The Dataverse® Project
Community Meeting

5-7th June 2023 Braga Portugal

DCM23 Poster Submission

Dataverse Community Meeting 2023, June 5-7, 2023, Braga, Portugal

São João Health Data Repository - The Firsts Steps

Author(s)

- Maria João Campos, Centro Hospitalar Universitário São João, Portugal
- Afonso Pedrosa, Centro Hospitalar Universitário São João, Portugal
- Maria Fernanda Gonçalves, Centro Hospitalar Universitário São João, Portugal
- Ana Azevedo, Centro Hospitalar Universitário São João, Portugal
- Luis Antunes, FCUP, Universidade do Porto, Portugal
- José Carvalho, KEEP SOLUTIONS, Portugal
- Miguel Ferreira, KEEP SOLUTIONS, Portugal
- Manuel Monteiro, KEEP SOLUTIONS, Portugal

Poster abstract

This poster delineates the establishment of the São João Health Data Repository at Centro Hospitalar Universitário de São João (CHUSJ), a Portuguese university public hospital, employing Dataverse software.

From a management perspective, this being a service integrated in the healthcare context, it must conform to strict requirements when it comes to publication policies and data curation. Such processes aim to avoid the inadvertent disclosure of patients' sensitive information to unauthorized people.

Our objective is to not only publish data publicly but also to securely provide traceable data to researchers who have approved projects at CHUSJ, ensuring that the risk of subjects' re-identification remains below 1%. This dual approach facilitates open access to information while maintaining stringent privacy standards for data that may be sensitive.

The goals of providing this service are:

- To improve collaboration and knowledge sharing within the healthcare network.
- To promote Open Science and the use of open infrastructure.
- To ensure data is properly managed, documented, and preserved for long-term use.

The service is based on a dockerized version of Dataverse with additional components and integrations such as a preview component and statistical dashboards supported by Apache Superset.
The Health Data Repository allows healthcare providers to upload and share clinical research data securely with persistent identifiers.
All datasets made available to the general public have been submitted to a risk analysis procedure that reduces the chances of patient re-identification, even when data has been anonymised. This risk analysis methodology, which is being developed in the context of CHUSJ, has already been approved by the national data protection authority (https://www.cnpd.pt/umbraco/surface/cnpdDecision/download/122003).
The Health Data Repository is still in its infancy but it intends to showcase health data that will impact healthcare and also society as a whole.
The first dataset to become available aims to provide a set of data in a shared data model with attribute-based access in order to respond more swiftly and efficiently to the continuing high number of access requests for clinical research purposes about COVID-19.
This approach allows for a wiser allocation of human resources currently assigned to this function (information system teams and data protection officers), the mitigation of the impact on data subjects' rights and improvement of data quality for research purposes.

Keywords

Data, Hospital, COVID-19, Implementation.

Related projects or initiatives

https://superset.apache.org.

https://github.com/gdcc/dataverse-previewers

Community meeting topics

- Metrics & usage statistics;
- Sensitive Data;
- Integrations & External Tools;
- FAIRness & interoperability
- Preservation