Carpé Data: Supporting Serendipitous Data Integration in Personal Information Management

ABSTRACT

While people naturally draw from diverse information sources in the course of routine decision-making tasks, complementary support for managing heterogeneous data in digital Personal Information Management (PIM) tools remains poor. In this paper, we present an initial investigation of user-driven, ad-hoc data integration approaches for PIM, towards enabling greater use of the emerging ecosystems of structured data on the Web. 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. These observations led to our design of 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 understood the interaction mechanisms introduced, and made more carefully-justified decisions during subjective-choice tasks when using DataPalette than when using a control interface, drawing information from more data sources in the process.

Author Keywords

Personal Information Management, Personal Data Management, User-driven data integration

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. This information has emerged from many channels, from the APIs now offered by most of the major retailers, apps, and social media sites, to the massive publishing of datasets by governmental and other institutions, to even the emergence of new kinds of data sources, such as wearable and smartphone-based bio-sensors. A primary goal of making much of this data available has been to give end-user citizens the ability to make more informed decisions pertaining to their health, wealth, and well-being [16]. This data is predominately used by specialised app developers, journalists and other "data specialists" but remains inaccessible to citizens because of the available tools. For example, the most commonly used structured personal information management (PIM) tools either 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 [3].

We hypothesise that end-users will benefit from tools that allow them to effectively browse, use and consume heterogeneous structured data from diverse sources. In this paper, we present an investigation into extending PIM tools to effectively use emerging ecosystems of personal data. We used a three-stage sequential exploratory mixed-method approach: first, we performed a qualitative analysis of a semi-structured interview pre-study to understand the types of tasks people performed using multiple information sources, and the processes that they relied upon to perform such tasks. Our results suggested that people rely on multiple and diverse sources, then singular, integrated ones for a number of reasons, including coverage, reliability, and for alternative perspectives.

Second, we conducted a structural analysis of various popular personal data sources available today to identify potential barriers to unifying data. Drawing upon definitions from the database integration and automatic schema matching literature, we characterised the types of heterogeneity exhibited across data sources in six domains: contacts, events, music, shopping, social networking and weather, 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.

Third, we developed DataPalette, an interface that enables serendipitous "data mixing," negating the need for people to write bespoke code to effectively combine and compare heterogenous data. It was principally designed to support people's data integration needs for data heterogeneity, using observations made from the previous studies. In particular, we designed a user-centric interface to facilitate simple, "light-touch" integration tasks of diverse, heterogeneous data sources. We then performed a usability evaluation of DataPalette, which revealed that most users were comfortable with the interactive integration mechanisms we devised, and that this interface effectively improved people's ability to perform multifaceted decision making tasks using multiple sources.

BACKGROUND

In this section, we motivate our work using three fields: PIM, data integration, and end-user mashups and toolkits. Our primary motivation stems from field studies in PIM, which foregrounded work to consolidate data from multiple sources and which still remains a challenge. Specifically, reconciling the differences that arise among representations when data is created by different people, at different times, 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 [10].

A result of this problem is that users can experience data fragmentation [15] across multiple web-based applications and services, and struggle to consolidate across them [2, 4]. In our own previously-conducted field study, (reference anonymised) we observed variations of manual "coping strategies" (or, if programmers, elaborate, custom-coded one-off solutions) to allow people to manage both online and offline data. Similar findings from other studies observed volunteers' "homebrew databases": manually-maintained information collated to handle the heterogeneous data requirements of groups needing to coordinate their activities [18].

A number of research PIM systems have demonstrated how users benefit from unified, integrated data models. SEMEX's [5] uniform query interface allows users to quickly trace references of people, places and things across files, folders and repositories. Similarly, Haystack allows users to visualise and customise how they worked with their data in a uniform, consolidated manner [13].

The database integration and Semantic Web [16] research communities have primarily focused efforts on purely-automatic approaches to the problem of data integration. Such approaches typically take the form of *ontology/schema-level matching* [9, 7], or *direct/instance-level* matching [17, 6]. We focus on *instance-level* matching, as schemas may not always be available, or easily accessible to users.

PRE-STUDIES: IDENTIFYING NEEDS AND CHALLENGES

We performed an exploratory sequential mixed-method study to help us design our contribution. For our first pre-study, we sought an updated understanding on the use of multiple information sources in information gathering on the Web. There were two driving questions behind this inquiry: first, were people increasingly relying only on a handful of "supersites" (Facebook, Wikipedia), or searching across distributed sources of information? If the former were true, merely integrating PIM tools with the small number of supersites would provide greater benefit than tackling the more challenging problem of integrating data from arbitrary sources on the Web. Second, why did people choose a single or multiple sources for information gathering? If a number of websites were used, why was this preferred over a single source?

The second pre-study focused on the characteristics of the data available from the kinds of sites used by the first pre-study's participants for the tasks described. The purpose of this follow-on was not to identify specific characteristics of particular sites in question, but to identify, in general, the degree of complexity of the data feeds currently available across a variety of domains, and typical integration problems that might arise when mixing this data. In the next two sections, we present our methodologies and results for our pre-studies.

PS.1 - Understanding Data Diversity in Everyday Tasks

For the first pre-study, we conducted a semi-structured interview to understand the reasons and ways that people drew information from diverse information sources. We interviewed 8 participants, asking them to identify tasks they performed recently that required use of multiple information sources, the kinds of sources used, and the tools used to manage the resulting information. We then inquired as to how each participant would go about planning various aspects of a large party, such as date, location and entertainment.

Results - Heterogeneous Data Tasks

The 8 participants, recruited informally via word of mouth, consisted of 7 males and 1 female. They were split evenly into two age groups of 18–25 and 26–32.

All participants reported regular use of multiple websites to accomplish tasks, examples of which including: shopping (food, consumer electronics, hotels and flights), choosing restaurants, job-seeking, selecting a university, seeking recipes, and finding answers to technical problems. When beginning a task, it was common for participants to use Google to discover several related websites, which would then be used in different ways depending on the task at hand. For shopping-oriented tasks, balancing price, quality and speed of delivery was the main aim. For product reviews, single sources were not trusted due to possible bias and incomplete feature coverage. To provide a full coverage of information, websites of different types were used, such as manufacturers websites for technical details, review aggregators for a range of opinions, and Google maps for location-based decisions.

Participants said they would organise a large social event by first conferring with friends about their ideas, preferences and recommendations. They selected venues, locations, restaurants and activities through a number of sources: their own prior experience, friends' recommendations, searching Google maps, dedicated reviews sites (such as Yelp for restaurants), and through web search. 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, or organised carpools.

Summary of Findings

All users experienced using multiple web sites in order to complete a task, and felt that the lack of existing integration between sites hindered their ability to make informed choices. Participants felt that they did not include particular aspects of a decision because the process to investigate them would be too difficult or take too much time.

PS.2 - Technical Challenges of Data Integration

Our analysis of structured data feeds started with identifying a set of candidate sources to examine. We consulted the

	source	average props/record	average depth/record	equivalent to	prop name inconsistent	structural inconsistency	model inconsistency
social networking (profiles)	google+	12	1.52	tw:5, li:4, fb:4	tw:3, li:2, fb:2	tw:0, li:1, fb:1	tw:0, li:0, fb:0
	twitter	39	1.37	li:4, fb:4	li:2, fb:1	li:0, fb:0	li:0, fb:0
	linkedin	18	2.42	fb:7	fb:5	fb:3	fb:0
	facebook	22	1.5				
calendars (events)	google calendar	15	1.46	eb:9	eb:6	eb:1	eb:1
	eventbrite	32	1.43				
retail (products)	milo	5	1.2	z:2, s:3, e:1, a:3	z:2, s:2, e:1, a:3	z:0, s:0, e:0, a:2	z:0, s:0, e:0, a:0
	zappos	6	1	s:2, e:2, a:3	s:2, e:2, a:2	s:0, e:0, a:0	s:0, e:0, a:0
	shopping.com	11	3.74	e:2, a:8	e:2, a:8	e:0, a:0	e:0, a:0
	etsy	31	1.13	a:11	a:10	a:1	a:2
	amazon	108	1.96				
contacts	google contacts	7	1.46	yp:2	yp:2	yp: 0	yp: 0
	yellowpages	25	1				
music (songs/ tracks)	echonest	4	1.83	so:3, r:3, sp:3	so:3, r:3, sp:3	so:0, r:0, sp:0	so:0, r:0, sp:0
	soundcloud	42	2.14	r:12, sp:7	r:11, sp:6	r:1, sp:1	r:1, sp:0
	rdio	27	2	sp: 7	sp: 6	sp: 1	sp: 1
	spotify	10	2.44				
weather (forecasts)	weatherbug	8	2	wu: 5, y: 5	wu: 5, y: 5	wu: 5, y: 4	wu: 0, y: 0
	wundergound	52	1.38	y:12	y:12	y:7	y:3
	yahoo weather	22	1.58				

Table 1. *PS.2 Analysis of Data Feeds*: Structural statistics about data sampled from each API organised by category. The first column lists the category and type of data record that was examined. Columns 3 and 4 list the average number of properties per data record, and number of levels from root to leaf. The remaining fields show results from a pairwise analysis of data records from each of the data sources with those of the others in each category; values are prefixed with the first letter(s) of the target schema the row schema was compared against. For example "Tw:5" in the 5th column of Google+ row indicates that Facebook and Twitter records had 5 equivalent fields.

ProgrammableWeb API directory[1] list of most popular data feeds for this purpose. To increase the variety of 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 feed, 3–5 "typical" records of a single type were obtained; for social networking sites this was user profiles, for retail sites this was product info, etc. For each such data record, two simple complexity metrics were computed: the *average width* and *average depth*. Determining the width required first finding the appropriate level for comparison; for example, Amazon's retail product "Item" root record had few attributes about the actual product compared to its "ItemAttributes" sub-property, which was used instead. The average depth was simply the mean depth to each leaf node from the chosen comparison level.

The third metric measured degree of overlap among sources in each category. This was done by first mapping *equivalent properties* between schemas, which was done 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 make matching decisions.

Once established, equivalent properties were examined firstly for naming inconsistencies. Then, ignoring property names, we examined whether the property's values were structurally compatible. If both values consisted of the data type and same equivalent subfields, they were considered comparable; otherwise they were considered *structurally inconsistent*. Finally, we examined equivalent fields which were incomparable for reasons other than structural or terminological differences. These *modelling differences* arose from a variety of causes, including, but not limited to: measurement unit inconsistencies, measurement method (e.g., dew point vs relative humidity), differences in scale and granularity, and mismatches in what was being modelled ("offer" versus "listing").

Results

Table 1 shows the raw results of the process described. A number of features were quite striking about the data; First, while most of the data records were small and relatively simple, nearly one in each category produced substantially more complex records than the rest. In particular, Amazon, Soundcloud, Twitter and Weather Underground stood out in their respective categories. Looking at equivalent properties, we found overall little overlap among records; that is, except for the smallest records (which had few properties to begin with), most of the data records had more unique properties than ones in common with others. While some of these 'unique records' were primarily for internal, service-specific use (such as service-specific IDs) others were useful properties that were simply absent. Only Spotify, out of the 4 music providers, included information about the album featuring a given song. Such omissions show us that these data APIs reflect each provider's own specific needs, and suggests that integrating information from multiple sources may allow more complete and useful representations to be assembled.

The second finding is that equivalent properties were only rarely given the same name. Thus, terminological heterogeneity was prevalent across properties and records we observed. Structural heterogeneity, was slightly rarer; the most common case consisted of a leaf property from one data source modelled as a subtree in the other. We also encountered two examples where data was nested in textual descrip-

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

tion fields, which hints of their intended use in feed readers rather than in structured data applications. Examples of modelling differences were rarer than those of structural heterogeneity, with the few we found mostly being attributed to concepts on one site not mapping perfectly to those of another site, and differences in units of measurement.

We also examined the source schemas of data sources for enumerated value types. There was substantial terminological inconsistency among such value ranges of such types, but also cases where sources had values not represented in the other source's ranges. For example, "gender" on Facebook allowed only "male" and "female" for values, while Google+ included a third option,"other". Few string representations of enumerated value sets matched across data sources perfectly.

In summary, while data records ranged in complexity, no source subsumed the rest; all contained unique properties. Moreover, the most common types of heterogeneity exhibited were terminological and value-related, which occurred vastly more than structural or modeling issues.

DATAPALETTE : AN INTERFACE FOR DATA MIXING

At a high level, the overall goal of DataPalette was to empower end-users to mix arbitrary structured data feeds on-the-fly while solving their particular information task(s) — be them sense-making or decision-making related — in order to more effectively accomplish these tasks. Unlike Mashup makers, including Yahoo Pipes!, the goal was to facilitate such integration without a need for an explicit separate mapping or integration step. Instead, DataPalette's approach is most directly inspired by "lightweight" visual, user-interactive data collection, consolidation and alignment tools such as Potluck [14] and Cards [8].

The two pre-studies provided valuable insight towards this goal that allowed us to guide 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, and identified the kinds of sources most commonly used. PS.2, meanwhile, revealed that the most common kinds of heterogeneity across data feeds were of the terminological structural varieties. As these were the "simpler" kinds of heterogeneity, this gave us optimism that an interactional approach would be feasible. In the following sections, we describe the design process we used to create the prototype of DataPalette, followed by a detailed description of the design of the prototype and justification.

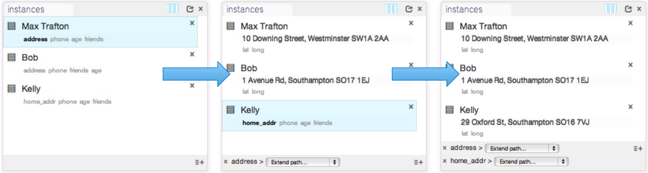
Design Process

We followed a five-phase spiral model consisting of planning, designing, mocking-up, prototyping, and testing to derive the design of DataPalette. This process was used because of the high degree of initial uncertainty associated with how we'd achieve effective data mixing by end-users. 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 develop-test-iterate rapidly through seven iterations of this process.

In order to feel familiar to new users, we borrowed the overall "look and feel" from OS X Finder-style file managers. In the DataPalette workspace, shown in Figure 1, the user can freely drag and drop entities from data sources (on the left) into windows that they create, resize, arrange and label. 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. An instance's properties are displayed beneath it; selecting a property creates a *path selection*, which dereferences all of the instances in a group that contain the selected property and displays their corresponding values. For example, when a user selects the *manufacturer* property for a product, all products in the same group with a manufacturer property will be dereferenced, causing the manufactures to be displayed alongside their corresponding products. When such a path selection is applied, the properties displayed are updated to be the set of properties of the *terminating value of the path*, so that this value can be dereferenced further. For example, once selecting a manufacturer, a user may want to know the product's manufacturer's reputation and selects their "reputation score" property, causing the previous path selection to be extended. This process is illustrated in Figure 2.


instances
Max Trafton
address phone age friends
Bob
address phone friends age
Kelly
home_addr phone age friends

Max Trafton
10 Downing Street, Westminster SW1A 2AA

Bob
1 Avenue Rd, Southampton SO17 1EJ
lat long
Kelly
home_addr phone age friends
× address > Extend path...

Max Trafton
10 Downing Street, Westminster SW1A 2AA

Bob
1 Avenue Rd, Southampton SO17 1EJ

Kelly
29 Oxford St, Southampton SO16 7VJ
lat long
× address > Extend path...
× home_addr > Extend path...

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 cannot be dereferenced by the chosen path selection, an additional path selection can be created. As with the first path selection, it causes all matching instances to be dereferenced. Each group can carry an arbitrary number of path selections, and each path selection can be extended continually by selecting successive properties of values. This achieves the ability to *link slide* as introduced by Parallax [12], but across sets of heterogeneous items by supporting multiple parallel dereferenced paths simultaneously. Multiple selection group-dereferencing results in two key benefits. First, that users can quickly compare all instances that have the same structure, and second, that instances with different property names and be quickly consolidated.

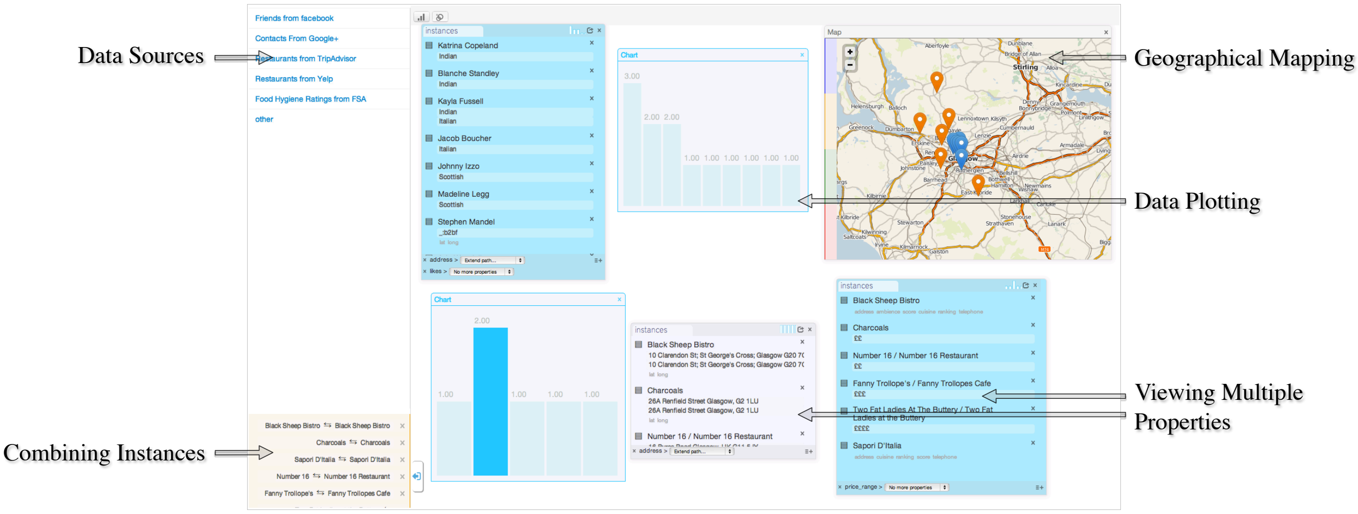

Data Sources
Geographical Mapping
Data Plotting
Viewing Multiple Properties
Combining Instances
Friends from facebook
Contacts From Google+
Restaurants from TripAdvisor
Restaurants from Yelp
Food Hygiene Ratings from FSA
other
instances
Katrina Copeland
Indian
Blanche Standley
Indian
Kayla Fussell
Indian
Italian
Jacob Boucher
Italian
Johnny Izzo
Scottish
Madeline Legg
Scottish
Stephen Mandel
_:b2bf
lat long
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
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
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

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.

Coreference consolidation: Drag-and-Drop Same-as

A major challenge with integration from multiple sources is coreference consolidation, or combining representations that co-refer to the same entity into one. For example, consolidating one's friends from Facebook, Google+, and LinkedIn, one might quickly want to de-duplicate and combine friends that use multiple services so that they each have one representation. Although we considered a way to do this automatically, PS.2 revealed no way to reliably identify 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.

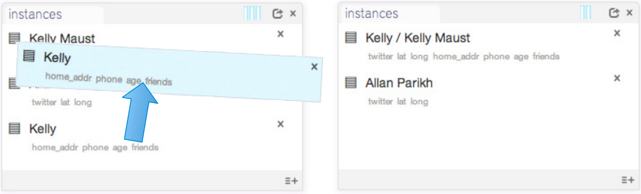


Figure 3. ***Coreference consolidation with drag-and-drop Same-As*** **- When two instances represent the same entity, users can drag one on top of the other, and they are combined - effectively displaying the union of properties of both source entities. This can be un-done by deleting the Same-As relationship 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 on top of another to specify that two things should be combined. This action also caused the *same-as display* in the lower left of the screen to display this new combination 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. Although general, this can be tedious for groups of items. Thus, to support more efficient combination, we extended this to allow entire groups to be dropped on other groups. When this is done, DataPalette attempts to identify matching pairs among items in the dragged and target groups by comparing each pairs' labels. Only exact matches (modulo white space/underscores) were considered matches and automatically combined. The rest are added to the target group as distinct elements, and may be explicitly 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 quickly visually compare aspects of instances, we created a charting tool and integrated a mapping feature. To make these tools suitable for fast use, e.g. data exploration, we designed these tools to automatically configure display parameters instances or groups of were dropped on them, rather than making users manually specify them. In the current prototype, the DWIM behaviour of the chart automatically sets whether the plot is a histogram representation (counts of values) or numeric value display based upon the types of the dereferenced values being plotted; values that are non-numeric are counted, while numeric values are displayed. 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

Since it can get confusing 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), causes all other visible representations of that instance to lightly glow, so that the user can quickly identify all the places that a single instance is represented.

EVALUATION

In order to determine whether the interaction affordances introduced in DataPalette (henceforth DP) allow users to perform serendipitous data mixing on real data feeds by end-users, we devised a within-subjects A-B trial lab study which we describe next.

Methodology

We first describe our methodology, starting with the three hypotheses we set out to test:

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) lab-based A-B 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 each of two tasks, but were told that if they finished earlier they could tell us and move on to the next condition. Prior to A conditions, participants were trained on how to use DP using a five-minute instructional video.

In the control condition (B), 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 web sites pertaining to the candidates they were asked to choose from. Participants were told they were free to use the web sites and spreadsheets however they wished.

We based our two tasks on choosing a university and picking a restaurant for a large social event, because in PS.1 the majority of people stated that they had used multiple websites for these tasks. We used two different variables for both tasks, where the restaurant was located (Glasgow and Cambridge) and which subject people chose (sport science and history). For the restaurant task, the cities Glasgow and Cambridge were chosen because they were geographically distant to participants, and would therefore minimise likelihood of prior familiarity with the restaurants or location. For the University selection task, two university courses were selected at random from the complete university guide[2], and the six top-ranked universities for each subject were selected as candidates. We selected six because it enables us to scope our study, we found through two pilot studies that six data sources provided a sufficient challenge for users to explore the data and required them to use DP's main features. We manually collected the key data points from the websites for use in DP and into Excel spreadsheets (which 40% of the participants in PS.1 said that had previously used to compare data).

Task	Data Sources
Three universities they would apply to study 1) Sport Science, and 2) History	www.thecompleteuniversityguide.co.uk, www.ucas.ac.uk, unistats.direct.gov.uk, www.timeshighereducation.co.uk/world-university-rankings
Restaurant they would book for 12 friends in 3) Cambridge, and 4) Glasgow	www.yelp.co.uk, www.tripadvisor.co.uk, ratings.food.gov.uk, facebook and google plus.

Table 2. For each type of task we used two variables, which affected which data we used from the relevant data sources.

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 protocol to the participant and to answer any questions; and the role of the observer was to observe without interacting with the user, and to take notes. Both the facilitator and observer were trained on the purpose of the study, their role and how they could and couldn't influence the study. Explicitly both the facilitator and observer where trained in identifying processes that people used 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 run over 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 populated a spreadsheet summarising quantitative metrics pertaining to how much various features of the tool were used (see Table 3 for specific fields). After the transcriptions were complete, we categorised people's comments made during the study, suggests in their exit survey, and clustered them to common themes.

Participant Recruitment

We recruited participants through adverts posted near the University campus and e-mails to student mailing lists. We screened participants to be at least 18 years of age, but did not filter on any other criteria. We took the first 20 that signed up and that successfully specified time slots that were available. Participants were offered a small gift certificate from an online retailer for their time.

[2] The Complete University Guide:
`www.thecompleteuniversityguide.co.uk`

RESULTS

We received 26 responses to our call for participants, from which we selected the first 20 (ten of which 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 the University, we ended up getting a large number of non-native English speakers. However, since all students and staff had passed requisite proficiency tests, we did not think this would substantially impact results. On average, each study took 40 minutes to complete. The information in Table 3 summarises the captured metrics for each participant.

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

The second metric was the number of factors the participant considered in making their 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 used to gauge the task performance is 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 demonstrated that interface choice only approached significance ($F(1, 19) = 0.04$ $p < 0.15$).

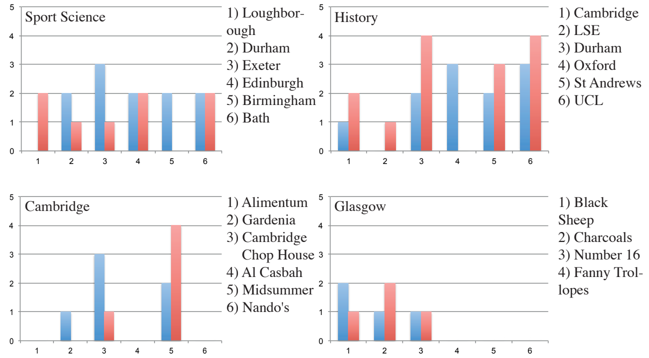


Figure 4. ***Choice picks per participant trial*** **- Histogram of number of times each restaurant/university was chosen in DP interface (A condition) in blue/left bars, compared to the Excel/website (B condition) in red/to the right.**

Effort: External cognition

Although we did not have an explicit measure of effort for this experiment (such as NASA's TLX [11]) 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 their choice on a histogram for each trial, per task and dataset for both conditions, visible in Figure 4. As can be seen there is considerable variability and no discernible difference in variability between conditions.

Strategies: Successive Elimination vs Tallying

As described in later this section, we noticed several different strategies people used to evaluate each candidate selection. One common strategy was *successive elimination* - to look at a single dimension at a time, starting with the most important, ruling out candidates that did not meet the required minimum for that value. However, when there is no clear "most important" aspect, this "greedy-style" can result in suboptimal decisions. Another strategy, which we called *tallying the pros and cons*, kept all candidates around, and counted up the advantages and disadvantages of each, potentially weighing each factor by its perceived importance. This strategy is less vulnerable to getting stuck in local maxima than the former.

To understand whether the interface influenced choice of strategy, we measured the number of candidates the participant considered *at final result*. This is a strong marker for use of each strategy; people using 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

Participants	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
gender	f	m	f	f	f	f	f	m	f	f	m	f	m	f	m	m	m	m	m	m
A condition first task	y	y	y	n	n	y	n	n	y	y	n	n	y	y	n	n	n	n	n	n
A Condition: task	4	2	3	4	3		2	1	4	1	4	1	3	2	3	2	1	4	3	2
paper used	n	n	n	n	n	n	y	n	y	y	n	y	n	n	y	n	n	n	n	n
charts used	1	2	2	0	2	4	5	0	0	0	2	3	2	0	0	0	3	3	4	1
maps used	1	1	2	2	2	1	1	0	1	0	1	1	2	1	0	1	0	0	2	1
same as used	4	3	3	6	0	2	0	2	1	0	6	5	6	2	7	0	6	7	7	7
# data sources used	4	3	4	4	4	3	4	3	4	4	3	3	5	4	5	4	4	5	5	4
chosen restaurant/uni	4.1	2.4	3.5	4.3	3.3		2.4	1.4	4.2	1.6, 1.3, 1.2	4.2	1.4, 1.5, 1.3	3.2	2.1, 2.3, 2.6	3.5, 3.3	2.6, 2.3, 2.5	1.6, 1.5, 1.3	4.1	3.3	2.4, 2.5, 2.6
# factors in choice	6	5	4	4	2	4	8	8	6	3	4	8	5	4	7	7	4	5	5	5
B Condition:task	3		4	3	4	2	1	2	2	3	2	3	1	4	1	4	2	2	1	3
paper used	n	n	n	n	n	n	y	n	y	y	n	y	n	y	y	n	n	y	n	n
number of spreadsheets used	5/5	4/4	2/5	1.5	2/5	5/5	4/4	2/4	3/4		2/4	4/5	2/4	4/5	2/4	2/5	4/4	4/4	0	5/5
number of websites used	1	1	4	3	3	0	2	3	2		1	2	3	1	2	3	2	0	4	0
chosen restaurant/uni	3.5		4.2	3.3	4.3	2.1	1.4, 1.2, 1.6	2.1	2.3, 2.6, 2.5	3.5	2.2, 2.3, 2.6	3.5	1.1	4.1	1.6	4.2	2.6, 2.3, 2.5	2.3, 2.5, 2.6	1.1, 1.3, 1.4	3.5
# factors in choice	4	3	2	3	1	6	5	7	7	5	4	7	6	3	6	4	3	6	4	5

Table 3. Data collected from each of our trials, for each participant.

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 ($M = 4.95; SD = 2.26$) $F(1, 19) = 5.323; p < 0.05$.

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 count accidental use of multi-path, such as happened a few times 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).

Use of Visualisation Features

To determine to what extent visualisation tools were used in condition A, 13 (65%) used charting tools at least once, while a majority (15, 75%) used the map. Determining the impact these had on performance, we performed a multiple regression on 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.

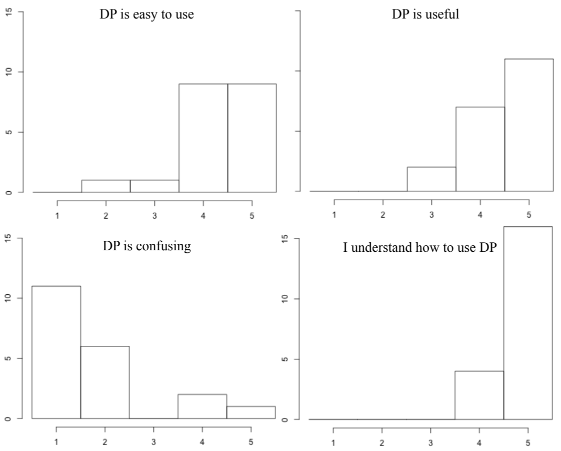


Figure 5. ***Survey results*** **- Answers to questions on a 5-point Likert scale, from 1-*strongly disagree* to 5-*strongly agree*.**

Condition B: Results and Observations

Participants were given the option of using web sites and/or a spreadsheet to make their choices, using the same data as condition A. There was mixed use of websites and the spreadsheet, 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 web sites, while one avoided Excel entirely. The uses of Excel for external cognition ranged from collation behaviour to tallying annotations, to colour-coding cells to allow easy identification of "pros" and "cons". The complete data collected is given in Table 3. In general, users struggled with both the websites and with the spreadsheet, although could generally find the information they required after a short period of frustration.

Survey Results

When asked to rate the DP interface, all participants responded that they felt they understood the system, with sixteen (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. On the statement "DP is confusing", 11 strongly disagreed, 6 disagreed, and 3 agreed. Revisiting the time trials, we discovered that the 3 participants who rated DP as confusing completed the trials statistically significantly slower than the rest of the group. Survey results were visualised in Figure 5.

In the exit survey, the users noted that they would use an interface like DataPalette in the future, for finding a property to rent/buy, choosing a job and purchasing electronic devices. However, one participant said that they would not use it for things like deciding on a restaurant because "it would cheapen the experience" by removing the element of spontaneity of choice, but would use it for making more important decisions. In order to improve our solutions 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 re-visit our study hypotheses introduced in Methodology, using observations from results to support views on each. We follow this with a discussion of 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 end-user citizens would understand the interface mechanisms of DP (h1). The second, h2, was that DP enables people to mix and integrate heterogeneous data in the process of performing their tasks. The third, h3, was that using DP facilitated task completion (h3). Using real data obtained from typical data sources identified during our pre-study (PS.1), we tested the DP prototype and its integration interaction mechanisms with "real" users, who were students, staff and alumni of our University who had varying backgrounds and levels of expertise with computers.

We find 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 their feedback, specifically feature requests and desired capabilities. Several participants requested search functionality, sort and filter functionality, the ability to display multiple properties for a single instance simultaneously, and visualise multiple attributes 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 have been 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 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 DP improved task performance (h3) was difficult given the small sample size of our study and large number of factors influencing each person's performance. Tasks were completed on average slightly faster than the control interface, but not statistically 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 people 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 the control interface, suggesting that decisions may have been made considering a greater number of factors. Participants entertained more possibilities for longer in the DP trial, while participants were more prone to eliminate candidates early in the baseline interface. Finally, people took fewer paper notes in the DP condition than 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 "pain points" in the design of the DP interface. One such pain point was window layout and screen real estate management; participants were prone to opening up a large number of windows, including maps and charts. When their screen started to fill up, we found that participants spent a lot of time arranging windows, and second, that people tended to forget what each window represented – "what was this again?". We plan to address these issue in the next version of DataPalette.

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 modeling-oriented heterogeneity cases observed in DP.2. Extending the "Do What I Mean" capabilities of the visualisation features to be unit-of-measure-aware we think would be tremendously valuable; a core problem, however, stems from the fact that few sources explicitly annotated data values with such units.

Study Limitations

Our study had several limitations. First, the small sample size of our study (20 participants) meant that we could not confirm some of the effects observed with statistical significance. Second, giving 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 some participants unlimited time on tasks of similar complexity. A second departure from a realistic task 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 it. 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 said in Voltaire's poem *La Bégueule*, "the best is the enemy of the good", we believe that the quest for complete approaches to data integration have, perhaps, caused "ad-hoc" approaches to be overlooked. Based upon the results of our investigation, however, we see a number of benefits to lightweight, user-driven, interactive approaches over automatic and bespoke ones. Just as the *user-subjective approach* to PIM [1] showed that subjective attributes and labels were more useful than physical ones, we have shown that a user-driven approach to data integration, as achieved with DataPalette, affords users flexibility to combine and keep separate data at will, and may facilitate decision-making tasks. Moreover, through our pre-studies and DataPalette evaluation, we have shown that much of heterogeneity exhibited by structured data feeds available on the Web today can be managed reasonably through easy to use, and intuitive interactive approaches for end-users.

REFERENCES

1. Bergman, O., Beyth-marom, R., and Nachmias, R. The user-subjective approach to personal information management systems. *Journal of the American Society for Information Science and Technology 54* (2003), 872–878.
2. Bergman, O., Beyth-Marom, R., and Nachmias, R. The project fragmentation problem in personal information management. Proc. CHI '06 (2006), 271–274.
3. 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.
4. 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* (2004).
5. Cai, Y., Dong, X. L., Halevy, A., Liu, J. M., and Madhavan, J. Personal information management with SEMEX. In *SIGMOD '05* (2005), 921–923.
6. Castano, S., Ferrara, A., and Montanelli, S. Matching ontologies in open networked systems: Techniques and applications. *Journal on Data Semantics V* (2006), 25–63.
7. Doan, A., Madhavan, J., Dhamankar, R., Domingos, P., and Halevy, A. Learning to match ontologies on the Semantic Web. *The VLDB Journal 12*, 4 (2003), 303–319.
8. 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.
9. Euzenat, J. An API for ontology alignment. *The Semantic Web–ISWC 2004* (2004), 698–712.
10. Halevy, A., Rajaraman, A., and Ordille, J. Data integration: the teenage years. In *VLDB* (2006), 9–16.
11. 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.
12. Huynh, D., and Karger, D. Parallax and companion: Set-based browsing for the data web. In *WWW Conference* (2009).
13. Huynh, D., Karger, D., and Quan, D. Haystack: A Platform for Creating, Organizing and Visualizing In-formation Using RDF, 2002.
14. 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 (2008), 274–282.
15. Jones, W., Karger, D., Bergman, O., Franklin, M., Pratt, A., and Bates, M. Towards a unification and integration of PIM support, 2005.
16. Shadbolt, N., Berners-Lee, T., and Hall, W. The Semantic Web Revisited. *IEEE Intelligent Systems 21*, 3 (2006), 96–101.
17. Suchanek, F., Abiteboul, S., and Senellart, P. PARIS: probabilistic alignment of relations, instances, and schema. *Proceedings of the VLDB Endowment 5*, 3 (2011), 157–168.
18. Voida, A., Harmon, E., and Al-Ani, B. Homebrew databases: complexities of everyday information management in nonprofit organizations. Proc. CHI '11, ACM (New York, NY, USA, 2011), 915–924.