July 6th, 2018

Bitergia Privacy Statement

1. Introduction

Free and Open Source software (or FOSS) is more and more prevalent, and public and private entities are using Open Source software and components in their IT infrastructure on a daily basis.

However, one of the perceived and real risks of open source software projects (other than company or foundation sponsored projects) is the voluntary basis for participating in the project, and thus the risk of a lack of continuity, dynamism, a lack of response to new security threats, etc. i.e. on the whole, the project may not be sustainable in the long run.

The risk can replicate itself, as a self-fulfilling prophecy, because if third parties perceive that a project is not dynamic or sustainable, they will not be willing to use or participate in the project, thus holding back or withdrawing support which the project might have had which would have made it stronger – and the project's community is smaller and less dynamic, key shortcomings in terms of sustainability.

In the proprietary world, when a company is considering purchasing a license or investing in a software project, it can obtain information regarding the software and its management through both the company's own information offerings, and also through independent information such as the company's accounts, references and case studies. This is often not the case in a FOSS project, where there are marketing or administrative and finance departments to provide clear information about the software and its ongoing viability.

Companies interested in using, contributing or funding or otherwise participating in an open source software project thus have a legitimate interest in understanding the nature and dynamics of the projects they are interested in, the level of interaction, commitment, the quality of project management and – e.g. – bug-fixing, as these aspects can determine the quality, longevity, adaptability and security of the software they are investing in (whatever their type of contribution or participation).

Since the software development industry has declared open source as key for their success, companies and organizations are becoming more and more engaged in open source communities to get their support and attraction. And they are more and more dependent on projects, whose software is becoming part of critical infrastructure (the SSL story being typical of this situation).

2. Bitergia Analytics Service

The key element of a FOSS project is the community behind the projects. Bitergia Analytics helps companies and organizations to understand not just the software of the FOSS project, but also the community behind it, through quantitative analysis of their communities, in terms of people involved, activity done, processes and performance (e.g. time to solve or attend tasks).

Bitergia's analytic services looks at key questions such as:

- Who is contributing to open source software projects? (not just code – also any conversation or discussion in a forum or any issue submitted, are helpful for a project, by solving bugs or being even potential seeds for new features, new products, etc.)
- Who is behind an open source project?
- How much/what/where are they contributing?

- Who is driving successful projects? (For whatever definition of “successful” defined by the analysts)

These metrics are used to provide objective analysis of the dynamics and “KPIs” of a project, which can support investment and/or technology adoption decisions, which in turn will affect (positively) the life and health of the project as more (or less) entities become interested and support the project.

In this manner, Bitergia Analytics processing helps communities and organizations (from non-profit to profit oriented entities) to understand open source development, by having metrics and data about the projects. These metrics reveal the maturity, dynamics, sustainability, and even security of the project, which are important factors for third parties to understand to take the decision to use the software or not. As we have noted above, FOSS projects do not have a marketing department to provide quantitative or qualitative information about the project, information that is usually essential for a software user (enterprise, public administration) to take a decision on using this software

To illustrate the work we do, we define the project ecosystem which describes a set of relevant projects, where each project is defined by some custom metadata and a set of data sources (git, github, bugzilla, etc.), and we retrieve data from data sources defined by a list of repositories:

- Source Code Management itself (Git, Baazar, Mercurial)
- Issues / Tasks Management (Bugzilla, GitHub Issues, JIRA, Maniphest, Redmine)
- Source Code Review (Gerrit, GitHub Pull Requests)
- Mailing lists / Forums (Askbot, Discourse, Gmane, Hyperkitty, mbox files, NNTP, Pipermail, Stack Exchange, including StackOverflow)
- Continuous Integration such as Jenkins
- Synchronous Conferencing (including IRC, Slack, Telegram)
- Wikis (Confluence, MediaWiki, etc.)
- Meetings Management (Meetup)
- Other sources (DockerHub, RSS)

On the basis of this public data, metrics and KPIs are defined from the data model described before, using bucketing, aggregation, filtering, and counting methods available in the toolkit. Some basic metrics available are:

- Aggregated and evolution over time of activity, people, organizations, etc.
- Time to respond (answer, solve, merge).
- Diversity (items status, organizations, gender, etc.).
- Network and relationship levels

One of the available analyses is **Bitergia Identities Manager**, which is based on SortingHat (https://github.com/chaoss/grimoirelab-sortinghat) technology, and studies information on contributions so as to

- Merge multiple contributors identities in a single unique identity (for right contribution attribution). This is because contributors might have several identities in the same data source (multiple git emails, for example) and across the diverse set of ecosystem’s data sources.
- Affiliate contributors to the companies and organizations they are working or have worked for.
- Set country and gender contributor’s information

So, all in all, Bitergia provides a means for its customers, based on public data, to reduce the risk for companies wishing to use, contribute and invest in open source software projects, revealing in an objective manner the basic health and sustainability of the project through analyzing the public data of the project and applying its specific know-how to provide metrics on projects.

Any of these projects can be made public by the customer. For example, if the customer hires Bitergia to analyze project 'X', and wants project's community to comment, ask or report issues about the dashboard Bitergia Analytics is providing for them, we recommend to open the Support project in their GitLab group

3. Privacy Issues

What are the implications for privacy in our work? Bitergia aims to be fully compliant with data protection laws, and respect individual's privacy to the fullest extent. And we wish to support our customers to be the same.

In the course of carrying out this analysis for our customers interested in understanding the dynamics, sustainability and health of a project in which they are interested, Bitergia collects data on the project which is associated to the names of the contributors, available through the information made public on what one may call the "Software Development and Support Systems" (SDSS for short, here). Thus the analysis involves processing of "*personal data*", regulated in the EU by the General Data Protection Regulation and other laws. Bitergia itself is based in Madrid, Spain, and many of the contributors to FOSS projects are EU citizens, so the GDPR is fully applicable.

The GDPR introduces several (new) specific data protection requirements to comply with, so that Bitergia may carry out the analyses that its Clients ask it to do in compliance with the legal framework, the most important of which are (a) to have a **legal basis** for processing any personal data involved in the analyses (Art. 6) and (b) to **inform** data subjects of any processing (Arts. 13 and 14).

Please note that Bitergia is also carrying out a full GDPR compliance exercise to ensure that it is compliant in all the many other aspects, such as the implementation of adequate technical and organisational measures to protect personal data within its control.

From a technical "privacy" perspective, we take the view that our customers determine which projects are analysed, with what tools, and for what purpose. Our customers are thus Data Controllers, under the terminology of the GDPR, and Bitergia is Data Processor, accessing and processing this data during the course of provision of its services for our customers. And these GDPR requirements must be met by Bitergia's customers, responsible for commissioning the study that Bitergia is providing.

To support our customers in this, we are publishing here our analysis of our engagements from a privacy perspective.

a. Legal basis

There are several bases under the GDPR for processing personal data, the best known of which is consent of the data subject, but that is not the only basis. Bitergia does not process any special category of data (as defined by GDPR), so "express" consent (in GDPR terminology) is not required.

- **Consent**: certain online source code repository and systems supporting, or related to, SDSS and anticipating the data protection issue, provide in their "terms of use" that the personal data of developers made available in the repository are made public, and may be accessed and compiled by third parties. Users consent to this. However, this may not be case for all repositories that we are asked to analyse, and the scope of the consent, in order for it to comply with the GDPR requirements (freely given, specific, informed and unambiguous) is not always clear.
- **Legitimate Interest**. Failing obtaining (prior) consent from developers (which could be obtained if required, in certain cases), the analysis that Bitergia is asked to carry out for our Customers is better based on another ground, which is the *legitimate interest* of the Customer in carrying out the study and – during the course of this study – processing personal data relating to the FOSS project. Our Customers hire our services when they are contemplating a decision to participate, use or contribute to a FOSS project, and we believe they have a strong legitimate (business and technical) interest to analyse and understand the attributes, health and sustainability of the target project and its technologies. To achieve this understanding of the project attributes, the processing of the data – linked to identified or identifiable persons – that is made publically available on SDSS is necessary, and this involves collecting and processing the personal data (name, surname, nickname, email address) of the contributors and other developers involved in the project. This data has already been made publically available on the SDSS (with the consent of the data subject, see above), and taking this into account, together with the nature of the data (just personal contact data), we find that this legitimate interest is not overridden by any interests or fundamental rights and freedoms of the developers as data subject, as commented below.

GDPR requires us to carry out a careful study of this legitimate interest, balancing it against the fundamental rights and freedoms of the "data subjects" (in this case, contributors, release managers, users, and other persons in the FOSS project ecosystem). What's more, while our access and processing of developer personal data for our customers may not fall within the original purpose for which the developer published his/her data on the SDSS (which we understand to be management of the SDSS and attribution of authorship, recognition of the developer's / committer's contribution), we consider that the analysis we carry out for our Customers on the attributes, health and sustainability of the projects out is compatible (in the meaning of Art. 6.4 GDPR) with the original purpose of disclosure, considering the elements that the GDPR establishes to be taken into account (Article 6.4).

Thus we argue that our Customer, acting as Data Controller, has a legitimate interest in analysing the data accessed during a Bitergia analysis, for gaining insight on different aspects related directly or indirectly to different aspects of software development in the analysed FOSS projects, including:

- Sustainability and resiliency of the projects.
- Performance, including the performance and efficiency of the many processes related to software development.
- Community, including aspects such as diversity, involvement, onboarding and exiting.

This interest is due to, depending on the case, one or more of the following reasons:

- Interest in supporting FOSS projects with specific advice about key points learned from the analysis of the data.
- Interest in potentially using the products (programs) produced by FOSS projects.
- Interest in potentially collaborating with or contributing to FOSS projects.

- Interest in analyzing the situation in the market of the products produced by FOSS projects, including how they relate to others, and how they interact in the wider FOSS ecosystem.

We need to weight this interest against the fundamental rights and freedoms of the developers and other persons who data is processed during the course of our work, taking into account the following factors:

- The Data Subjects have expressly made their Personal Data public on the target source code sites
- The terms of use of those Sites provide that the Data Subjects permit access and compiling of that Personal Data, and in some cases expressly permit data analytics
- The data that is accessed is contact data relating to the Data Subject as an amateur or professional software developer, and does not impinge on their personal / domestic life.
- No other Personal Data is accessed or processed
- Less intrusive measures are not possible, as although certain reports and analyses of Bitergia are anonymous or pseudonymous, the identification of the developers is sometimes a key factor in understanding the characteristics of the FOSS Project under analysis.

Looking at this interest, we have analysed the following aspects:

Purpose

- *Who benefits from our analysis?* There are at least two potential beneficiaries: our customers, companies or organizations managing or willing to use or to invest in these open source projects, because they can take decisions based on the data made public by the project in order to manage, use or invest in the project. And secondly, the open source community itself, because it can have a better view of the status of the project, from the data they already control, and a new way to show its transparency.
- *What is the scope of legitimate interest?* There is an interest of Bitergia Customers to do the analysis based on use of data subject personal information like name, email and affiliation because this is the only way to relate people, and the organizations they are working for, with the activity they have done in the project. From that relation, Bitergia is able to provide metrics about the activity and quantitative performance in the project as a whole, and individual contributors . Bitergia clients also have a legitimate interest, as noted above, to understand the characteristics of the project (sustainability, etc.), in order to take decision about using the software, or participating in the project.

Necessity

- *Is the processing of personal data necessary?* While some of our analyses are anonymous or pseudonymous, without identification data we cannot relate authors/contributors and the activity done. And, since that relation has been made explicitly public by the author in the data source, we think it’s the best choice to use that data source as a resource.
- *Can the interests/objectives be achieved in any other way?* No, as far as we know. Bitergia gathers personal information from open source projects development tools, where contributors agree to make public their activity and associated metadata, that includes their names, emails and/or usernames. For example, a code change in an

open source repository, is made public by publishing the code change itself, the name and email of the author of the change, and the date. So, the only way to get, just basic things, like the number of code changes done by each person of the project, is to process their personal information to ensure activity is unequivocally associated with the right people. We can pseudonymise some of the data, to minimise the personal data processed during the course of the analysis, but most of the data has value if it is qualitative (i.e. nominal) as opposed to purely quantitative (i.e. anonymised or pseudonymised). But we cannot segment the data, as this would cancel out any value in the analytics.

- *Is legitimate interests a targeted and proportionate way of achieving your purpose?* Yes, because we are using the information made public by the data subject. It would be nearly impossible to obtain informed consent. The use of the personal data is targeted at only creating these analytics and the personal data is not used in any other manner.

Balance

- *Would people expect you to use their data in this way?* We believe so (also, see "compatible purpose comment below). The SDSS, platforms used for open source development, already inform their users that their data is public and can be compiled by third parties. For example:
  - https://help.github.com/articles/github-privacy-statement/#public-information-on-github
  - https://about.gitlab.com/terms/ or https://about.gitlab.com/privacy/, where it says: "[...] GitLab is an open source project and collaborative community, as well as a company. This means that many portions of our Websites, including information you voluntarily provide, will be public-facing for the open sharing of innovative developments, ideas, and information that makes our collaborative community so great.[...]"
  - Atlassian, authors of collaboration tools, says in their https://www.atlassian.com/legal/privacy-policy: "[...] Collaboration: As a natural result of using Atlassian Services, you may create Content and grant permission to other Atlassian users to access it for the purposes of collaboration. Some of the collaboration features of Atlassian Services display your profile information, including Personal Information included in your profile, to users with whom you have shared your Content. Where this information is sensitive, we urge you to use the various security and privacy features of the Atlassian Services to limit those who can access such information. Your sharing settings may make any Information, including some Personal Information, that you submit to the Atlassian Services visible to the public, unless submitted to a restricted area.[...]"

Bitergia checks all the data sources they track to ensure open source development contributors are informed about these privacy policies. Even custom data sources maintained by organizations.

  - For example, Eclipse Foundation Bugzilla informs (http://www.eclipse.org/legal/termsofuse.php): "[...] As a service to you, Eclipse software may offer functionality whereby you can store certain information on Eclipse Foundation's computer systems, so that it is accessible to you from various devices that you use. Examples include, but are not limited to, the User Storage Service and the Automated Error Reporting Initiative.

> PLEASE BE AWARE THAT ALL INFORMATION THAT YOU STORE WITH ECLIPSE WILL BE PUBLICLY ACCESSIBLE ON THE INTERNET IN UNENCRYPTED FORM AND WITHOUT ANY ACCESS RESTRICTIONS. Do not store passwords, any personally identifiable information, any confidential business information, or anything else that you do not want to be generally and publicly available. By using this functionality, any information that you store (the “Stored Information”) will be subject to the CC0 1.0 Creative Commons license, where, for purposes of that license, the Stored Information shall constitute the “Work” and you shall be the “Affirmer”. What that means is that anybody who accesses your information on the Internet has a worldwide, unrestricted, royalty free, irrevocable, perpetual, non-exclusive license to use, make, reproduce, prepare derivative works of, publicly display, publicly perform, transmit, sell, distribute, sublicense or otherwise transfer the Stored Information without any obligation to you, including any obligation of attribution.[...]”

- *Does the processing have a minimal privacy impact on the individual(s)?* Yes, and we provide safeguards measures to ensure people privacy if requested. By design, Bitergia doesn’t provide any additional information beyond the information already been made voluntary public by the open source project contributors. Even so, Bitergia is able to pseudonymise the output of the data processing (dashboards and reports) to avoid listing personal data like name, emails, etc. Each data subject personal data gathered from the public data sources, is stored with a unique hash code, so, during output production phase, instead of data subject name, Bitergia might use data subject hash code if requested.

- *How does the processing benefit the individual?* The same way that making that information public originally in the data source has benefited them (the individuals): to claim that certain activity has been done by the individual, enhance their reputation, provide evidence of their proactivity and participation in the community

- *Is any individual likely to find the processing intrusive or raise objections?* No, as far as we know. The processing doesn’t imply any interaction with the individual. It actually collects data from the logs and traces voluntary left in the tools and publicly published as part of the contributions made. So, there is no intrusion into the tools used by the individual. Regarding objections that might raise, we don’t expect any objections since Bitergia only provides quantitative data, not qualitative, nor any subjective conclusions about the data we analyse. Even when Bitergia might “qualify” contributors (for example, as core, regular or occasional), it’s based on quantitative data, and it is not a “bad” or “good” evaluation.

- *Do the rights and freedoms of the individual override your interests?* We believe not. We are processing personal data made voluntarily public by the contributors, so we are not overriding any privacy rights and freedoms of the individuals. If the individual wants to be kept in private or hiding, we might do that by pseudonymising their personal data, and we are implementing processes so that this can be respected *a priori*.

- *Are you processing high-risk, special category, children or confidential information*? No.

- *Is the processing for any marketing or other extrinsic purposes?* We are not using the data subject for any marketing or other purposes not specifically related to analysing the dynamics and characteristics of the open source project.

One particular aspect that needs to be put in the balance when determining the legitimate interest of the Customer, is the potential for profiling of developers and other data subjects

in the projects. Profiling is understood as "*valuate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements*". While many of these aspects (health, location, etc.) are not determined, the Bitergia analysis could reveal performance, interests, reliability and/or certain behavioural patterns. These aspects only relate to the interaction of the data subjects and the project/software development (personal aspects intrinsic to the context). They are no external factors or profiling for external purposes other than this interaction with the project , such as marketing, promotions, or other privacy intrusive activities (extrinsic aspects). In the balance, while Bitergia respects Data Subjects' right to object to profiling activities, we believe that this profiling potential does not outweigh the legitimate interest of our Customers to understand a project where it is investing, adopting the software, participating or sponsoring, and the people involved. On top of this aspect, our analysis of contributor's roles in FOSS projects can reveal their merit and dedication, and lead to positive externalities such as public recognition and even hiring.

In any event, Bitergia, Customer will work with the project itself to inform the Data Subjects of the processing and provide an opportunity to object. Based on Bitergia's experience, we provide our customers several options (protocols) - prior to delivery of the results to our Customers, so that the Data Subject's right to remain anonymous or pseudonymous may be respected, and so that the right to object to any potential profile analysis may be exercised PRIOR to the analysis or to delivery of the results to the Customer.

b. Compatible purpose

To reinforce the legitimate interest analysis, we also consider that Bitergia's processing for our customers of personal data made available in the source code repositories is a compatible secondary purpose of processing, in relation to the original purpose for data collection by the source sites (attribution of authorship, community recognition – which has already been consented to), in accordance with Art. 6.4 GDPR, having taken into account the following factors:

a) any link between the purposes for which the personal data have been collected and the purposes of the intended further processing;	The original purpose of publishing the data is to identify the contributors and other persons involved in FOSS project. The purpose of Bitergia's processing for its Customers is the same, to identify and indeed recognise (authorship, merit, effort) those developers in the projects that it analyses.
b) the context in which the personal data have been collected, in particular regarding the relationship between data subjects and the controller;	The original context for collecting the data was the SDSS (software development and support systems) and the FOSS project within that SDSS, where the data subject is the developer, and the original data controller is the SDSS and/or the FOSS project itself. There has been no further processing of data in another context – i.e. Bitergia's processing is carried out in relation to the same context (the FOSS Project). Bitergia may enrich this data with other public data made available by the same Data Subject in other SDSS, but overall, the context is the same (FOSS development), and the purpose for Bitergia processing this data is also the same: to analyse the health and sustainability, and performance of those same FOSS project

c)	*the nature of the personal data, in particular whether special categories of personal data are processed, pursuant to Article 9, or whether personal data related to criminal convictions and offences are processed, pursuant to Article 10;*	The personal data is normally limited to user name, name and sometimes email address, which is basic data that the Data Subject has voluntarily made public in the SDSS. There is no special category of data nor any data relating to criminal activities.
d)	*the possible consequences of the intended further processing for data subjects;*	The purpose of the processing is to determine the sustainability and resilience of the FOSS project where the Data Subject has published his/her data, performance of the project processes and community aspects. I.e. the consequences of the processing by Bitergia are making more visible and understandable the dynamics of the FOSS project. The raw data about the project is already public, so there is no "further disclosure" of the personal data – Bitergia's clients already have access to the original data, just not in the format and manner that Bitergia presents its results. Due to the analytics carried out on the data, there is potentially a profiling of the data subject, as the analysis of the project activity my reveal and be used to evaluate "*certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, …., personal preferences, interests, reliability, behaviour, …. …* ". See analysis above
e)	*the existence of appropriate safeguards, which may include encryption or pseudonymisation*	Bitergia implements strict security safeguards in relation to its processing of all data from the FOSS projects it analyses. As regards, pseudonymisation, certain reports and analyses of Bitergia are anonymous or pseudonymous, but the identification of the developers is sometimes a key factor in understanding the characteristics of the FOSS Project under analysis.

c. Information

Under the requirements of GDPR, a reasonable effort should be made to inform the data subjects about the processing, and this includes the name and contact of the data controller, the legal basis of processing, retention periods, and importantly data subject rights (access, cancellation, etc.).

Usually this must be provided immediately on collection or within at most 30 days. Given our experience in the matter, Bitergia can work with the customer, and support them whenever possible, in this information effort. We have considered whether clients can take advantage of the waiver of this obligation (art. 14.5) however given the circumstances (access to developer emails, project mailing list, etc.) it would be hard to argue that the effort of providing this information is disproportionate. Bitergia will support Clients in providing this data, and we have provisions below to this effect.