A Personal Data Framework for Exchanging Knowledge about Users in New Financial Services

Beatriz San Miguel, Jose M. del Alamo and **Juan C. Yelmo**[1]

Abstract. Personal data is a key asset for many companies, since this is the essence in providing personalized services. Not all companies, and specifically new entrants to the markets, have the opportunity to access the data they need to run their business. In this paper, we describe a comprehensive personal data framework that allows service providers to share and exchange personal data and knowledge about users, while facilitating users to decide who can access which data and why. We analyze the challenges related to personal data collection, integration, retrieval, and identity and privacy management, and present the framework architecture that addresses them. We also include the validation of the framework in a banking scenario, where social and financial data is collected and properly combined to generate new socio-economic knowledge about users that is then used by a personal lending service.

1   INTRODUCTION

Tailored and customized features are increasingly becoming more popular in IT services. These adjust offers and functionalities of services to the user preferences, interests and personal needs, generally going beyond functionality of the service itself and thus, improving it. In the banking sector, it is not an exception and for some time now new players have appeared to offer financial services based on personalization and recommendations.

Traditionally, banks have been early adopters of new technology solutions, but mainly following a bank-centric approach that users are rarely able to notice [1]. IT companies and new service providers have leveraged this gap to offer user-centric financial services. For example, on-line payment is one of the most competitive areas into which IT companies such as PayPal, Google or Apple, have entered. Moreover, many financial services related to crowdfunding, lending clubs, investment recommendations, financial aggregators that allow the management of personal finances, the comparison or recommendation of banking products, etc. have transformed the traditional ways of financial organizations, or have even created entirely new ones.

These innovative financial services create new opportunities, but also potential threats in the industry. It is vital for banks to understand the new directions and develop threats into new opportunities and returns. In this sense, most of these new financial services require personal data and financial information about users in order to know them better and then, offer and improve services. Here banks possess inherent competitive advantages, since they have a large amount of customer data, transaction information, and the capabilities to enable financing and secure services [2] and [3].

Well aware of this situation, in 2014 the Center for Open Middleware (COM), a joint technology center created by Santander Bank and Universidad Politécnica de Madrid, launched a pilot project intended to research, analyze and evaluate new potential opportunities and applications around personal data. Specifically, the project aims to establish a framework that allows the sharing and use of personal data among companies, and the creation of knowledge about users, while allowing users to manage and control their flow of personal information, defining who access which data and why.

In this paper we introduce the aforementioned framework which has been called the Personal Data Framework (PeDF). The PeDF includes mechanisms for gaining access to personal data from several heterogeneous data sources, and integrating them to facilitate their analysis and processing to produce and infer new knowledge about users. This information can be provided to new financial service providers that, as new players, do not have sufficient personal data to offer their services. On the other hand, there are currently tensions related to the use of personal data, causing privacy and trust concerns in users. In this context, the European public sector is attempting to regulate and evolve the existing legislation to strengthen individual rights in relation to the uses of their personal data and their privacy, while boosting digital and personal data economy [4]. Therefore, the framework includes the necessary tools to involve users in the management and control of their personal information.

The remainder of the paper is organized as follows. First, Section 2 includes the technological background for each issue that covers the PeDF related to personal data: collection, integration, retrieval, and identity and privacy management. Then, Section 3 describes the PeDF architecture, and Section 4 includes the PeDF validation that we have conducted in the financial context. Finally, we present related work in Section 5, and conclude the paper by highlighting conclusions and future directions in Section 6.

2   TECHNOLOGICAL APPROACCHES

The PeDF acts as an intermediate entity between service providers and individuals to allow the former to share and exchange existing personal data and new knowledge obtained from them which cannot be done unilaterally, while enabling users to retrieve a global view of their personal information and decide who can access which data and why. To make it possible, the PeDF has to include mechanisms for gaining access to personal data that are scattered across different service providers (data sources). When the data sources supply personal data to the PeDF, it has to be able to integrate them. This integration must allow the PeDF to provide

[1] Center for Open Middleware, Universidad Politécnica de Madrid, Spain, email: {beatriz.sanmiguel, jose.delalamo, juancarlos.yelmo}@centeropenmiddleware.com

personal data and knowledge obtained from these data to service providers (referred to as data consumers). All of the above has to be controlled by the user and thus, it requires the PeDF to include identity and privacy management solutions.

In summary, the PeDF covers four main technological issues: personal data collection, integration, retrieval, and identity management and privacy. Next, we will present the background associated with each issue, detailing its technological solutions.

2.1 Personal data collection

Data sources can be classified into two main categories in relation to personal data access: public or private, but one source can be categorized as both, depending on the personal data concerned.

The public data sources contain personal data that are accessible in an equitable way for any entity in the public network. On the other hand, in the private data sources, the personal data can only be accessed by authorized entities. We can think of numerous examples of personal data sources, such as social networks, instant messaging services, mobile applications, and many other service providers specialized in a specific user domain such as education, banking, or e-commerce. As an illustrative example, a social network can act as a public or private data source depending on the user configuration.

There are different technologies that allow third parties to collect the personal data from data sources. For the public ones, the so-called Internet bots, spiders, or web crawlers are the most representative. These are software solutions that automatically search, access and retrieve public information on the Internet.

As regards private data sources, there are several mechanisms based on user consent that allow third parties to access the protected personal data. One of the easiest ways is the method based on data files. This kind of files contains personal data created by a user in a specific data source and can be exported by users. For example, Google allows its users to access their personal data, downloading different files[2]. The main problem associated with this solution is that it requires extra work for the users, since they have to be actively involved to download their files, carrying out manual tasks. Moreover, files can be easily manipulated to change their content, and therefore, the security mechanisms are weak. In order to solve this problem, a set of programming functions, protocols, and standards has appeared to automate the process: data sharing Application Programming Interfaces (APIs).

APIs have become the de facto mechanism for sharing and exchanging personal data, since they allow different software applications to communicate and interact directly [3]. They offer code-based access to different functionalities and services to third parties by abstracting their implementation details. On the Internet, the Representational State Transfer (REST) [5] architectural style has recently emerged as the favorite for implementing APIs. It is based on the Hypertext Transfer Protocol (HTTP) to allow connectivity, but it does not specify the syntax of messages. The individual messages and interfaces are designed according to the suppliers' semantic. For example, Facebook and Twitter include different APIs (Graph API[3] and REST APIs[4], respectively) to read and write their user personal data, which are based on the HTTP for communication, and JavaScript Object Notation (JSON) [6] for

data interchange. Although the same protocol and language still apply, there are differences, since the suppliers' API use different syntax and semantic to refer to the same data.

In a nutshell, there is no unified API specification, each API contains its own description, which can be poorly documented, and therefore, understanding each one is challenging. There are some initiatives to solve the associated API problems, such as the OpenSocial standards [7] that include a set of open APIs that developers can use to gain access to user personal resources hosted by different providers who have implemented them. We can find a few related solutions in the social network services, such as [8], that proposes a framework to integrate the interaction with different social APIs.

2.2 Personal data integration

Data integration is an old field of research that aims at combining data from different sources and providing them in a unified view [9]. Over time, many solutions have been proposed [10], but two main approaches regarding storage can be followed:

- Centralized way. The personal data is retrieved from external data sources, saved, and stored in a central repository. This is a replication of the personal data stored by data sources and thus, maintaining and updating the replicated data is a key issue. It must incorporate techniques to carry out a periodical refreshing of personal data, or even better, mechanisms that allow the detection of data changes in real time. Despite the aforementioned, it has clear benefits related to availability and timeliness. Furthermore, it facilitates data analysis and processing.

- Decentralized way. Here, there is a central directory or registry and a distributed data storage. It entails little or no storage since personal data is maintained and stored by each external data source. However, personal data access is more complex and generally less efficient than the previous way because recovering data is carried out on the fly and there can be source access limitations.

The two mechanisms are complementary since the central repository of the first way can be considered as an extra storage point for the decentralized solution. Furthermore, both solutions face the challenges of corresponding personal data at different data sources, and giving them a common definition. The former entails the development of algorithms and mapping techniques that (semi)automate the correspondence process to eliminate manual tasks. On the other hand, the common definition of personal data involves establishing a standard to represent the personal data.

There is no standard or a generally adopted representation for personal data, neither the structure (format of the representation), nor even the semantic (meaning of the content). We can find many proposals for standards and proprietary solutions to define each personal data category, almost as many as there are service providers. One of the most promising solutions for integrating all these discrepancies is the use of ontologies.

An ontology is an engineering artifact made up of a vocabulary that describes a certain reality, and a set of explicit assumptions regarding the intended meaning of the vocabulary terms [11]. It enables a common understanding of a specific domain to be shared across a wide range of service providers, adding interoperability, consistency, reusability, and many other advantages [12].

Over time, many ontologies have been proposed for diverse domains including healthcare, molecular biology, or web searching. There are general ontologies describing concepts (e.g., object, process and event) that are the same across different domains, such as the Suggested Upper Merged Ontology (SUMO) [13]. Additionally, there are more specific ontologies (namely domain ontologies) that represent the particular concepts of a domain. In the social network field, the Friend of a Friend (FOAF) ontology [14] includes the main terms to describe people, the links between them and the things they create and do on Internet. In the financial industry, the Financial Industry Business Ontology (FIBO) [15] is an ongoing definition of financial industry terms such as contracts, product/service specifications and governance compliance documents. SUMO also includes domain ontologies for finance and economy.

Finally, there are different methodologies and languages for defining your own ontologies, such as those described in [16]. One of the most popular languages is the Web Ontology Language (OWL) [18] that is part of the W3C technology stack. OWL allows the definition of concepts and the complex and rich relationships between them.

2.3 Personal data and knowledge retrieval

Personal data can be offered to third entities, and even more interestingly, these data can be analyzed and processed to obtain knowledge that cannot be achieved unilaterally by service providers. The process for producing this knowledge is referred to as user modelling in the literature [19].

Traditionally, user modelling is a one-sided process in which service providers autonomously collect personal data and then generate user models that satisfy their business needs in a specific domain. A user model is understood as the interpretation of a person in a specific context for an organization. It includes what the organization thinks the user is, prefers, wants, or is going to do, and comprises mainly derived and inferred data. The user model can be used to recommend new contents or services, personalize user interaction, or predict user behavior, among others.

There are different techniques to create user models, choosing one or another depends on what information is been stored and the final application of the model. Next, we point out some of the approaches that can be taken.

2.3.1 Vector-based models

Here, a user is represented by a set of feature-value pairs. The features can be items or concepts of a domain, such as products of a shop, or links on a web site. Each of them has associated a value (usually, a boolean or real number) that indicates the attitude of a user to this feature. For example, the value can indicate whether a user has searched for a product or the number of visits to a link.

There are other approaches similar to this one such as keyword-based, bag of words, or user-items rating matrix [20], which consider only words or terms interesting to users with or without an associated value, or historical user ratings on items, respectively.

This approach is one of the simplest since its implementation and retrieval is quite easy. It has been used by nearly every information retrieval system [21]. However, it is difficult to share with other data consumers because the features and values can be misinterpreted. Moreover, there is a lack of connection between concepts and it does not help in modelling users for other contexts.

2.3.2 Stereotypes

Stereotype modelling [21] attempts to cluster all possible users of a system into different groups, namely stereotypes. Each user that belongs to the same stereotype is treated like the rest of the members of the group so his or her individual features are not considered. Typically, the data used in the classification is a demographic that users have to provide, for example in a registration form.

The main goals of this modelling approach are to define the stereotypes of a system and to implement the trigger techniques that provide mapping from a specific user to one stereotype. These include different clustering analyses, machine-learning techniques and reasoning among others [22]. There is an obvious disadvantage of this approach and it lies in the limited personalization and individualization of users, besides the difficulty in recovering new user models from the existing ones.

2.3.3 Classifier based models

Classifier systems [23] use information about items or the domain together with user data as an input to generate a custom response to the user. These can be implemented using different machine learning methods and the user model is represented as the particular model structure of the used classifier. For example, there can be user models based on decision trees, association rules, or Bayesian Networks. This approach, like the previous ones, has difficulties in retrieving and sharing user models since it is very limited and is based on solving specific tasks.

2.3.4 Semantic user modelling

Semantic technologies have appeared as a way to solve communication problems, and interoperability issues among systems, and to provide and facilitate reusability, reliability, and a common specification [12]. Semantic user modelling [20] is based on using ontologies that model a user or a specific domain using a rich network where terms are connected by different kinds of links that indicate its relations [24].

Using ontologies solves the polysemy problem and facilitates to retrieve and share user models between entities. There are different languages and techniques that allow the extraction of data from ontologies. For example, the SPARQL Protocol and Resource Description Framework (RDF) Query Language (SPARQL) and the accompanying protocols [25] make possible to send queries and receive results from semantic data (expressed as RDF information), e.g., through HTTP. Moreover, new relations between concepts and thus, about user features, can be inferred from ontology representation. Particularly, reasoner engines [16] are software components that allow autonomously the discovery of new knowledge from ontologies. Generally, they employ their own rules, axioms and appropriate chaining methods. We can find stand-alone reasoners, such as Pellet[5], or reasoners included in different semantic frameworks as for example, Protégé[6] and Jena[7].

---

[5] http://clarkparsia.com/pellet
[6] http://protege.stanford.edu/
[7] https://jena.apache.org/

2.4 Identity Management and Privacy

Identity management commonly refers to the processes involved in the management and selective disclosure of personal data, either within an institution or between several entities, while preserving and enforcing both privacy and security requirements. There are different approaches to implementing identity management, mainly: network-centric and user-centric approaches [26].

Network-centric approaches are based on agreements between service providers that establish trust relationships. Each service provider maintains its own personal data but users can link (federate) isolated accounts that they own across different providers to be recognized within the federated domain. Technological standards for identity federation include the OASIS Security Assertion Markup Language (SAML) [27] and the Kantara Initiative[8].

On the other hand, user-centric approaches highlight user empowerment in the governing of their personal information. Generally, there is a third entity that is in charge of providing user identity to service providers and the user is in the center of the transactions, managing the sharing of personal data. Examples of this approach are [28]: OpenID, OAuth 2.0, and OpenID Connect. Most of the social-based APIs for personal information sharing rely on OAuth 2.0, as for example the Facebook Login API[9]. It introduces a third role to the traditional client-server authentication/authorization model: the resource owner. Following this model, the client (who is not the resource owner, but is acting on his behalf) requests access to resources controlled by the resource owner, but hosted by a container i.e. the online social network. OAuth 2.0 allows the service provider to verify the identity of the client making the request, as well as ensuring that the resource owner has authorized the transaction without revealing their credentials.

Identity management technologies also contribute to privacy management by allowing users to decide on the sharing process. However, this is not enough, as any system managing personal information must abide by the privacy and data protection legal framework in place, and thus fulfill a set of requirements derived from the legal principles. For example, in Europe the main principles include lawfulness collection and processing; gathering specific, informed and explicit consent from data subjects; purpose binding; necessity and data minimization; transparency and openness; rights of the individual; and, security safeguards [29].

The state of the art includes a plethora of technological solutions, each addressing a specific privacy concern, and globally referred to as Privacy Enhancing Technologies (PETs) [29]. However, adding PETs on top of an existing system does not solve all privacy requirements, and thus there is a general consensus on the need to introduce Privacy by Design (PbD) approaches when developing systems i.e. considering privacy issues from the onset of a project and through its entire lifecycle [30].

All the aforementioned technologies facilitate the access and management of personal data. However, user-centric solutions allow users to control and manage their personal data directly, bringing a better user-experience.

3 FRAMEWORK ARCHITECTURE

As described in the previous section, there are many solutions and specific technologies to handle the design and implementation of the PeDF. We have proposed a comprehensive architecture for the PeDF that considers different approaches for personal data collection, integration, retrieval, and identity and privacy management, regardless of the specific technologies and implementations. Figure 1 represents this PeDF architecture where we can distinguish its modules, and its relationships with different external data sources, data consumers, and the user.
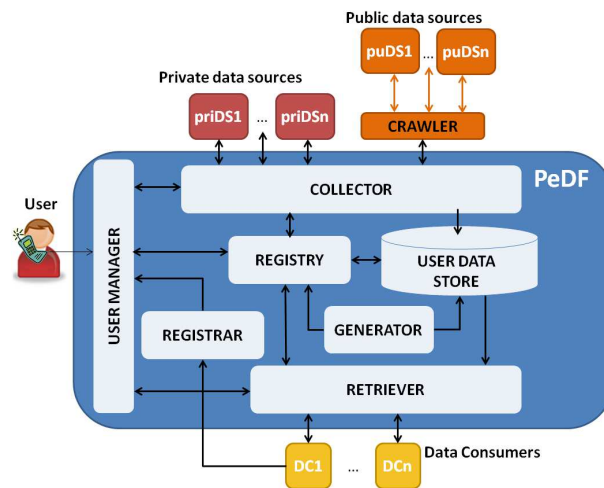

Figure 1. Personal Data Framework architecture

Firstly, we have considered that there are diverse existing data sources (private or public), and crawlers on the Internet that can be linked with the PeDF to gain access to user personal data. This data source-user association can be carried out by the user through the User Manager module, or by data consumers via the Registrar module but the latter requires user consent.

Once the data sources are linked, the Collector module is in charge of obtaining personal data from them and these data have to be integrated. We have proposed two complementary approaches to carry out this integration. One is based on collecting and storing personal data, which requires a User Data Store module. The other method is based on indexing personal data, which entails a Registry module that identifies which personal data can be accessed and where they are stored.

Moreover, we have provided the PeDF with the ability to supply personal data and user models to data consumers through a Retriever module. The creation of user models entails the incorporation of different components that extract knowledge from personal data. These components have been grouped together in a main component namely Generator.

Summarizing, the PeDF incorporates seven modules:

1. User Manager. It is a vertical module that allows users to interact with PeDF to sign in, activate the incorporation of new data sources, and check and manage authorizations for access to their personal data and user accounts. It implements an identity management infrastructure and privacy solutions.

2. Registrar. This module allows data consumers to ask for the incorporation of new data sources in order to include new

---

personal data in the PeDF. It interacts with the User Manager module to obtain the user consent.

3. Collector. This module is in charge of obtaining personal data from external data sources, checking user authorization. It can also include crawlers' components that get personal data from public data sources.

4. Registry. It allows the PeDF to store pointers to external personal data that the PeDF is able to recover from data sources.

5. Generator. It comprises a set of components that allow PeDF to obtain user models from personal data. These implement different techniques of user modelling to uncover user needs, preferences, interests, etc.

6. User Data Store. It is a central repository that stores the personal data that is obtained from external data sources or by the Generator module. It contains different interfaces that allow the updating and refreshing of personal data.

7. Retriever. This module is in charge of communicating with data consumers who are interested in obtaining personal data and user models of a specific user. It interacts with the User Manager module to check user consent and with the Registry or User Data Store to retrieve the personal data requested.

4 FRAMEWORK VALIDATION

We have validated the PeDF in a banking scenario which considers a person-to-person payment service namely PosdataP2P, and the social network Facebook as data sources. Moreover, it includes a financial service called FriendLoans that uses user models from the PeDF to offer its users recommendations about microloans. It is an integration effort to provide user models that fulfill individual business needs of third entities. We have focused our work on a centralized integration based on semantic technologies, which improve the user modelling process. Moreover, we have validated the PeDF with five beta testers from our research group.

Figure 2 represents our validation to the PeDF. Here, we can observe the two private data sources (PosdataP2P and Facebook), the data consumer (MicroLoans), the user and the main PeDF modules that we have validated: User Manager, Collector, User Data Store, Generators, and Retriever.
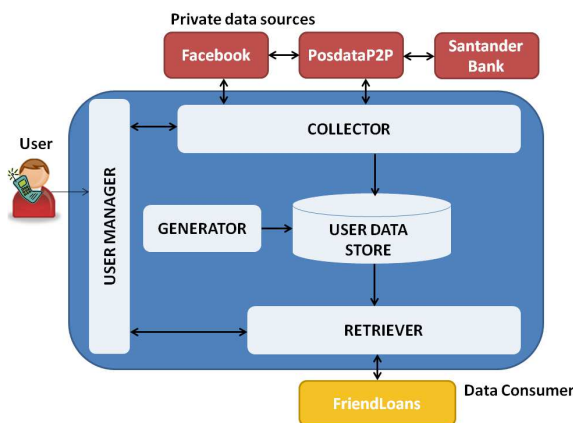

Figure 2. Personal Data Framework validation architecture

4.1 External data sources

We have considered two private data sources for PeDF validation: PosdataP2P service, and the social network Facebook.

PosdataP2P service [17] is an innovative financial service developed within the context of a COM project. It allows Santander University Smart Card (USC) holders to make payments to or request money from friends, using alternative social channels such as texting systems e.g. Telegram, or online social networks e.g. Facebook or Twitter.

The USC is a smart card issued by over 300 universities in collaboration with Santander Bank. It is used by 7.8 million people worldwide to access university services, such as libraries, control access (for example, to computers, campus, sports pavilions, etc.), electronic signature, discounts at retailers, etc. It can be also used to gain access to Santander Bank financial services, working as a credit/debit card linked to the holder's saving account.

To use PosdataP2P service, USC holders have to activate the service first, providing their USC information. Then, they choose the social channels that they want to use to carry out financial transactions. Having done that, students can start making financial transactions by simply posting messages to their friends within their enabled social channels (Figure 3).

Figure 3. PosdataP2P screenshot using Facebook as a channel

The PosdataP2P service generates financial data on USC holders, which is properly recovered by the PeDF in real time. Specifically, the PosdataP2P has an interface to notify financial transaction to PeDF.

The PeDF also obtains demographic and social data from Facebook with user consent. It is based on the Facebook Login and the Facebook Graph API as mentioned in Section 2.

4.2 A Personal Socio-Economic Network

The PeDF validation applies a centralized approach where personal data obtained from external data sources are stored in a central repository. Specifically, it is based on a semantic modelling and storing, and an ontology, namely the Personal Socio-Economic Network (PSEN).

The PSEN represents the exchange of money between people and user social data. We have considered the reusing of existing ontologies, which is a must to allow semantic and syntactic

interoperability. Thus, we have identified the FOAF ontology as the best alternative for representing people in a social network context and the SUMO's financial ontology (using the OWL version) for representing the financial concepts. We have also extended them and linked the different socio-economic concepts. The nomenclature that we have used to represent the PSEN concepts is based on SUMO terms so it can be easily related to the upper ontology.

Briefly, the PSEN includes the main terms to describe people, the relationships between them, and the financial data and activities carried out between them (Figure 4). We represent people as the *Person* class from FOAF and we use the corresponding FOAF properties to describe their user's demographic information: *firstName, lastName, gender, age, birthday,* and *mbox* (omitted in Figure 4 for the sake of simplicity). We also made use of the *Online Account* class from FOAF that allows the modelling of different web identities or online accounts of a person. We have extended it to include online payment and banking accounts. The former is devoted to service providers that allow users to carry out payment operations through the Internet, such as PosdataP2P service. It has associated a *BankCard* or a *Financial Account* class from the SUMO financial ontology that denotes where the payment will become effective. These classes have a relationship (namely, *cardAccount*) since a *BankCard* is always associated with a *FinancialAccount.* On the other hand, the *Online Banking Account* class represents online banking services including financial institutions, such as Santander Bank.

To model user economic activities, we have defined a *SocialInteraction* class within the PSEN ontology. It includes three main properties: *timestamp, channel* and *patient.* The *timestamp* and *channel* properties indicate when and where the social interaction happens respectively, and *patient* designates an *Entity* that participates in the social interaction, i.e. the money exchange. The *SocialInteraction* class also has two subclasses: *Transaction* and *Communication* that have *Payment* and *Request* subclasses correspondingly. These are related to a *hasPayment* link that indicates whether a request for money has been paid.
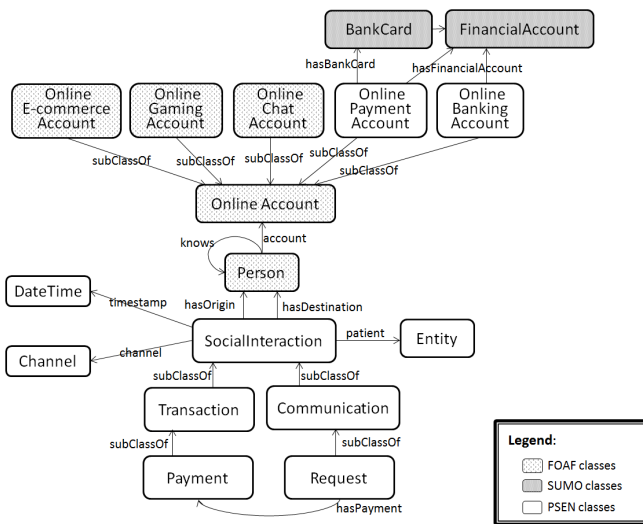

Figure 4. Personal Socio-Economic Network definition

In Figure 4, the rounded rectangles characterize the main concepts and the edges indicate the relationships between two classes. We have distinguished the terms of the different ontologies with darker rectangles indicated in the legend of the figure.

4.3 Knowledge retrieval

We have validated the retrieval of user knowledge through the FriendLoans service, which is based on friendsourcing [31]. It is a form of crowdsourcing where the user's social network is mobilized to achieve a specific objective. Specifically, FriendLoans relies on the PSEN data to offer financial recommendations on microloans to raise money from friends. It has been implemented as a web application in which authenticated users can ask for money from their friends. Basically, a user accesses to the service, indicates the money needed (Figure 5 at the top) and the service provides a list of prospective borrowers who are trusty, available, and solvent enough to lend (Figure 5 at the bottom). Figure 5 shows an example of the FriendLoans service for a user called Maria who needs 200€ from her friends
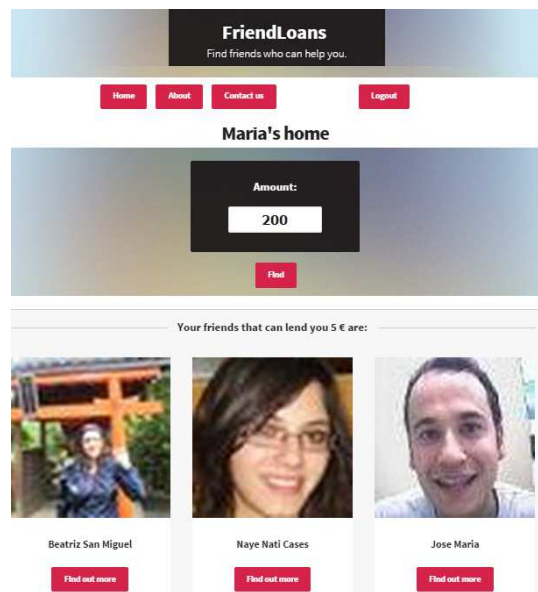

Figure 5. Screenshot of FriendLoans for a user called Maria

Generating a list of friends for a user requires user models that are unknown to FriendLoans, but can be retrieved from the PeDF. The PeDF has incorporated two mechanisms that allow data consumers to ask for user financial relationships and other banking information, all with the consent of the user. Specifically, the PeDF abstracts a set of SPARQL sentences and calls the reasoners which obtain and derive additional knowledge from the PSEN.

The SPARQL sentences obtain personal data and user models directly from the PSEN which can be used by FriendLoans. This information does not derive facts or inferences under the PSEN data, just data contained in it. For example, the list of friends for a specific user, if a person has carried out payments or requests for money, if a person has received money, if a person has requests for money and no associated payments, etc.

As regards the reasoners, they include the mechanisms that allow the extration of derived data. For this, we have implemented four custom rules that detect: 1) whether a user knows another user *A*; 2) whether a user owes money to a user *A*; 3) whether a user has

received a payment greater than X euros; and 4) whether a user has requests for money with greater amount of money than Y euros. In the rules, the user A and the amount of money X and Y can be indicated by FriendLoans to give recommendations to its users. In this way, for the example shown in Figure 5, A will be the authenticated user Maria who needs money from her friends, X and Y could be at least 200€ or the amount wanted by FriendLoans. The results obtained from executing these rules are a set of users that fulfill all conditions. This set is not ordered since the order of execution of the rules is not predictable in the reasoner. However, the PeDF has implemented an algorithm that orders the results including tags that indicate the prioritization.

The next program listing shows an example of a rule that tags the results as the most important ones (it is indicated by the tag *isFirstFor*) for the user Maria (specified by the second line of the rule). The conditions of the rule are: 1) a user who has debts with Maria (defined in a function called *hasDebtWith*), and 2) a user has not requested an amount of money greater than 5€ with other people (defined in a function called *possibleProblem*).

```
[isFirst:
(?Maria psen:isTarget "true"^^xs:Boolean)
(?person psen:hasDebtWith psen:Maria)
noValue(?ecAct          psen:possibleProblem
"true"^^xs:Boolean)
-> (?person psen:isFirstFor ?Maria)]
```

4.4 Identity management and privacy

We have based our identity management infrastructure on OAuth 2.0, as it has become the de facto standard to gain access to personal data on the Web. The User Manager includes the component that manages the interaction with external sites.

Users can currently link their accounts on the PosdataP2P service and Facebook to the PeDF. The process works as follows: when a user activates a data source (i.e. Facebook), he is then redirected to the service provider site to grant the PeDF the required level of authorization. If successful, the data source delivers a token that allows access to the user profile.

As regards privacy, the PeDF has been designed to observe European privacy and data protection principles following a privacy-by-design approach. The User Manager is also the key component here, since it provides users with an identity and privacy dashboard allowing them to 1) grant/revoke consent to the collection, processing and disclosure of their personal data, 2) check the PeDF privacy policies, 3) manage the personal data known and stored by the PeDF, their sources, and the details on the disclosures to third parties as well as exercising their right to access, rectify, erase or block personal data. At the same time, the User Data Store implements security safeguards to avoid and mitigate privacy threats derived from malicious attackers or unwitting users. Finally, as regards the data minimization principle, the use of reasoners allows third parties to be limited and allows justified users to be able to query and retrieve that specified and agreed to by the data subject.

5 RELATED WORK

The PeDF is an ambitious solution that covers four main technological challenges related to personal data: collection, integration, retrieval, and identity and privacy management. These have been widely analyzed separately over time in different contexts, and we can find many researchers addressing each of them in depth. For example, the previously cited literature [10] includes a study into data integration in business environments, or [32] presents the user modelling techniques, its challenges and the state-of-the-art research, focusing on ubiquitous environments.

We can find aligned systems that attempt to solve the same issues as the PeDF in the personal data context. For example, the so-called data brokers [33] are companies that collect personal data on individual (generally, from public data sources), and resell them to or share them with third parties. These systems are focused on data collection and integration, but individuals are generally unaware of their activities. Otherwise, there are a number of companies and projects within the initiative called Personal Cloud[10]. It advocates the creation of safe places where users have complete control of their data. The associated solutions address the definition of a new interaction model between users, service providers, and devices, where clouds connect voluntarily to services which use stored personal data. They focus on identity management, encryption, data storage, cloud computing, as well as other user modelling works related to reputation. Closely related to these, there are different identity management systems [34] that implement end-user solutions with the goal of making personal data available only to the right parties, establishing trust between parties involved, avoiding the abuse of personal data, and making these provisions possible in a scalable, usable, and cost-effective manner. These latter solutions do not generally include user modelling techniques.

On the other hand, there are also specialized systems, namely Generic User Modelling Systems [35] that can serve as a separate user modelling component to different service providers. They address issues related to data representation, inferential capabilities, management of distributed information, or privacy. However, they focus on the reuse of technological user modelling components rather on the reuse of the personal data and user models themselves. Finally, there are solutions referred as Personal Data Store, Personal Data Locker, or Personal Data Vault that roughly describe the same concept. Generally, these solutions are based on a central place where the user can save and manage all their personal data, including data such as text, passwords, images, video or music [36]. These solutions have an end-user approach.

To summarize, the aforementioned solutions are rather diverse from one another, and each of them focuses on a main objective (i.e., personal data collection, identity management, and data storage). Our work is an integration effort to provide an end-to-end solution that aims at incorporating the best solutions for each issue. Our first approach is based on integrating social and financial data. To the best of our knowledge, this is the first effort in this context.

6 CONCLUSIONS AND FUTURE WORK

In this paper we have presented a comprehensive framework intermediating between users and organizations to support the seamless integration of personal data from several, distributed sources and generating advanced knowledge on users, to be shared with interested third parties, all supervised by the users who control and manage the flow of their personal data. The framework includes components for personal data collection, integration, and retrieval, as well as users' identity and privacy management.

The framework has been validated in a financial context, integrating social information from Facebook and a person-to-person payment service, to generate knowledge useful for a personal lending application.

Our future work includes advancing on the design of the privacy-preserving elements required to minimize the personal information retrieved by the data consumers while keeping it useful enough to fit their business needs. These developments will comprise advanced privacy enhancing technologies for attribute-based credentials and database privacy.

ACKNOWLEDGEMENTS

REFERENCES

[1] I. Barri, T. Loilier, M. van Rijn, A. Stolk, and H. Vasiliadis, 'Open innovation in the financial services sector - Why and how to take action', Technical report, GFT Technologies AG, (2014).

[2] J. P., Moreno, Harvard Business Publishing, *Banks' New Competitors: Starbucks, Google, and Alibaba.* https://hbr.org/2014/02/banks-new-competitors-starbucks-google-and-alibaba/

[3] Open Data Institute and Fingleton Associates, 'Data Sharing and Open Data for Banks', Technical report, (2014).

[4] European Commission, *Protection of personal data.* http://ec.europa.eu/justice/data-protection/

[5] R. T. Fielding, *Architectural Styles and the Design of Network-based Software Architectures,* Ph.D. dissertation, University of California, 2000.

[6] Internet Engineering Task Force (IETF), 'The JavaScript Object Notation (JSON) Data Interchange Format', Proposed Standard RFC 7159, (2014).

[7] W3C, *OpenSocial Foundation Moves Standards Work to W3C Social Web Activity*, http://www.w3.org/blog/2014/12/opensocial-foundation-moves-standards-work-to-w3c-social-web-activity/

[8] G. Gouriten and P. Senellart, 'API Blender: A Uniform Interface to Social Platform APIs', *CoRR,* abs/1301.2086, (2013).

[9] M. Lenzerini, 'Data integration: A theoretical perspective', in *Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '02, pp. 233–246, New York, NY, USA, (2002). ACM.

[10] P. Ziegler and K. R. Dittrich, 'Three decades of data integration - all problems solved?', in *18th IFIP World Computer Congress (WCC 2004)*, Volume 12, Building the Information Society, pp. 3–12, (2004).

[11] N. Guarino, 'Formal ontology and information systems', in *FOIS98*, pp. 3–15, Trento, Italy, (1998). IOS Press.

[12] M. Uschold and M. Gruninger, 'Ontologies: Principles, methods and applications', *Knowledge Engineering Review*, 11(2), 93–136, (1996).

[13] I. Niles and A. Pease, 'Towards a standard upper ontology', in *Proceedings of the International Conference on Formal Ontology in Information Systems, FOIS01*, pp. 2–9, New York, NY, USA, (2001). ACM.

[14] D. Brickley and L. Milller, 'Foaf vocabulary specification 0.99', Namespace Document - Paddington Edition, (2014).

[15] Object Management Group, *Financial Services Standards.* http://www.omg.org/hot-topics/finance.htm

[16] L. Yu, *A Developer's Guide to the Semantic Web*, Springer, 2011.

[17] B. San Miguel, J. M. del Alamo, J. C. Yelmo, 'Creating and Modelling Personal Socio-Economic Networks in On-Line Banking' in *7th International Workshop on Personalization and Context-Awareness in Cloud and Service Computing, PCS 2014,* pp. 177–190, (2015) Springer [In press].

[18] World Wide Web Consortium (W3C), 'OWL Web Ontology Language', W3C Recommendation, (2004).

[19] N. P. de Koch, *Software engineering for adaptive hypermedia systems: reference model, modeling techniques and development process*, Ph.D. dissertation, Ludwig Maximilians University Munich, 2001.

[20] S. Gauch, M. Speretta, A. Chandramouli and A. Micarelli, 'User profiles for personalized information access', in *The Adaptive Web*, eds., P. Brusilovsky, A. Kobsa, and W. Nejdl, Springer-Verlag, (2007).

[21] P. Brusilovsky, and E. Millán, 'User Models for Adaptive Hypermedia and Adaptive Educational Systems', in *The Adaptive Web*, eds., P. Brusilovsky, A. Kobsa, and W. Nejdl, Springer-Verlag, (2007).

[22] J. Kay, 'Lies, damned lies and stereotypes: Pragmatic approximations of users', in *Proceedings of the Fourth International Conference on User Modeling*, pp. 175–184, Hyannis, MA, (1994). ACM.

[23] M. Montaner, B. López, and J. L. de la Rosa, 'A Taxonomy of Recommender Agents on the Internet', *Artificial Intelligence Review*, 19(4), 285–330, (2003).

[24] S. Sosnovsky, and D. Dicheva, 'Ontological technologies for user modelling', *International Journal of Metadata, Semantics and Ontologies*, 5(1), 32–71, (2010).

[25] World Wide Web Consortium (W3C), *SPARQL Current Status.* http://www.w3.org/standards/techs/sparql#w3c_all

[26] J. M. del Alamo, M. A. Monjas, J. C. Yelmo, B. San Miguel, R. Trapero, and A. M. Fernandez, 'Self-service privacy: User-centric privacy for network-centric identity.', in *Trust Management IV. 4th IFIP WG 11.11 International Conference on Trust Management, IFIPTM 2010,* pp. 17–31, Morioka, Japan, (2010). Springer Berlin Heidelberg.

[27] OASIS, *OASIS Security Services (SAML) TC.* https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=security

[28] O. Manso, M. Christiansen, and G. Mikkelsen, 'Comparative analysis - Web-based identity management systems', Technical report, The Alexandra Institute, (2014).

[29] G. Danezis, J. Domingo-Ferrer, M. Hansen, J. H. Hoepman, D. L. Metayer, R. Tirtea, and S. Schiffner, 'Privacy and Data Protection by Design – from policy to engineering' Technical report, European Union Agency for Network and Information Security (ENISA), (2014).

[30] A. Crespo García, N. Notario McDonnell, C. Troncoso, D. Le Métayer, I. Kroener, D. Wright, J. M. del Álamo and Y. S. Martín, 'D1.2: Privacy and Security-by-design Methodology', Technical report, PRIPARE (2014).

[31] M. S. Bernstein, D. Tan, G. Smith, M. Czerwinski, and E. Horvitz, 'Personalization via friendsourcing', *ACM Transactions on Computer-Human Interaction (TOCHI)*, 17(2), 6:1–6:28, (2008).

[32] J. Kay T. Kuflik, and B. Kummerfeld, 'Challenges and solutions of ubiquitous user modeling', in *Ubiquitous Display Environments*, eds., A. Krüger and T. Kuflik, Springer Berlin Heidelberg, (2012).

[33] E. Ramirez, J. Brill, M. K. Ohlhausen, J. D. Wright, T. McSweeny, 'Data Brokers – A Call for Transparency and Accountability', Technical report, Federal Trade Commission, (2014).

[34] E.Bertino and K. Takahashi, *Identity Management: Concepts, Technologies, and Systems*, Artech House, Inc., 2010.

[35] A. Kobsa, 'Generic user modeling systems', *User Modeling and User-Adapted Interaction,* 11(1-2), 49–63, (2001).

[36] M. Sabadello, 'Startup Technology Report – Phase One: Acquiring, Storing, Accesing and Managing Personal Data', Technical report, Personal Data Ecosystem Consortium, (2014).