ISOLDE Collaboration Open Data Policy and Advice

Approved by ISCC – 15th June 2023

This document sets out an "Open Data Policy for the ISOLDE Collaboration", followed by advice to spokespersons of ISOLDE proposals on important aspects of a data management plan.

OPEN DATA POLICY FOR THE ISOLDE COLLABORATION

The ISOLDE Open Data Policy reflects values that have been embodied in the CERN Convention for more than sixty years and formed the basis for CERN's own Open Science Policy[1].

Curating data and making it available in a responsible fashion using the FAIR Guiding Principles[2] maximises the scientific potential of the data and fulfils collective moral and fiduciary responsibilities of CERN member states. There are many potential opportunities to reuse data that include reanalysis and reinterpretation of physics results, education and outreach activities, analysis with the aim of technical or algorithm development, and additional physics research.

ISOLDE is a facility rather than a single experiment. Groups of users develop beam line installations that they use to perform individual experiments with specific radioisotope beams. The beam line installations are owned and operated by the users who control the data collection. Each experiment is done by different subsets of the ISOLDE Collaboration and each individual experiment has its own spokesperson.

This scientific culture is different from the operation of large experiments at CERN and so this document describes an approach towards openness and preservation of data that is suitable for the ISOLDE Community, based on the CERN Open Science Policy and informed by the Policy for LHC Experiments[3].

Experiments at ISOLDE are undertaken with a diverse range of instruments and methods, addressing topical issues in a wide range of different disciplines including nuclear structure, reactions, and astrophysics; fundamental interactions, atomic and molecular physics; material science and condensed matter physics; and aspects of chemistry, biology, and medicine. As a result, the data produced by ISOLDE users vary significantly in form and complexity from experiment to experiment. Different analysis methodologies are employed, leading to a variety of different intermediate data structures and to many different types of results.

Given this breadth of activity and the mode of operation of the facility, there can be no single approach to curation of ISOLDE data and so each individual experiment spokesperson carries the responsibility for developing an effective data management plan for the specific data that they create. A flexible approach is consistent with the range of ISOLDE science and ensures that any requirements imposed by their specific academic institutions or funding agencies can also be met easily by the spokesperson.

Notwithstanding that flexibility, each spokesperson (after discussion with the participants, as required) should ensure each experiment has a data management plan, before the experiment is undertaken, covering at least the following points:

[1] *https://openscience.cern/policies*
[2] *https://www.go-fair.org/fair-principles*
[3] *https://cds.cern.ch/record/2745133/files/CERN-OPEN-2020-013.pdf*

- Data from an ISOLDE experiment should be made freely available, after a specified time, in a suitable form and at a practicable level of abstraction. The form, level of abstraction and when to share data are decided by the spokesperson.
- These data should be stored securely in any trusted data repository that provides a persistent data identifier with which the data can be cited.
- Sufficient contextual information or meta data should be made available, or referenced, to enable the reuse of the shared data.
- Data should be made available in a way consistent with the Creative Commons 0 waiver. However, the spokesperson may wish to give expectations on the use of the data consistent with scientific norms.

This policy is written on the assumption that ISOLDE experiments do not involve any personal data; in the unlikely event of the need to handle such data, additional provisions are needed, and the spokesperson should seek advice in the initial planning stages of the project.

In order that the ISOLDE Facility can locate shared data and demonstrate compliance with CERN guidance and other external requirements, the spokesperson will be required to supply a short summary of their data management plan when making beam requests for their experiment in the form of a short questionnaire.

ADVICE TO SPOKESPERSONS

The ISOLDE Open Data Policy describes the general approach to Open Data approved by the ISCC. This following advice is intended to help spokespersons of INTC proposals to develop a Data Management Plan (DMP). Indeed, addressing the following points would constitute a DMP. The text in italics after each point is NOT intended to be prescriptive, but instead intended to provide advice or suggestions that can be tailored to specific needs of the spokesperson.

Specify the type of data produced and in what format will it be stored and shared?

During the initial analysis, the raw event data is often processed to create intermediate data structures (e.g., spectra, matrices, etc.) and final observable results are obtained (e.g., cross sections, level schemes, energies, intensities, etc). These are often interpreted to give model-dependent quantities (e.g., spectroscopic factors, moments, transition rates, strength functions, etc.). Both final observable results and model-dependent quantities should be given in an open-access research publication.

For the underlying data, a pragmatic decision needs taking over what level of abstraction the data should be made available. This balances the benefit of the reanalysis of the data against practical issues, such as file sizes and complexity. Often raw event-by-event data is impractical due to difficulties in storage volume. Moreover, in some experiments, the recreation of intermediate data structures might only have limited benefits. In these circumstances, sharing spectra or similar intermediate data structures might be a more practical way forward. It might help in decision making to think about the data might be valuably reused and how to facilitate that reuse.

What documentation or metadata is needed for the data to be used in the future?

To reuse data, additional information is often needed, e.g., experimental setup, calibrations, data formats etc. This is often given in associated publications or theses – if these are open access, a reference to them might be an easy way to provide necessary information. If providing spectra, the format should be specified (for common formats like ROOT or Radware, this might be done via a reference). For 1D spectra, ascii might be the most straightforward. If providing raw event data, the format should be specified, or software given to read it. If specialist software is required to reuse the

data, this should be made publicly available or an indication of how it may be obtained, for example, if it is commercial software.

How will the data be stored and shared?

During the initial analysis phase, this might be local storage where the usual advice about backing up is appropriate (although use of a repository from the start might be more secure). For long-term storage and to share the data more widely, there are many different trusted data repositories that might be used. It is best practice to use a repository where the data is tagged by a Digital Object Identifer (DOI), which can be used to identify the data set and might be given in publications to publicise the data (this is required by some funding agencies).
You may be required to use a repository that your institution or funding agency provides, or you may already have your own favourite. If not, Zenodo (https://zenodo.org) is a good solution – it is intuitive to use and quick to set up, assigns a DOI to each dataset, and also can be used as safe private storage until the data is made public.

When will the data be shared?

A decision needs to be made about when the stored data is made publicly accessible – that timescale will set the proprietary period in which you have sole use of the data. For example, this could be a set period or date after the experiment, or it could be after all your own analyses have been published.

Who will implement the DMP?

The most efficient way to implement the DMP is often for the person performing the direct analyses of the data (e.g., student or PDRA) to collate and upload data and information to the chosen repository; they have first-hand knowledge and should be able to quickly perform the necessary tasks. But whoever is chosen must be made aware of any relevant formalities (e.g., informing the experimental collaboration, any credit and acknowledgement requirements, etc.) and the experimental spokesperson should ensure that the archiving has been properly completed.

What expectations should be placed on the data reuse?

CERN policy mandates the use of the Creative Commons 0 waiver, which gives unfettered use of data. But you may want to ask data users to follow some appropriate scientific norms. These may include the need for appropriate acknowledgement of the original collaboration; the need to make sure any third-party publications are clearly distinguishable from the original collaboration; a disclaimer that any scientific claims made in such publications are the responsibility of their authors and not the original collaboration; or similar requests.