Understanding Personal Data as a Space – Learning from Dataspaces to Create Linked Personal Data

Laura Dragan[1], Markus Luczak-Roesch[1], and Nigel Shadbolt[1]

University of Southampton, UK
lcd@ecs.soton.ac.uk, m.luczak-rosch@soton.ac.uk, nrs@ecs.soton.ac.uk

Abstract. In this paper we argue that the space of personal data is a dataspace as defined by Franklin et al. We define a personal dataspace, as the space of all personal data belonging to a user, and we describe the logical components of the dataspace. We describe a Personal Dataspace Support Platform (PDSP) as a set of services to provide a unified view over the user's data, and to enable new and more complex workflows over it. We show the differences from a DSSP to a PDSP, and how the latter can be realized using Web protocols and Linked APIs.

Keywords: Personal Information Management, Dataspaces, Linked Data

1 Introduction

In their 2005 paper [7] Franklin et al. introduced Personal Information Management (PIM) as one of the two usage scenarios for what they called *dataspaces*. They define a *dataspace* as an "abstraction for data management" across heterogeneous collections, and propose the design and development of a suite of basic services, collectively named DataSpace Support Platforms (DSSPs), to solve in a general way the recurring data management chalenges identified for heterogeneous collections of data: search and query, integration, availability, recovery, access control, evolution of data and metadata. At that time, PIM included mostly desktop data, with the extension to remote file storage, and few Web services and mobile devices, but since then, the notion of personal data evolved, as did the online services available to users. Personal data shifted from being desktop centric to the Web, and the types and amount of personal information that were being captured increased and diversified greatly. An important consequence of this shift was that the data, while still personal, moved out of the users' control, and under the control of the many organizations providing Web services, like Facebook, Runkeeper, Amazon, etc.

Not being in full control of its data is one of the key distinguishing characteristics of a DSSP. In this paper we argue that the space of personal data *is* a dataspace as defined by Franklin et al., and we describe the characteristics of a Personal Dataspace Support Platform (PDSP). We show the differences from a DSSP to a PDSP, and how the latter can be realized using Web protocols and Linked APIs.

2 Background and Related Work

In this section we start by detailing the characteristics of dataspaces as originally defined. Then, we present existing related work which describes the Web of Data as a dataspace, followed by related work in the field of Personal Information Management.

2.1 Characteristics of Dataspaces

Franklin et al. describe dataspaces as having the following distinguishing characteristics [7]: C1) Support *all* the data in the dataspace, without leaving out any. This requires handling a wide variety of formats, systems, and interfaces. C2) Is *not* in full control of the data it handles, as this data might be accessible and changed though other, native intefaces. C3) May offer varying levels of service, returning *best-effort* or approximate answers, depending on the available data sources. C4) Must provide the tools to create better integration between data sources, in a *pay-as-you-go* manner. This includes the creation of links and mappings between data from unconnected sources.

Dataspaces are modeled [7] as a set of *participants* and *relationships*. The participants are the data sources which are available in the dataspace, regardless of the types of data, data access, and available services they provide. Relationships are all references between any two or more of the participants, describing mappings at the schema or instance level, but also rules describing the transformation of the data from one participant into a different form, which results in two distinct participants – the original data source and the transformed replica. The dataspace should be able to model any kind of relationship between any of its participants.

Franklin et al. propose a suite of key services, under the name of DataSpace Support Platform (DSSP) from which we highlight the following: (a) access to data, (b) cataloging, (c) browsing, (d) search and querying, (e) monitoring and event detection. The goal of a DSSP is to provide a set of basic functions over the heterogeneous data in a dataspace, to support the incremental building of more complex and specialized services. While dataspaces are "not a data integration approach, they are more of a data co-existence approach", information can be integrated in a "pay-as-you-go" fashion, as described by Madhavan et al.[9]. Heath and Bizer adopted this terminology to highlight the evolutionary nature of Linked Data on the Web, which we briefly describe below.

2.2 The Web of Data

What has started as a Web of Documents, has evolved into a Web of Data. The Web architecture is well suited to solve data integration problems at very large scale in an evolutionary fashion. By means of the Resource Description Framework (RDF) a graph-based data representation is provided, which, in combination with the use of stable HTTP URIs, allows for integrating data without the need for fully-fledged a priori schema alignment. The support for such flexible, pay-as-you-go type of data integration [9] makes the Web of Data "a realization of the dataspaces concept on global scale"[2].

2.3 Personal Information Management

Personal Information Management stemmed as a field of study in response to the problem of information overload. Many methods and tools have been developed to enable us to organize information better, so that we can easily find it and re-find it when needed. However, the growing number of applications designed to aid the management of the information had the side effect of the information being trapped in unconnected repositories, and incompatible (proprietary) data formats, which in turn led to duplication of data and increased effort required for organizing.

The Semantic Desktop [4] systems aimed to solve the fragmentation problem by marrying PIM with Semantic Web technologies like common representation languages and ontologies. Systems like Gnowsis [11], IRIS [3], SEMEX [6], X-COSIM [8], Nepomuk [1], use predefined ontologies to classify the resources they manage. These ontologies vary from general to detailed, from fixed to user editable and from monolithic to layered and modular. Some of the systems use a central data store for all the personal semantic information extracted from non-semantic applications, while others propose the replacement of the non-semantic tools with semantic counterparts. They all provide enhanced search and query functionality, as well as facetted browsing and link traversal for "follow-your-nose" exploratory browsing. Dittrich et al. [5] define the concept of a Personal Dataspace, as containing the entire personal information of a user, and define an architecture for a Personal Dataspace Management System (PDSMS), the equivalent of a DSSP. However, the iMemex system described is more similar to the vision of the Semantic Desktops.

With the shift to the Web, much of the personal information residing online consists of the usual PIM elements: documents, email, and calendar, but is amplified by social networks, multimedia, personal annotations, browsing history, activity feeds, location feeds, and much more, all out of the control of the user.

Personal Data Stores are a solution to give users back some control over their personal information. Some come in the form of cloud services which collect and integrate data from numerous other services on behalf of the users, and some are systems that can be installed locally or on servers under the administration of the users themselves. Many Personal Data Store solutions rely on the principles and standards of the Web architecture, as they provide a flexible platform to realise highly interoperable network-based systems.

3 Personal Data as a Dataspace

Personal Information Management is cited as the first of two scenarios to illustrate dataspaces by Franklin et al. [7], and at that time it fulfilled three out of the four characteristics of dataspaces: PIM contained *all* the personal data (C1), it supported *pay-as-you-go integration* (C4), and would return *best effort* results to queries, based on the available data (C3). Since then we have seen much advancement in the area of PIM, including the development of Semantic Desktops, which aim to solve the data integration problem on the desktop, and Personal Data Stores which bring personal data back under users' control. However, with the shift to the Web, the Semantic Desktops no longer manage *all* the user's data, and the remaining characteristic becomes true for personal dataspaces – the information contained is *not under the full control* of the DSSP (C2).

Based on these characteristics we argue that the space of personal data, as it is now, matches the definition of a dataspace, by fulfilling the characteristics described above. The personal dataspace contains heterogeneous data, is distributed, controlled by many different service providers, unlimited in relation to the number and types of information that it can contain, and most importantly, is centered around the user whose data it contains. This last characteristic is what differentiates a personal dataspace from a generic dataspace.

In the next sections we describe in more detail the components of a personal dataspace, and the services that a Personal Dataspace Support Platform (PDSP) could provide.

3.1 Participants and Relationships

In our application to personal data, all the data sources that collect, store and manage personal information are participants in the dataspace. In the past, the desktop contained most of the personal data, so the participants in the personal dataspace would have been the individual applications – file system and manager, email clients, calendar, task manager, etc. Semantic Desktops unified most of the desktop information under a single uniform data representation and storage, regardless of the application it comes from. Thus, such a Semantic Desktop would be seen as a single participant in the personal dataspace. However, desktop applications whose data was not included in the Semantic Desktop can be separate participants, even if they reside on the same desktop. Another set of participants to the dataspace consists of the online applications that collect and store information about the user. This set can include shopping history and wish lists, activity and food tracking, trip planning, emails and calendaring, blogging and micro-blogging, social networks, etc. Mobile applications which track the user movements and activities are also participants in the dataspace.

One of the key characteristics of a dataspace is that it contains *all* the data in the space. Applied to personal data, this means that the personal dataspace must contain *all* the data about the person at the centre of the space. However, despite the ever growing number of applications and devices which collect and deal with personal data, only those systems which handle data about *the* central user, are participants in the personal dataspace of that user. Let's assume we have two people, Alice and Bob, each using fitness tracking applications: Alice uses Runkeeper and Fitbit, and Bob uses Fitbit and Endomondo. The Fitbit application will be a participant in the personal dataspaces of both Alice and Bob, while Runkeeper will be a participant only in Alice's personal dataspace.

Going further with the example, while Fitbit is a participant in both users' personal dataspaces, it participates in Alice's dataspace only with Alice's collected data, and respectively in Bob's dataspace with Bob's collected data. So, while under the control of the same organization (in this example Fitbit), the participants are in fact distinct for the two different users. We can make the distinction better by specifying that in the case of a Web application where the user has an account, the participant to that user's dataspace provides in fact only the view that the user's account in that application has over the data. Some service providers might collect more information on the users than they make available back to the users – for example the browsing history in online shops, and unpublished posts in social networks. Such information, while it is "personal data" is not available in the personal dataspace.

Another consequence of the personal dataspaces revolving around a single person's data is that the personal dataspace of a user can act as a participant in another user's personal dataspace, by means of sharing information between the two. This possibility requires access control management in the PDSP, a service described in the following section.

The relationships between the participants can be explicit and formally expressed – like for example when the user manually connects two applications and allows them to share data completely or partially. For example, our user Alice can explicitly allow access to her Fitbit data from her Runkeeper account, for better monitoring of her daily activity. Relationships can be also implicit, when two participants manage the same type of data, like for instance social connections or fitness tracking, but without actually sharing any of the data itself. For example if Bob would also buy and use a Nike+ Fuelband, both his Fitbit and the Fuelband would record activity and movement information, like number of steps, distance, and calories, but independently of each other and without exchanging any data. Such relations can be made explicit by providing a mapping between the schemas, and thus enabling better data integration outside of the systems holding the data. The mappings can be created in the pay-as-you-go fashion, as needed and as available. The PDSP must support the creation of such mappings and the import of existing mappings from third parties.

3.2 Services and Functions

Like a DSSP, a Personal Dataspace Support Platform (PDSP) should provide a set of services to access the data contained in the dataspace, and allow basic handling of this data, to support the development of more specialized services on top of the platform.

We list below the services a PDSP should provide, although different PDSPs can choose to implement just a subset, or add more services. Some of the services here were identified in [7] for a DSSP, and are relevant for personal dataspaces as well. In their description of the iMemex, Dittrich et al. [5] divide the list of services in two layers, the *Physical Data Independence Layer* (PHIL), which is closer to the data layer, and the *Logical Data Independence Layer* (LIL), closer to the applications which are built on top of the platform. Although we also define services which rely on other services, some closer to the data, and some closer to the user and applications, we do not define specific layers, because the dependencies and relative positions of the services could change over time, we do not restrict their relative positions, nor possible interactions.

Data access The data available from the participants is heterogeneous and dynamic. Accessing it requires support for multiple Web APIs, formats, and protocols, as well as high tolerance to errors and unknowns. This service must have available all the information required to be able to access the data in all the participating data sources, thus relying on the identity management and cataloging services.

Cataloging Depending on the kinds of APIs offered by the participants, the catalog can contain alternative access points for different formats – for example HTML or JSON, or capabilities provided by the participant – for example SPARQL or REST. The catalog should also contain a schema for the data it contains. The VOID[1] vocabulary is one possible way of describing the participants exposing RDF data. The catalog does not have the function of storage or archive, but rather of a meta-data store containing information about the data available.

Indexing Search is one of the essential services of a PDSP, and it requires a reliable indexing service. Some of the participants might already contain an index over their data, and allow direct access to it, or mediated by a search and query function, but the PDSP should provide a uniform API over all the participants. Specialized indexes are recommended for special types of data like time and location. In the context of personal information the time dimension is important for reminding [10]. With the growth of location services and activity tracking, geographical information becomes also an important dimension, which can enable useful aggregations and filters.

Mapping The PDSP should provide a way for incremental accumulation of schema mappings between its participants as to enable the pay-as-you-go data integration. The mappings can be manually created by the user, automatically generated, or imported. They can be partial, when not all the possible types of data from a participant are mapped. The mappings are data themselves, and as such they can be shared with other users' PDSPs.

Integration Pay-as-you-go data integration is one of the key functions of a dataspace. In the PDSP it is provided by the mapping service, which allows the incremental creation of a directory of mappings between the data structures provided by various participants. Deeper integration can be done by allowing the creation of relations between resources from different participants.

Monitoring The PDSP must be aware of the changes in the data and in the state of the participants, and update its catalog and indexes accordingly. The monitoring service can use event detection mechanisms, subscriber APIs, or any other feature provided by the data sources.

Browsing, search, and query Browsing, search, and query services should allow the users to explore their dataspace, through keyword search as well as

[1] http://www.w3.org/TR/void/

structured query, by iteratively refining and restricting a domain. These services should work uniformly across the dataspace, regardless of the structure and mode of access of the particular participants, as this allows the user to have a unified and uniform view into the data, independently of where the data shown comes from. The user should be able to use these services on data as well as on metadata. As with the indexing service, some participants may already provide a search interface.

Identity management As described in the previous section, some of the participants which provide Web APIs to access the users' personal information require that the users are authenticated, or that the application used to access the data on their behalf is authorised. The method for authentication and authorisation could vary, and it is recommended that the PDSP supports all or most of the existing methods. The account information and access tokens must be securely stored.

Access control In combination with the data access and the identity management service, the PDSP becomes a gateway to the user's entire personal information, thus it needs to ensure that the privacy and security of the data is maintained. Additionally, in the case of one user's PDSP becoming a participant in another user's dataspace, access must be controlled so that only the authorised part of the personal data is disclosed.

Annotation Annotation refers to the creation of metadata, and in PIM it is an important feature. The PDSP should support annotation of any type of resources, data, metadata, or participants. Annotations can include data from multiple participants. Provenance is a special type of annotation, and all annotations are data, which can be shared, annotated, queried, etc.

Update Some participants may support updating data through Web APIs, although not all, and not all types of data. In cases when the update cannot be propagated to the source, the service could support saving local versions of the updated data. This can however lead to conflicting versions of the same data, which is why provenance information and change logs are important. There are cases when updating data cannot be done at all, for example when it comes from sensors whose readings cannot be changed. In these cases annotations can replace the update.

The architecture of the PDSP that we envision does not provide any applications, although it does not prohibit them either. We see applications as an extra layer on top of the PDSP, and not at the core of the architecture as part of the foundational layers. We propose that the applications, as well as higher level services, are built on top of the PDSP, using the APIs provided by its services.

We consider storage as a higher level service, thus not part of the core suite of services provided by the PDSP. We do not require that the PDSP fetches and keeps duplicates for the personal data already stored by the participating data sources. The way the storage of data is handled is one of the main differences between a PDSP and Personal Data Stores.

4 Conclusion

A personal dataspace is the space of a user's entire personal data, regardless of where it is located, on the desktop or on the Web; integrated in a Semantic Desktop, or fragmented across many Web platforms. Following the original definition of dataspaces by Franklin et al., we define the logical components of a personal dataspace, and we describe a Personal Dataspace Support Platform as a set of services to provide a unified view over the user's data.

Acknowledgments

This work was supported by the EPSRC Theory and Practice of Social Machines Programme Grant, EP/J017728/1.

References

1. Bernardi, A., Decker, S., van Elst, L., Grimnes, G.A., Groza, T., Handschuh, S., Jazayeri, M., Mesnage, C., Moeller, K., Reif, G., Sintek, M.: The Social Semantic Desktop: A New Paradigm Towards Deploying the Semantic Web on the Desktop, chap. 118, pp. 2279–2303. IGI Global (2009)
2. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. International Journal on Semantic Web and Information Systems 5(3), 1–22 (2009)
3. Cheyer, A., Park, J., Giuli, R.: IRIS: Integrate. Relate. Infer. Share. In: Proceedings of the 1st Workshop on The Semantic Desktop (2005)
4. Decker, S., Frank, M.: The social semantic desktop. In: Workshop on Application Design, Development and Implementation Issues in the Semantic Web (2004)
5. Dittrich, J.P., Blunschi, L., Farber, M., Rene, O., Shant, G., Karakashian, K., Antonio, M., Salles, V.: From Personal Desktops to Personal Dataspaces: A Report on Building the iMeMex Personal Dataspace Management System. In: Proceedings of BTW 2007 (2007)
6. Dong, X., Halevy, A.: A Platform for Personal Information Management and Integration. In: Proceedings of the Second Biennial Conference on Innovative Data Systems Research (CIDR2005) (2005)
7. Franklin, M., Halevy, A., Maier, D.: From databases to dataspaces: A new abstraction for information management. SIGMOD Rec. 34(4), 27–33 (Dec 2005)
8. Franz, T., Staab, S., Arndt, R.: The X-COSIM Integration Framework for a Seamless Semantic Desktop. In: Proceedings of the 4th International Conference on Knowledge Capture (K-CAP2007) (2007)
9. Madhavan, J., Jeffery, S.R., Cohen, S., Dong, X.L., Ko, D., Yu, C., Halevy, A.: Web-scale Data Integration: You Can Only Afford to Pay As You Go. In: Proceedings of CIDR2007 (2007)
10. Ringel, M., Cutrell, E., Dumais, S., Horvitz, E.: Milestones in Time: The Value of Landmarks in Retrieving Information from Personal Stores. In: Proceedings of the International Conference on Human Computer Interaction (2003)
11. Sauermann, L.: The Gnowsis Semantic Desktop Approach to Personal Information Management. Ph.D. thesis, Fachbereich Informatik der Universität Kaiserslautern (2009)