MPRI Web Data Management Project

Serge Abiteboul Amélie Marian David Montoya Pierre Senellart

April 22, 2026

1 Overview of the project

Personal data is now pervasive as digital devices are capturing every part of our lives. Data is constantly collected and saved by users, either voluntarily in files, emails, social media interactions, multimedia objects, calendar items, contacts, etc., or passively by various applications such as GPS tracking of mobile devices, records of utility usage, financial transactions, or quantified-self sensors. Everywhere users go, everything they do, they leave a digital trace that acts as a digital memory of their past actions, interactions, and whereabouts. Having a personal "memex", a digital system to supplement one's memory, as envisioned by Vannevar Bush last century, is now close to possible.

Recent applications have started exploiting this information to help users with their day-to-day activities and in their daily decision making. For instance, *intelligent personal assistants* such as TripIt or Wipolo now automatically extract flight details from a booking confirmation email to raise a notification on a user's smartphone a few hours before departure if a flight is to be delayed.

On the other hand, this wealth of *personal information* has also infamously been exploited by large companies for financial gain: search engines capitalize on user queries and web usage to improve their ad sales and search results; social networks profit from the social interactions of their users, online stores learn from past sales to recommend new products. More recently, it has been reported that the government monitors and mines personal information from a wide variety of services for national-security purposes. While many benefit from information produced by users, the users themselves have a difficult time accessing, searching, and learning from their own data.

The overall objective of this project is to give data back to their owners, and, more specifically, to be able to align information about events (a person being at a location at a given point of time) from various personal data sources (emails, social networking sites, personal files, location tracking, etc.), and possibly across individuals.

A generic application could use the aggregated data to present a visual timeline of a person's activities: places visited, people met, work done, commuting time and itinerary, goods bought; as well as answer simple queries: "When did I last see Alice?", "What is the number of the florist I last bought flowers at?".

2 Project directions

For this project, you ought to take one of the following two directions:

1. Data fusion of personal data from different sources. For instance, suppose our calendar has information about a meeting on November 10th 2014 at "Le Train Bleu" but does not mention the person we are meeting. This information might be available through an email from Paul who confirmed the appointment time two days earlier. In this case, an application should connect these two information items together and present an aggregate view of this event. Data from several individuals can be used to detect and present *coincidences* in creative ways, e.g., "At one point of time T, two persons X and Y were at the same location L".

2. Journey generation : a random generator of personal data from a set of parameters. That is, given an appropriate input (e.g., training data, dictionary, numeric parameters), the generator should be able to render a chronological sequence of events, in interaction with different persons on different activities at different places, and centered around the life of one person. The generator might also

generate "fingerprint" data, such as event creation time, acknowledgment times, and source of knowledge (email, calendar, social network...).

3 Data model

Generated data should conform to the RDF model, using the `http://schema.org` vocabulary. Students can create their own classes and relations, as long as they provide an appropriate alignment to this vocabulary. In other words, represented classes, instances and relations should be *explicitly* related to the `http://schema.org` vocabulary through `df:type , \veb` rdfs:subClass , `dfs:subP`opertyOf or `wl:sameAs relatins`. Students are expected to represent events as instances of `http://schema.org/Event` or its subclasses for which answers to the following meta-questions are available, listed in order of importance: *When?, Where?, Who?, What?, How?*

For every *Event*, an answer to *When?* is **required**, and at least one other information. It is strongly encouraged to provide the *Where*, as this will allow spatiotemporal representation of personal data. *Who?* is meant to capture people who directly related to the event (organizer, attendees, etc.) or second-degree relationships (friend of an attendee). *What?* might capture the topic and related topics of an event. *How?* should capture any relevant context that enriches data around the event: is Alice a friend from university or high school? Was an invitation by email, or via a social network?

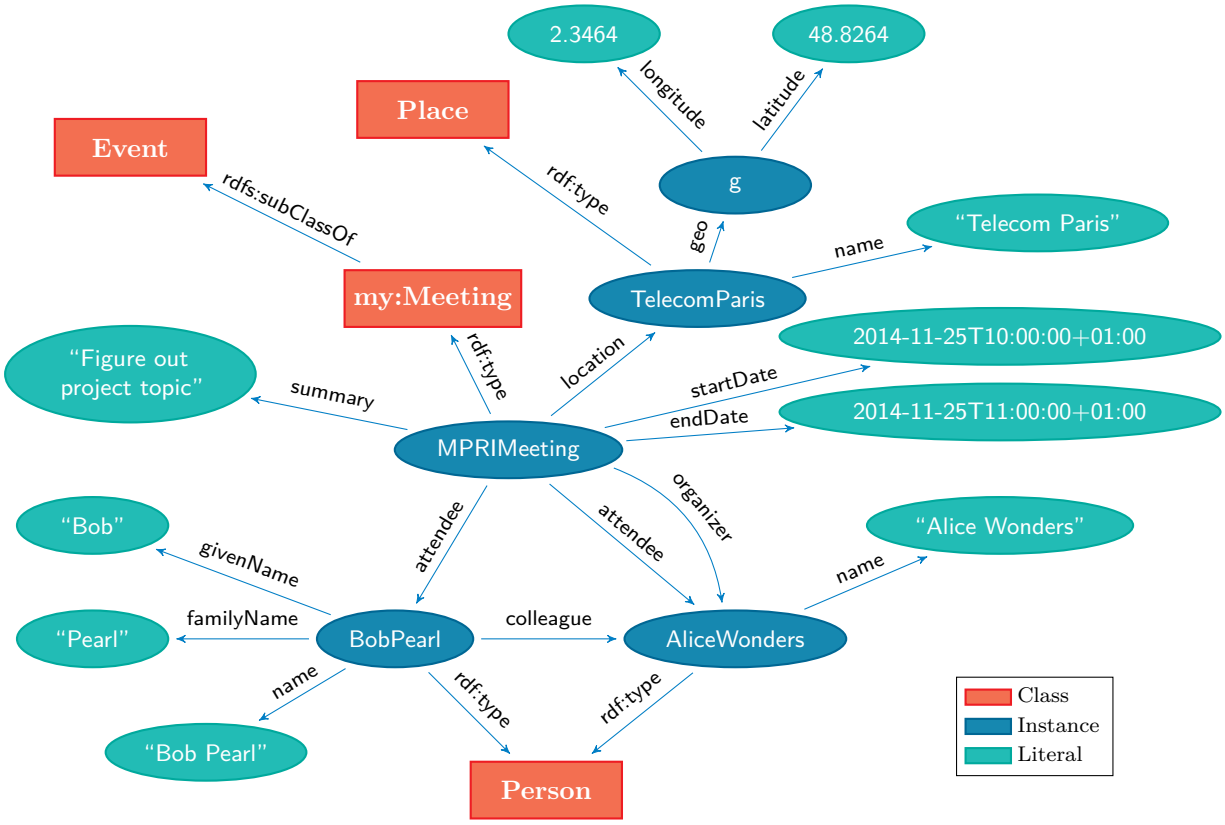

Figure 1: RDF data inferred from Bob's personal data and represented using `http://schema.org` vocabulary. MPRIMeeting was extracted from Bob's Google calendar. Bob's first and last names were extracted from his Facebook profile, and it was inferred that Alice is his colleague from LinkedIn data. The latitude and longitude of *TelecomParis* were extracted from the GeoNames database.

To extract data from Web applications, we recommended using public APIs whenever possible, standardized application protocols if necessary (e.g., IMAP/POP for emails), and fallback to Web scraping otherwise. For the following Web applications, we encourage you to use reuse the code at `https://github.com/ameliemarian/DigitalSelf/`: *Dropbox, Facebook, Foursquare, Gmail, Google*

Calendar, Google Contacts, Google Plus, LinkedIn, Twitter. While this is not a mandatory requirement, students are expected justify themselves if they choose otherwise.

The DigitalSelf code provides a simple data extractor, called Neemi, that uses public APIs from each Web application. For each Web application, an API client id and secret key has to be manually set up, and then the user is expected to authorize Neemi to extract their personal data via Neemi's interface. Once setup, Neemi can extract data and store it raw in a MongoDB database as a set of JSON documents. Figure 2 shows an example of event extracted from Facebook.

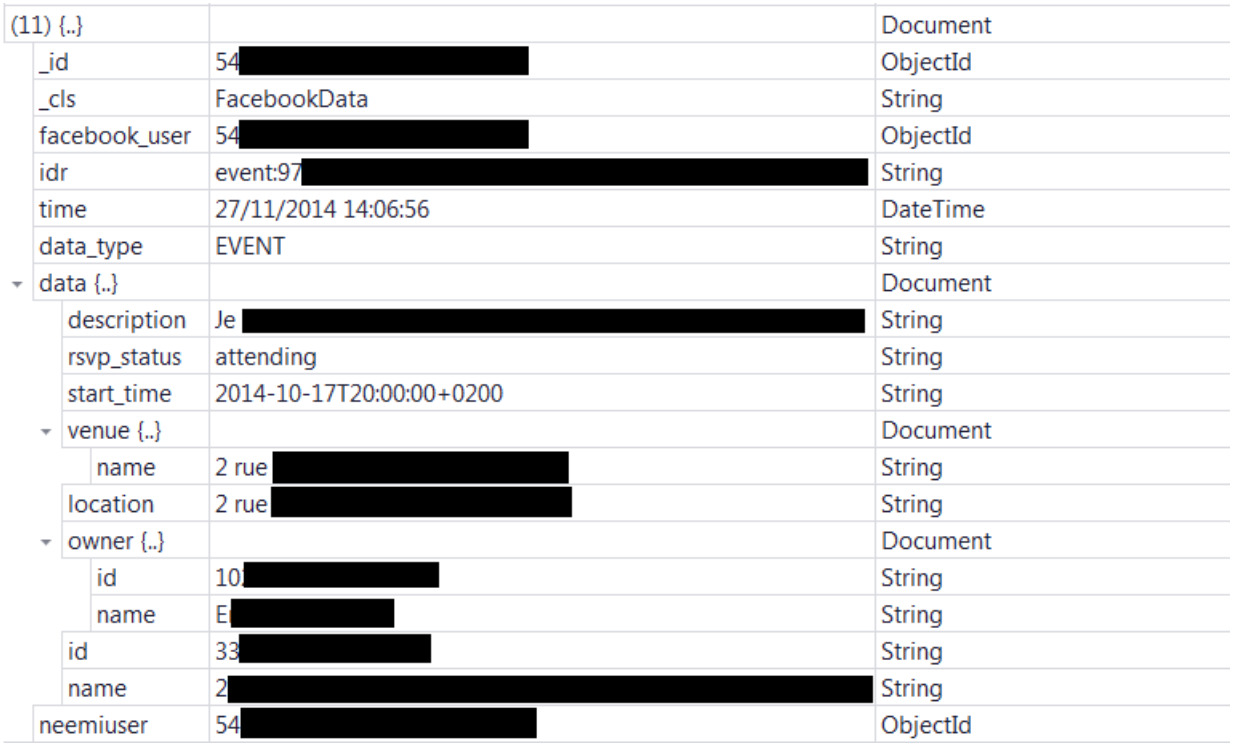

Figure 2: Facebook EVENT extracted by Neemi and represented in JSON format

Neemi is a work in progress, and students may encounter some bugs – bug reporting or fixing is appreciated. Additional contributions to the Neemi project (such as adding support for a new data source) are also appreciated.

4 Organization

Individual contributions can be made by individual students, or in groups of two. Projects chosen by different groups of students can integrate to each other (and are encouraged to do so). The contribution each student or group of students will work needs to be identified and communicated to Pierre Senellart by **January 4**. Groups will defend their contributions, by giving an overall presentation and showing a demonstration of their system, on **February 22**. A very short report (less than a page) is required. Students will also need to hand out an archive of their code.

5 Evaluation

The project is expected to be an implementation project; for some particular aspects, such as data fusion, coincidence detection, and data generation, contributions that are more at the algorithmic level will be accepted, but implementations of these algorithms are still required. The following elements will be particularly valued when evaluating a group's work:

- Applicability of the component to the overall "Personal Data fusion" or "Journey generation" problems;
- Re-usability of the component developed in other scenarios, e.g., by integrating it into the Neemi platform, or by developing re-usable standalone open-source software;
- Impact, wow effect of the demonstration;
- Integration with other groups' contributions;
- Initiative, creativity.

In the case components from different groups of students interact with each other, presentations can be combined, but each group is requested to emphasize its own particular work.

6 Questions

Questions related to the Neemi software should be directed to Amélie Marian <`amelie@cs.rutgers.edu`>. General questions about the project should go to Pierre Senellart <`pierre.senellart@telecom-paristech.fr`>.

References

[1] S. Abiteboul, B. André, and D. Kaplan. "Managing your digital life". In: *Commun. ACM* 58.5 (2015), pp. 32–35.