Proceedings of the Federated Conference on Computer Science and Information Systems pp. 621–624
DOI: 10.15439/2019F328
ISSN 2300-5963 ACSIS, Vol. 18

MOPS – A feasibility study for working with GPS and sensor data in a medical context

Christof Meigen
Leipzig Research Center for Civilization Diseases – LIFE Child
Email: cmeigen@life.uni-leipzig.de

Mandy Vogel
Center for Pediatric Research Leipzig, University Hospital for Children and Adolescents
Email: vogel@medizin.uni-leipzig.de

Jan Bumberger
Helmholtz-Centre for Environmental Research - UfZ
Email: jan.bumberger@ufz.de

***Abstract*—New kinds of data collection like GPS-tracking, wearable sensors and mobile apps impose both technical and privacy challanges for medical research. In the MOPS study (*Machbarkeitsstudie für Ortsbezogene Parameter und Sensordaten* – feasibility study for geocoded parameters and sensor data) we provided 10 participants with a newly developed app and sensors for various physical and environmental parameters. We want to explore the feasibility of the recently established Medical Research Platform (MRP) of the Medical Faculty of the University of Leipzig and similar platforms for this kind of data collection and processing.**

After briefly describing the Medical Research Platform we report on the technical set-up of the MOPS project in this setting and first practical experiences.

I. Introduction

TECHNICAL advances in the last decade – especially the ubiquity of smartphones – have made new kinds of data collection feasible for research. Sensors for many physical parameters are now comfortably wearable. Public facilities and initiatives for Open Data make more and more datasets publicly available, and it has become easy to – for example – link individual GPS data to public land use or noise maps. Software to work with this data is also freely available.

Medical research in particular is rooted in a tradition with strong focus on ethical and privacy consideration[2], and on long-term reproducibility on the results from raw data. That may sound trivial at first, but in practice it means getting explicit approval from an ethics board for each specific data collection, and storing all your raw data and analysis scripts for at least 10 years according to Good Clinical Practice.

A. Privacy Issues

The collection of vast amounts of data about an individual raises serious questions about privacy. Long gone are the days that just using a pseudonym for each participant was viewed as a sufficient protection against re-identification. GPS data reveals your home and work address, answers to questionnaires might be matched against social media profiles and high-resolution sensor data from physical parameters contain highly specific individual patterns. Ever-present timestamps can be used to identify events and may themselfes contain sensitive information.

An innocent looking data point like `{lat: 51.30175, long: 12.3775013, timestamp: "2019-05-20 17:32:12"}` could already be proof that the person this record refers to is an alcoholic (it's the time and place of an AA meeting).

For research, German law requires the "separate storage" of identifying data and research parameters (§27(3) BDSG). While this already means you should be using pseudonyms and store the names and contact information of study participants in a different database (or at least a different database table), it has been interpreted in the past as a requirement to also store research parameters with higher risk of re-identification like MRI data or genetic data separated *from each other* using different pseudonyms and a separate system (ideally managed by a trusted third party) to store the connection between these pseudonyms[6]. As discussed, tracking data from sensors certainly falls into the same category and should be treated accordingly.

B. Reproducibility

In a setting where research datasets are basically CSV files with one row per participant and one column per variable, and the – previously commited to – analysis plan consists of a few well-understood statistical tests, reproducibility can be achieved by archiving a few data files in a text-based format. The description in the publication is often sufficient to redo the calculations.

Nowadays, however, data is often requested on the fly from various data sources, and sophisticated software packages (with many dependencies on other packages) are used to process the data.

Reproducing results – especially many years later – requires archiving not only data in various formats, but also scripts to automatically obtain the published results from the data. For this, archiving the exact software environment used for analysis is often necessary due to changes in packages, incompatibility of new versions or deprecation of features. This is also in line with more recent general requirements for scientific data management like the FAIR data principles[5]. It is however at odds with workflows that rely on online services (which might become unavailable or return different results), specific versions of proprietary software (whose licence key might expire) and in general complex systems of interacting parts.

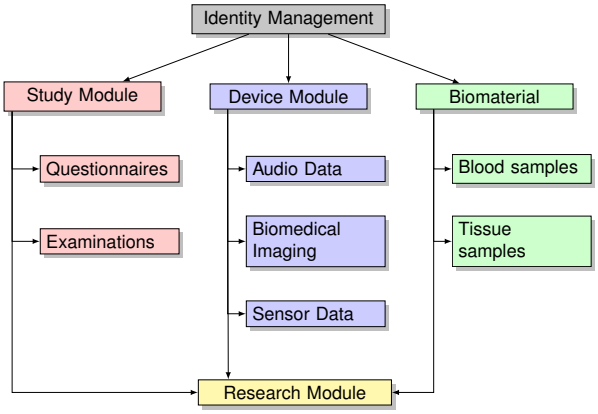

Fig. 1. Main Modules of the Medical Research Platform

C. Scope of the MOPS study

With the MOPS study, we build on previous considerations on combining sensor data and publicly available data with classical approaches from epidemiological research laid out by Kirsten et. al. [1].

We use a small sample set (n=10) to test the feasibility of the recently established Medical Research Platform (MRP) of the Medical Faculty of the University of Leipzig for this kind of study.

II. The Medical Research Platform

A. Overview of the Modules

The study management software REDCap[3] has been used extensively at the Medical Faculty of the University of Leipzig since 2013 to conduct studies, especially those that are not part of a drug approval process. Electronic case report forms (eCRF's) are easy to set up and the system provides excellent support for various data management tasks.

In light of the new General Data Protection Regulation (GDPR) the use of REDCap was re-evaluated in 2018 and confirmed as a platform for future research projects – but it was amended by a separate ID- and Consent-management system to store identifying information and connection between pseudonyms separate from each other and from the REDCap system. Additionally, nextcloud-based file archives have been set up to separately store data from devices and to archive research datasets, and the LabCollector LIMS has been installed to track biomaterial samples.

A data protection concept was drafted to define the various modules of the Medical Research Platform (see Figure 1 – the Biomaterial module is not used in the MOPS study) and to describe what data (under which pseudonyms) is stored in each module, what the interfaces between the modules are and especially how and when the re-pseudonymisation is performed and how access rights are granted.

B. ID Management

The ID Management solution LEIM was developed in coordination with the Data Integration Center of the University of Leipzig Medical Center to serve as a flexible model implementation for a *Patient Identifier Cross-reference Manager*, a *Consent Management System* (accessible only via API) and a separate Web application for Contact Management. It's main focus is the smooth integration with the other components of the Medical Research Platform and an ongoing adaptation to the concepts and interfaces established by the Data Integration Center as part of the German Medical Informatics Initiative (SMITH). Currently LEIM does not use a so-called *PID-Generator* like the Mainzelliste[4], but instead the pseudonymisation service generates a random Contact ID (*KID*) for use in the contact management and links it internally to the (also randomly generated) *PID*.

Fig. 2. The main screen of the MOPS app

III. The MOPS study

A. Overview

In the MOPS project we equip participants with an app for GPS and mood tracking (see Figure 2), a wristband (bodymonitor.de) to collect physiological parameters like skin conductivity and temperature, and a sensor for environmental parameters like air humidity and temperature (FreeTec, Model NC-7004-675). Participants are asked to wear these devices for at least 24 hours, preferably longer.

It is a feasibility study to establish technical and organisatorical processes and to investigate the usability of the Medical Research Platform for this kind of data collection and processing.

Data	Source	Nr of records
GPS location data	App	286 805
Stress & Movement	BodyMonitor	4 416 083
Mood	App (manual)	270
Transportation	App (manual)	196
Temperature/Humidity	FreeTec Sensor	2 504
Questionnaires	Interview	20

TABLE I
DATA COLLECTED IN THE MOPS STUDY

B. Organisatorical preliminaries

In order to conduct the study we submitted a study protocol to the Ethics committee. The study protocol did not only include the precise description of collected data, the aim of the study, description of recruitment process but also a data protection impact assessment detailing the risks for the participants in case of data leaks.

C. Collected data

We collected data from 10 participants with the wristband, MOPS-App and questionnaires. Two of the participant used an additional sensor for surrounding temperature and humidity. An overview of the collected data points can be seen in Table I.

After collecting the data, participants were interviewed for their experiences. Most people found the wristband and the notifications of the app annoying or slightly annoying, while 7 out of 10 participants used Google Services to improve the accuracy of the location information, thereby transfering all the location data collected in the MOPS study also to Google.

Data cleaning and matching the various time-related data is still ongoing, but the overall functioning of the data collection was verified during a piloting phase (see Figure 3).

IV. PRACTICAL EXPERIENCES WITH THE MRP

A. LEIM for ID and consent managament

For contact management, ID management and consent management a solution called LEIM (*Leipziger Einwilligungs-und Identitätsmanagement* – Leipzig Consent and identity management) had been developed based on the experience with similar tools especially at the German Center for Neurodegenerative Diseases.

The process of setting up a new study, defining the Informed Consent form and the types of Pseudonyms used is done through a simple web-based interface. Pseudonym types are defined by a simple declaration of fixed parts (most often prefixes) and random parts. The pseudonymisation service makes sure that random parts are created by an cryprographically secure pseudorandom number generator and that pseudonyms are unique within a study.

We defined the following pseudonyms: a Study Identification Code (*SIC*), an ID for the MOPS-App, an ID for the wristband, an ID for the sensor, and a pseudonym for the research dataset (*PSN*). In addition, each participant gets assigned a leading person identifier *PID* in the pseudonymisation service of the Identity Management and an contact identifer *KID* in the contact management part of the Identity Management.

After that, a study- and role-specific token was created which allows requests to the pseudonymisation service of LEIM from other modules in order to map pseudonyms and to request consent status. Access to identifying information (names, addresses) is not possible through these requests.

As a last step, the project ID in the REDCap system has to be entered to allow easy referrals from contact management to data entry forms in REDCap (which requires the appropriate role for the user in both systems).

B. REDCap for study data

In the MOPS study we collect only basic sociodemographic data – age and gender – as well as body height and weight in the study module. The pseudonyms for the devices (which are entered in the app, and stored on the the wristband and attached to the sensor) are stored only in LEIM, not in REDCap. Data entry and data export workes flawlessly in REDCap, as expected.

C. Nextcloud for device data and research data archive

Nextcloud is an open-source, self-hosted file share and collaboration platform. It stores the uploaded files on the host file system, but provides checksums and versioning for all files, fine grained access control, access through a web interface as well as mounting it as a network drive (through WebDAV) and even a client for local synchronisation (not currently used in the Medical Research Platform). Using nextcloud is absolutely straightforward for any computer user.

V. PRACTICAL EXPERIENCES WITH TECHNOLOGIES USED IN MOPS

Apart from the technical platform already provided by the Medical Research Platform we also explored various tools specific to the collection of sensor data.

A. Flutter for App development

While there is certainly no shortage of mobile apps, especially with respect to monitoring anything health/fitness related, we wanted to explore how difficult it is to create a simple app where we could completely control the data collection and data transfer process.

Having a background in web development and scripting languages, we evaluated mobile frameworks that provided a familiar development workflow, especially having no (re-)compile times, a simple dynamic language and an accessible set of simple cross-platform widgets. We chose Flutter (flutter.io) and were able to create an app with GPS tracking, uploading, notifications to regularly enter the current mood status and mode of transportation within 3 days of work, resulting in little more than 500 lines of code.

The only compromise we made was not using GPS tracking in the background (i.e. when the app is not running), as this is currently only available as a commercial plugin.

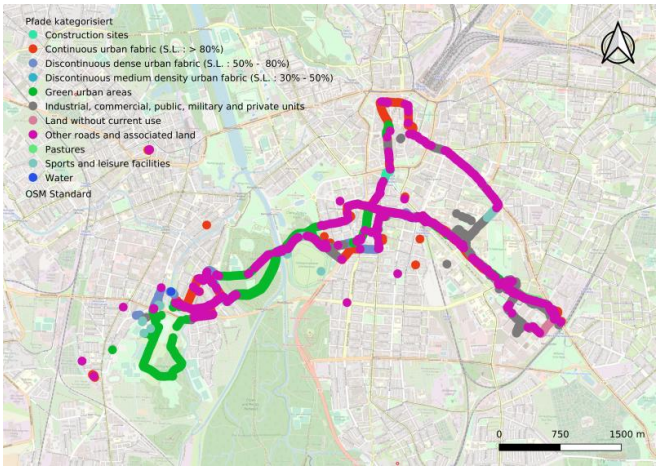

Fig. 3. A path recorded by the MOPS app, categorized by land use using the GeoEtiology PostGIS database

B. PostGIS and QGIS for Geodata

As part of the long-term GeoEtiology project a PostgreSQL/PostGIS database was set up containing geocoded information about the Leipzig region from various sources. This includes noise maps, the road network, information about land use, places of interest (restaurants, schools), trees etc. Information about the social characterisation of a large number of addresses was purchased from the *SINUS Markt- und Sozialforschung GmbH*.

For the MOPS study, the database acted as an (rather large) part of the data processing pipeline in the device module. We loaded the data temporarily into the database, ran some queries involving the various spatial information stored there, and exported the aggregated results, see for example Figure 3 for a dataset from piloting where a path was categorized according to land use.

Although we repeatedly encountered queries that required some restructuring, additional indices and usage of materialized views to perform well, the overall experience with the GeoEtiology PostGIS database has been splendid.

C. Guix for reproducible environments

The Medical Research Platform curently provides no special environment for scientific computing, instead, researchers are encouraged to ensure the reproducibility of their postprocessing and analysis scripts themselves by defining the required software environment alongside the scripts.

There are various lightweight aproaches for different programming languages – virtualenv for Python, packrat for R, the Manifest file for Julia – but having whole virtual environments of all the software tools used has in the past been limited to container-based solutions like Docker.

We instead defined Guix environments[7] using the concept of channels, where specific versions of all used software (including R and Python packages, but also R and Python themselves) can be defined. All dependencies down to the operating system kernel are then tracked and set up. The channel definitions are simply checked into the version control system alongside the scripts. Switching between environments is instantanious, and reproducing environments on a different computer might require downloading packages but is otherwise guaranteed to reproduce the exact same results. While we used Guix environment for data processing scripts, we do currently not run the GeoEtiology database in such an environment.

VI. Conclusion

Conducting the MOPS study on the Medical Research Platform is an interesting experience. The usage of different pseudonyms for different kinds of data and re-pseudonymisation for research datasets with the help of a separate ID management might seem cumbersome and overly cautions at first, but in practice it works well using the API of the LEIM pseudonymisation service. Generally, privacy concerns should be taken seriously and addressed at every step of the data processing.

Consequently working within Guix environments to ensure reproducibility seems doable in daily practice at least for the currently rather limited set of postprocessing and analysis scripts.

We hope to soon report on analysis results from the dataset collected in the MOPS study.

Acknowledgment

The authors wish to thank Prof. Dr. Antje Körner, PI of the GeoEtiology project, for her support in using the GeoEtiology infrastructure. The activities were co-financed by the research project *Smart Sensor-based Digital Ecosystem Services* (S2DES, 2016-2020), funded by the European Social Fund (ESF; Grant Agreement No. 100269858)

References

[1] Kirsten T., Bumberger J. et al. (2017), "Research in Progress on integrating Health and Environmental Data in Epidemiological Studies," In *Abramowicz W., Alt R., Franczyk B. (eds) Business Information Systems Workshops. BIS 2016. Lecture Notes in Business Information Processing, vol 263. Springer, Cham*, doi: 10.1007/978-3-319-52464-1_32

[2] World Medical Association. (2001), "World Medical Association Declaration of Helsinki. Ethical principles for medical research involving human subjects", Bulletin of the World Health Organization, 79 (4), 373 – 374. World Health Organization. http://www.who.int/iris/handle/10665/268312

[3] Paul A. Harris, Robert Taylor, Robert Thielke, Jonathon Payne, Nathaniel Gonzalez, Jose G. Conde (2009), "Research electronic data capture (REDCap) – A metadata-driven methodology and workflow process for providing translational research informatics support", J Biomed Inform. 2009 Apr;42(2):377-81, dio: 10.1016/j.jbi.2008.08.010

[4] M Lablans, A Borg, F Ückert (2015), "A RESTful interface to pseudonymization services in modern web applications", BMC Med Inform Decis Mak. 2015 Feb 7;15:2. doi: 10.1186/s12911-014-0123-5.

[5] Wilkinson MD, Dumontier M, Jan Aalbersberg I, et al. (2016), "The FAIR Guiding Principles for scientific data management and stewardship", Sci Data. 2016(3), article number 160018, doi:10.1038/sdata.2016.18

[6] Pommerening K, Drepper J, Helbing K, Ganslandt T (2014), "Leitfaden zum Datenschutz in medizinischen Forschungsprojekten – Generische Lösungen der TMF 2.0", ISBN: 978-3-95466-123-7

[7] Courtès L, Wurmus R (2015): "Reproducible and User-Controlled Software Environments in HPC with Guix" In *Euro-Par 2015: Parallel Processing Workshops*, 579–591, Springer International Publishing, Cham, doi: 10.1007/978-3-319-27308-2_47