The Journal of Open Source Software

Castellum: A participant management tool for scientific studies

Tobias Bengfort[1], **Taib Hayat**[1], **and Timo Göttel**[1¶]

1 Max Planck Institute for Human Development, Berlin, Germany ¶ Corresponding author

DOI: 10.21105/joss.04600

Software

- Review
- Repository
- Archive

Editor: Olivia Guest

Reviewers:

- @samhforbes
- @htwangtw

Submitted: 30 June 2022
Published: 21 November 2022

Summary

`Castellum` is a web application aimed mainly at the lifecycle of human research studies by supporting participant management, pseudonym service, recruitment, study appointments and study execution.

The main design focus of `Castellum` is compliance with the European Union's (EU) General Data Protection Regulation (GDPR) and with IT security. Additionally, it is designed to be adaptable to other workflows or processes at different research institutions.

Statement of Need

The web app `Castellum` provides a clear structure for handling the data of all study participants. Contact information, personal data relevant to a study (for example, age, highest degree qualification, ...) and process information (for example, existing consents or last study participations at the institution) are stored in `Castellum` itself. `Castellum` was developed in order to ensure this data is managed in accordance with the applicable standards of the GDPR.

Scientific data, on the other hand, is stored separately and uses study-specific pseudonyms that were generated by `Castellum`.

Overview Key Features

`Castellum` allows assigning individual study-specific pseudonyms to the participants of each study to ensure that the actual scientific data can be stored separately from the contact data in another system (on a different server or file system).

In addition, `Castellum` allows for multiple pseudonym spaces ("domains") to be created in studies, so that several pseudonyms can be used per participant in a study.

The specific and expandable assignment of pseudonyms is intended to prevent the pooling/merging of scientific data across studies (or groups of people) without an appropriate legal basis.

During the execution of a study, it is technically ensured that only users with the corresponding study-specific access role can access the contact data in `Castellum` via a domain specific pseudonym search. As soon as the data collection phase of a study is marked as finished in `Castellum`, this possibility expires. From then on, only the data protection coordinator role is able to access study participations.

Additionally, `Castellum` allows deleting the pseudonym space of a study independent of the study status. After the deletion of a pseudonym space, it will be impossible for any user to connect pseudonymized scientific data to the contact data in `Castellum`. This process

corresponds to the usually paper-based procedure in human science of destroying the coding list at a certain point after data collection.

Moreover, Castellum keeps track of the legal bases to keep information about a participant. When all legal bases have expired, the participant is listed in a data protection dashboard. This allows data protection officers (or staff) to handle or rather delete participant data accordingly.

Roles and Rights Management

Users can be imported via LDAP groups. Users are assigned to different roles in Castellum. There are five global roles: Data protection coordinator, Principal subject manager, Receptionist, Study approver, Study coordinator. Two of these roles (Data protection coordinator, Principal subject manager) allow comprehensive access to participant data. These roles should only be assigned to a very small group of people. Within a study recruitment, recruiters are granted access to the contact data of persons who match the set filter criteria of the study. Conductors get access to the contact data and pseudonyms of persons who are marked as participating in the study. While users with the role of Principal subject manager are shown all matching records in the results list of participant search, the study-specific subject managers only receive search results for participants of their affiliated studies.

In addition to user management with roles, Castellum controls access to datasets via privacy levels. Different privacy levels are assigned to users, participants and attributes. This ensures that users, regardless of the roles assigned to them, can only view records that correspond to their privacy level.

Castellum allows that data is distributed across two databases (contact data can be kept separated from all other data). The actual research data is **never** stored in Castellum. The linking to this data is regulated by pseudonyms, which can only be resolved via appropriate rights in Castellum.

Two-factor Authentication

To protect individuals' personal data from compromise or weak passwords, Castellum uses a standalone two-factor authentication (2FA) for Django that was also developed by a Castellum team member (Bengfort, 2022). Users must enter an additional code before they can log into Castellum. Currently, any generic time-based one-time password (TOTP, (M'Raihi et al., 2011)) application or FIDO2 hardware security token is supported.

Considerations on the Production Environment

The application should be hosted locally at a research facility and it is recommended to make it only available on the intranet; i.e., direct access via the internet should not be possible to minimize security threats. The access to the web interface should always be encrypted (SSL).

Access to the server should be restricted to a small group of administrators of the corresponding institute. It is recommended that users should undergo recurring training on GDPR compliant usage of Castellum and its suggested interaction with participants.

It remains to mention that regular database backups of Castellum have implications for data protection: After a participant has been deleted in Castellum, their personal data will still be stored in a backup. Therefore, it is recommended to keep backups only for a short amount of time (e.g., 14 days) and inform participants about the date when all their data, including backups, will be deleted.

State of the Field

While there are several commercial providers offering (proprietary) online software services around participant management it is hard to find open-source packages that tackle these topics in full scope. For example, the MOSAIC project (MOSAIC Team, 2022) is offering a set of tools that could be used to set up an ecosystem to support study execution. However, as described by the authors (Bialke et al., 2015) these modules are strongly focused on managing participant data for a single (often large-scale) study that needs a third party as fiduciary to hold the actual contact data of participants. In fact, a sister project we are in contact with (Huscy Team, 2022) is covering this modular approach as well. Yet another approach is presented by the ORSEE tool (ORSEE Team, 2020): It covers a *self-booking* recruitment approach for studies that needs potential participants to sign up for studies that they find interesting (Greiner, 2015).

Although these are important application scenarios, they often fail to support research facilities that are running many studies in parallel and want to proactively recruit participants out of an in-house pool of interested potential participants. Such institutions want to invite potential participants who are a perfect fit for a study to guarantee a random sample as possible. This makes it important to allow staff members to define study filters and contact according potential participants at ease but still respecting GDPR requirements. From our point of view, Castellum is currently the only turn-key open-source solution for GDPR compliant participant management in that manner.

Acknowledgements

The `Castellum` project is financed by internal funds of the Max Planck Society.

References

Bengfort, T. (2022). Django-mfa3: An opinionated django app that handles multi factor authentication (MFA) via FIDO2, TOTP, and recovery codes. In *GitHub repository*. GitHub. https://github.com/xi/django-mfa3

Bialke, M., Bahls, T., Havemann, C., Piegsa, J., Weitmann, K., Wegner, T., & Hoffmann, W. (2015). MOSAIC – a modular approach to data management in epidemiological studies. *Methods of Information in Medicine*, *54*. https://doi.org/10.3414/ME14-01-0133

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association*, *1*. https://doi.org/10.1007/s40881-015-0004-4

Huscy Team. (2022). Huscy modules. In *Bitbucket repository*. Bitbucket. https://bitbucket.org/huscy/

M'Raihi, D., Machani, S., Pei, M., & Rydell, J. (2011). TOTP: Time-based one-time password algorithm. In *RFC*. Internet Engineering Task Force. https://doi.org/10.17487/rfc6238

MOSAIC Team. (2022). MOSAIC tools. In *GitHub repository*. GitHub. https://github.com/mosaic-hgw

ORSEE Team. (2020). Online recruitment system for economic experiments (ORSEE). In *GitHub repository*. GitHub. https://github.com/orsee/orsee