MPRI Web Data Management Project

Serge Abiteboul 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 three 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. Personal data 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. Novel applications on top of a personal information management system. Assuming an existing solution for collecting, integrating, or simulating personal data, you should come up with novel ways of exploiting, fusing, enriching, visualizing, or querying this data.

3 Data model

Generated data should conform to the RDF model, using the `http://schema.org/` vocabulary (see that URL for documentation), as well as extensions from the `http://thymeflow.com/personal#` vocabulary described in Figure 1. 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 whenever possible to the `http://schema.org` and `http://thymeflow.com/personal#` vocabularies through `df:type` , \veb rdfs:subClass , `dfs:subPopertyOf` or `wl:sameAs` `relatins`.

To extract data from Web applications, we recommended using public APIs whenever possible, standardized application protocols if necessary, and fallback to Web scraping otherwise.

Whenever convenient, we encourage students to build their code on top of the Thymeflow Personal Information management system[1], that already includes (between others) connectors and extractors for IMAP email accounts, Google Calendar, Google Contacts, Google Location History. Thymeflow also provides *enrichers* that are able to integrate contacts from different sources, to match calendar events and location histories, etc. A full description of the Thymeflow system can be found in [2].

Please note that Thymeflow is a work in progress, and students may encounter some bugs – bug reporting or fixing (to Pierre Senellart, or directly through GitHub) is appreciated. Additional contributions to the Thymeflow project (such as adding support for a new data source) that can eventually be merged into the project are appreciated. Help can be provided to set an instance of Thymeflow up.

[1] https://github.com/thymeflow/

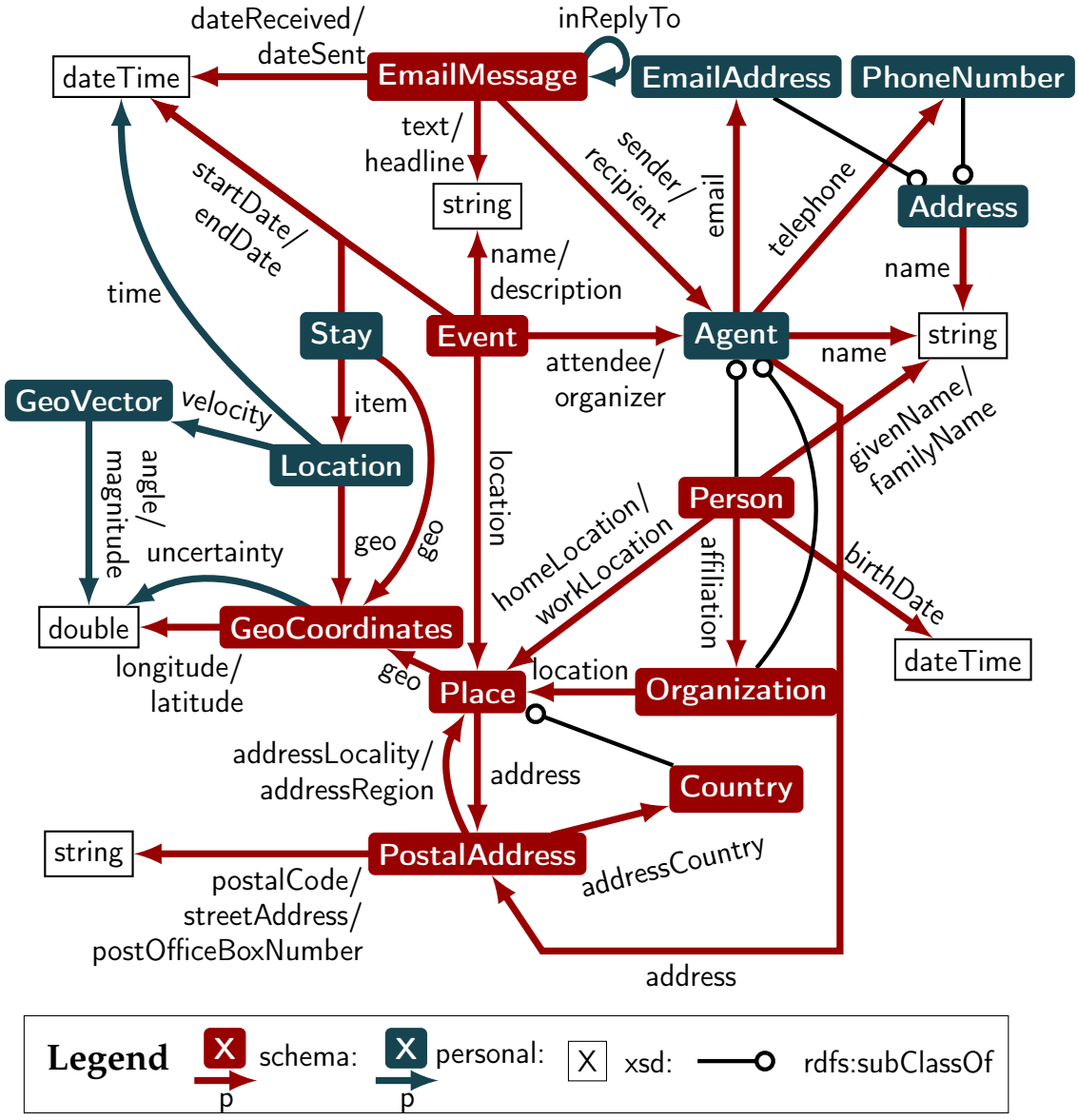

Figure 1: RDF data model, using the RDF prefixes `schema:` for `http://schema.org/` and `personal:` for `http://thymeflow.com/personal#`

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 15**. Groups will defend their contributions, by giving an overall presentation and showing a demonstration of their system, on **February 26**. 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", "Personal data generation", or "Novel applications" problems;
- Re-usability of the component developed in other scenarios, e.g., by integrating it into the Thymeflow 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 Thymeflow software or about the project should go to Pierre Senellart <pierre.senellart@ens.fr>.

References

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

[2] D. Montoya, T. Pellissier Tanon, S. Abiteboul, P. Senellart, and F. Suchanek. *Thymeflow, An Open-Source Personal Knowledge Base System*. Tech. rep. https://thymeflow.com/publications/2016-12-thymeflow-tech-report.pdf. 2016.