MrData: An iRODS Based Human Research Data Management System

iRODS User Group Meeting
6 July 2022

Blake G. Fitch
blake.fitch@tuebingen.mpg.de
The MrData Project for Human Research Data Archival

MrData Project -- Blake G. Fitch -- MPI KYB

6 July 2022
Overview

Castellum is a human subject database system developed at the Max Planck Society. Its main goals are:

- GDPR compliant data protection and security
- Flexibility so it can be used in different organizations

Features

Subject management
- Castellum is a central place to collect references to all data related to a subject, e.g., so it can be deleted on request

Pseudonym service
- Contact details are stored in Castellum so all other databases can work with pseudonyms instead.

Recruitment
- Castellum allows you to find potential subjects from an existing pool using study specific filters

Appointments
- You can manage appointments for experiment sessions

Application Programming Interface (API) (optional)
- Available API for exportable attributes. (pseudonyms, etc) which are needed to integrate with a data management system

What Castellum is not

Castellum does not store scientific data. It manages information about studies and subjects required for recruiting and provides subject pseudonyms for external use.

Mixed Use Metadata: A Challenge For Human Research

Certain bits of metadata are required for both:
• recruiting activities
• scientific search and catalog

Rules for Mixed Use Metadata:
• GDPR Personal metadata can only be kept in Castellum
• A given bit of metadata should be acquired from a single, authoritative source
• Mixed use metadata can only be sourced from Castellum

Selected, GDPR safe, subject attributes move one-way from Castellum to MrData by admin domain automation.

Uses Castellum API for safe data access.

6 July 2022

MrData Project -- Blake G. Fitch -- MPI KYB
Experiment Registration

• Associates StudyID and Subject pseudonym with the ExperimentID
• The ExperimentID is entered in place of a ”Patient Name” on the MRI scanner
• This avoids contaminating MRI scanner data files with personal information about the subject...
• but does require maintaining this information for the life of the data
• This association provides search mixed use metadata
• An admin operation can remove all data for a subject

Experiment Registration Web Pages: Request/Response

Register an Experiment

<table>
<thead>
<tr>
<th>Register an Experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment Owner (campus user id)</td>
</tr>
<tr>
<td>scnameuser</td>
</tr>
<tr>
<td>Subject Pseudonym (from Castellum)</td>
</tr>
<tr>
<td>2H4UY2</td>
</tr>
<tr>
<td>Study ID (from Castellum)</td>
</tr>
<tr>
<td>22</td>
</tr>
<tr>
<td>Scanner</td>
</tr>
<tr>
<td>Siemens 9.4T</td>
</tr>
<tr>
<td>Scan Type</td>
</tr>
<tr>
<td>Human Scan</td>
</tr>
<tr>
<td>Experiment URL</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Experiment Description (NO GDPR VIOLATIONS)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Submit</td>
</tr>
</tbody>
</table>

Form Capture Confirmation

<table>
<thead>
<tr>
<th>Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment Registration</td>
</tr>
</tbody>
</table>

New Experiment ID (Save! Needed at Scanner!)

24KOC-STAP

Copy to Clipboard

Go To Main Page

These two web pages are implemented as a small Python Flask app in the Forms container.

Note: Prior to using Castellum, other web pages for Subject and Study registration existed.
MrData MRI Research Investigator Workflow

MRI Scientific Workflow
- Use Castellum web gui to define a Study and acquire a StudyID
- Use Castellum web gui to recruit Subjects to participate in a Study
- Use Castellum web gui to acquire a Pseudonym for each Subject in a Study
- Register an Experiment using Pseudonym and StudyID, get an ExperimentID
- At MRI Scanner, Operator enters their LDAP user name and the ExperimentID
- MRI data is imported, anonymized, and archived based on info in ExperimentID
- MRI data is then accessed via iRODS (as a network share or via Python API, etc)

Python data import function
- verify experiment info
- anonymization process
- store data / metadata

GPFS Storage (3PB)
MrData: Infrastructure Environment and Microservice App

MrData Docker Host Boundary

Siemens MRI Console

GPFS

DICOM

TWIX

EXP DATA

MRI Data Landing Zone

s94t2irods

DICOM

s94t2irods

TWIX

s94t2irods

EXP DATA

DavRODS

iRODS

DavRODS

Archived MRI Data ➔

← Scientific Results

ExperimentID ➔

Pseudonym, StudyID

ExperimentID ➔

 metadata

StudyID ➔

metadata

StudyID ➔

ExperimentID + metadata

StudyID + metadata

StudyID + metadata

StudyID + metadata

SciRents

Scientific Results

Investigator

Workstation

Recruiting

StudyID & Pseudonym

Castellum Virtual Machine

Castellum

-to-

iRODS

Forms

GitLab

EXP DATA

TWIX

DICOM

MrData Project -- Blake G. Fitch -- MPI KYB

6 July 2022
MrData Component Services Live in Separate Containers

This is a Docker based micro-services based architecture

- **Pros**
  - Integrates home grown and external services
  - Single Docker host (currently)
  - Ansible deployed docker containers for all services
    - More robust than Docker Compose, easier than K8s, etc
    - Possible to test and redeploy individual services (via Ansible)
    - Independent test infrastructure can be deployed to Vbox VM
    - Full system CI/CD possible though not nearly completed
    - Extensible within the Docker ecosystem
  - Cons
    - micro-services, docker, ansible – the usual stuff

6 July 2022

MrData Project -- Blake G. Fitch -- MPI KYB
Data layout in the iRODS Data “VAULT” on Linux FS

iRODS internal path:
/MRDataZone/home/mrdata/echtdata/studies/<StudyID>/experiments/<ExperimentID>/<data_type>/

```
[mradmin@mrdata01 ~]$ sudo find /zfs-pool/MRDATA/prod/irods/IRODS_VAULT/home/mrdata/echtdata/studies/43/experiments -maxdepth 2
/zfs-pool/MRDATA/prod/irods/IRODS_VAULT/home/mrdata/echtdata/studies/43/experiments
/zfs-pool/MRDATA/prod/irods/IRODS_VAULT/home/mrdata/echtdata/studies/43/experiments/HGZD-0037
/zfs-pool/MRDATA/prod/irods/IRODS_VAULT/home/mrdata/echtdata/studies/43/experiments/HGZD-0037/MR_9.4T_TWIX
/zfs-pool/MRDATA/prod/irods/IRODS_VAULT/home/mrdata/echtdata/studies/43/experiments/HGZD-0037/EXPDATA
/zfs-pool/MRDATA/prod/irods/IRODS_VAULT/home/mrdata/echtdata/studies/43/experiments/HGZD-0037/MR_9.4TDICOM
/zfs-pool/MRDATA/prod/irods/IRODS_VAULT/home/mrdata/echtdata/studies/43/experiments/V5U4-TJ2B
/zfs-pool/MRDATA/prod/irods/IRODS_VAULT/home/mrdata/echtdata/studies/43/experiments/V5U4-TJ2B/EXPDATA
/zfs-pool/MRDATA/prod/irods/IRODS_VAULT/home/mrdata/echtdata/studies/43/experiments/V5U4-TJ2B/MR_9.4T_TWIX
/zfs-pool/MRDATA/prod/irods/IRODS_VAULT/home/mrdata/echtdata/studies/43/experiments/V5U4-TJ2B/MR_9.4TDICOM
[mradmin@mrdata01 ~]$ 
```
Questions?
WebDav Interface: See what’s in iRODS via browser login

iRODS internal ‘ipath’: /MRDataZone/home/mrdata/echtdata/studies/<StudyID>/experiments/<ExperimentID/>
MrData view from MacOS Finder (iRODS mounted as a file system)