An architecture for privacy-enabled user profile portability on the Web of Data[*]

Benjamin Heitmann
Digital Enterprise Research Institute
NUI Galway
Galway, Ireland
benjamin.heitmann@deri.org

James G. Kim
Biomedical Knowledge Engineering Laboratory
Seoul National University
Seoul, Republic of Korea
jgkim@bike.snu.ac.kr

Alexandre Passant
Digital Enterprise Research Institute
NUI Galway
Galway, Ireland
alexandre.passant@deri.org

Conor Hayes
Digital Enterprise Research Institute
NUI Galway
Galway, Ireland
conor.hayes@deri.org

Hong-Gee Kim
Biomedical Knowledge Engineering Laboratory
Seoul National University
Seoul, Republic of Korea
hgkim@snu.ac.kr

ABSTRACT

Providing relevant recommendations requires access to user profile data. Current social networking ecosystems allow third party services to request user authorisation for accessing profile data, thus enabling cross-domain recommendation. However these ecosystems create user lock-in and social networking data silos, as the profile data is neither portable nor interoperable. We argue that innovations in reconciling heterogeneous data sources must be also be matched by innovations in architecture design and recommender methodology. We present and qualitatively evaluate an architecture for privacy-enabled user profile portability, which is based on technologies from the emerging Web of Data (FOAF, WebIDs and the Web Access Control vocabulary). The proposed architecture enables the creation of a universal "private by default" ecosystem with interoperability of user profile data. The privacy of the user is protected by allowing multiple data providers to host their part of the user profile. This provides an incentive for more users to make profile data from different domains available for recommendations.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering; H.3.4 [**Systems and Software**]: Distributed system

General Terms

Recommender Systems, Data Integration, Privacy, Architecture, Web of Data, Linked Data, Social Web

Keywords

WebID, Web Access Control, user profiles, FOAF, RDF

1. INTRODUCTION

Personalised recommendations have proven themselves to greatly enhance the user experience of searching, exploring and finding new and interesting content [8] on social websites like Facebook[1] and Last.fm[2]. However, in order to provide an attractive and successful recommendation service, appropriate data and knowledge is required, depending on the domain of the service and the algorithm used [2].

While most recommender systems collect profile data from their own users, an alternative approach is to share user profile data among an ecosystem of sites, thus enabling cross-domain personalisation. In such an ecosystem one site typically has the role of the *hub site*, which provides the main entry point for the whole ecosystem and stores the user profile data. Prominent social networking sites like Facebook and Twitter[3] are such central hub sites. *Third party services* can provide value-added and personalised services for the user of an ecosystem. Examples include TweetMeme[4] which shows the most popular links on Twitter, and the Flickr[5] integration for Facebook which posts pictures uploaded on Flickr to the user's Facebook activity stream. *Users* stay in control of their profile data, as their profile is stored on the central hub site and the user can specify which services can access their profile data. If a service e.g. spams the user with messages then the user can revoke access for the service.

While the creation of such ecosystems provides powerful incentives for users to allow the sharing of their profile data

[*]The work presented in this paper has been funded in part by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Líon-2) and by Korean Small and Medium Business Administration under Grant No. 0420-2009-0061.

HetRec '10, September 26, 2010, Barcelona, Spain
Copyright 2010 ACM 978-1-4503-0407-8/10/09 ...$10.00.

[1] `http://www.facebook.com`
[2] `http://www.last.fm`
[3] `http://www.twitter.com`
[4] `http://tweetmeme.com`
[5] `http://flickr.com`

between different services, it also leads to user lock-in and social networking data silos: User profiles are not portable between systems, connecting to users from a different system is not possible and the user can not evade changes to the terms of service. In this paper we propose an alternative: Instead of creating ecosystems around closed networking silos, we propose to create ecosystems around portable user profiles. These user profiles can be moved between social services or they can be hosted by the user themselves.

We present an architecture which describes how to combine existing infrastructure of the Web of Data and existing standards for decentralised identity management in order to achieve privacy-enabled user profile portability. Building on work by Hollenbach, Presbrey and Berners-Lee [14], our architecture describes how to combine Linked Data, WebIDs and the Web Access Control (WAC) vocabulary: Linked Data [5], and the Friend-of-a-Friend (FOAF) and Semantically Interlinked Online Communities (SIOC) vocabularies allow the description of domain independent user profiles [6]. WebIDs [19] securely connect a user identity to the information in a user profile and can be used for authenticating a user. The WAC vocabulary [14] allows the user to authorise third party services for accessing different parts of his profile information.

This architecture allows users to benefit from the privacy that is provided by centralised and closed social networking ecosystems as well as from the portability that is provided by the decentralised and open Web of Data. User profiles and activity stream data can be securely shared with any third party that supports the architecture. User profiles can be hosted by social networking sites or they can be self hosted by the user. There is no lock-in to any specific social networking site or ecosystem. We provide a qualitative evaluation of the presented architecture based on the evaluation framework for privacy-enhanced personalisation suggested by Wang and Kobsa [20]. In addition we describe how the architecture applies to a use case from the e-Health domain.

The contributions of this paper are: (i) an architecture for privacy-enabled user profile portability, (ii) a list of the requirements for privacy-protected sharing of profile data for the purpose of data mining and recommendations, (iii) a use case which shows how to apply the described architecture to the e-Health domain.

The rest of the paper is structured as follows: Section 2 introduces the emerging Web of Data and the challenges for recommender systems in acquiring user profile data. Section 3 describes related work in identity management, distributed social networks and privacy-enhanced personalisation. Section 4 lists the requirements for our architecture based on an existing evaluation framework for privacy-enhanced technologies. Section 5 describes our architecture and the roles and communication pattern of it's participants. We also describe its application in an e-Health use case. Finally sections 6 and 7 provide a discussion of our work and a conclusion to the paper.

2. BACKGROUND

In order to provide an attractive and successful recommendation service, sufficient data and knowledge is required. While the Web of Data can provide sources of data and knowledge for recommendation services, it currently does not provide the means for creating an ecosystem around privacy-enhanced and portable user profiles.

In this section we first explain why acquiring data and knowledge for a recommender system is a challenge. Then we introduce the Web of Data as a source of public data and knowledge. Then we show that users have the expectation of privacy, when it comes to making their profile available on the Web. As the Web of Data currently does not provide the means for users to control access to their profile data, an important incentive for users in sharing their profile data is missing.

2.1 The challenge of acquiring data for recommendations

Recommender systems require three components to provide recommendations [8]: (1) background data, which is the information the system has before the recommendation process begins, (2) input data, which is the information provided about the user in order to make a recommendation, and (3) the recommendation algorithm which operates on background and input data in order to provide recommendations for a user.

In order to provide relevant recommendations appropriate background data is required, depending on the domain of the service and the algorithm used [2]. The high entry barriers of providing good recommendations are caused by the problem of acquiring data and knowledge to provide the background data for the recommendation algorithm.

The *data acquisition problem* [2] is characterised by three complementary challenges: (a) The new item problem: To provide good recommendations for any item, the recommendation algorithm needs information about the item. If a new item has been added, then no information about user preferences has been collected for the item. This makes it challenging to provide collaborative recommendations for new items. (b) The new user problem: In order to personalise the recommendation, the recommendation algorithm needs a user profile. For collaborative recommendations new users are a challenge because the user has no profile of preferences connecting him to items. Together, the new-item and new-user problems are known as the ramp-up or cold-start problem. (c) The sparsity problem: If the number of ratings is low compared to the number of items in the background data then it will be hard to match other users or item profiles, and this will lead to ineffective recommendations.

In addition, for knowledge intensive recommendation approaches, knowledge bases fitting the recommendation scenario are required. This adds the *knowledge acquisition problem* [8], which is characterised by the high effort of knowledge engineering: In order to provide recommendations in knowledge intensive recommendation scenarios, a knowledge base about the recommendation domain and the users needs to be acquired.

2.2 The Web of Data as a data source for recommender services

The term Linked Data refers to a set of best practices for publishing and connecting structured data on the Web [5]. Taken together, all linked data constitutes the Web of Data. While the World Wide Web provides the means for creating a web of human readable documents, the Web of Data aims to create a web of structured, machine-readable data.

In order to acquire the necessary data and knowledge for recommender systems, external data sources can be used. The Web of Data can provide access to such external data

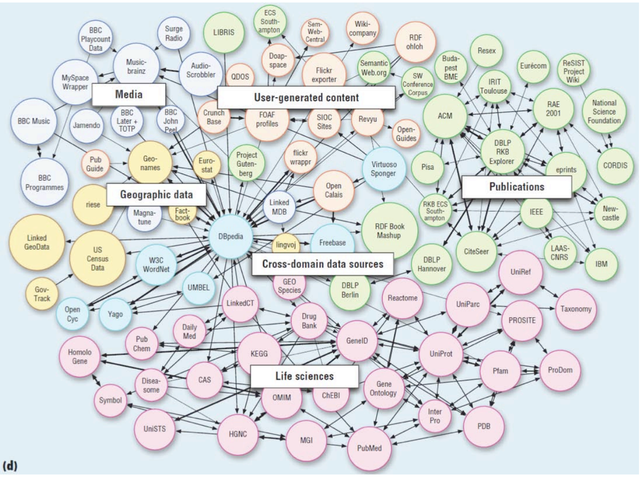

Figure 1: Overview of Linking Open Data sources as of July 2009 with source types shown

sources [13]. However, the Web of Data does not provide the means for creating an ecosystem around privacy-enhanced and portable user profiles. In other words the Web of Data is "public by default", whereas many users have come to expect the privacy of a "private by default" medium. Because of this, an important incentive for users to share their profile data via the Web of Data is currently missing.

The Web of Data utilises technologies from the Semantic Web technology stack: the Resource Description Framework (RDF) provides a graph based data model and the basic, domain independent formal semantics for the data model [11]; the SPARQL Query Language allows querying RDF data with graph patterns, and provides basic means for transforming RDF data between different schemata. In addition, technologies from the World Wide Web provide the fundamental infrastructure: Uniform Resource Identifiers (URIs) are used to provide globally unique identifiers for the data, and the HyperText Transfer Protocol (HTTP) is used for accessing and transporting the data. RDF Schema and OWL allow the definition of vocabularies, taxonomies and ontologies, which provide the basis for shared domain semantics between applications.

In order to build a single Web of Data, all data providers have to follow the same guidelines for publishing their data and connecting it to other data sources. These guidelines are provided by the Linked Data principles [5], which specify how to use the different standards of the Web of Data together:

1. Use URIs as names for things (and e.g. persons, places).

2. Use HTTP URIs so that people can look up and access those names via HTTP.

3. When someone looks up a URI, provide useful information, using the standards RDF and SPARQL.

4. Include links to other URIs, so that data about more things can be discovered.

The Linked Data principles have been adopted by an increasing number of data providers, especially from the Linking Open Data community project[6], which makes free and public data available as linked data. The nucleus of the linked data cloud is formed by DBpedia[7], which extracts RDF from wikipedia topic pages, and thus provides URIs and RDF data about topics from any domain.

Social websites provide a big contribution to the linked data cloud, by making information about their users available. This data is modelled after the principle of object centred sociality [6]: persons are not only directly connected to other persons, but also indirectly via objects of a social focus. In this way a community is connected to each other not only via direct links from person to person, but also via their links to e.g. music from an artist. Such data uses the Friend of a Friend (FOAF) vocabulary for describing users and their connections to interests and other users, and the Semantically-Interlinked Online Communities (SIOC) vocabulary for describing user generated content on forums, weblogs and Web 2.0 sites, as described in [6].

FOAF and SIOC provide the means for putting user profiles and data about object centred sociality on the Web of Data. However this data is usually "public by default", as the Web of Data does not currently allow the user to specify which services can access which parts of a user profile.

However, as the recent Facebook privacy backlash [12] has shown, users are not comfortable with the assumption that all of their profile data is publicly accessible. Facebook launched in 2004 with a strong "private by default" policy, however it tried to move away from this in 2010 towards a "public by default" policy. This move caused a backlash with many high profile users cancelling their account. This led to Facebook reverting their policy and now suggesting more private settings again. See [7] for more background.

In order to provide incentives for users to share their profile with different recommendation services on the Web of Data, it is necessary to provide the means for controlling the access to the profile data. This will allow users to move towards a "private by default" policy, ultimately leading to more user profiles being available to recommendation services.

3. RELATED WORK

Privacy and personalisation are currently at odds [20]. In order to provide a personalised experience it is necessary for a website to have access to data about that same user. However users can feel reluctant in sharing their personal data with a website, as they fear their data can be misused or traded with unknown entities.

Different approaches to manage the identity and the profile data of the user have been suggested both inside of the recommender systems community as well as from the industrial Web standards community in general. In this section we will first provide a short overview of privacy-enhanced personalisation approaches. Then we will introduce the most prominent current standards for identity management and decentralised social networking.

3.1 Privacy-enhanced personalisation

According to Wang and Kobsa [20] the existing approaches for enhancing personalisation to enable user privacy fall into these categories:

[6] `http://preview.tinyurl.com/LOD-community`

[7] `http://dbpedia.org`

Pseudonymous personalisation allows users to remain anonymous towards a personalised system, whilst enabling the system to still recognise the user in different sessions so that it can cater to the user personally. This also allows the user to keep different parts of his online activity on the same service apart (e.g. professional use and private use), as used by Arlein et. al [3].

Distributed personalisation, in which either the storage of the user data is distributed or the computation of the recommendation. This enables better privacy for the user, as each user controls the storage and the distribution of his own data. Miller et. al [18] propose a peer-to-peer algorithm called PocketLens. The algorithm first aggregates data from the direct neighbours in the P2P network and then generates an item-to-item similarity model based on this data. Then peers incrementally share the item-to-item model and use this to update their own model. This model then can be used to generate recommendations for a user.

Cryptography enhanced methods for personalisation treat the privacy-preserving computation of recommendations as secure multiparty computation problem, where users and different websites jointly conduct computations based on their private data without the need to trust each other. In order to achieve this the user data can be transmitted in an already encrypted state [9], it can contain randomised errors which cancelled out during the computation, or it can use obfuscation or aggregation to hide a single users preferences amongst the data of a group of users [4].

Of these privacy-enhanced personalisation approaches, the work of Arlein et. al [3] from 2000 is most similar to our contributions. Arlein presents an architecture for so-called "global customisation", which enables third party services (called "merchants") to collect data about a user and share it between sites using a "profile database". Users are represented in the architecture as personas which are stored on a "persona server". Every user can have multiple personas, and the privacy of the user is protected because the profile database only knows about personas without being able to link personas to real users.

While our architecture has similar goals in allowing the sharing of user profile data for recommendations while maintaining the privacy of the user at the same time, we use different technologies for this which were not available in 2000. By combining existing standards and infrastructure to achieve these goals our architecture can be easily integrated into the emerging Web of Data from the start.

3.2 Decentralised social networking standards

Outside of research on privacy enhanced personalisation, standards have been developed for identity management and decentralised social networking. Some of these standards such as OpenID are widely used by now, while other standards are missing industry adoption, depending on the maturity of the standard. We introduce the most prominent industry standards for authentication, authorisation and profile data exchange in the following.

OpenID[8] is a standard for decentralised authentication of a user. It provides a way to prove that an end user owns an identity URL without passing around the password of the user to a third party service. OpenID is completely decentralised meaning that anyone can choose to be a third party service ("consumer" in OpenID terminology) or hub site ("identity provider") without having to register or be approved by any central authority. Users can pick which hub site they wish to use and preserve their identity as they move between hub sites. As of December 2009, there are over 1 billion OpenID enabled accounts and approximately 9 million sites have integrated OpenID consumer support, such as Google and Yahoo.

OpenID provides the means to decouple identities from real users, thus enabling pseudonymous personalisation. However OpenID is not well suited for machine agents and it requires a large overhead in terms of the number of HTTP connections which are required to gain access to a secured resource [19].

OAuth[9] specifies a protocol for decentralised authorisation of resource access. It specifies how a user can authorise a third party service (called "client" in OAuth terminology) to access parts of his profile (his "resources") on a hub site ("resource owner"). Instead of using a user's password to access parts of his profile at a hub site, third party services obtain access tokens which are used to access the protected resources. In addition to removing the need for users to share their password, users can restrict access to a limited part of their data and they can limit access duration.

OAuth is used by two of the most popular current social websites (Twitter and Facebook) to authorise third party services for accessing data from user profiles. OAuth complements OpenID, as it provides delegation of authorisation on top of the authentication through OpenID identity URLs. However this thight integration also means that OAuth requires a multitude of HTTP connections, which will lead to scalability problems for decentralised authorisation on the Web of Data.

OpenID itself does not provide any mechanism to exchange profile information, although it is possible to link an OpenID identity to a profile data such as a vCARD document. However **OpenID attribute exchange**[10] provides a protocol accessing profile data and provides a data model for storing it. The OpenID attribute exchange standard has not reached industry wide adoption, as it specifies a very limited vocabulary for expressing a user profile and it does not allow easy extensions to this vocabulary.

4. REQUIREMENTS

In order to arrive at an architecture for privacy-enabled user profile portability for the purpose of making recommendations, we collected requirements for the architecture. The requirements are based on the evaluation framework for privacy enhancing technologies (PETs) by Wang and Kobsa [20, 15]. We first outline their list of general privacy principles and the main areas in which users have privacy concerns. Then we describe our non-functional requirements for the architecture which are informed by the emergence of the Web of Data.

4.1 Privacy principles

As part of their evaluation framework Wang and Kobsa identify [20] the main privacy principles which motivate the creation of privacy enhancing technologies. These privacy principles are grouped as follows:

Anonymity related principles from the security literature.

[8] `http://openid.net/specs/`

[9] `http://tools.ietf.org/html/draft-ietf-oauth-v2`

[10] `http://openid.net/srv/ax/1.0`

These include anonymity, pseudonymity, unobservability, unlinkability and deniability.

Privacy principles from privacy laws, regulations and recommendations. These principles have been identified from a review of 40 international privacy laws, and they include limiting the collection of data, specifying the explicit purpose of the collection, limiting the use of the data for specific functions of the service, and informing the user of onward transfer of the data to third part services, as well as asking the user for his consent.

Human-computer interaction for the purpose of enabling privacy. These include asking the user for his privacy preferences, allowing the user to negotiate by giving him multiple privacy choices and the usability of the service by e.g. not requiring installation of new infrastructure software from the user.

4.2 Privacy concerns

Wang and Kobsa also identify [15] the main privacy concerns which users have regarding the impact of technology on the previously identified privacy principles. The concerns of users about their privacy fall roughly into three areas:

The *protection of identity*: Users want to control who can identify them, or who can link their identities on the Web back to their official and legal identity. This corresponds to enabling and protecting the anonymity related principles.

Control over the user's data: In addition to controlling their identity, users value the ability to control who can access which parts of their profile data. For our purposes this not only includes mostly static information like name, gender and location, but also the highly dynamic activity stream of the user and all the multimedia resources associated with the user. This area is affected by enabling and protecting the privacy principles collected from laws, regulations and recommendations.

Human-computer interaction: Different aspects of enabling the protection of identity and the control of a user over his data depend on the user interface of a service. For instance, by law a user needs to be informed of the data which is collected of him, which is a task of the UI. In addition the UI has the task to community all of the possible privacy settings without e.g. hiding some of them in a complicated menu structure. This area is affected by enabling and protecting the privacy principles collected from laws and those principles collected from human-computer interaction research.

4.3 Non-functional requirements

In addition to the requirements imposed by the privacy concerns, we have identified non-functional requirements for our architecture. While these do not directly impact the privacy of the user, it is necessary to take them into account to make large scale adoption of the architecture possible.

Universality: The World Wide Web was perceived to be one universal space[11], where any resource can be connected to any other resource. However for user identities on the current generation of social websites this is not true, as one user can only be connected to other users from the same social network. Therefore our architecture should enable universality of user profiles.

Scalability: An architecture for portable user profiles should scale well for the number of users in total, as well as for the number of hub sites and third party services. Users might have one or multiple profiles. Hub sites might be created just for one user if the user decides to host his user profile himself, or they might contain data from millions of users. Third party services might access data from any number of hub sites or individually hosted user profiles.

Reuse of infrastructure: Deploying an architecture for portable user profiles should not depend on new backend infrastructure or on new client software on the side of the users. Ideally existing infrastructure from the World Wide Web such as Web servers using HTTP and URIs should be reused. In addition technologies from the emerging Web of Data can be extended for user profile portability.

5. ARCHITECTURE

In order to enable users to share their profiles with different ecosystems while maintaining their privacy at the same time, it is necessary to define an architecture for privacy-enhanced user profile portability. This architecture prescribes the standards as well as the roles and the communication pattern between the different participants. By implementing this architecture, all *individual* participants agree on the same technical principles, which in turn allows the architecture to guarantee the identified requirements on a *global* level.

In this section we first describe the Semantic Web standards and technologies which provide the foundation for our architecture. Then we describe the roles performed by participants of the architecture and introduce a use case grounding it in the e-Health domain. Based on the use case we describe the required communication pattern of the participants, followed by a qualitative evaluation against the identified requirements from the previous section.

5.1 Foundation standards

Hollenbach, Presbrey and Berners-Lee [14] suggest using FOAF, WebIDs and the Web Access Control (WAC) vocabulary to enable access control in collaborative environments on the Web of Data. This allows integrating with existing infrastructure thereby extending the Web of Data in a natural way.

The *FOAF vocabulary* allows the description of domain independent user profiles [6]. FOAF provides properties to describe all of the details which are usually contained in a social networking profile or on a personal homepage. In addition a FOAF profile provides a container for other information from different domains. For instance, this information could use the SIOC vocabulary to list the content which the user has generated on his blog, on his twitter stream and the comments on different forums.

WebIDs [19] securely connect a user identity to the information in a user profile and can be used for authenticating a user. A WebID consists of two parts: (1) an SSL certificate which contains a link to (2) the URI from which information about the user can be obtained. The data which is obtained from the URI is associated in return with the SSL certificate, as it lists the public key which is associated with the private key contained in the SSL certificate.

The *Web Access Control (WAC)* vocabulary [14] allows the user to authorise third party services for accessing different parts of his profile information. Each private resource is tied to an Access Control List (ACL) resource. The ACL resource can say which agents or groups of agents have ac-

[11] http://www.w3.org/DesignIssues/Axioms.html

cess rights to the resources it governs, so the content of the ACL resource can be considered as a whitelist (i.e., "*private by default*").

Users are granted full access to the ACL resources for their profile so users can *read*, *write*, and *control* their whitelists as well as the profiles themselves. In other words, users can update their profiles and they can also give third party services access to all or parts of their profile data.

5.2 Roles

The interplay between FOAF, WebIDs and the WAC vocabulary requires the participants to perform one of three roles: profile storage services, data consumers and user agents.

Profile storage services roughly correspond to the hub sites in current profile sharing ecosystems. They provide the storage for the user profile or parts of it, and they secure the access to the profile data by following the rules from the ACL about a profile. In addition they provide a user interface for changing and maintaining the ACLs from the WAC metadata by e.g. adding or removing read rights for data consumers. Profile storage services can be either self hosted by the user or they can be hosted by a social networking site.

Data consumers correspond to any type of third party service which is accessing user profile data in current ecosystems. Each consumer has its own WebID, which identifies the service every time it is accessing profile data from a profile storage service. This allows the storage to determine if the access is granted to the consumer.

User agents manage the different identities of a user. Each identity is represented by a WebID, which is used for authenticating the user towards profile storage service or data consuming services.

5.3 Use case: personal health records

In order to illustrate how the standards of FOAF, WebIDs and WAC vocabulary enable privacy preserving profile portability, we will ground the architecture in a use case from the e-Health domain.

The emergence of Personally Controlled Health Record (PCHR) platforms such as Google Health[12], Microsoft HealthVault[13], and Dossia[14] leads to "tectonic shifts in the health information economy" [17]. PCHR platforms enable patients to import and manage their health data, and third party services to securely access it and to provide added value. For instance, patients can get recommendations of clinical trials matching their condition, or they can exchange experiences with other patients having the same disease. However, while being central hub sites, these platforms also contribute to the data silo problem and the privacy of health data is required to be protected by law.

In order to meet these requirements, our proposed architecture can be applied to this domain, creating a Personal Health Application (PHA) ecosystem around portable health data without compromising privacy. As a result, third party services like TrialX[15] and PatientsLikeMe[16] can utilise patient (user) profiles with their consent, including their health data, to provide personalised information recommendations.

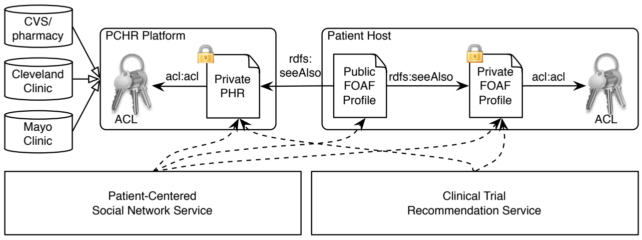

Figure 2: Architecture of the use case

In this use case, the patient profile is divided into three different resources (Figure 2): (a) *Public* FOAF profile, which is also used for WebID authentication and self hosted by the patient. This resource contains the public key of the certificate and pointers to the other resources, as well as the same kinds of public information social networking sites often provide, such as name, gender, and date of birth. (b) *Private* Personal Health Record (PHR) hosted by a PCHR platform. This contains the same kinds of information which would traditionally be found in a patients' medical record. This includes a patients' health conditions, medications, and laboratory results. The PHR is described by using the Health Level 7 Clinical Document Architecture (CDA), which is a widely used standard for exchanging patient health data. CDA specifies the semantics of PHR as well as the structure, so it can be converted into RDF triples and semantically queried [16]. (c) *Private* FOAF profile, which contains a list of friends who share the same or a similar disease with the patient. This patient-centered social network can be used for exchanging experiences about the treatment or the symptoms between patients having the same or similar diseases. This resource is also self hosted by the patient.

The different profiles can be implemented either as separate URIs and documents, which are discoverable via the main public profile, or they can be accessible via the same WebID, depending on the credentials of the WebID associated with the requesting user [1].

As the PHR is a rich source of information about a patient, it will be a valuable asset for third party services to provide personalised information recommendations. For example, a service similar to TrialX can utilise health data in the profile to give the patient recommendations on which clinical trials are matched to him/her. To this end, the patient has to grant the service access to the PHR in his/her profile by adding a new entry to the ACL.

In the same way, a service like PatientsLikeMe can gain exclusive access to the private FOAF profile which contains the social network of patients who suffer from the same or a similar disease. This kind of service can then give recommendations on care options based on the treatment of similar patients without compromising the privacy of the patient (e.g., information about the disease will not be accessible to the work related social network of the patient).

5.4 Communication pattern

In order for user agents, data consuming services and profile storage services to participate in the architecture, they need to interact with each other according to their role. In the following we describe an example of the resulting com-

[12] http://health.google.com

[13] http://www.healthvault.com

[14] http://www.dossia.org

[15] http://trialx.com/

[16] http://www.patientslikeme.com/

munication pattern as applied to the use case in section 5.3. The communication pattern is illustrated in figure 3.

1. The patient goes to the third party service using an HTTPS connection. During the SSL handshake, the patient sends his/her WebID to the service. The service verifies the identity of the patient by comparing the public key in the certificate with the one in the public FOAF profile.

2. The third party service welcomes the patient, and asks for permission to access the patient's PHR.

3. When the patient say "yes" to the request, he/she is getting sent to the PCHR platform wherein his/her PHR resides, then the PCHR platform authenticates the patient via WebID and authorises him/her to add a new entry to the ACL for his profile.

4. The PCHR platform asks the patient to confirm the new ACL entry, which specifies that the third party service denoted by the user specified WebID can now access to the patient's PHR.

5. Then the patient confirms the creation of the ACL entry.

6. When the third party service tries to access to the patient's PHR data, the PCHR platform verifies the WebID of the consuming service and checks the ACL of the patient's PHR data. If the permission of the consuming service to access the patient's PHR data is verified, access is granted and the consuming service can perform read operations on the data.

7. After the service gets the patient's PHR data, it compares the patient's health data with clinical trials and recommends the matched ones to the patient.

5.5 Qualitative evaluation

In order to evaluate the presented architecture, we now describe how it meets the identified requirements from section 4:

Protection of identity: Users can choose to use multiple identities, each identity being represented by a unique WebID. Each time a user interacts with a data consuming service his user agent can allow him to choose which WebID to use. In this way pseudonymity, unobservability and deniability of the user identity are supported. None of the identities need to be tied to a real world identity, thus supporting anonymity. Data consumers should not be able to link user identities, however user profile storage services need to be trusted in order to maintain unlinkability of user identities. Self hosting of a user's profiles however can impact the protection of his identities negatively, if the server can be easily linked to a real world identity.

Control over the user data: The user stays in control of his profile data, as the portability of user profiles allows him to move his profile freely between storage services or even to host the storage of his profiles on his own server. Lock-in to a specific ecosystem or to a specific storage service should not be possible, as the open standards of RDF, FOAF and SIOC are used for describing the profile.

Human-computer interaction: Services can provide an easy user interface for managing the ACL of a user profile. A user interface example is given in [14]. The proposed architecture does not require the user to install or understand new software. WebIDs can be used in contemporary Web browsers such as Firefox, which support the installation of user generated SSL certificates.

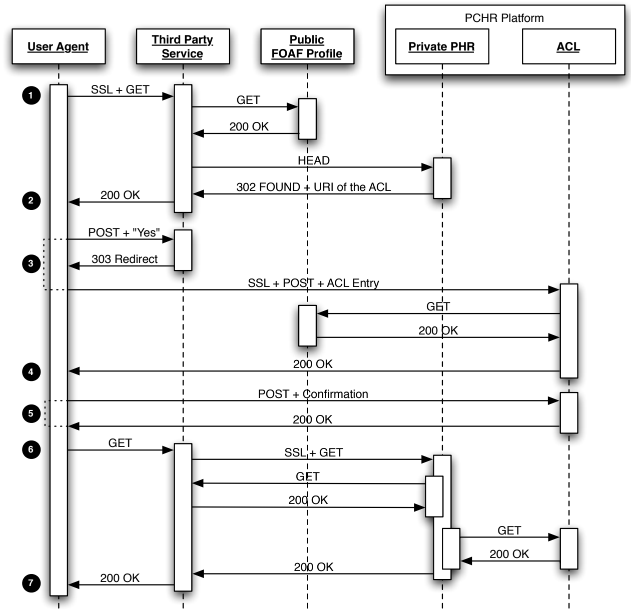

Figure 3: Sequence diagram depicting the communication pattern between the user agent, the PCHR platform and the third party service

Non-functional requirements: The presented architecture allows any user agent, profile storage service or data consumer to participate in one universal ecosystem, as all participants will support the same standards and implement the same communication pattern. The architecture is scalable, as there are no bottlenecks or central points of failure, due to the decentralised nature of the used standards. For profile storage and data consumption existing standards and infrastructure from the World Wide Web and the Web of Data, such as HTTP and RDF are reused, thus making future adoption by service providers easy.

This shows that the proposed architecture allows us to create an ecosystem in which users can protect their identity, they have control over their own data and the interaction with the technology is easily understandable. The architecture also has the properties of universality, scalability and reuse of existing infrastructure. This allows users to benefit from an ecosystem which provides privacy and security while enabling portability of user identities at the same time.

6. DISCUSSION

To be able to provide accurate recommendations by seamlessly combining multiple sources of information is a key objective of recommender system research. However, it invariably raises questions about data ownership and control. By seeking better data fusion techniques, are we tacitly acknowledging that we will ride roughshod over user privacy in order to build richer user models? Furthermore, we need to ask who will own and control the user models and how we may prevent them from being abused. The recent resistance to Facebook's change in its terms of service suggest that Web users are sensitive to how their data is used.

We argue that innovations in reconciling heterogeneous data sources must also be matched by innovations in architecture design and recommender methodology. The problem is simply not one of creating a richer, centralised database on which to create innovative new models, but of designing flexible recommender systems that can cater to differing degrees of data access and control.

Previous research has established the importance of maintaining user trust [10] for successful adoption of personalisation technology. To build systems in which engendering user trust is a fundamental principle requires researchers to solve the challenge of data integration for recommendation systems by algorithmic, architectural and policy-based innovations.

In this paper, we do not address data integration by examining new data fusion techniques. Instead we point to the Web of Data as an example of how heterogeneous data sources can be linked by current standards. However, the Web of Data still requires techniques to allow data providers control access to their data. As such, our focus is a simple personalisation architecture, created from existing standards, that enables user profile portability and cross-domain recommendation while also allowing the user control of his data. The focus of current ecosystems is on locking data into a central hub site, and not on empowering the individual user. Our proposed architecture enables a universal ecosystem which is built around the user profile, by specifying an interoperable way to share and protect the profile data at the same time. Extending the infrastructure of current ecosystem is feasible, as our presented architecture naturally builds on existing standards of the World Wide Web and the Web of Data, such as HTTP and RDF.

7. CONCLUSION

In this paper we addressed the problem of preserving user privacy while seeking to integrate multiple personal information sources. The default architectural solution requires a centralised hub with users reliant upon the good will of the service provider to 'do no evil'. We argue that in order to maintain user trust and support the challenge of heterogeneous data integration must be addressed by algorithmic, architectural and policy-based innovations. As such, we presented an architecture for privacy-enabled user profile portability based on existing standards. We described the requirements for privacy-enabled sharing of profile data for data-mining and modelling. Finally, we illustrated our approach with a use case from the e-Health domain.

8. REFERENCES

[1] F. Abel, J. De Coi, N. Henze, A. Koesling, D. Krause, and D. Olmedilla. Enabling Advanced and Context-dependent Access Control in RDF Stores. In *International Semantic Web Conference*, 2007.

[2] G. Adomavicius and A. Tuzhilin. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-art and Possible Extensions. *IEEE transactions on knowledge and data engineering*, 17(6):734–749, 2005.

[3] R. Arlein, B. Jai, M. Jakobsson, F. Monrose, and M. Reiter. Privacy-preserving Global Customization. In *Proceedings of the 2nd ACM conference on Electronic commerce*, pages 176–184. ACM, 2000.

[4] S. Berkovsky, Y. Eytani, T. Kuflik, and F. Ricci. Privacy-enhanced Collaborative Filtering. In *Proc. User Modeling Workshop on Privacy-Enhanced Personalization*, 2005.

[5] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.

[6] U. Bojars, A. Passant, J. Breslin, and S. Decker. Social Network and Data Portability using Semantic Web Technologies. In *Workshop on Social Aspects of the Web*, 2008.

[7] D. Boyd and E. Hargittai. Facebook Privacy Settings: Who cares? *First Monday*, 15(8), 2010.

[8] R. Burke. Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002.

[9] J. Canny. Collaborative Filtering with Privacy. In *IEEE Symposium on Security and Privacy*, 2002.

[10] R. Chellappa and R. Sin. Personalization versus Privacy: An Empirical Examination of the Online Consumers Dilemma. *Information Technology and Management*, 6(2):181–202, 2005.

[11] S. Decker, S. Melnik, F. Van Harmelen, D. Fensel, M. Klein, J. Broekstra, M. Erdmann, and I. Horrocks. The Semantic Web: The roles of XML and RDF. *IEEE Internet computing*, 4(5):63–73, 2000.

[12] D. Fletcher. How Facebook Is Redefining Privacy. Time Magazine, May 2010.

[13] B. Heitmann and C. Hayes. Using Linked Data to Build Open, Collaborative Recommender Systems. In *AAAI Spring Symposium: Linked Data Meets Artificial Intelligence*, 2010.

[14] J. Hollenbach, J. Presbrey, and T. Berners-Lee. Using RDF Metadata to enable Access Control on the Social Semantic Web. In *Workshop on Collaborative Construction, Management and Linking of Structured Knowledge*, 2009.

[15] A. Kobsa. Privacy-enhanced Web Personalization. In *The adaptive web*, pages 628–670. Springer, 2007.

[16] H. Liu, X. Q. Hou, G. Hu, J. Li, and Y. Q. Ding. Development of an EHR System for Sharing - A Semantic Perspective. *Studies in Health Technology and Informatics*, 150:113–117, 2009.

[17] K. D. Mandl and I. S. Kohane. Tectonic Shifts in the Health Information Economy. *The New England Journal of Medicine*, 358(16):1732–1737, 2008.

[18] B. Miller, J. Konstan, and J. Riedl. PocketLens: Toward a Personal Recommender System. *ACM Transactions on Information Systems (TOIS)*, 22(3):437–476, 2004.

[19] H. Story, B. Harbulot, I. Jacobi, and M. Jones. FOAF+SSL: RESTful Authentication for the Social Web. In *Workshop on Trust and Privacy on the Social and Semantic Web*, 2009.

[20] Y. Wang and A. Kobsa. Technical Solutions for Privacy-Enhanced Personalization. *Intelligent User Interfaces: Adaptation and Personalization Systems and Technologies*, 2009.