, USA.

[15] Marian Jureczko and Lech Madeyski. 2010. "Towards identifying software project clusters with regard to defect prediction," in Proceedings of the 6th International Conference on Predictive Models in Software Engineering (PROMISE '10). ACM, New York, NY, USA, Article 9, 10 p.

[16] B. Boehm. 1981, "Software. Engineering Economics," Prentice Hall.

[17] Donald Reifer, Barry W. Boehm, and Sunita Chulani. 1999. "The rosetta stone: Making COCOMO 81 estimates work with COCOMO II," Crosstalk. The Journal of Defense Software Engineering, pp. 11-15.

[18] Barbara Kitchenham, Shari Lawrence Pfleeger, Beth McColl, and Suzanne Eagan. 2002. "An empirical study of maintenance and development estimation accuracy," J. Syst. Softw. 64, 1, pp. 57-77.

[19] Jane Cleland-Huang, Raffaella Settimi, Xuchang Zou, and Peter Solc. 2006. "The Detection and Classification of Non-Functional Requirements with Application to Early Aspects," in IEEE RE.

[20] Tim Menzies. 2007. "Text Mining PITS issue reports," Technical Report, West Virginia University.

[21] Gregory Gay, Tim Menzies, Misty Davies, and Karen Gundy-Burlet. 2010. "Automatically finding the control variables for complex system behavior," Springer Science and Business Media, LLC'10.

[22] S. Wagner. 2009. “A Bayesian Network Approach to Assess and Predict Software Quality Using Activity-Based Quality Models, in Proceedings of the International Conference on Predictor Models in Software Engineering (PROMISE ’09). ACM Press.

[23] M. Morisio, M. Ezran, and C. Tully. 2002. “More Success and failure factors in software reuses,” IEEE Transactions on Software Engineering, vol. 28, 4, pp. 340-357.

[24] Bart Massey. 2005. “Longitudinal Analysis of Long-Timescale Open Source Repository Data,” in Proceedings of the PROMISE Workshop.

[25] Jia Zhou, Kendra Cooper, and I-Ling Yen. 2006. “QoS Data Collection: An Approach to Assist Predictable QoS Behavior Modeling in Component Based Development,” in Proceedings of the PROMISE Workshop.

[26] ISO 25000. 2010. “System and Software Engineering – System and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models,” International Organization for Standardization, Geneva, Switzerland.

[27] Yeong-Seok Seo, Kyung-A Yoon, and Doo-Hwan Bae. 2009. “Improving the Accuracy of Software Effort Estimation Based on Multiple Least Square Regression Models by Estimation Error-Based Data Partitioning,” in the 16th Asia-Pacific Software Engineering Conference, pp. 3-10.

[28] Jacky Keung and Barbara Kitchenham. 2008. “Experiments with Analogy-X for Software Cost Estimation,” in the 19th Australian Conference on Software Engineering, pp. 229-238.

[29] Faheem Ahmed, Salah Bouktif, Adel Serhani, and Issa Khalil. 2008. “Integrating Function Point Project Information for Improving the Accuracy of Effort Estimation,” in 2nd International Conference on Advanced Engineering Computing and Applications in Sciences, pp. 229-238.

[30] Leandro L. Minku and Xin Yao. 2011. “A Principled Evaluation of Ensembles of Learning Machines for Software Effort Estimation,” PROMISE ’11, Proceedings of the 7th International Conference on Predictive Models in Software Engineering, September 20-21, Canada.

[31] Daniel Rodríguez, Jesús C. Riquelme, Roberto Ruiz, and Jesús S. Aguilar-Ruiz. 2009. “Searching for Rules to Find Defective Modules in Unbalanced Data Sets,” 1st International Symposium on Search Based Software Engineering, pp. 89-92.

[32] Cagatay Catal and Banu Diri. 2007. “Software Fault Prediction with Object-Oriented Metrics Based Artificial Immune Recognition System,” Springer-Verlag Berlin Heidelberg, pp. 300-314.

ISSN 2277-3061

Author' biography with Photo

Alain Abran holds a Ph.D. in Electrical and Computer Engineering (1994) from the École Polytechnique de Montréal (Canada) and Master's degrees in Management Sciences (1974) and Electrical Engineering (1975) from the University of Ottawa. He is a Professor and the Director of the Software Engineering Research Laboratory at the École de Technologie Supérieure (ÉTS) of the Université du Québec (Montréal, Canada). He has over 15 years of experience in teaching in a university environment, and more than 20 years of industry experience in information systems development and software engineering. His research interests include software productivity and estimation models, software engineering foundations, software quality, software functional size measurement, software risk management, and software maintenance management. He has published over 300 peer-reviewed publications and he is the author of the book "Software Metrics and Software Metrology" and a co-author of the book "Software Maintenance Management" (Wiley Interscience, Ed., & IEEE-CS Press). Dr. Abran is co-editor of the Guide to the Software Engineering Body of Knowledge – SWEBOK, and he is the chairman of the Common Software Measurement International Consortium (COSMIC).

Laila Cheikhi is a Professor at Computer Science and Systems Analysis School (ENSIAS, Rabat, Morocco). She received a M.Sc. (2004) from University of Montréal and Ph.D. (2008) from ETS, University of Quebec at Montreal, and Both in software engineering. She has over eight years of experience in computer engineering at the Ministry of Finance of Morocco. Her research interests include software quality models, software metrics, software engineering ISO standards, software product and process quality, software engineering principles and data analysis.