MPRI Web Data Management Project

Aligning Personal Data

Serge Abiteboul Amélie Marian 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.

This wealth of Personal Information has 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 back to users their data, 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.

One example of application could be to use personal data to infer, and visually display, a timeline of a person's location through the past months or years. Another example would be to integrate personal data from several individuals to detect *coincidences* of the form "At the point of time T, two persons X and Y were, independently, at the same location L." while preserving the *privacy* of the personal data handled. You can be creative and come up with your own applications.

As it is a fairly ambitious goal, the project can be decomposed into subprojects handled by different groups of students, each individual group contributing to one or several parts.

2 Data Model

The main concept of interest is that of an *event* that encompasses one or several *persons*, a *date*, and a *location*. In order to facilitate inter-group integration, students are expected to

use the `http://schema.org/` vocabulary, under its RDF representation[1]. In particular, the following concepts will be used (the list is non-limitative):

- `http://schema.org/Event`
- `http://schema.org/Person`
- `http://schema.org/Place`
- ISO 8601[2] date literals

The use of custom vocabulary is fine, as long as no equivalent vocabulary exists in `schema.org`. See also `http://schema.rdfs.org/` for how to translate `schema.org` microdata into RDF.

3 Topics

Students should cover several of the following aspects:

- For one particular personal data source (e.g., email headers, email content, Facebook, Twitter, Dropbox files, personal filesystem, Google Calendar, contact lists, LinkedIn, Google Location history, TripAdvisor reviews, Vélib' history, IP connection logs, banking data, instant messaging logs, etc.), collect and extract the data from this source. Some of these sources are covered by the `https://github.com/ameliemarian/DigitalSelf` software[3], whose use is encouraged, and which can also be extended to cover other sources. Generating new personal data yourself (e.g., via your own geolocation collection software) is also possible.
- Mapping this data to the data model presented above. In some cases (e.g., extracting data from plain text), this is a highly non-trivial step.
- Storing this data in a way appropriate for querying, large-scale storage, querying, etc.
- Align the instances (place names, person names, etc.) retrieved by using personal data collection on several student's personal data (e.g., the two members of a group).
- Integrate with other components developed by other groups.
- Build an application on top of the personal data retrieved and integrated, such as location timeline or coincidence detection.
- Maintain the privacy of personal data in multiple-individual applications.

Personal data should (as much as possible) **not** be shared with other people; software components to extract one's own personal data can be shared, however.

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

[1] `https://dvcs.w3.org/hg/htmldata/raw-file/default/ED/microdata-rdf/20120107/index.html`

[2] `http://en.wikipedia.org/wiki/ISO_8601`

[3] This is student code. It may be unstable and we offer no guarantee – however, we do welcome improvements and extensions.

so), but cannot overlap in functionality. The contribution each student or group of students will work needs to be identified and communicated to Pierre Senellart by Monday, January 6th. Groups will defend their contributions, by giving an overall presentation and showing a demonstration of their system, on Monday, February 24th. 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, especially privacy or coincidence detection, 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 "Coincidence from Personal Data" problem;
- Re-usability of the component developed in other scenarios, e.g., by integrating it into the DigitalSelf project, 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

All questions will be directed to Pierre Senellart, at `pierre@senellart.com`.