A Historical Dataset for the GNOME Ecosystem

Mathieu Goeminne, Maëlick Claes, and Tom Mens
Software Engineering Lab, COMPLEXYS research institute, UMONS, Belgium

***Abstract*—We present a dataset of the open source software ecosystem GNOME from a social point of view. We have collected historical data about the contributors to all GNOME projects stored on git.gnome.org, taking into account the problem of identity matching, and associating different activity types to the contributors. This type of information is very useful to complement the traditional, source-code related information one can obtain by mining and analyzing the actual source code. The dataset can be obtained at https://bitbucket.org/mgoeminne/sgl-flossmetric-dbmerge.**

I. Introduction

The historical and empirical study of open source software (OSS) *ecosystems* is a relatively recent but fast-growing research domain. An important characteristic of such ecosystems, at least according to our definition [15], is the fact that they are made up of a set of software projects sharing a community of users and *contributors*. A well-known example is GNOME. Its constituent software projects are designed to work together in order to constitute a complete software desktop environment. The GNOME projects are developed by a developer community that is spread across the world. We have observed that it is not uncommon for a contributor to be actively involved in many projects at a time [16]. In addition to this, the type of activity a contributor is involved in may change from one person to another. For example, a very important activity involves internationalization (localization and translation), which is globally managed via the web application Damned Lies[1] for all GNOME translation teams.

Many tools and datasets have been proposed to analyse a software project's history, but few are available at the level of the ecosystem because of the additional level of difficulty involved. It does not suffice to simply consider the union of all project histories belonging to the same ecosystem. Because some projects may have contributors in common, and some contributors may be involved in different projects over time, this information needs to be explicitly represented at the ecosystem level. The same is true for the types of activity of an ecosystem's contributor, and how this varies over time, and over the different projects he is involved in.

In this paper, we present the process we have used to create a dataset containing the historical information related to contributors to the GNOME ecosystem. Our database and the tools and scripts used to created it can be found on a dedicated Bitbucket repository[2].

In contrast to many other datasets, we do not focus on source code, since a significant amount of files committed to GNOME's project repositories do not even contain code (e.g., image files, web pages, documentation, localization and many more). Such type of information is often ignored in MSR research while it is very relevant to understand which types of activities contributors are involved in. For GNOME we observed, for example, that a significant fraction of the community is working on internationalization instead of code [16].

II. Motivation

An important motivation for creating a historical dataset for analysing contributors to the GNOME ecosystem was inspired by the many OSS repository mining studies that have used GNOME as a case study [2], [13]. In 2009 and 2010, GNOME was part of the MSR Mining Challenge, which lead to many contributions [1], [5], [8], [9], [11], [12], [14].

Of specific interest, in the context of software ecosystem research, are the social interactions in the community of contributors. Following a holistic approach, [7] estimated effort and studied developer co-operation and co-ordination in GNOME, based on the version control repositories and mailing lists. Similarly, [4] developed an advanced measure of individual developer contribution based on the source code repository, mailing lists and bug tracking systems, and applied the measure to a number of GNOME projects. [6] studied six GNOME projects in order to understand how contributors join, socialize and develop within GNOME. [10] studied relations between the GNOME contributors by means of social network analysis.

In our own previous work [15], [16] we used the dataset presented in this article to statistically analyse the specialization of workload and involvement of GNOME contributors across projects and activity types, and we explored to which extent projects and contributors specialise in particular activity types.

III. Methodology

To obtain a historical GNOME dataset, we first collected the projects supported by the GNOME git repositories[3]. At the extraction date (8 January 2013), GNOME totaled 1,315,997 commits in 1,418 projects. The URLs of the git repositories were parsed with a Scala script. Next, we used CVSAnalY2[4] to analyse the source code repository of each project. CVSAnalY2 extracts and analyses data from repositories using the version control systems CVS, SVN and Git. The tool populates a FLOSSMetrics[5]-compliant database containing each commit

[1] http://l10n.gnome.org

[2] https://bitbucket.org/mgoeminne/sgl-flossmetric-dbmerge

[3] http://git.gnome.org/browse/

[4] http://github.com/MetricsGrimoire/CVSAnalY

[5] http://flossmetrics.org/

978-1-4673-2936-1/13 © 2013 IEEE

MSR 2013, San Francisco, CA, USA

in the repository, the name and e-mail of its committer and author (if any), the commit date and comment, and the files touched by the commit.

CVSAnalY2 can store the results of multiple extractions into the same database by associating an analysed commit to a source code repository listed in a database table. The tool merges accounts of committers and authors with the same name into a single *meta-account* representing the same person. If this person used different e-mail addresses, only a single e-mail address is preserved. This solution is not acceptable for our purposes because it leads to a loss of information. For example, an empirical study based on the analysis of the e-mail addresses would produce incorrect results. More importantly, it is dangerous to merge accounts based on a person's name only, since different persons may have identical names, and since there may be different names corresponding to the same person (e.g., if initials are used instead of the full name, or if nicknames were used). Merging accounts based on e-mail address is more reliable, since an e-mail address is supposed to belong to a single person (with some exceptions such as generic mailing lists) but then again we saw different people using sometimes the same address or people not filling the e-mail field.

We therefore modified CVSAnalY2 so that it doesn't merge two accounts automatically. Instead, we added an extra table to the CVSAnaly2 database schema to represent account merging. First, we filled this table automatically using an identity merging approach based on [3]. Next, we manually verified and corrected the proposed identity merges in an incremental way to remove homonymous developers and aliases, by searching information on the project's webpages as well as on the personal and professional webpages of developers. The merging resulted in a reduction of 11,094 distinct identities to only 5,923 unique persons. The results of the merge may still contain a limited number of incorrectly merged accounts but this is not accurately quantifiable since we are not aware of any official document associating the different accounts of a given developer.

A second important feature of our dataset is the association of activity types to each GNOME project commit. Using regular expressions based on the file extension and file path we determined whether files corresponded to a particular activity (such as coding or internationalisation). For example, code-related files can be identified based on a combination of their file extension (e.g., `.c`, `.h`, `.cpp`, `.py`) and file path (e.g., files contained in the `src` directory). For more details on the activity types and the regular expressions used for computing them we refer to [16] and our Bitbucket repository.

IV. Database Schema

The schema of the database used to contain our dataset is illustrated in Figure 1 using an entity-relationship diagram. The schema is similar to the one used by FLOSSMetrics. The center table, *scmlog*, represents the list of commits for each git repository. Each *scmlog* entry contains general information related to a commit such as the revision number, repository, date, message. Two particularly relevant fields are the committer and author. The first one is the person who committed the changes to the repository while the second is the person that actually made the changes.[6] The *actions* table stores the set of modifications of each commit. For more information on the other tables we refer to help documents and man-pages of CVSAnalY.

To allow us to analyze software ecosystems, we added three tables (depicted in red) to the FLOSSMetrics schema: *people_merged*, *identity_merging* and *commits_activities*.

Table *commits_activities* provides additional information on commits. This table associates with each commit the type of contributing activity (e.g., *code*, *i18n*, *image*) and the intensity of this activity type, computed as the number of files touched for this activity type. For example, if a commit touched two C code files (corresponding to the *code* activity), and one PNG file (corresponding to the *image* activity), then the table will have two entries associated to this commit. The first one with type *code* and intensity 2, the second one with type *image* and intensity 1.

Tables *people_merged* and *identity_merging* store the merged contributor accounts. An account, stored in the *people* table, is characterized by a name and an e-mail address. A real person (*i.e.* a contributor) is stored in the *people_merged* table and is characterized by a single name. The *identity_merging* table is used to link each account from the *people* table to a real person from the *people_merged* table.

V. Use Case

To illustrate the added value of our dataset, we present a short use case involving *active* GNOME contributors and the projects they are involved in. We consider a contributor as being *active* during a given period interval if he contributed at least one modification to a project's repository during this period. Similarly, we call an active contributor a *coder* if he touched at least one code file during that period. We chose for *6-month* time periods, since this duration corresponds to the time between two GNOME releases. We extracted all GNOME data since the beginning of activity recording (January 1997) until December 2012, thus obtaining 32 consecutive 6-month intervals.

Figure 2 shows the impact of identity merging. On average, it reduces the number of contributors by 25.4%. In 2004, the number of active merged contributors during the first and second semester is 618. Without identity merging, those numbers are 930 and 1011. While the merged number of contributors has remained stable between both periods, not using identity merging would have (incorrectly) shown an increase. Figure 2 also compares the number of active contributors (after merging) against the number of coders. On average the fraction of coders is 77.9%, but this fraction appears to be decreasing over time.

Table I illustrates in a different way the impact of merging contributor identities or considering only coders, by looking

[6] The author and committer fields are identical for version control systems (like CVS) that do not distinguish between both roles.

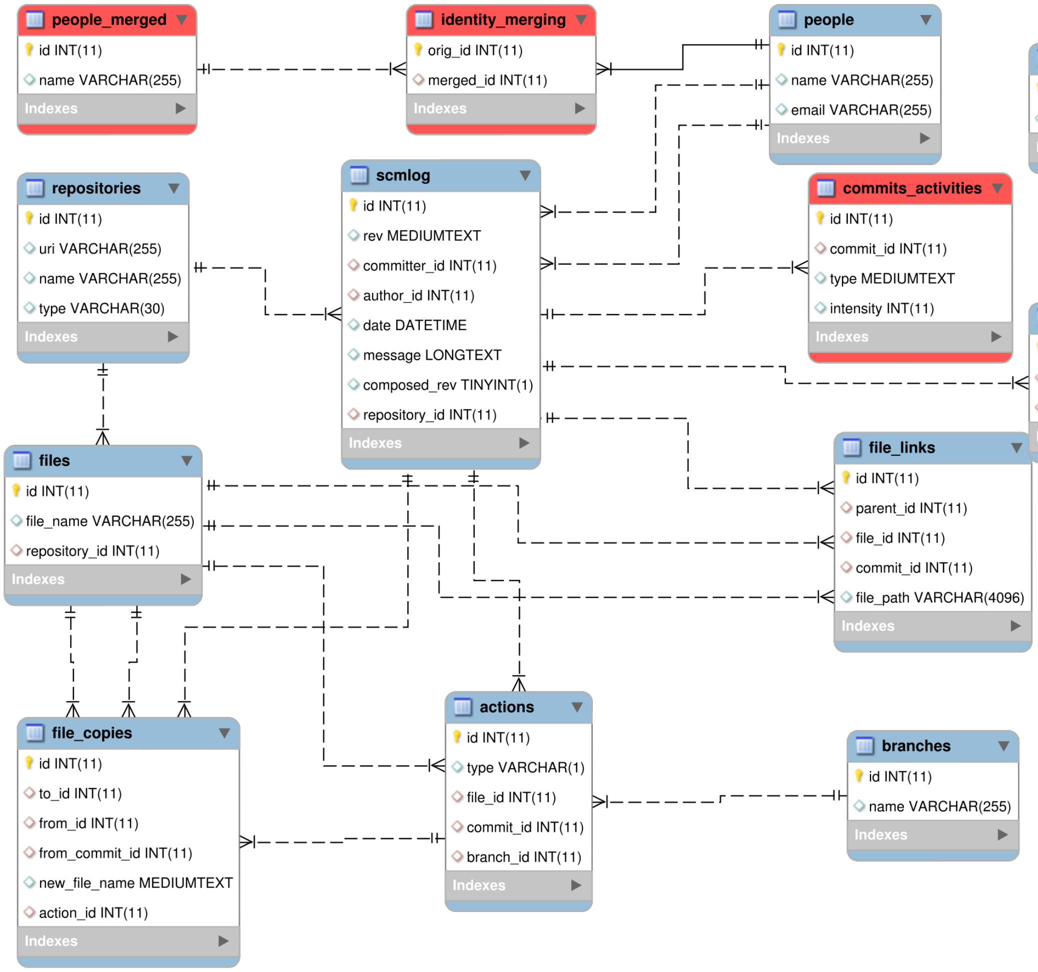

Fig. 1. Entity-relationship model of the FLOSSMetrics-compliant database schema for software ecosystems.

TABLE I
GNOME ALL-TIME TOP 10 CONTRIBUTORS AND CODERS.

	non-merged contributors	merged contributors	coders
1	Matthias Clasen	Matthias Clasen	Matthias Clasen
2	Kjartan Maraas	Kjartan Maraas	Michael Natterer
3	Sven Neumann	Murray Cumming	Sven Neumann
4	Matthias Clasen	Sven Neumann	Bastien Nocera
5	Joseph Sacco	Christian Persch	Alexander Larsson
6	George Lebl	Bastien Nocera	Murray Cumming
7	Owen Taylor	Michael Natterer	Christian Persch
8	Martin Baulig	Alexander Larsson	Owen Taylor
9	Christian Persch	Jorge Gonzalez	George Lebl
10	Murray Cumming	Owen Taylor	Jeffrey Stedfast

at the all-time top 10 of contributors with the highest number of commits. Before identity merging, Matthias Clasen appears twice in the top 10. Murray Cumming is listed 10th, while after merging he suddenly becomes 3rd (he used 18 different accounts). Kjartan Maraas is the second most active contributor, while he does not even appear in the top 10 of most active coders (because he is mainly a translator).

VI. Conclusion

We have provided a dataset for storing the history of contributors (and contributions) of the GNOME OSS ecosystem over a 16-year timespan, taking into account identity merging. The FLOSSMetrics-compliant database schema we used for this purpose explicitly represents the contributors participating in GNOME's projects, and takes into account distinct activity types (such as coding). As such, this database can be used as primary data source for the empirical study of the evolution of collections of GNOME projects, their contributors, and the activity types they are involved in. For empirical studies requiring source code related information, our dataset needs to be complemented with other data sets.

Acknowledgment

This work is partially supported by research projects AUWB-08/12-UMH-3 and AUWB-12/17-UMONS-3 two *Actions de Recherche Concertée* financed by the *Ministère de la Communauté française - Direction générale de l'Enseignement non obligatoire et de la Recherche scientifique*, Belgium.

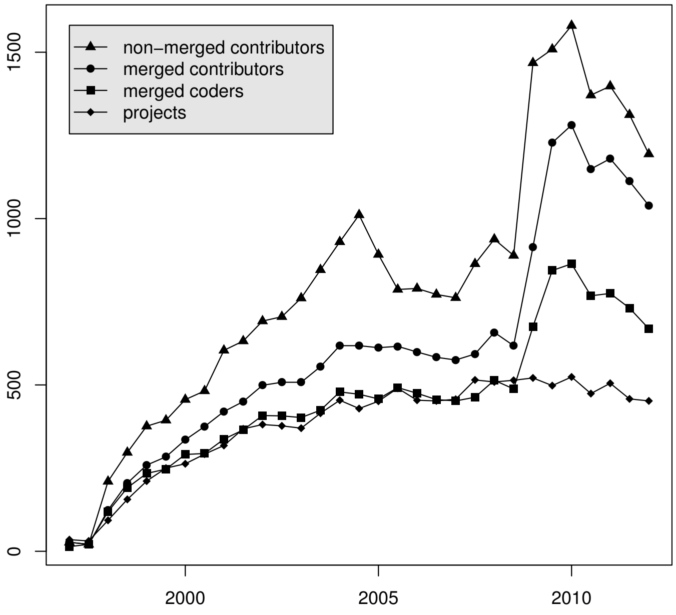

Fig. 2. Evolution, per 6-month interval, of number of active GNOME projects, active contributors (before and after identity merging) and active coders (after identity merging).

REFERENCES

[1] J.R. Casebolt, J.L. Krein, A.C. MacLean, C.D. Knutson, and D.P. Delorey. Author entropy vs. file size in the GNOME suite of applications. In *MSR*, pages 91–94. IEEE, 2009.

[2] D.M. German. The GNOME project: a case study of open source, global software development. *Software Process*, 8(4):201–215, 2003.

[3] M. Goeminne and T. Mens. A comparison of identity merge algorithms for software repositories. *Science of Computer Programming*, 2011.

[4] G. Gousios, E. Kalliamvakou, and D. Spinellis. Measuring developer contribution from software repository data. In *MSR*, pages 129–132. ACM, 2008.

[5] A. Hindle, I. Herraiz, E. Shihab, and Z. M. Jiang. Mining challenge 2010: FreeBSD, GNOME Desktop and Debian/Ubuntu. In *MSR*, pages 82–85. IEEE, 2010.

[6] C. Jergensen, A. Sarma, and P. Wagstrom. The onion patch: migration in open source ecosystems. In *SIGSOFT FSE*, pages 70–80, 2011.

[7] S. Koch and G. Schneider. Effort, co-operation and co-ordination in an open source software project: GNOME. *INFORM SYST J*, 12(1):27–42, 2002.

[8] J. Krinke, N. Gold, Y. Jia, and D. Binkley. Cloning and copying between GNOME projects. In *MSR*, pages 98–101. IEEE, 2010.

[9] E. Linstead and P. Baldi. Mining the coherence of GNOME bug reports with statistical topic models. In *MSR*, pages 99–102. IEEE, 2009.

[10] L. Lopez-Fernandez, G. Robles, J. Gonzalez-Barahona, and I. Herraiz. Applying social network analysis techniques to community-driven libre software projects. *INT J Information Technology and Web Engineering*, 1(3):27–48, 2006.

[11] B. Luijten, J. Visser, and A. Zaidman. Assessment of issue handling efficiency. In *MSR*, pages 94–97. IEEE, 2010.

[12] M. Lungu, J. Malnati, and M. Lanza. Visualizing GNOME with the small project observatory. In *MSR*, pages 103–106. IEEE, 2009.

[13] S. Neu, M. Lanza, L. Hattori, and M. D'Ambros. Telling stories about GNOME with Complicity. In *VISSOFT*, pages 1–8. IEEE, 2011.

[14] H. Schackmann and H. Lichter. Evaluating process quality in GNOME based on change request data. In *MSR*, pages 95–98. IEEE, 2009.

[15] M. Goeminne T. Mens. Analysing ecosystems for open source software developer communities. In *Software Ecosystems: Analyzing and Managing Business Networks in the Software Industry*. Edward Elgar, 2013.

[16] B. Vasilescu, A. Serebrenik, M. Goeminne, and T. Mens. On the variation and specialisation of workload: A case study of the Gnome ecosystem community. *Empirical Software Engineering*, 2013.