

# Carpé Data: Supporting Serendipitous Data Integration in Personal Information Management

**1st Author Name**  
Affiliation  
Address  
e-mail address  
Optional phone number

**2nd Author Name**  
Affiliation  
Address  
e-mail address  
Optional phone number

**3rd Author Name**  
Affiliation  
Address  
e-mail address  
Optional phone number

## ABSTRACT

### Author Keywords

### ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI):  
Miscellaneous

### General Terms

Human Factors; Design; Measurement.

## INTRODUCTION

In recent years, an unprecedented quantity and variety of information has been made available as structured data on the Web, through APIs, datasets, and data feeds. This information includes access to data previously hidden behind services and applications, such as retailer product catalogues, information not previously directly released to the public, such as open government data or records of personal financial transactions, and new kinds of data generated by emerging kinds of data sources, such as wearable and smartphone-based biosensors.

While a primary goal of opening up such direct access to information is the hope that end-user citizens will make more informed decisions pertaining to their health, wealth, or well-being [], use of this data beyond specialised app developers, journalists and other data specialists remains minimal. Our hypothesis is that the main impediment towards end-user access to this data is a mismatch between the capabilities of tools currently used to satisfy daily information needs, and those required to effectively browse, use and consume heterogeneous structured data from diverse sources. For example, while much of this data represents a many and varied set of things, activities, and descriptions at various granularities, most commonly used structured personal information management (PIM) tools either can only manage a small, fixed set of data types (such as digital calendaring tools and to-do list managers) or provide little or no support for structured data at all, such as text editors and sketching/drawing tools [].

If PIM tools were extended to enable end-users to make effective use of the emerging ecosystem of personal data, how would this impact personal information practice? This paper presents our initial results towards answering this question through a multi-method investigation of the personal data integration problem. We start with a small interview study examining information tasks people perform that rely on multiple sources of information, and how such needs are currently satisfied. Our results suggest that people prefer to rely upon multiple, diverse sources to singular integrated ones for a number of reasons, including coverage, reliability, and seeking out more suitable alternatives. Second, we conduct an analysis of various popular personal data sources available today to identify barriers to effective unification of data from each. We draw upon definitions and work from the database integration systems literature to characterise types of heterogeneity exhibited across data sources in three popular domains: social networking, shopping, and dining, finding that terminological heterogeneity dominates the simpler data feeds (including social networking and restaurant recommendations), while structural issues pervade the more complex schemas of online retailers' product catalogues.

Based upon the data integration needs revealed by people in the interviews, and the kinds of heterogeneity observed in the data feeds, we embarked on a user-centric design exercise towards an interface to facilitate simple, "light-touch" integration tasks of diverse, heterogeneous data sources. We describe our initial foray into this exercise, an interface we call DataPalette, which focuses on the most common types of heterogeneity to enable serendipitous "data mixing" - a form of simple integration sufficient to let people easily and effectively combine and compare information sources without the need to write code. Our usability evaluation of DataPalette reveals that most users were comfortable with and that this interface is effectively expedites the most sophisticated of the typical kinds of "data integration" tasks people need to perform while planning personal,

## BACKGROUND

In this section, we touch on motivating work for DataPalette from three fields: database integration, personal information management, and end-user mashups and toolkits.

The primary motivation for this work stems from field studies in personal information management, where letting users effectively consolidating and work with data from multiple sources has remained a challenge for over a decade. The

problem of data fragmentation ([7], for example, arises, in part, from a failure of PIM tools to adequately support the effective consolidation of data across multiple applications, services and platforms (e.g., [1], [2]). In our own previous field study, (reference anonymised) we observed several different variations of manual “coping strategies” (or, if programmers, elaborate, custom-coded one-off solutions) that people devised to merge in data from external sources – be them online, offline, or other people. Similar findings from other studies include Volda et al.’s observations of volunteer coordinators’ “homebrew databases”, manually-maintained information assemblages concocted to handle the widely varied and heterogeneous data collection requirements such groups needed to coordinate their activities [8].

It became clear that effectively coping with data heterogeneity would be a central focus to letting people effectively draw upon multiple, arbitrary sources of data when consulting the theoretical data integration literature. Alon Halvey describes heterogeneity as an inevitable by-product of the distributed processes that are used to author it:

[The problem stems from the fact that] we are trying to integrate data systems that were developed for slightly (or vastly) different business needs. Hence, even if they model overlapping domains, they will model them in different ways. Differing structures are a byproduct of human nature ?people think differently from one another even when faced with the same modeling goal. [?]

In this paper, we take the position that there needs to be a difference in focus between the data integration approaches discussed in the databases literature and that of PIM, because of corresponding differences in requirements. In database systems, the overall objective is usually to allow multiple large databases to be accessed as if one. This requires essentially full mapping of all data types and objects and well-defined semantics concerning where and how data gets converted. The upfront costs of such complete integration are typically high, but are typically justified in the long-term, high-volume use cases imagined. In typical sensemaking scenarios of PIM, however, people might need to combine information from multiple sources for a one-off task; meaning that any upfront cost would be fully consume the time and effort a person has allocated for their task. Moreover, while a fully, uniform query capability might be nice, such capability is likely of less overall importance for a single quick task than simply being able to quickly and easily identify and cope with differences.

A number of prototype PIM systems have demonstrated data integration capacities. For example, SEMEX [3] was a PIM environment that provided users a uniform query interface over the contents of their personal information environments, across heterogeneous files, folders and repositories. Similarly, Haystack was a personal information manager that allowed users to manipulate and managing their data in a uniform, consolidated manner [5]. Pertaining to each system’s ability to integrate information from arbitrary new data sources, however, both systems generally required new sources to either already conform to a pre-specified set of ontologies, or have the user provide a manual specify such a

mapping translation themselves.

Systems that have focused solely on the end-user-driven, interactive reconciliation of data heterogeneity have come out of the web “data mash-up” community. Yahoo Pipes! is a visual programming language aimed at letting novice programmers easily create data-transformation workflows, but remained rather difficult to use and poorly suited for one-off information tasks. Mash-up makers such as Intel Mashmaker [], QDEWiki, and TX2

DataPalette’s approach is most directly inspired by “lightweight” visual, user-interactive data collection, consolidation and alignment tools such as Potluck [6] and Cards [4].

## PRE-STUDY 1 - UNDERSTANDING HETEROGENEITY FOR EVERYDAY TASKS

### Method

Our pre-study investigated the types of tasks people perform online that required heterogeneous data sources, and which tools they would use to organise a large social event involving 15 or more people. We interviewed 8 participants, using a general interview guide approach based loosely on a set of questions (see Appendix ?). This approach allowed both the participant and interviewer a degree of freedom in conversation and each interview could adapt to the participants experiences, while collecting the same general areas of information. Each interview was recorded and notes taken, the interviewer was trained so that they were knowledgeable about the importance of the study and to minimise bias.

### Results and Discussion

The participant demographic consisted of 8 people, 7 of which were male and 1 female. They were split evenly into two age groups of 18-25 and 26-32. All but one participant used social network sites very regularly, and they either logged in multiple times a day, left a website page open or used dedicated application for their social networking sites updates. Those that used social networking websites primarily use Facebook. All of the participants used Twitter to listen to broadcasted tweets, and 6 out of 8 people regularly tweeted.

#### Question 1: Tasks

All participants had experience in using multiple websites to complete tasks. They listed example tasks such as shopping (including food, consumer electronics, hotels and flights), searching for jobs, choosing recipes, study and work. The biggest focus for shopping focused tasks was balancing price, quality and speed of delivery. In general, people did not trust a single websites for reviews because they said that they could be biased and did not always review the features they were interested in. One participant said that video reviews of items were really important because they showed the aesthetics and scale of the item. In general for initial research into organising a task it was typical to Google keywords to find related websites to their tasks. They identified that they used different website for different jobs, such as manufacturers website for technical details, review aggregators for a

range of opinions, and Google maps for location based decisions. Social networks were also identified to have different functions, such as different groups of people belong to each and share different types of opinions. One participant said that they shared their outlook and gmail calendars, but only for work events and did not record social activities on them.

#### *Question 2: Large Social Event*

All of the participants would organise a large social event by talking to their friends, about their ideas, preferences and recommendations. The tools they would use to acquire this information included face-to-face, phone calls, email, Skype, Facebook events. The method chosen depended on the time scale required to organise an event, in general if there was a short time scale people would speak in person or on the phone. Otherwise, people would set up a FaceBook event page to discuss ideas with their friends. If their friends were not on Facebook they would email them. All but one participant said they would not use Doodle to organise day and time of an event, because they felt that people did not fill in the form and it was more suit to organising work event not social.

Most people felt that being assertive and posting their decisions about the date and time on Facebook meant that organising events were more successful than trying to gather a consensus. They expected that their friends would voice any objections if they could not attend or like the restaurant or activity selected. They selected venues, locations, restaurants and activities based on their own knowledge, friends recommendations, Google maps, reviews through Googling and review sites such as tripadvisor. People stated that the cost, price and easy to get to locations were the most important factors when choosing a venue. They considered how to get there and looked up train or bus times on the Web. One participant organised carpools informally and at last minute via Facebook or text messages.

When organising an event the weather was not that important because it they did not trust its accuracy for events plan further in the future than one week. Also, their friends preferences in price and food choices were not a priority; they would try to be inclusive but only if it meant small changes to the plan. Choosing food and timings were often made on-the-fly, by either with the use of mobile applications or by walking past a location.

#### **Prestudy 1 Summary of Findings**

People said that they would like a website that could support the evolution of an event, that allowed them to post drafts of an event. They felt that organising an event should be a process that changes over time and did not need rushing. Three of the participants said that they would like a recommendation system for places to eat and activities. Another strong requirement was that it had to be ubiquitous so that everyone could use it because they wanted a single place to communicate with their friends. While Facebook was the most popular social networking site for organising events, half of the participants felt that it was not the best solution and would prefer a collaborative environment, which could be wiki based.

## **PRESTUDY 2 : TECHNICAL CHALLENGES OF INTEGRATING THE DATA**

### **Method**

### **Results**

### **Analysis**

#### **Prestudy 2 Summary of Findings**

#### **WEBBOX INTERFACE - DESIGN OF THE DATAPALETTE**

Design Goals, What are the problems observed needed to address What kind of interactions can be designed to solve them?

Interaction features multiple paths selection sameas interactions visualisation of partial results

Design process

### **METHODOLOGY**

Lab based usability study,

Design of Study How we designed the tasks, subjective task that requires multi dimensions How we got data How we recruited participants - International student list

Our aim was XXX, To accomplish thies XX participants were recruited to the lab for a 1 hour session

Latin square dance! 4 factors Interface or not Task (restaurant or uni) Dataset (glasgow / cambridge; history or sports) Participant

Hypotheses Interaction-method specific hypotheses h1.1 - Do people understanding the problem of co-reference, and the need to reconcile coreference problems? h1.2. Does the ability to do drag + drop combination for co-reference reconciliation effectively solve this problem? h2.1- Do people understand the problem of structural differences in data? h2.2 - Does multipath selection let people effectively deal with this – and work with collections of heterogeneous items?

Are these interaction methods sufficient to perform common tasks involving heterogeneous data sources? Does supporting these simple techniques facilitate task completion?

### **RESULTS**

How many native english speakers, only 8 were all enrolled at university all past proficiency test.

### **DISCUSSION**

What we found

What we didnt find (evidence for)

What we might look at next – follow-on work We focused on terminological heterogeneity – didnt address structural heterogeneity, semantic heterogeneity We focused on finite (small) sets Streaming data (how do select future data –) Scale - small data, what are the facilitates required for large data Collaborative - how might interaction support multiple users collaborating?

Limitations found from study... what is the take away from the limitations

## CONCLUSION

## ACKNOWLEDGMENTS

## REFERENCES

1. Bergman, O., Beyth-Marom, R., and Nachmias, R. The project fragmentation problem in personal information management. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, CHI '06, ACM (New York, NY, USA, 2006), 271–274.
2. Boardman, R., and Sasse, M. A. Stuff goes into the computer and doesn't come out: a cross-tool study of personal information management. In *Proc. CHI '04*, ACM Press (2004).
3. Dong, X., and Halevy, A. Y. A Platform for Personal Information Management and Integration. In *Second Biennial Conference on Innovative Data Systems Research (CIDR)* (2005), 119–130.
4. Dontcheva, M., Drucker, S. M., Salesin, D., and Cohen, M. F. Relations, cards, and search templates: user-guided web data integration and layout. In *Proceedings of the 20th annual ACM symposium on User interface software and technology*, UIST '07, ACM (New York, NY, USA, 2007), 61–70.
5. Huynh, D., Karger, D., and Quan, D. Haystack: A Platform for Creating, Organizing and Visualizing Information Using RDF, 2002.
6. Huynh, D., Miller, R., and Karger, D. Potluck: Data mash-up tool for casual users. *Web Semantics: Science, Services and Agents on the World Wide Web* 6, 4 (Nov. 2008), 274–282.
7. Jones, W., Karger, D., Bergman, O., Franklin, M., Pratt, A., and Bates, M. Towards a unification and integration of pim support, 2005.
8. Volda, A., Harmon, E., and Al-Ani, B. Homebrew databases: complexities of everyday information management in nonprofit organizations. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, ACM (New York, NY, USA, 2011), 915–924.