Dataset of Developer-Labeled Commit Messages

Andreas Mauczka, Florian Brosch, Christian Schanes, Thomas Grechenig
Vienna University of Technology
{andreas.mauczka, florian.brosch, christian.schanes, thomas.grechenig}@inso.tuwien.ac.at

TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

Summary

Current research on change classification centers around automated and semi-automated approaches which are based on evaluation by either the researchers themselves or external experts. In most cases, the persons evaluating the effectiveness of the classification schemes are not the authors of the original changes and therefore can only make assumptions about the intent of the changes. To support validation of existing labeling mechanisms and to provide a training set for future approaches, we present a survey of source code changes that were labeled by their original authors. Seven developers from six different project applied three existing classification schemes from current literature to enrich their own changes with meta-information, so the intent of the changes becomes more evident. The final data set consists of 967 classified changes and is available as an SQLite database as part of the MSR data set.

Step 1: Selection of Developers and Projects

Step 2: Assembly of commit data and creation of survey forms

Step 3: Providing instructions and guidance

Step 4: Aggregation of the data into a single data source

Projects

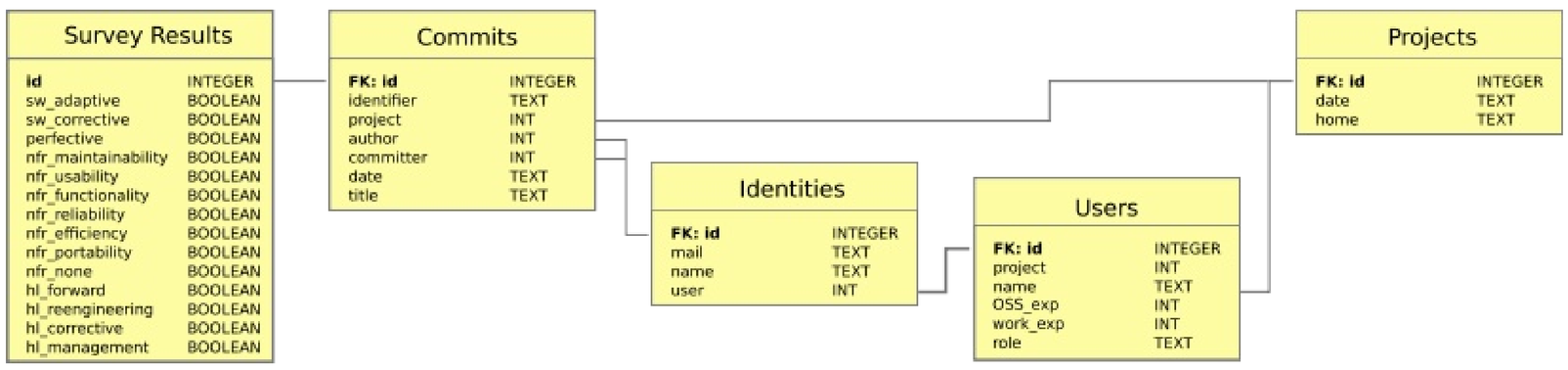
	Description	LOCs
Vala	A source to source compiler used by GNOME. » http://www.vala-project.org	236.071
Valadoc	The documentation generator for Vala. » http://www.valadoc.org	50.709
Drupal Search API	A framework for creating searches on any entity known to Drupal, using any kind of search engine. » https://www.drupal.org/project/search_api	21.696
TapiJI	A set of smart tools that integrate into the Eclipse IDE for Java developers with the goal to reduce effort of Internationalization. » http://code.google.com/a/eclipselabs.org/p/tapiji/	19.611
MyLyn	Task and application lifecycle management framework for Eclipse. » http://eclipse.org/mylyn/	76.464
DeltaSpike	A number of portable CDI extensions that provide useful features for Java application developers. » http://deltaspike.apache.org/	35.202

Developers

	Role	Classified Commits
Luca Bruno	**Maintainer**	116
Evan Nemerson	**Maintainer**	194
Florian Brosch	**Maintainer**	200
Thomas Seidl	**Maintainer**	118
Martin Reiterer	**Architect**	123
Kilian Matt	**Developer**	81
Mark Struberg	**Architect**	135
		Total: 967

Data Assembly

SubCat, a tool for automated repository analysis, was used to extract and transform the raw data from the various open source projects into an SQLite database. The survey forms for the participants of the study were exported from this database which has been stripped of all entities and attributes not relevant for the survey. The returned survey forms could be easily reimported into the database into a seperate table that now contains the survey results.



Classification Schemes

Swanson's Maintenance Tasks

A customized classification scheme based on Swanson's maintenance tasks that fits the open source development life cycle.

- Corrective Tasks
- Adaptive Tasks
- Perfective Tasks

NFR Labeling

A classification schema based on non-functional requirements (NFR) a commit addresses. It is based on the ISO9126 quality model and was proposed by Hindle et al.

- Functionality
- Reliability
- Usability
- Efficiency
- Maintainability
- Portability

Software Evolution Tasks

A classification schema based on activities during software evolution in open source projects, as defined by Hattori and Lanza.

- Forward Engineering
- Re-Engineering
- Corrective Engineering
- Management

Classification Overview