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

**Max Van Kleek, Daniel A. Smith, Heather S. Packer, Jim Skinner, Nigel R. Shadbolt**
Web and Internet Science Research Group, Electronics and Computer Science
University of Southampton, Southampton, UK
{emax, ds, hp3, js40g09, nrs}@ecs.soton.ac.uk

## ABSTRACT

The information processing capabilities of humans enable them to opportunistically draw and integrate knowledge from nearly any information source. However, the integration of digital, structured data from diverse sources remains difficult, due to problems of heterogeneity that arise when data modelled separately are brought together. In this paper, we present an investigation of the feasibility of extending Personal Information Management (PIM) tools to support lightweight, user-driven mixing of previously un-integrated data, with the objective of allowing users to take advantage of the emerging ecosystems of structured data currently becoming available. In this study, we conducted an exploratory, sequential, mixed-method investigation, starting with two pre-studies of the data integration needs and challenges, respectively, of Web-based data sources. Observations from these pre-studies led to *DataPalette*, an interface that introduced simple co-reference and group multi-path-selection mechanisms for working with terminologically and structurally heterogeneous data. Our lab study showed that participants readily understood the new interaction mechanisms which were introduced. Participants made more carefully justified decisions, even while weighing a greater number of factors, moreover expending less effort, during subjective-choice tasks when using DataPalette, than with a control set-up.

### Author Keywords

Personal information management; end-user data integration; mash-ups; sensemaking with data

### ACM Classification Keywords

H.5.2. Interaction styles: Direct manipulation; H.2.5. Heterogeneous Databases: Data translation; H.5.m. Information Interfaces and Presentation: Miscellaneous

## INTRODUCTION

In recent years, an unprecedented quantity and variety of information has been made available as structured data on the Web. The explosion of structured data APIs and downloadable data sets has arrived from many different sectors, including retail, banking, social networking, opinion and recommendation sites, mobile and desktop apps, as well as from new kinds of sensors and devices such as wearable activity and bio-sensors. Beyond commercial apps and services, governments have embraced releasing data as a way of providing transparency and accountability to their citizens, causing a huge push to release "open data" to the public, at scales ranging from the smallest local councils to entire national departments.

A common goal for the release of such data has been to provide end-users with the ability to make more informed decisions pertaining to their health, wealth, and well-being [17]. Thus far, however, this data has been predominately used by app developers, journalists and other data specialists in one-off data analyses, apps and investigations. Hence, as yet, these data still remain far from making a positive impact upon the activities of ordinary people. A core reason for this, we posit, is a lack of suitable tools for accessing and interrogating these heterogeneous data sources, as well as a means by which these disparate information sources can be effectively brought together and cross-referenced. For example, most personal information management (PIM) tools only provide the capacity to manage small, pre-fixed sets of data types (such as calendar appointments, address book entries, tools and to-do list items). Alternatively, they may provide little or no support for structured data, as exhibited in word processors, text editors, and sketching/drawing tools [2].

In this paper, we present an investigation of ways that PIM tools might be extended to allow data to be brought together in an ad-hoc fashion from arbitrary structured data sources, so that end-users can effectively harness knowledge emerging from the vast ecosystems of data. For this purpose, we have employed an exploratory, mixed-method approach with three sequential stages. First, in a pre-study, we have conducted a series of semi-structured interviews in order to understand the various types of tasks people perform using multiple information sources, and the processes that they rely upon to perform such tasks. Our results suggest that people typically draw upon multiple, diverse sources for a number of reasons, including increasing the breadth of information, assessing the reliability of information, and gaining multiple perspectives.

Second, we have carried out a structural analysis of data records from personal data sources to identify the most com-

*CHI 2013*, April 27–May 2, 2013, Paris, France.

mon kinds of problems likely to arise from integrating data from these multiple sources. For this, we have examined data sources from six common domains: contacts, events, music, shopping, social networking and weather. We have discovered, in our analysis, that while terminological heterogeneity is the most common issue for unifying simpler data records (such as address book contacts and social network profiles), structural differences are exhibited among more complex schemas, such as online retailers' product catalogues.

Third, we have developed DataPalette, an interface that enables serendipitous "data mixing," eliminating the need for people to write bespoke code to effectively combine and compare heterogeneous data. DataPalette facilitates basic integration of diverse, heterogeneous data sources using simple interface gestures. A usability evaluation of DataPalette has revealed that most users are comfortable with its interactive integration mechanisms, and that it effectively improves people's abilities to perform multi-faceted decision-making tasks using multiple sources.

## BACKGROUND

Our work has been informed by three fields of inquiry, comprising PIM, database integration, and end-user mashups and toolkits. Our primary motivation stems from field studies in PIM, where the need for consolidating data has been observed. This is an area in which many challenges remain, specifically, in reconciling the differences that arise among representations when data is created by different people, at different times, and for different purposes. As described by Alon Halevy:

> 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 modelling goal [11].

The benefits to the end-user in conquering data integration have been demonstrated in PIM prototype systems that employ integrative data models. For example, the uniform query capabilities provided by SEMEX [3] and SIS [7], show the speed, efficiency and re-findability gained when a user is able to quickly trace and cross-reference information about people, places and things mentioned throughout files, folders, e-mail, and other information repositories. Similarly, Haystack [14] demonstrates that, when a single, consolidated data model is used across applications, the results include increased efficiency, the elimination of information fragmentation, and reduced effort, with the user being able to view and re-use information easily in multiple task contexts.

Because of the multi-faceted challenges of data integration, contrasting approaches have been investigated in different research communities. In the fields of database integration and the Semantic Web [17], for example, automatic, machine-learning approaches, at either the level of *ontology/schema matching* [9, 5], or *instance matching* [18, 4] have been pursued. In contrast, the end-user programming community has preferred user-driven approaches, in which the user orchestrates the process of reconciliation at various levels of specificity. Systems that use this approach include "mash-up makers" (such as Mashmaker, [8], Marmite [16], and Vegemite [19]) and visual programming language environments such as Yahoo Pipes [10].

DataPalette's approach exhibits similarities with the user-driven methods just mentioned, in that the user performs data reconciliation using concrete instances as examples, but also avoids the need for any sort of programming. Previous systems utilising this approach have included "Cards, Relations and Templates" [6] and Potluck [15] both of which permit the user to reconcile instances from different schemas into a uniform representation using drag-and-drop gestures to specify relations among fields. DataPalette extends this approach to a data workbench environment which allows people to easily consolidate and compare instances by their properties.

## PRE-STUDIES: IDENTIFYING NEEDS AND CHALLENGES

In the first pre-study, we aimed to get an updated understanding of the use of multiple information sources in information gathering on the Web. Several basic questions drove our inquiry. First, were people increasingly relying only on a handful of "super-sites" (Facebook, Wikipedia), or searching across distributed sources of information? If the former were true, merely integrating PIM tools with the small number of super-sites would provide greater benefit than tackling the more challenging problem of integrating data from an array of arbitrary sources on the Web. In a related vein, why did some people choose a single source for information gathering, while others selected multiple sources? What advantages were gained in consulting multiple sources?

The second pre-study focused on the characteristics of the data available from the selected sites. The purpose of this exercise was not to identify specific characteristics of particular sites, but to establish general characteristics across a variety of domains, so that typical integration problems that might arise when mixing these data could be identified. In the next two sections, we present our methodologies and the results of our pre-studies.

### PS.1 - Understanding Data Diversity in Everyday Tasks

In the first pre-study, we held semi-structured interviews to better understand the reasons that people drew information from single or diverse information sources. We interviewed 8 participants, asking them to identify tasks they had performed recently, which had required the use of multiple information sources. This was followed by determining the kinds of sources which were used, and the tools that were needed to manage the resulting information. We then inquired how each participant would go about planning a hypothetical social event entailing details such as scheduling a date, finding a suitable location and selecting the appropriate entertainment.

#### *Results - Heterogeneous Data Tasks*

The 8 participants, recruited informally via word of mouth, consisted of 7 males and 1 female, ranging in ages of 18–32.

All participants reported regular usage of multiple websites to accomplish tasks, examples of which included: shopping, choosing a restaurant, job-seeking, selecting a university, seeking a recipe, or finding answers to technical problems. When initiating a task, it was common for participants to use Google to discover several related websites, which would then be explored in different ways, depending on the task at hand. For shopping-oriented tasks, a balance of price, quality of merchandise, and speed of delivery was sought. For product reviews, single sources alone were not trusted due to possible biases, or incomplete coverage of products' desired features. To benefit from the fullest coverage of information, many different websites of different types were consulted, such as manufacturers' websites for technical details, and review aggregates for a range of opinions.

In response to our question of how participants would plan a large social event, most stated they would first confer with friends to get their ideas, preferences and recommendations. Then, they would select the venue, locations, restaurants and activities through a number of sources: their own prior experience, friends' recommendations, searching Google maps, and dedicated reviews sites (such as Yelp for restaurants). The participants cited the cost and ease of accessing a location as the most important factors when choosing the venue.

### Summary of Findings

All of the participants used multiple websites in order to complete their tasks, and all felt that the lack of existing integration between sites hindered their ability to make informed choices. Participants prioritised the sources they used by both the ease and immediacy of accessing them. As a result, sources that took much effort or time to investigate were often excluded.

## PS.2 - Technical Challenges of Data Integration

For Pre-Study 2, our analysis of structured data feeds started with identifying a set of candidate sources to examine. We consulted the ProgrammableWeb API directory[1] list of most popular data feeds for this purpose. To examine the many varieties of personal data sources, we selected from 5 categories: social network services, retailers, online event calendars, music sites, and weather, selecting 2–5 sources from each, for a total of 20.

For each source, 3–5 typical records of a particular type were obtained from each service's API or feed. For social networking sites, user profiles were the type of record examined, while for retail sites, it was product information; for event tracking sites, we focused on event time/date information; for music sites, on song listing information; and for weather sites, forecast records were selected for our study. For each such data record, two complexity metrics were computed: the *average width*, corresponding to the number of properties, and the *average depth*, corresponding to the number of nested structures. In addition, a third metric measured degree of overlap among the sources in each category. This was accomplished by first mapping *equivalent properties* between schemas, performed manually by creating alignment tables for all properties of data sources in each category. Each row comprised a property of a record of one data source, and all its closest matching properties from the others. We used both property names and attribute descriptions (in documentation, if available) to determine which properties were semantically equivalent.

Once established, equivalent properties were examined, first, for naming inconsistencies. Then, disregarding names, we examined whether the property's values were structurally compatible. That is, if both values consisted of the same literal data type, or had one or more equivalent subfields, they were considered compatible. Finally, we examined incompatibilities not covered by either naming or structural differences alone. These *modelling differences* arose from a variety of reasons, including, but not limited to, measurement unit inconsistencies, measurement method disagreements (e.g., "dew point" versus "relative humidity"), differences in scale and granularity, and mismatches in what was being modelled (e.g., "offer" versus "listing").

### Results

While most of the data records examined in PS.2 were small and relatively simple, nearly one in each category was substantially larger and more complex. In particular, Amazon, Soundcloud, Twitter and Weather Underground were the most complex in each of their respective categories. Mapping equivalent properties among the instance records for each category revealed little overall overlap, except among the smallest records (which had few properties to begin with). This finding suggested that consolidating such records would substantially increase the amount of information held by each. Some of the unique properties were primarily for internal use (such as service-specific IDs). Some, however, contained useful properties which were simply absent in alternate sources. For example, among the 4 music data providers examined, only Spotify listed the source album that featured a particular song. Such omissions revealed that data APIs often were reflective of each provider's own specific needs, necessitating the integration of information across multiple sources to create more complete and useful representations.

A second observation in PS.2 was that equivalent properties were only rarely given the same name. Thus, terminological (or naming) heterogeneity was prevalent across nearly all records observed. Structural inconsistencies were rarer, with the most common case occurring where a record that represented a value as a simple literal was expanded out in another as an entire sub-structure. This most commonly occurred for fields with common string serialisations, such as dates, prices, and intervals. Meanwhile, examples of modelling differences were even rarer than those of structural heterogeneity, with the few we found mostly being attributed to concepts from one source not mapping perfectly to those of another source, for example, due to differences in methods of measurement, or measurement units.

To examine value consistency, we compared source schemas provided in the API and feed documentation of the data providers. There was substantial terminological inconsis-

[1] ProgrammableWeb: **http://www.programmableweb.com/apis/directory**

tency among the value names for enumerated types, with exact matches occurring less than 10% of examined value pairs. There were also range inconsistencies, corresponding to cases where values represented in one record were not represented as valid values in the range of the equivalent property in the target schema. A simple example of this was the "gender" property on Facebook, which allowed only "male" and "female" values, while Google+ included a third option,"other".

In summary, we found in our PS.2, that while data records ranged in complexity, no single source subsumed any others; all contained unique properties. Moreover, the most common types of heterogeneity exhibited were terminological and value-related, which occurred vastly more often than structural or modelling issues.

## DATAPALETTE : AN INTERFACE FOR DATA MIXING

The primary goal of DataPalette (henceforth DP) was to empower end-users so that they were able to mix arbitrary, structured data feeds on-the-fly, while solving particular information task(s) more effectively. Unlike mashup makers, such as Yahoo Pipes!, the aim was to facilitate such integration without the need for a separate step involving programming, scraping, shaping, mapping or integration.

The two pre-studies above provided valuable insights towards this goal, which, in turn, informed our design process. PS.1 confirmed the need for data integration, by showing that people preferred to rely upon multiple sources of information for making important decisions; PS.1 also identified the kinds of sources most commonly used. PS.2, meanwhile, revealed that the most common kinds of heterogeneity across data sources were of the terminological or structural varieties. As these were the "simpler" kinds of heterogeneity, this gave us optimism that an interaction-based approach would be feasible.

In the following sections, we describe the design process of creating the prototype of DP. This is followed by detailed descriptions of the prototype, with justifications grounded in the pre-studies.

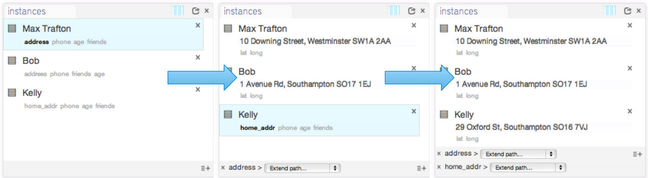
### Design Process

We followed a five-phase spiral model, consisting of planning, designing, mocking-up, prototyping, and testing, to derive the final design of DP. This multi-phase process was used, due to the high degree of initial uncertainty associated with how users would achieve effective data mixing. We approached this design challenge through alternative generation and elimination; alternatives were designed, drawn up and evaluated by colleagues (and in later iterations, non-expert potential users), who critiqued the alternatives. The ease with which designs could be prototyped in HTML5 with modern web frameworks allowed us to perform this process of develop-test-iterate rapidly through seven iterations.

In order to allow new users to be comfortable with our interface, we adopted the look and feel from OS X Finder-style file managers. In the DP workspace, shown in Figure 1, the user would be able to freely drag and drop entities from data sources (on the left) into windows that they have created, resized, arranged and labeled. Unlike file managers, the basic unit of data was not a file, but an *instance* — corresponding to a single data record, JSON object, RDF instance, or any small bit of structured data. In order to make working with relational graphs manageable, an early decision we made was to break up RDF-like relational data into discrete instances with properties. This was done at time of import, and could be done for every major Web data feed type (JSON, XML, RDF) without any loss of generality.

*Multi-path selection: link-sliding for heterogeneous sets*

Typically, users are interested in viewing and comparing the *properties* of instances; for example, a user might be keen to examine a product's price, rating, manufacturer and so on. To support such comparisons, an instance's properties would be displayed beneath it; selecting a property would create a *path selection*, which would dereference all of the instances in a group that contained the selected property and displayed their corresponding values. For example, when a user selected the *manufacturer* property for a product, all products in the same group with a manufacturer property would be dereferenced, causing the manufacturers to be displayed alongside their corresponding products. When such a path selection was applied, the properties displayed were updated to be the set of properties of the *terminating value of the path*, so that this value could be dereferenced further. Having selected a manufacturer, a user might want to be informed of its reputation. By selecting the "reputation score" property, the previous path selection would then be extended. This process is illustrated in Figure 2.


instances
address phone age friends
address phone friends age
home_addr phone age friends
lat long
x address > Extend path...
x home_addr > Extend path...

Max Trafton
Bob
Kelly

Max Trafton
10 Downing Street, Westminster SW1A 2AA
Bob
1 Avenue Rd, Southampton SO17 1EJ
Kelly

Max Trafton
10 Downing Street, Westminster SW1A 2AA
Bob
1 Avenue Rd, Southampton SO17 1EJ
Kelly
29 Oxford St, Southampton SO16 7VJ

**Figure 2. To see properties, users click on the name of the property. When instances do not have a particular property, additional property names can be added.**

When instances could not be dereferenced by the chosen path selection, an additional path selection could be created. As with the first path selection, this would cause all matching instances to be dereferenced. Each group would be able to carry an arbitrary number of path selections, and each path selection could be extended continually by selecting successive properties of values. This would achieve the ability to *link slide* multiple instances simultaneously, as introduced by Parallax [13], extended to support sets of heterogeneous items through multiple parallel simultaneous paths. Multiple-path group-dereferencing would result in two key benefits. First, users would be able to quickly compare all instances that have the same structure, and second, instances with different property names would be easily consolidated.

*Coreference consolidation: Drag-and-Drop Same-as*

A major challenge with integration from multiple sources is coreference consolidation, or combining into one, representations that co-refer to the same entity. For example, when consolidating one's friends from Facebook, Google+, and

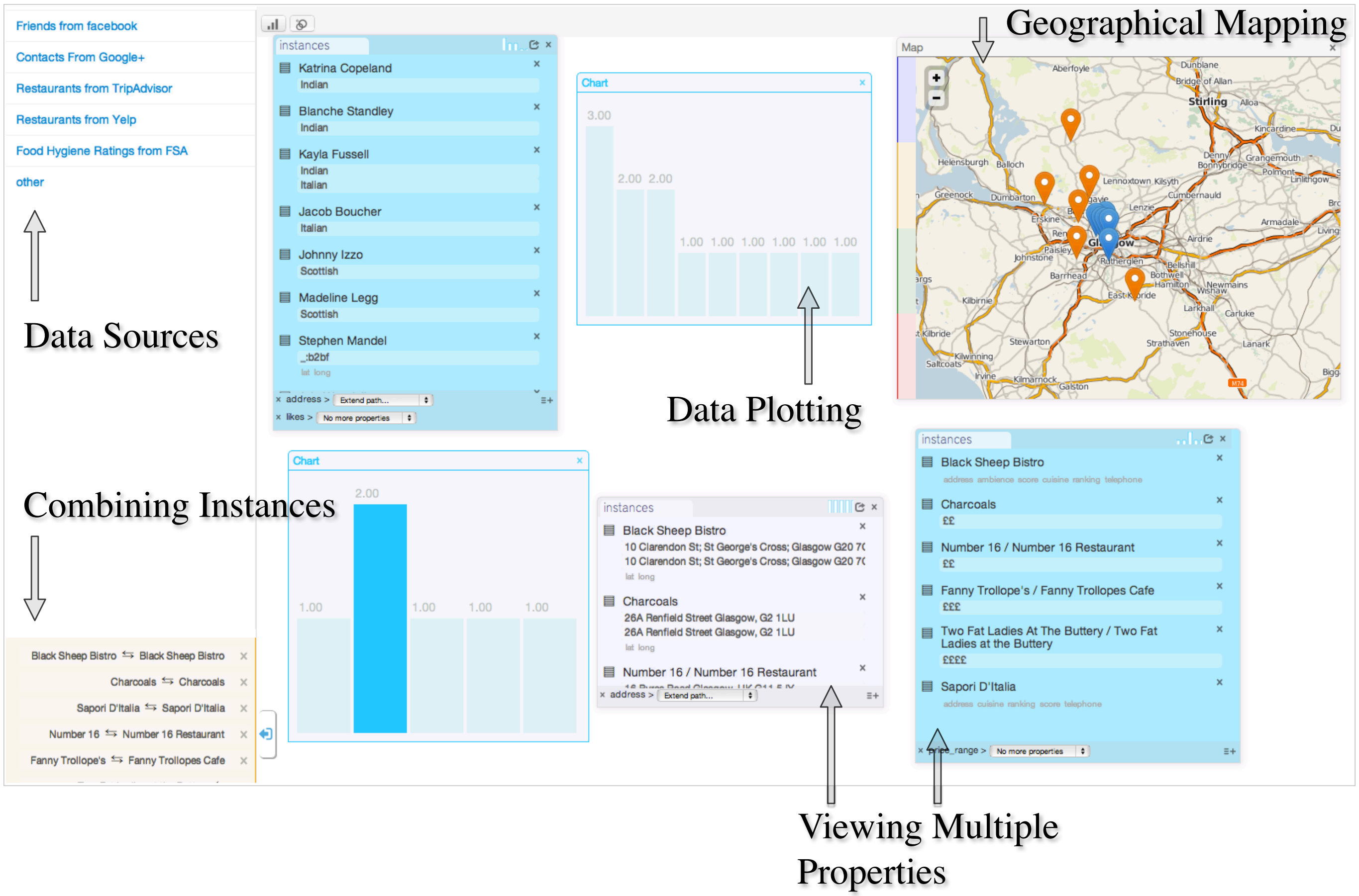

Geographical Mapping
Data Sources
Data Plotting
Combining Instances
Viewing Multiple Properties
Friends from facebook
Contacts From Google+
Restaurants from TripAdvisor
Restaurants from Yelp
Food Hygiene Ratings from FSA
other
instances
Katrina Copeland
Indian
Blanche Standley
Indian
Kayla Fussell
Indian
Italian
Jacob Boucher
Italian
Johnny Izzo
Scottish
Madeline Legg
Scottish
Stephen Mandel
_:b2bf
address >
Extend path...
likes >
No more properties
Chart
3.00
2.00
2.00
1.00
1.00
1.00
1.00
1.00
1.00
Map
Black Sheep Bistro ⇆ Black Sheep Bistro
Charcoals ⇆ Charcoals
Sapori D'Italia ⇆ Sapori D'Italia
Number 16 ⇆ Number 16 Restaurant
Fanny Trollope's ⇆ Fanny Trollopes Cafe
Chart
2.00
1.00
1.00
1.00
1.00
instances
Black Sheep Bistro
10 Clarendon St; St George's Cross; Glasgow G20 7(
10 Clarendon St; St George's Cross; Glasgow G20 7(
lat long
Charcoals
26A Renfield Street Glasgow, G2 1LU
26A Renfield Street Glasgow, G2 1LU
lat long
Number 16 / Number 16 Restaurant
address >
Extend path...
instances
Black Sheep Bistro
address ambience score cuisine ranking telephone
Charcoals
££
Number 16 / Number 16 Restaurant
££
Fanny Trollope's / Fanny Trollopes Cafe
£££
Two Fat Ladies At The Buttery / Two Fat Ladies at the Buttery
££££
Sapori D'Italia
address cuisine ranking score telephone
price_range >
No more properties

**Figure 1. DataPalette Workspace, a file manager-inspired Drag and Drop interface for structured data mixing. Collapsible displays of data sources are listed on the left, from which instances or entire sets of items can be dragged to form a window (group) on the workspace.**

LinkedIn, one might wish to eliminate duplication and combine friends' profiles across multiple services. Although we considered ways to do this automatically, PS.2 revealed no viable method to reliably identify the co-referring instances; due to a lack of standard URIs across social networks, identifiers used by each network were system-specific and inconsistent. Thus, we sought to instead make it user-driven, and easy to perform manually.

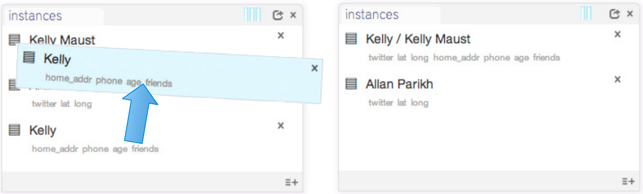


**Figure 3.** ***Coreference consolidation with drag-and-drop Same-As*** **- When two instances represent the same entity, users can drag one on top of the other, and they are combined - effectively displaying the union of properties of both source entities. This can be un-done by deleting the Same-As relationship in the view in the lower-left corner of the workspace.**

To remain consistent with the drag-and-drop interactions used throughout the interface, we introduced the ability to simply drag one instance directly on top of another, to specify that these two items should be combined. This action also caused the *Same-As display* in the lower left of the screen to indicate this new combined relationship. The display also acted as a mechanism to delete/undo combinations, which reverted representations of instances back to their original, separate forms. An illustration of this interaction is provided in Figure 3. Since individually dragging and dropping items could be tedious for large groups of items, we extended this to allow entire groups to be dropped on to other groups. DP would then identify matching pairs among items in the dragged and target groups by comparing the labels of items in each group, respectively. Only exact matches (excepting white space and underscores) were considered to correspond to label matches, and these were automatically combined. The rest were added to the target group as distinct elements, which could be manually combined by hand.

### *Enumerated-type value consolidation*

PS.2 revealed that enumerated values of properties were often inconsistent, e.g., "Casual dining" vs "Relaxed" for a restaurant's "atmosphere" property. To reconcile and consolidate corresponding values from different data sources, we extended the drag and drop support of co-reference combination to also allow users to drag and drop to combine enumerated literal values as well.

*"Do What I Mean" Visualisation*

To make it easier for people to visually compare aspects of instances, we created a charting tool and mapping feature. To ensure that these tools were suitable for rapid utilisation, (e.g., data exploration), they were designed to automatically configure display parameters based on the instances or groups that were dropped onto them. In the current prototype, this "Do What I Mean" behaviour automatically would set whether the plot was a histogram representation (counts of values) or a numeric value display, based upon the types of the dereferenced values being plotted. Hence, non-numeric values would be shown on a histogram, while numeric values were displayed as a bar or line chart. Similarly, the map was made to detect address strings and latitude/longitude pairs, geocoding them where appropriate, and determining the optimal bounding box to display all items at maximum zoom.

*Model-based brushing*

To avoid the confusion of having a single instance represented across several groups, we provided *universal brushing* across the interface. Hovering over any representation of any instance (an instance block, a map marker or a histogram bar), would cause all other visible representations of that instance to lightly glow, enabling the user to instantly identify all the places that a single instance was represented.

## EVALUATION

In order to determine whether or not the interaction affordances introduced in DP allowed users to perform serendipitous data mixing on real data sources, we devised a within-subjects controlled lab study, which we describe next.

### Methodology

We began with three hypotheses pertaining to people's ability to use DP, and its impact on task performance:

**h1. Usability -** People understand how to use DP.

**h2. Data Integration -** People can effectively mix heterogeneous data with DP.

**h3. Task completion -** Use of DP improves people's ability to perform the task.

*Study and task design*

To test these hypotheses, we designed a within-subjects (repeated measures) controlled lab study in which participants were asked to perform predefined subjective-choice tasks in two separate timed trials, once with the DP interface (condition A) and once with a baseline interface (condition B). Each task was paired with one of two datasets, depending on condition, which were counterbalanced in a full-factorial design to eliminate potential ordering, task, and dataset biases.

The participants were seated at a standard OS X desktop equipped with a 22-inch monitor with a standard keyboard and mouse. Each participant was given a maximum of 10 minutes to complete each of two tasks, but were told that if they finished early, they could inform us and move on to the next condition. Prior to the A condition trials, participants were trained on how to use DP using a five-minute instructional video.

In the control (B) condition trials, participants were given a web browser with tabs open to all of the data sources they were provided for use, as well as a spreadsheet containing the data taken from websites pertaining to the candidates they were asked to choose from. Participants were told they were free to use the websites and spreadsheets however they wished.

We established that our two tasks would entail choosing a university and selecting a restaurant for a large social event. These tasks were selected because in PS.1, the majority of people stated that they had used multiple websites for such tasks. We introduced a dependent *dataset* variable for each task. For the University selection task, this variable corresponded to the course for which to select a University, between history and sports science, and corresponding university candidates for each course. For the restaurant selection task, the variable corresponded to the city and respective restaurants located therein; for this experiment, Glasgow and Cambridge were chosen because they were geographically distant from participants, and would therefore minimise prior familiarity.

For each of the two tasks, the final candidates from which participants made their selection were pre-selected by us. For the university task, the six top-ranked universities for each subject according to The Complete University Guide were selected as candidates. We selected six in order to limit the difficulty of the task given the fixed time allocated for each trial. In addition, for each course, we provided participants with the profile of a hypothetical student applying to the course, for whom they should base their choice. This profile included the student's entrance admissions tests scores (corresponding to A/AS-levels), location of the student's hometown, and personal preferences pertaining to how far they wanted to be from home, tuition they could afford, gender balance, and social-athletic-academic balance.

For the restaurant task, we selected six restaurants at random from Yelp's list of popular eateries for each of the two cities. In addition, for each of the two cities, we generated 12 social network profiles for hypothetical friends who lived in either city, to accompany the participant to the restaurant. Each friend's profile expressed their street address and preference for favourite cuisine type which could be taken into consideration by participants during the venue selection process.

For the B conditions, we prepared a spreadsheet listing key statistics for each of the relevant candidates (e.g., universities or restaurants) using data from of a pre-selected set of 6 popular data sources (university and restaurant guides), and collated them into Excel spreadsheets.

### Data Collection and Analysis

Each study was overseen by a facilitator and an observer: the role of the facilitator was to explain the study's protocols to the participant and to answer any questions; the role of the observer was to monitor the study and record observations. Both the facilitator and observer were trained on the purpose of the study, their respective roles, and how they could or couldn't influence the study. Explicitly, the facilitator and the observer were trained to identify processes that participants might use during a task, so that they could thematically categorise them. Initially, we identified the following themes: organisation of data, eliminating candidates, co-referencing, and visualisations. The studies were conducted over a period of 4 days. At the end of each day, the facilitators and observers met to discuss the studies' data and to revisit protocols, if necessary. During the lab study, we recorded the audio and the participant's actions on screen. We asked the participants to follow a *think-aloud* protocol as they worked, so that we could understand the reasoning behind their actions. At the end of the study, the participant completed a short exit survey.

After the studies were complete, we generated transcripts of comments made by participants during the think-aloud protocols. At the same time, we created a spreadsheet summarising quantitative metrics pertaining to how much of the various features of DP were used. After the transcriptions were completed, we categorised comments made by the participants during the study, together with suggestions from their exit survey, and clustered them to common themes.

### Participant Recruitment

We recruited participants through adverts posted near the Southampton University campus and via e-mails to student mailing lists. We screened participants to ensure they were at least 18 years of age, but did not filter candidates based on any other criteria. We received 26 responses to our call, from which we accepted the first 20 applicants (10 of whom were female). Seven were non-students, comprised of university staff and alumni; the remainder were students, 8 studying computer science, and 5 studying other subjects. Due to the large population of international students and staff at our university, we ended up with a large sample (12) of non-native English speakers; however, since all had passed requisite proficiency tests, we did not think this would substantially impact the results. On average, each study took 40 minutes to complete, after which participants were offered a small gift certificate from an online retailer for their time.

## RESULTS

### Task performance

The following four sections describe the metrics we used to measure task performance.

*Efficiency: Time taken*

Participants spent an average of 8.9 minutes performing the tasks. Condition A (DP) took slightly less time ($M = 8.7$min, $SD = 1.79$) per trial than condition B ($M = 9.0$min; $SD = 1.44$); however, a 2-way repeated measures ANOVA test of time taken by interface condition, blocked by participant ID, revealed that interface condition did not have a significant effect on task completion time. This analysis also showed no significant effect of participant ID on task completion. Comparing tasks, the restaurant selection task took on average slightly less time ($M = 8.8$min,$SD = 1.58$) than the university trials ($M = 8.9$min,$SD = 1.66$), but this difference was not significant.

*Thoroughness: Factors weighed in final choice*

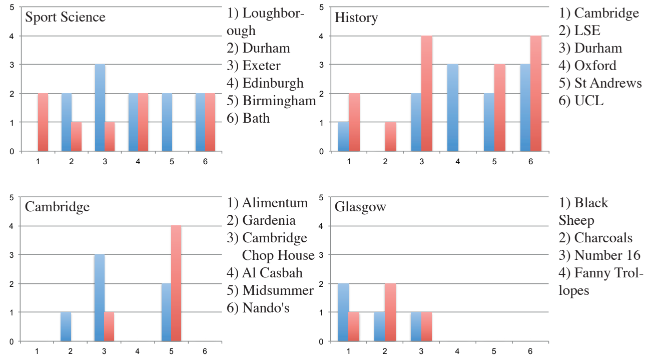


**Figure 4.** ***Choice picks per participant trial*** **- Histogram of number of times each restaurant/university was chosen in DP interface (A condition) in blue/left bars, compared to the Excel/website (B condition) in red/to the right.**

The second metric related to the number of factors each participant considered in making his or her final choice. To determine this, we asked participants to explain why they made their choice(s), and recorded the number of distinct data dimensions mentioned. A 2-way ANOVA test of the effect of interface and participant ID on the number of factors mentioned revealed a significant effect; post-hoc analysis with Tukey-Scheffé adjustments indicated that participants in the DP condition mentioned significantly more factors ($M = 5.5; SD = 1.7$) than those in the control condition ($M = 4; SD = 1.73$), $F(1, 19) = 3.95; p < 0.10$. A strong significant effect was also observed between participant ID and number of dimensions mentioned, meaning some participants mentioned significantly more factors than others ($F(19, 19) = 4.53; p < 0.001$).

*Diversity: Data sources consulted*

The third metric we employed to gauge the task performance was the number of different data sources consulted during the trial; our rationale for this metric was that more informed decisions derived from more thorough consideration of available data. In this metric, participants used an average of 88% ($SD = 0.13$) of data sources provided in the DP case, compared to 80% ($SD = 0.18$) in the control condition, although an ANOVA test blocking on participant ID demonstrated that interface choice only approached significance ($F(1, 19) = 0.04\ p < 0.15$).

*Effort: External cognition*

Although we did not have an explicit measure of effort for this experiment (such as NASA's TLX [12]) the use of paper notes provided an indicator of how much information participants had to keep track of during the study. We found that participants took notes less on average (25%) with the DP interface than the standard interface (35%), a difference which approached significance in a 2-way ANOVA blocked by participant ID ($F(1, 19) = 8.22; p \approx 0.10$).

*Candidate selection*

Since the tasks were subjective-choice, we could not evaluate the 'correctness' of answers. However, to determine whether there was a difference in variability of answers between interface conditions, we plotted the choices on a histogram for each trial, per task and dataset for both conditions, visible in Figure 4. As can be seen, there was less agreement among universities than restaurants, although little agreement overall in both conditions. This is likely due to the many subjective factors involved in such choice tasks.

### Strategies: Successive Elimination vs Tallying

As described later in this section, we noticed several different strategies the participants used to evaluate each candidate selection. One common strategy was *successive elimination*, that is, to regard a single dimension or factor at a time, starting with the most important, and ruling out candidates not meeting the minimum requirements for that value. However, when there was no clear aspect that was "most important", this "greedy" strategy can result in suboptimal decisions. Another tactic, which we called *tallying the pros and cons*, was to consider all candidate choices together, totalling up the advantages and disadvantages of each, potentially weighed by its perceived importance. This strategy was less vulnerable to getting stuck in local maxima than the former.

To understand whether the interface influenced the choice of strategy, we measured the number of candidates the participant considered at their final choice. This was a strong marker for use of each strategy; people who used the successive elimination strategy arrived with only 1–2 candidates at decision time, while those that tallied still maintained the entire starting set.

We found that the use of DP strongly influenced people to keep all candidates around, while participants in the B condition eliminated choices early. An ANOVA test demonstrated a significant effect between condition and number of candidates maintained; a post-hoc analysis with Tukey-Scheffé adjustments confirmed that the number of candidates used in the final choice for participants in the DP condition was greater ($M = 5.85; SD = 0.45$) than the control condition ($M = 4.95; SD = 2.26$) $F(1, 19) = 5.323; p < 0.05$.

### Condition A: Use of Data Integration Features

To determine whether participants used the data integration features of DP, we looked at use of the *multi-path selection* and *Same-As* features of the system. Pertaining to path selection, all participants readily used the path selection process to select values. Five participants (25%) deliberately used multi-path selection, defined as selecting more than one active path per group at the same time to display common values across heterogeneous items. (We did not take into account accidental uses of multi-path, such as when a participant had an already active path and wanted to switch to a separate path accidentally).

Participants used the drag-and-drop *Same-As* capability extensively. Sixteen (80%) used *Same-As* as least once; among these participants, they used *Same-As* an average of 4.6 times per trial. We counted a single "use" as a single drag-and-drop of an individual item, or a drag-and-drop operation of an entire group onto another (we did not differentiate enumerated value consolidation from entity consolidation, as these are not discernibly different from the interface perspective).

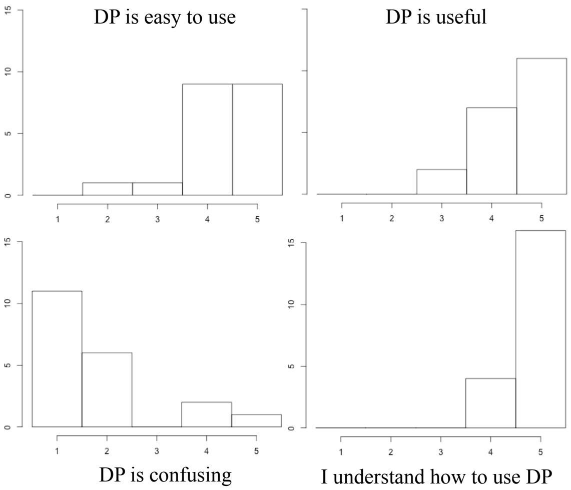


**Figure 5.** ***Survey results*** **- Answers to questions on a 5-point Likert scale, from 1-*strongly disagree* to 5-*strongly agree*.**

### Use of Visualisation Features

In our determination of the extent to which visualisation tools were used in condition A, 13 (65%) used charting tools at least once, while a majority (15, 75%) used the map. To assess the impact use of these features had on performance, we carried out a multiple regression on the time taken, with the number of uses of charting and mapping as variables. We found significant effects of both charting and mapping variables on time taken, ($R^2_{adj} = 0.04; F(2, 17) = 3.79; p < 0.05$), with coefficients ($Y_0 = 8.91; \beta_{charts} = 0.49; \beta_{maps} = -1.05$) demonstrating that charting positively influenced time taken, while the use of maps led to shorter times.

### Condition B: Results and Observations

In condition B, participants were given the option of using websites and/or a spreadsheet containing the same data as condition A, to make their choices. A mixed usage of websites and the spreadsheet was observed, and we measured the fraction of the trial time spent in Excel versus looking at the sites in a 5-value range (0,25,50,75,100%). Six individuals used Excel entirely (100%) without consulting the websites, while one avoided Excel entirely. The uses of Excel for external cognition ranged from collation behaviour, to tallying annotations, to colour-coding cells for making easier the identification of "pros" and "cons". In general, users struggled with both the websites and with the spreadsheet, although they could usually find the information they required after a short period of frustration.

### Survey Results

When asked to rate the DP interface in the exit survey, all participants responded that they felt they understood the system, with sixteen out of the twenty participants (80%) responding that they agreed *very strongly*. Eighteen (90%) reported that they thought DP was *easy*, as well as *useful*, with one person being neutral on each. As to the statement "DP is confusing", 11 (55%) strongly disagreed, 6 (30%) disagreed, and 3 (15%) agreed. On reviewing the time trials, we discovered that the

3 participants who rated DP as confusing completed the trials significantly more slowly than the rest of the group. (See Figure 5.)

The users also noted that they would use an interface like DataPalette in the future, for example, in finding a property to rent/buy, in choosing a job and for purchasing electronic devices. Interestingly, one participant indicated that he/she would not use it for decisions such as selecting a restaurant, because "it would cheapen the experience" by removing the element of spontaneity of choice; they would use it, however, for making more "important" decisions. As for ways to improve DP, the participants suggested making the origin of the information clearer, particularly when combining multiple data sources. The participants also requested the ability to sort instances, as well as explanations of ratings and scoring systems (however some of these are not even transparent on websites). Participants also requested standard features not present in our prototype, such as window management options (specifically, arranging and minimising windows), and the ability to undo operations.

## DISCUSSION AND LIMITATIONS

In this section, we first revisit the three hypotheses introduced in Methodology, incorporating observations from the study's results to support views on each. We follow this with a discussion of the limitations regarding the design of the interface, the state of the prototype, and the design and execution of the study. Finally, we discuss our current and follow-on plans for continuing this line of research.

### Does DataPalette enable data integration?

We set out to test three hypotheses with the DataPalette study. The first was that users would understand the interface mechanisms of DP (h1). The second, h2, was that DP would enable people to mix and integrate heterogeneous data in the process of performing their tasks. The third, h3, was that using DP would facilitate task completion.

The study produced substantial support for h1. All participants reported that they understood the tool, and that it was easy to use. All participants managed to use the interface to view, organise and collate data without running into major roadblocks or confusion, and all effectively were able to compare multiple attributes of heterogeneous data items directly. More than half of the participants used DP's visualisation features (charting or map) – often several times during the trial – suggesting that these features were useful and usable as well.

Additional evidence that participants had a solid understanding of the system was reflected in their feedback, particularly in their requests for features and desired capabilities. Several participants requested search functionality, sort and filter functionality, the ability to display multiple properties for a single instance simultaneously, and the ability to display multiple attributes together in a 2-or-3 dimensional visualisation. Perhaps the most interesting suggestion was also the most common – the ability to selectively view the provenance of information after instances were combined. For example, after his trial, P15 said:

> Properties are tricky. Sometimes you don't care [which data source] a property is from, like "address" or "phone number" - for these the way it's done now is fine. But in some cases you need context of where the properties come from in order to know what it really means - like for "rating", it makes a big difference whether you're talking about "Yelp rating" or some random reviewer's rating. You also don't know what typical ratings are, whether 5 stars are much better than 4, etc.

Pertaining to h2, all participants were able to work with multiple sets of heterogeneous data effectively. With respect to integration specifically, a majority (80%) of the participants successfully and deliberately integrated data using drag-and-drop SameAs capabilities. While single-paths were readily used, multi-path features were not as widely used (only by 4 participants). We believe that participants may not have realised that creating multiple path selections per group was possible, instead assuming it to be similar to the single "Current path" state per window of most file managers.

Assessing whether or not DP improved task performance (h3) was difficult, given the small sample size of our study, and the large number of factors influencing each person's choice. Tasks were completed on average slightly faster than the control interface, but not significantly. However, more data sources seemed to be consulted (we use "seemed" carefully because these findings approached significance) during trials with DP over the control interface, which meant that participants were looking at more diverse information than the control conditions. We also found that participants justified their choices with a greater number of factors in DP than with the control interface, suggesting that their decisions may have been made considering a greater number of factors. Participants entertained more possibilities for longer in the DP trial, while they were more prone to eliminate candidates early in the baseline interface. Finally, people took fewer paper notes in the DP condition than with the baseline, suggesting that there was less need for external cognitive support in DP.

### Design Limitations

In addition to the aforementioned feature requests, we observed a number of potential areas for improvement in the design of the DP interface. One such area pertained to window layout and screen real-estate management: participants were prone to opening up a large number of windows, including maps and charts. When their workspaces started to fill up, we found that participants spent a lot of time arranging windows, and that they tended to forget what each window represented. We plan to address these issues in the next version of DP.

A key design decision we made at the outset was to target only the most common forms of heterogeneity observed in PS.2. To expand DP's integration capabilities, we are considering a number of new interaction affordances for facilitating field combination and splitting, and unit of measure reconciliation. This would cover a majority of the structural and modelling-oriented heterogeneity cases observed in PS.2. Extending the "Do What I Mean" capabilities of the visualisation features to be unit-of-measure-aware, we believe, would be valuable to users. A core problem with realising such an capability, however, is that few sources explicitly annotated data values with measurement units.

### Study Limitations

Concerning limitations of the study, the small sample size of our study (20 participants) meant that we could not confirm some of the effects observed with statistical significance. Second, providing participants a time limit of 10 minutes on trials may have affected the strategies they used to make their choices. For this reason, in our follow-up study, we will give participants a greater amount of time for tasks of similar complexity. A departure from realistic tasks was the way in which we pre-determined the data sources for participants to use. We did this in order to focus participants' efforts on looking at and working with their choices, rather than retrieving them. However, as supporting effective retrieval and serendipitous discovery will be important for DP to be useful in practice, we plan to look at supporting these stages in the future.

## CONCLUSIONS

As expressed in Voltaire's poem *La Bégueule*, "the best is the enemy of the good", we believe that the quest for fully-automatic, complete approaches to data integration have caused simpler, manual approaches to be overlooked. Based upon the results of our investigation, we have seen a number of benefits to interactive, lightweight approaches over automatic and bespoke methods, at least for a frequently-occurring class of information gathering for decision-making tasks.

In particular, the results suggest that the key affordances for the suitability of such tools are flexibility, simplicity, and ease of use. Just as work in the "user-subjective approach" to PIM (e.g., [1]) showed that personal, user-specified attributes were ultimately the most useful, granting users the flexibility to combine or separate information items as needed allowed the tool to support the varied strategies participants employed to compare and examine their options. Second, the low cost of use (in terms of time and effort required) meant that integration could be done (and undone) as appropriate throughout the "inner loop" of users' exploratory processes. The reason this was significant was that this process itself determined, incrementally, the information and sources that were ultimately considered for the final decision. Therefore, any integration approach that required sources to be specified *a priori*, or that was time or effort-intensive at each integration step, was simply not appropriate for this context.

Due to the promising findings of this study, we plan to continue investigating lightweight, *in-situ* approaches to supporting ad-hoc data integration during the course of sense-making and decision-making activities. In particular, we are examining gestural input methods for the purposes of supporting more fluid exploration and mixing. On the basis of feedback from reviewers, we have also begun considering semi-automatic and crowd-powered methods to help users quickly resolve challenging reconciliation problems they might encounter.

## ACKNOWLEDGEMENTS

The work presented in this paper was supported by the SOCIAM project, funded by the Engineering and Physical Sciences Research Council under contract EP/J017728/1. Dr Packer is grateful for funding from the EPSRC in the form of a Post Doctoral Award. We would like to thank our collaborators Wendy Hall, mc schraefel, Tim Berners-Lee, Kieron O'Hara, Susan Davies, Maraim Elbadri, Tom Smith, Yang Yang, Igor Popov, and Shauna VanKleek whose help and advice have shaped this work throughout.

## REFERENCES

1. Bergman, O., Beyth-marom, R., and Nachmias, R. The user-subjective approach to personal information management systems. *JASIST 54* (2003), 872–878.
2. Bernstein, M., Van Kleek, M., Karger, D., and Schraefel, M. Information scraps: How and why information eludes our personal information management tools. *TOIS 26*, 4 (2008), 24.
3. Cai, Y., Dong, X. L., Halevy, A., Liu, J. M., and Madhavan, J. Personal information management with SEMEX. In *Proc. SIGMOD '05* (2005), 921–923.
4. Castano, S., Ferrara, A., and Montanelli, S. Matching ontologies in open networked systems: Techniques and applications. *Journal on Data Semantics V* (2006), 25–63.
5. Doan, A., Madhavan, J., Dhamankar, R., Domingos, P., and Halevy, A. Learning to match ontologies on the Semantic Web. *VLDB Journal 12*, 4 (2003), 303–319.
6. Dontcheva, M., Drucker, S. M., Salesin, D., and Cohen, M. F. Relations, cards, and search templates: user-guided web data integration and layout. In *Proc. UIST '07* (2007), 61–70.
7. Dumais, S., Cutrell, E., Cadiz, J., Jancke, G., Sarin, R., and Robbins, D. C. Stuff i've seen: a system for personal information retrieval and re-use. In *SIGIR '03*, ACM (2003), 72–79.
8. Ennals, R., Brewer, E., Garofalakis, M., Shadle, M., and Gandhi, P. Intel Mash Maker: join the web. *SIGMOD Rec. 36*, 4 (2007), 27–33.
9. Euzenat, J. An API for ontology alignment. *Proc ISWC '04* (2004), 698–712.
10. Fagan, J. C. Mashing up Multiple Web Feeds Using Yahoo! Pipes. *Computers in Libraries 27*, 10 (2007), 10–17.
11. Halevy, A., Rajaraman, A., and Ordille, J. Data integration: the teenage years. In *Proc. VLDB '06* (2006), 9–16.
12. Hart, S., and Staveland, L. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Human mental workload 1* (1988), 139–183.
13. Huynh, D., and Karger, D. Parallax and companion: Set-based browsing for the data web. In *Proc. WWW '09* (2009).
14. Huynh, D., Karger, D., and Quan, D. Haystack: A Platform for Creating, Organizing and Visualizing In-formation Using RDF, 2002.
15. Huynh, D., Miller, R., and Karger, D. Potluck: Data mash-up tool for casual users. *JWS 6*, 4 (2008), 274–282.
16. Lin, J., Wong, J., Nichols, J., Cypher, A., and Lau, T. A. End-user programming of mashups with vegemite. In *Proc. IUI '09*, ACM (2009), 97–106.
17. Shadbolt, N., Berners-Lee, T., and Hall, W. The Semantic Web Revisited. *IEEE Intelligent Systems 21*, 3 (2006), 96–101.
18. Suchanek, F., Abiteboul, S., and Senellart, P. PARIS: probabilistic alignment of relations, instances, and schema. *Proc. VLDB '11 5*, 3 (2011), 157–168.
19. Wong, J., and Hong, J. I. Making mashups with marmite: towards end-user programming for the web. In *Proc. CHI '07*, ACM (2007), 1435–1444.