PROCEEDINGS OF SCIENCE

Dark Matter Data Center: Fostering Data and Information Sharing Within the Dark Matter Community

Heerak Banerjee[a,b,*] **and Nahuel Ferreiro Iachellini**[b]

[a] *Physik-Department and Excellence Cluster Origins, Technische Universität München, James-Franck-Straße 1, DE-85748 Garching, Germany*

[b] *Max-Planck-Institut für Physik, Föhringer Ring 6, DE-80805 München, Germany*

E-mail: heerak.banerjee@tum.de, ferreiro@mpp.mpg.de

The Dark Matter Data Center (DMDC) is an ORIGINS Excellence Cluster initiative, supported by the Max Planck Computation and Data Facility. It aims at bringing together the large amount of recorded data and theories pertaining to Dark Matter (DM) research in a unified platform, making it easily accessible for the community. The DMDC offers a repository where data, methods and code are clearly presented in a unified interface for comparison, reproduction, combination and analysis. It is a forum where Experimental Collaborations can directly publish their data and Phenomenologists the implementation of their models, in accordance to Open Science principles. Alongside the repositories, it also offers easy online visualization of the hosted data. It offers an online simulation of signal predictions for experiments using model data supplied by the users, all in a friendly web-based GUI. The DMDC also hosts guidance tools from the Collaborations illustrating the usage and analysis of their data through Binders that run online and support all popular programming platforms. It hosts a continuously growing compendium of ready-to-use, copy-pastable code examples for inference and simulations. It can also provide support and computational power for comparison of model and experimental observations as well as the combination of these results using modern and robust statistical tools through similar Binders.

41st International Conference on High Energy physics - ICHEP2022
6-13 July, 2022
Bologna, Italy

*Speaker

PoS(ICHEP2022)305

1. Introduction

The quest for Dark Matter (DM) and its nature has engendered theories that span nearly hundred orders of magnitude in mass scales with widely contrasting properties. It has also motivated decades of experimental efforts correspondingly different in the wide variety of their target masses, observables, technologies and interpretations (For a detailed review see, for example, Ref.[1]). As analyses of experimental data leading to exclusions or posteriors become more and more complicated, the final products also become more and more dependent on the specific methods employed. This labyrinth of theories and experiments not only makes their analyses and combination a daunting task, but also often leads to confusion and conflict regarding their implications.

The solution is to make experimental data and analyses public. This is also in line with the "Open Science" principles catching steam all over the world and in particular in the particle physics Community, where the technical details are thoroughly being addressed[2, 3]. It is in the interest of more efficient and rapid progress in research that not only the experimental data but all the analysis workflows are made public allowing for completely reproducible results. This applies as much to those engaged in theoretical research as it does to experimentalists. In order to ensure reproducibility and enhance accountability, it is essential that model implementations and the workflows be made public. While the first steps in ensuring these requirements for Open Science are already being taken, the Dark Matter Data Center (DMDC) aims to provide the DM community with a unified platform where all of this information can be readily available.

The Dark Matter Data Center is an ORIGINS Excellence Cluster initiative, supported by the Max Planck Computation and Data Facility (MPCDF). All public data from experiments connected to DM research are listed on the DMDC in an easily navigable interface. In addition to being a repository of public data, it also aims to be a repository of connected software and workflows. The DMDC also aims to provide online tools eg. recoil rate simulators, detector responses and comparison tools to aid in DM research directly online in the near future. We present here an outline of the facilities offered by the DMDC.

2. Functionality

The Dark Matter Data Center is designed to be useful for a wide range of users at varying stages in DM research. Its organization ensures that a user can get an overview of the larger experimental status right away. The datasets are organised by Collaborations and further by specific data releases. Each dataset is accompanied by a table of its salient features eg. target material, detector technology, fiducial volume, live time, citeable resources, etc. An example of how a dataset is represented on the website can be seen in Fig.1(a). Resources within the data release can be explored by clicking on the "Resources" button. Each resource is associated with a detailed description and sometimes with a brief note on its usage. Resources can be downloaded individually by simply clicking on their names or collectively as a tar.gz archive.

It is possible to visualize each resource within any dataset online before downloading it. Each dataset has its own visualization pane, an example of which is shown in Fig.1(b). This can be activated simply by clicking on the "Visualize" button. Different categories of resources can be visualized by selecting them from the drop-down menu on the upper left corner of the pane. Within

each category, different resources can be toggled to be visible by clicking on the respective legend entries.

DMDC hosts data usage tools in the form of JuPyter notebooks that can be run online by clicking on the "Launch Binder" button, as can be seen in Fig.1(a). These Binders provide to the public the connection between the datasets that are being made public and the final plots and exclusions published by the Collaborations. Without these connections it is often the case that very different final results are reached, even while using the same datasets due to subtleties in the data processing. These notebooks also provide users with many code snippets that are either released or verified by their authors and can be used directly in their projects as a starting-point. The DMDC is also a repository for model implementations. These implementations will also be made publicly available through Binders. This will ensure their long-term persistence and reproducibility. We already host public data from CRESST[4–8], XENON[9] and ANAIS[10], while more datasets are being added continuously.

2.1 Experimental Data

Experimental data hosted on DMDC are generally categorized in three genres:

1. Event Data: The actual unbinned/binned events observed by the experiment for each detector unit. Preferably, events should be listed in as many observables as available (photoelectrons, heat, timestamp, positions, etc.).

2. Background Information: Expected numerical background rates (or background model) for each detector unit in the same observables as the event data.

3. Detector Response: Any information required to generate a simulated signal in terms of the observables in the event data from a theoretical differential cross-section. This may include thresholds, cuts, efficiencies, survival probabilities, band coefficients, quenching factors and resolutions.

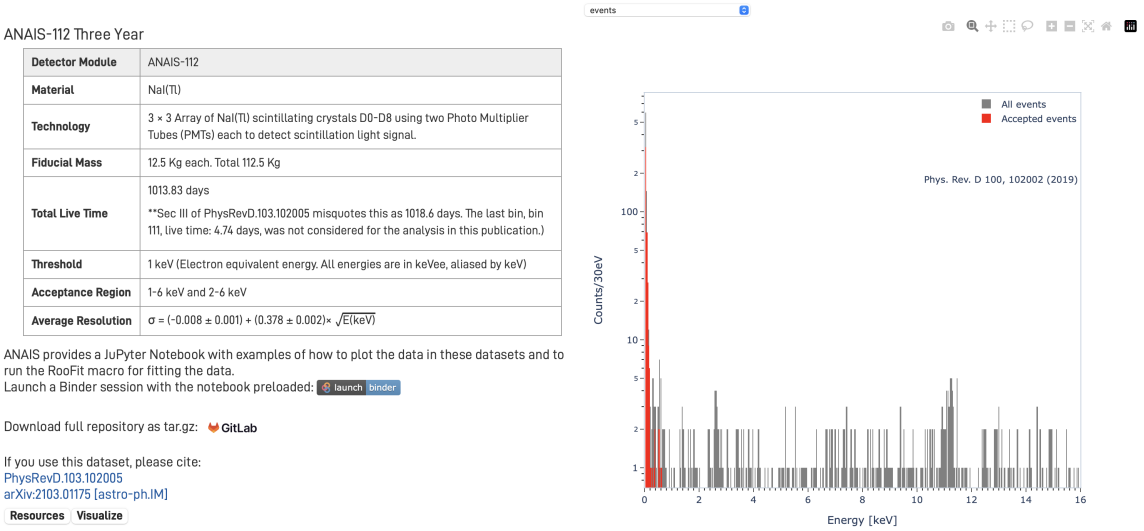
A full public data release is expected to include all three categories of information or their equivalent. The data files should preferably be in a platform independent format (like txt, csv or dat) with the headers for the columns in the first row.

2.2 Model Implementations

We can currently host only implementations of models that have an associated publication in a peer-reviewed journal. These implementations need to be self-contained systems running on a JuPyter notebook session, hence they should typically include:

1. Model Information: The routines for generating the phenomenological observables studied in the model (eg. relic densities, differential rates at experiments, etc.) in terms of model parameters.

2. Comparison Algorithms: The routines used to compare predicted observables with experimental observations, leading to posteriors/exclusions on the model parameter space.



(a) (b)

Figure 1: Fig.1(a) shows a sample dataset from the Dark Matter Data Center. One can check out and visualize the resources made available in the datarelease by clicking on the "Resources" or "Visualize" buttons. Fig.1(b) shows a sample visualization pane. Different categories of resources can be chosen to be visualized using the drop-down menu on the upper left corner. The plots are interactive. One can show/hide plots by clicking on the legend items in addition to zooming in, scaling, selecting or downloading them.

3. List of Dependencies: A comprehensive list of all external packages, softwares, etc. required to run the full workflow along with necessary citations. Eg. ROOT, MadGraph, CalcHEP, WIMPRates, MicrOmegas, etc.

4. The list of publications connected to the model implementation. The submission needs to be corresponded by one of the authors of these publications.

3. Database Management

The Dark Matter Data Center uses a GitLab interface to serve the datasets, visualizations and the Binders for use on the website. All the datasets hosted by DMDC are on the DMDC group in the MPCDF GitLab instance. The organizational structure of the database is shown in Fig.2. Each Collaboration with public data has its own subgroup. Each dataset is a project under the specific Collaboration subgroup. Each Collaboration can assign up to two members as maintainers for their subgroup. Considering that typical public datasets are not too large in volume, the GitLab hosting service is usually sufficient. However, it is possible to host volumes upto the order of $\sim$ 100 Gb on the Git Large File Storage (LFS), should the necessity arise. All datasets and workflows hosted by DMDC can be assigned a DOI, thus providing a platform where Collaborations can directly publish them. The GitLab interface provides a familiar and convenient framework for the maintenance of the database. This is typically done through maintainer accounts provided to the Collaborations in addition to support offered by the DMDC team.

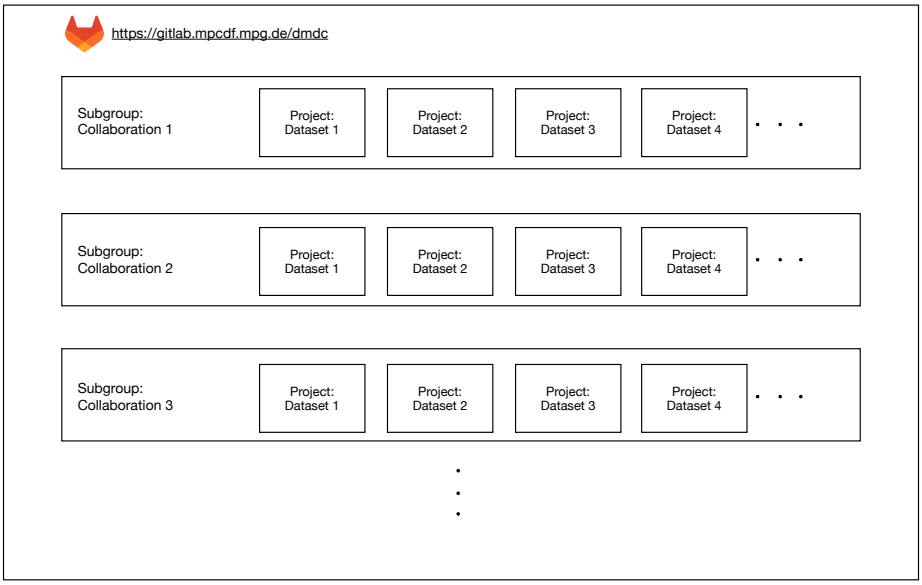


Figure 2: Organizational structure of the GitLab interface that forms the backbone of the Dark Matter Data Center

The visualizations for the dataset are hosted on the repository as html files served via GitLab Pages. Each project can have its own Page to host visualizations, which are then directly embedded into the main website. The maintainers for the subgroups can thus create their own visualizations and maintain them. The basic information about each dataset appearing on the website, as well as the descriptions for the resources are served by means of a json file on the respective repository. The maintainers can edit this file to control what appears on the website.

The online tools running as Binders are hosted on the BinderHub provided by MPCDF so the Collaborations no longer need to use public BinderHubs or their own servers to host them. These Binders can be served by simply uploading the necessary JuPyter notebook to the project repository. The system requirements for running the notebook eg. specific Anaconda environments or Linux platforms, additional softwares, specific python packages, etc. can be listed out in an "environment.yml" or a "requirements.txt" file. The DMDC team can then use the repository to create a Binder.

4. Conclusion

We have presented the Dark Matter Data Center as a unified platform for hosting experimental data, analysis workflows, software and model implementations connected to DM research. On the front end, it provides an intuitive user interface for browsing experimental progress in the field and gleaning detailed information about the experiments. The user can browse all the data releases of the experiments, read about them in detail, visualize the resources and then download them. It is possible even to test out ideas directly online using the virtual machines that include workflows

used by the Collaborations. On the back end, it uses a familiar GitLab interface for maintaining the database. The databases are handled by the respective Collaborations through maintainer accounts, providing them with control over what appears on the website. The maintainers of the subgroups are free to manage their resources. However, the DMDC team is available for helping them whenever required. All content on DMDC can be assigned with a DOI providing a single framework for the Collaborations where they can publish their data and workflows conveniently. DMDC will ensure greater visibility, persistence and reproducibility for both experimental and theoretical results and, hopefully, enhance synergy between the two. We welcome members of the DM community to contact us for submissions. The DMDC team is happy to provide assistance in preparing the datasets, visualizations and workflows/implementations. We hope to soon build a comprehensive catalog of experimental data, software and models in the field of DM research.

Acknowledgements

The Dark Matter Data Center is funded by the Deutsche Forschungsgemeinschaft (DFG - German Research Foundation) under Germany's Excellence Strategy - EXC 2094-390783311 (Cluster of Excellence - ORIGINS). The computational framework for DMDC is provided by the Max Planck Computation and Data Facility.

References

[1] Particle Data Group collaboration, *Review of Particle Physics*, *PTEP* **2022** (2022) 083C01.

[2] R. Ramachandran, K. Bugbee and K. Murphy, *From open data to open science*, *Earth and Space Science* **8** (2021) e2020EA001562.

[3] K. Cranmer et al., *Publishing statistical models: Getting the most out of particle physics experiments*, *SciPost Phys.* **12** (2022) 037 [2109.04981].

[4] CRESST-II collaboration, *Results on low mass WIMPs using an upgraded CRESST-II detector*, *Eur. Phys. J. C* **74** (2014) 3184 [1407.3146].

[5] CRESST collaboration, *Results on light dark matter particles with a low-threshold CRESST-II detector*, *Eur. Phys. J. C* **76** (2016) 25 [1509.01515].

[6] CRESST collaboration, *Description of CRESST-II data*, 1701.08157.

[7] CRESST collaboration, *Description of CRESST-III Data*, 1905.07335.

[8] CRESST collaboration, *First results from the CRESST-III low-mass dark matter program*, *Phys. Rev. D* **100** (2019) 102002 [1904.00498].

[9] XENON collaboration, *Light Dark Matter Search with Ionization Signals in XENON1T*, *Phys. Rev. Lett.* **123** (2019) 251801 [1907.11485].

[10] J. Amare et al., *Annual modulation results from three-year exposure of ANAIS-112*, *Phys. Rev. D* **103** (2021) 102005 [2103.01175].