Fusepool

STREP

FP7 – 296192

D2.2 Privacy preserving data gathering and curation

Deliverable Lead: Andrew Janowczyk

Author: Stephane Gamard, Andrew Janowczyk

1st Quality reviewer: Adrian Czerny

Deliverable nature:	Report (R)
Dissemination level: (Confidentiality)	Public (PU)
Contractual delivery date:	30 June 2013
Actual delivery date:	18 July 2013 (updated 30 Dec 2013)
Version:	1.4
Total number of pages:	14
Keywords:	Fusepool, information retrieval, user requirements, living lab, software architecture, data mining, linked data, social curation, machine learning, graphical user interface, data visualization, user engagement, hackathon

Abstract

Privacy is an important issue for any online platform, not only because individuals require a certain level of trust for the usage of such systems but also there are legal considerations which must be complied with. This deliverable is focused on reviewing European Law and instantiating an end user license and privacy model which ensures that the platform is entirely compliant regarding user privacy information and tracking data. This is a core component as we collect various information, similar to Facebook, to improve and enhance our system based on user feedback data.

Executive summary

The Data Protection Directive (95/46/EC) governs the European requirements for user data collection and utilization. It is focused on 7 major pillars: Notice, Purpose, Consent, Security, Disclosure, Access, and Accountability. In the attached report we review each of the individual facets and explain how the Fusepool platform meets or exceeds each of the requirements. Thankfully, tracking user information is not a new paradigm, so there are many possible frameworks already in place that we can merge in order to precisely cover our needs. To proceed, we reviewed not only the Directive but many end user licenses which are commonly used by leading platforms such as Facebook, Google and Microsoft. Using these we compiled an end user license and understood our responsibilities to modify development tasks so that the final process is legally compliant.

Document information

FP7 Project Number	296192	**Acronym**	Fusepool
Full Title	Fusepool – Information pooling for product/service development and research		
Project URL	http://www.fusepool.eu		
Document URL			
Project Coordinator	Michael Kaschesky		
EU Project Officer	Martina Eydner		

Deliverable	**No.**	D2.2	**Title**	Privacy preserving data gathering and curation
Work Package	**No.**	WP2	**Title**	Data sourcing and component integration

Date of Delivery	**Contractual**	M12	**Actual**	M12
Status	Version 1.4		Final X	
Nature	Prototype □ Report X Dissemination □			
Dissemination level	Public X Consortium □			

Authors (Partner)	Stephane, Gamard, Andrew Janowczyk			
Lead Author (WP Lead)	**Name**	Andrew Janowczyk	**E-mail**	andrew.janowczyk [at] fusepool.net
	Partner	Searchbox	**Phone**	+41 78 914 88 02

Version Log			
Issue Date	**Rev. No.**	**Author**	**Change**
10-6-2013	I	A. Janowczyk	Initial Version
4-7-2013	II	A. Janowczyk	Proofread and formatting adjustment
18-7-2013	III	A. Janowczyk	Integrate reviewer's comments
19-1-2014	IV	S. Gamard	Integrated reviewer's (Michael, Adrien Cze.) comments & Resubmission

Table of Contents

Abbreviations

Acronym	Full meaning
API	Application programming interface
BUAS	Bern University of Applied Sciences
DoW	Description of Work (part of the Grant Agreement)
ENOLL	European Network of Living Labs
GA	Grant Agreement
GEOX	Geox KFT
LL	Living Lab
MS	Milestone
OSGi	Open Services Gateway initiative
RAD	Rapid application development
RDF	Resource Description Framework
SEARCH	Searchbox SA
SME	Small and Medium Enterprise
SWOT	Strengths, Weaknesses, Opportunities, and Threats
TREPA	Treparel Information Solutions BV
WP	Work package
XEROX	Xerox SAS
XSLT	eXtensible Stylesheet Language Transformations
XML	Extensible Markup Language
OSS	Open Source Software
ML	Machine Learning
T&A	Terms & Agreements
ACL	Access Control List

1. Privacy preserving data gathering and analysis (Task T2.4)

There was only a single task associated with this deliverable. This task is described in the following subsection. The main purpose was to ensure that the platform that we are developing is both legally compliant and comfortably meets user expectations in regards to privacy and data collection.

In this deliverable we outline the matter of privacy as well as on how privacy is implemented and guaranteed within the Fusepool platform.

1.1. Specifications

The work on the 'Privacy preserving data gathering and analysis' concerns Task T2.4 in the DoW which this deliverable D2.2 reports on. No major changes to the task specification were required when carrying out the task. This task was broken down into several subtasks with subtask-deadlines that were defined and agreed in the first meeting of the Fusepool General Assembly the 25th of July 2012, about three weeks after the Fusepool kick-off meeting. The sub-tasks are defined as follows:

- T2.4a – Privacy-compliant user data profiling [3/2013 planned; **5/2013 actual**]
- T2.4b – Privacy protection best practices and Fusepool approach [4/2013 planned; **5/2013 actual**]

1.2. Background

Privacy concerns in online platforms is an especially sensitive topic, as many people feel very uncomfortable with unnecessary tracking of their online usage. The goal of this task was to address these requirements by identifying European laws that govern the minimum requirement needed to be in compliance.

The European Union directive which regulates the processing of personal data within the European Union is the Data Protection Directive (95/46/EC), described in great detail here: http://en.wikipedia.org/wiki/General_Data_Protection_Regulation. There are seven main principals that govern its implementation.

These seven principles, outlined below, are addressed by four distinct measures within the Fusepool platform. By enforcing the privacy measures directly within the architecture of the platform we show how the Fusepool platform meets or exceeds the necessary requirements to be legally compliant with the European Data Protection Directive (95/46/EC).

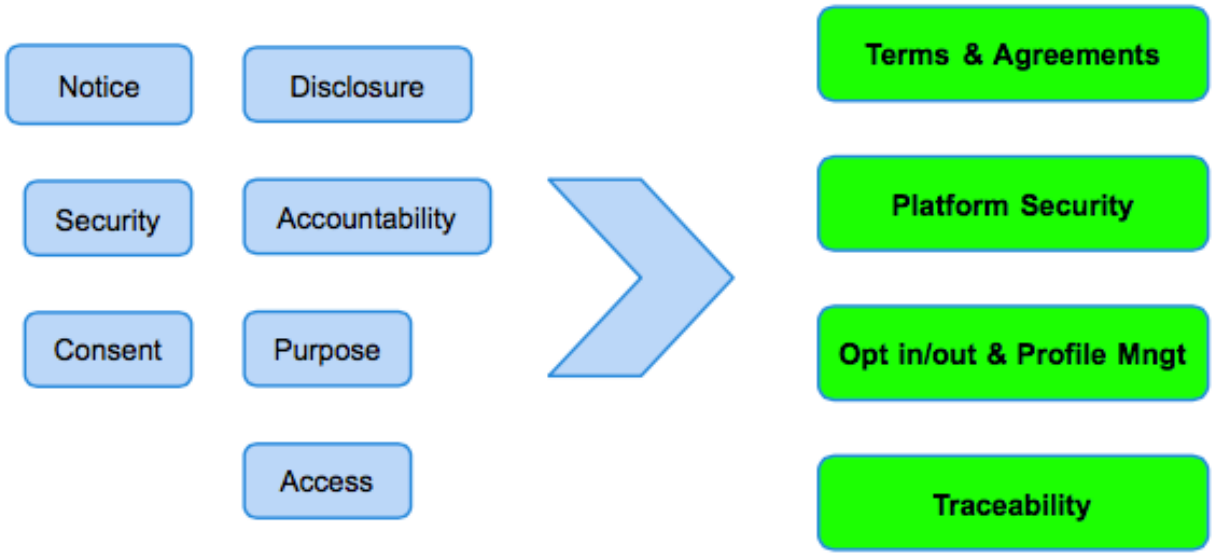
1.2.1. Privacy Fundamentals

By examining the Data Protection Directive we identified and complied with the following seven privacy fundamentals:

1. **Notice**: data subjects should be given notice when their data is being collected;
2. **Purpose**: data should only be used for the purpose stated and not for any other purposes;
3. **Consent**: data should not be disclosed without the data subject's consent;
4. **Security**: collected data should be kept secure from any potential abuses;
5. **Disclosure**: data subjects should be informed as to who is collecting their data;
6. **Access**: data subjects should be allowed to access their data and make corrections to any inaccurate data
7. **Accountability**: data subjects should have a method available to them to hold data collectors accountable for following the above principles.

1.2.2. Privacy Measures

By integrating four distinct privacy measures we can cover the seven fundamentals of privacy within the Fusepool platform. The detail for each measure is outline in the implementation segment of this deliverable and its implication toward the seven fundamentals of our privacy policy is revealed in the solution chapter of this deliverable.



The Fusepool platform will have specific Terms & Agreements (T&A) that cover the specific usage and handling of the information to be processed. More than a search engine, the T&A must reflect the crawling, user rights, administration and data-gathering for Machine Learning purposes.

Of course, registered users will have the option to opt-in/out with regards to usage of their collected data through their use of the platform. The profile management is also capable - for registered users - to account for all records given the inherent traceability implemented within the Fusepool platform (see 1.3.4).

And last but not least, we are striving for a very secure platform where user and data records are safe. The security component of the Fusepool platform has been considered since its first designs and we have spent numerous efforts at implementing the level of security that will satisfy our requirements deeply within the entire framework.

It is only the **conjunction** of these four measures that satisfies the privacy requirements put forth by the European Data Protection Directive (95/46/EC). This is the reason why we have acted early in the project and embedded some of these measures deep within the architecture of the Fusepool platform as to go above and beyond the "guideline principle" and ensure that these measures would not easily be circumvented by participants, developers and users of the Fusepool platform.

1.3. Implementation

The implementation of our privacy measures is not purely architectural and programmatic. We are now outlining the work done in order to make sure that each privacy measure stands up to its expectation within the Fusepool platform.

1.3.1. Terms & Agreements

The Terms & Agreements (T&A) measure is neither an architectural nor a programmatic measure. Its implementation is in the form of a contract between the user and the service provider (the Fusepool project in this case).

The T&A must cover and explain the seven privacy fundamentals to the user, as well as provide him with the means to either accept or deny the usage of the platform given the outlined terms.

There are a multitude of T&A drafts available, but given the nature of the Fusepool platform, we had to base our inquiry on similar services. Indeed, the Fusepool project is not just a search engine but more of a data-market place. The Fusepool platform is a service where users can **exchange** and leverage each other's **information**.

Given the nature of the platform we had to take into account the fact that users not only provide valuable insights for machine learning by gathering their activity, but they also put forth in the system data sources

that need to be protected. The privacy content and the span of the Fusepool Terms & Agreements would have to match expectation of the user as well as its data.

It is for these reasons that we looked closely to the European version of Google's Terms & Agreements as a starting point. Google supports multiple services, including Google Doc and Gmail which have similar interactions with the users as does the Fusepool platform:

- Google saves documents and ensures its privacy with its users (given the right access and privileges),
- Google uses the user's interactions with its services to enhance its performance, service quality and data-mining.

Of course, the sheer nature of the Google and the Fusepool entities is entirely different. This difference is grounded in the fact that Fusepool does not market, neither to itself nor to third parties, the user's collected information.

The key point in our T&A is that whatever is collected of the user's interaction with the platform stays within the platform for its own machine learning. No entities (programmatic, human or business) that do not belong to the Fusepool project will ever have access to the user's collected information.

As a result, we present our own implementation of the Fusepool's Terms and Agreements:

Fusepool Terms & Agreement

What information do we collect?

We collect information from you when you register on our site, respond to a survey or fill out a form.

When registering on our site, as appropriate, you may be asked to enter your: name, e-mail address, mailing address, phone number or other company information. You may, however, visit our site anonymously.

What do we use your information for?

Any of the information we collect from you may be used in one of the following ways:

1. To personalize your experience (your information helps us to better respond to your individual needs)
2. To improve our website (we continually strive to improve our website offerings based on the information and feedback we receive from you)
3. To improve customer service (your information helps us to more effectively respond to your customer service requests and support needs)
4. To administer a contest, promotion, survey or other site feature
5. To send periodic emails (The email address you provide may be used to send you information, respond to inquiries, and/or other requests or questions.)

Your information, whether public or private, will not be sold, exchanged, transferred, or given to any other company for any reason whatsoever, without your consent.

How do we protect your information?

We implement a variety of security measures to maintain the safety of your personal information when you enter, submit, or access your personal information. We offer the use of a secure server. All supplied sensitive/credit information is transmitted via Secure Socket

Layer (SSL) technology and then encrypted into our Database to be only accessed by those authorized with special access rights to our systems, and are required to keep the information confidential.

After a transaction, your private information (credit cards, id numbers, financials, etc.) will not be kept on file for more than 60 days.

Do we use cookies?

Yes (Cookies are small files that a site or its service provider transfers to your computers hard drive through your Web browser (if you allow) that enables the sites or service providers systems to recognize your browser and capture and remember certain information). We use cookies to help us remember and process the items in your search results, understand and save your preferences for future visits, keep track of advertisements and compile aggregate data about site traffic and site interaction so that we can offer better site experiences and tools in the future.

Do we disclose any information to outside parties?

We do not sell, trade, or otherwise transfer to outside parties your personally identifiable information. This does not include trusted third parties who assist us in operating our website, conducting our business, or servicing you, so long as those parties agree to keep this information confidential. We may also release your information when we believe release is appropriate to comply with the law, enforce our site policies, or protect ours or others rights, property, or safety. However, non-personally identifiable visitor information may be provided to other parties for marketing, advertising, or other uses.

Third party links

Occasionally, at our discretion, we may include or offer third party products or services on our website. These third party sites have separate and independent privacy policies. We therefore have no responsibility or liability for the content and activities of these linked sites. Nonetheless, we seek to protect the integrity of our site and welcome any feedback about these sites.

Can I edit my information?

As part of the European Union Directive (95/46/EC), all users of our site may make any changes to their information at anytime by logging into their control panel and going to the 'Edit Profile' page.

Children's Online Privacy Protection Act Compliance

We are in compliance with the requirements of COPPA (Children's Online Privacy Protection Act), we do not collect any information from anyone under 13 years of age. Our website, products and services are all directed to people who are at least 13 years old or older.

Online Privacy Policy Only

This online privacy policy applies only to information collected through our website and not to information collected offline.

Your Consent

By using our site, you consent to our online privacy policy.

Changes to our Privacy Policy

If we decide to change our privacy policy, we will post those changes on this page, and/or send an email notifying you of any changes.

1.3.2. Security

There are two levels of security within the Fusepool platform. In this measure we address both the inner security of the components of the Fusepool platform as well as the security of its data.

Fusepool platform security

The platform is based on Open Source Software (OSS) with state of the art security components. We did not directly create those components but merely integrated them (JASS for example for ACL, see next chapter) within the core of the Fusepool platform.

By relying on proven software that is maintained by large community and is widely used, we are confident that the platform is secured to its best possible standards. Given the nature of the components within the Fusepool platform (OSGI) we can ensure that each component complies with our implementation of the OSS security standards.

Fusepool data security

The Fusepool platform ingests and uses large amounts of information coming from multiple sources and organisations. It is to be expected that not all users have the same rights and privileges with regards to the source or organisation that some information comes from.

In order to insure that the right information is available to the right user with the right privilege we have implemented within the Fusepool platform a Restrictive Access Control List.

An Access Control List (ACL) is a list of permissions for a given entity. Permissions ranges from “can access” to “can edit”. As a matter of fact, within our implementation, a permission can be as fine-grained as a method of a service (equivalent to a single function in Java). The Entities for which permissions are granted are represented as a node within the Fusepool platform RDF graph. This node can represent a user, information or pretty much anything that is expressed in the knowledge graph of Fusepool.

By using a restrictive ACL we disable accessibility/editability to anything that does not directly belong to the user. Further access to individual users or to groups can be granted within the User-Management interface developed as part of the Fusepool Project.

1.3.3. Profile Management

As part of the User-Manager component of the Fusepool platform users can review and manage their profile. What is available to registered users is the option to opt-in/opt-out of the user-data collection. When a user opts-in for data-collection, he has the possibility to survey the information collected by the platform on his behalf as well as delete that information.

Of course, within the Profile Management part of the Fusepool platform a user can review the Fusepool T&A and if he so wishes un-subscribe from it. In the event of the user un-subscribing from the Fusepool platform while having user-data collected, all of his data will then be anonymized (equivalent to a non-registered user of the platform).

1.3.4. Traceability

Traceability is the key for multiple features within the Fusepool platform. The main usage of user-based data collection is to feed in the machine learning components of the platform as to provide a better, higher quality service to end users.

The traceability of what the platform does and how does a user use it is available at two distinctive levels. Logging of all activities, including access log, is common practice for all web platform. The activity log of the platform is anonymous, and can be used to mine certain user behaviours as well as track down bugs within the platform. The end user has no incident on the system logs and is not provided with a way to retrieve them or read them.

The second level of traceability is executed at the User Interface (UI) level. By implementing good practice within the Fusepool platform we have enforced that each action a given user undertakes to be registered and stored in the "annotation store" (AS). It is from that data collection that processes requiring Machine Learning takes advantage. The AS is purely available to software components within the Fusepool platform, and users have access to the collection of all of their interactions with the Fusepool platform thru the AS as part of the Profile Management.

Such recorded behaviour currently includes, but is not limited to, the user selection of hits after a search, the user labels and classification among other interactions.

Users that have not registered all share a common "anonymous" user object. As per stated in the T&A section of this deliverable, anonymous users also contribute to the annotation store but have no way to access and alter the data their usage of the platform will have generated.

1.4.Solution

We are confident that the implementation of the four measures outlined above yields above standard results with regards to user privacy as dictated by the Data Protection Directive (95/46/EC). As a solution we present a highlight on how the seven fundamentals of the Data Protection Directive (95/46/EC) are taken care of with regards to our implemented measures.

1.4.1. Notice

data subjects should be given notice when their data is being collected;

Upon registration the user is presented with an end **T&A** which dictates specifically what information of theirs will be retained. For all anonymous users, whose data is harvested (such as in ranking of search results, etc), a pop-up indicates that certain pieces of information have been stored on their behalf. The latter case is less serious since there is no possible way to identify an anonymous user.

1.4.2. Purpose

data should only be used for the purpose stated and not for any other purposes;

In the **T&A** we clearly indicate that the information being collected is used not only for machine learning algorithms - which can greatly improve relevancy of results in their favour - but also as a way to have SME collaboration via similar interests, of which the user must manually opt-in.

1.4.3. Consent

data should not be disclosed without the data subject's consent;

The consent is received by acceptance of the **T&A**, at this time the user can also opt-out of any data collection, freeing the system from the responsibility of maintaining and tracking the users actions entirely (from the **traceability** component).

1.4.4. Security

collected data should be kept secure from any potential abuses;

The entire Fusepool platform is bound by tight **security** at the architecture level. OSS standards and best practices for hosting have been set in place. Furthermore the **ACL** which binds content and user together ensures the privacy of user generated content.

1.4.5. Disclosure

data subjects should be informed as to who is collecting their data;

We have no desire to forward the user information to other parties, nor will we be selling it. This leaves Fusepool as the sole collector and user of the information, a point made very clear in the **T&A**.

1.4.6. Access

data subjects should be allowed to access their data and make corrections to any inaccurate data

With the **Profile Manager** the users will have the ability to edit incorrect information caused by either typos or aging information (such as a changed mailing address).

1.4.7. Accountability

data subjects should have a method available to them to hold data collectors accountable for following the above principles.

Given the traceability of user and content, as well as the ACL security, each action and delivery within the Fusepool platform complies with the accountability. It is always possible to know who did what and when. In the case of non-registered user, they do not have the permission to perform certain actions (enforced by the ACL) and hence cannot generate content for which they could not be accountable.

1.5. Conclusion & Future Work

The future work needed for this task is the placement of the end user agreement online, and ensure that users accept it before any of their information is tracked. In the long term, since laws and society are constantly changing, regular review of the policies should be undertaken to ensure continued compliance.

2. Outlook

The contributions described herein enable the Fusepool platform to become legally compliant in regards to data gathering and user monitoring. We are pleased to say that the above mentioned task was completed in an organized and well managed fashion. As discussed above, the task in this particular deliverable was completed on time and according to specification. The interesting part is of course seeing how the various existing platforms such as Facebook manage their own risk via licenses. At the end of the day, everyone seems to take the same approach and we've followed suit. With the task completed in good working order, the next task of implementing the requirements on the front-end are expected to go smoothly. This is of a critical nature as the next steps involve more complicated software and algorithms which will be what gives Fusepool its unique sales proposition.