Data Management

Cindy Kunkel*, Katja Heeschen*, Knut Behrends* and Ronald Conze§

Besides samples, data is one of the major products of a scientific drilling project. Data Management is therefore essential for both successful project completion and potential follow-up operations and publications. It is a long-term task that extends throughout the entire project and deserves a high level of attention from both the project management and the entire Science Team.

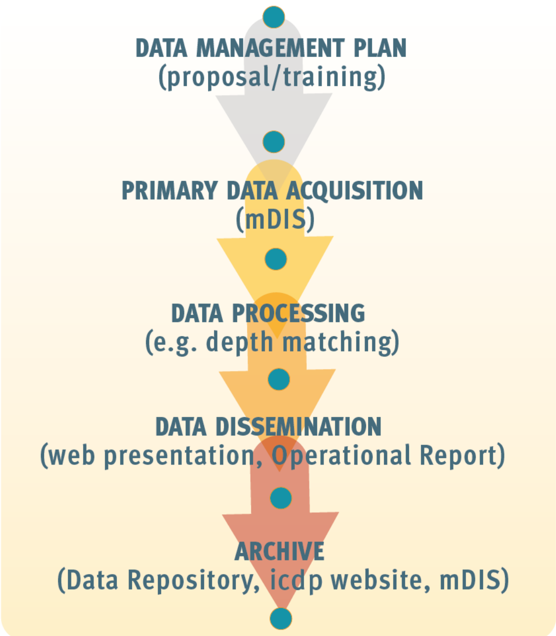

Fig. 4.3.1: Data Management Lifecycle phases of an ICDP drilling project

Ideally, all data gathered during the project forms a common basic data set (Operational Data Set), which is accessible for all Science Team members during the projects' drilling and sampling phases as well as for the whole science community at latest after the moratorium period. For that reason, ICDP provides the mobile Drilling Information System (mDIS), which is a database management application that ensures data integrity.

During the project lifetime Data Management is divided into five distinct phases, which are essential for any successful project and will be discussed in detail in this chapter and is shown in Fig. 4.4.1. The first phase is the development of the Data Management Plan (DMP), which should already be outlined in the Full Proposal. The second phase entails the primary data acquisition during drilling and laboratory operations. The third phase is the data processing, in which, for example, the measured core depth is matched with the depth gathered by geophysical logging measurements. The data gathered from the primary data acquisition and the data processing together form the Operational Data Set, which is in the fourth phase disseminated first to the science team and at latest after the moratorium period to the general public. The final data management phase comprises long-lasting activities such as archiving, ensuring that no data will be lost, and compiling key datasets produced by participants after the project ended.

Phase 1: Data Management Plan

The Data Management Plan, DMP should already be outlined in the full project proposal and it should address and document all lifecycle phases (Fig. 4.4.1). Principal investigators should describe in detail which financial means and resources will be needed during the operational phase for the data management. A dedicated person for data and also sample management in the project can be very useful. The DMP should include budgets for human resources, hardware, transport of devices such as the core scanner as well as the travel costs for the Data and Sample Management (mDIS) training course (s. Supplement S1). Within the training course, participants will develop the final Data Management Plan and the details for data acquisition.

Phase 2: Primary Data Acquisition

With the plan in place, the next phase is the data acquisition, which starts with the first shift on-site and continues until after the sampling party (Fig. 4.4.2). As long as the fieldwork proceeds, project scientists will collect data in a way that is commonly unique and project-specific.

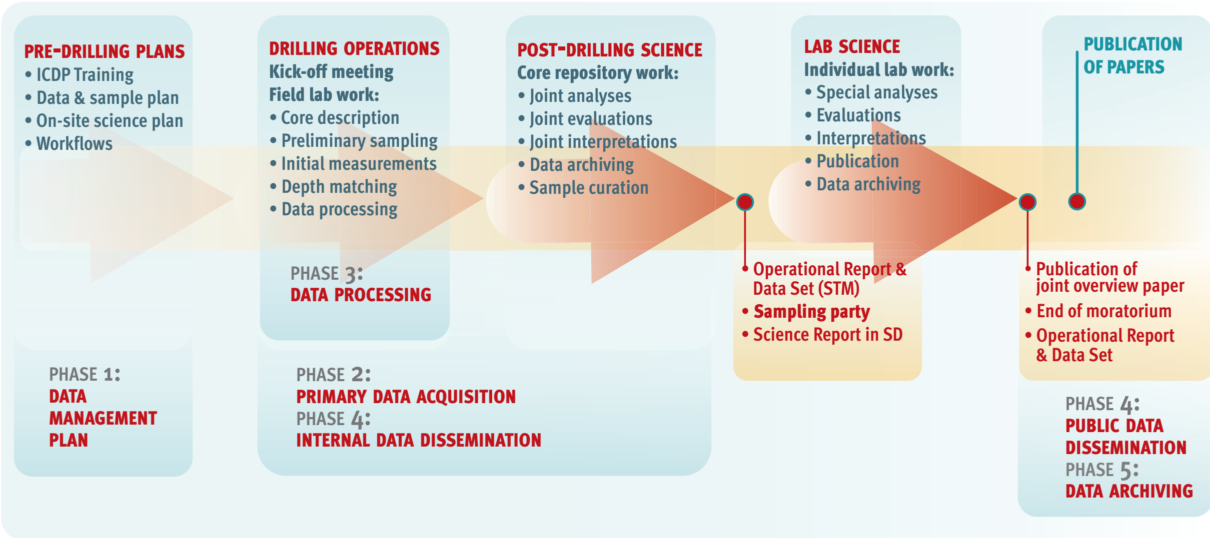

Fig. 4.3.2: Data Management Phases during an ongoing ICDP drilling project

The data comprises a multitude of drilling parameters as well as details on the recovered material, sampling and lithological descriptions.

Many drilling projects decide to limit the onsite workflow to capturing the technical parameters of the drilling operations and produce only essential reports, citing recovered sample material such as cores, cuttings, mud, fluids, and gases. Other drilling projects perform imaging and initial lithological descriptions onsite as additional part of the project documentation. Additional measurements for continuous petrophysical and/or geochemical properties can be included. If sampling is allowed for reasonable special cases, these samples must be tracked.

A basic architecture of a typical data acquisition and workflow model is shown in Fig. 4.4.3. If data acquisition such as core logging, scanning and description is not possible in the field, it can be carried out simultaneously or subsequently under more reliable conditions in the laboratory or core repository.

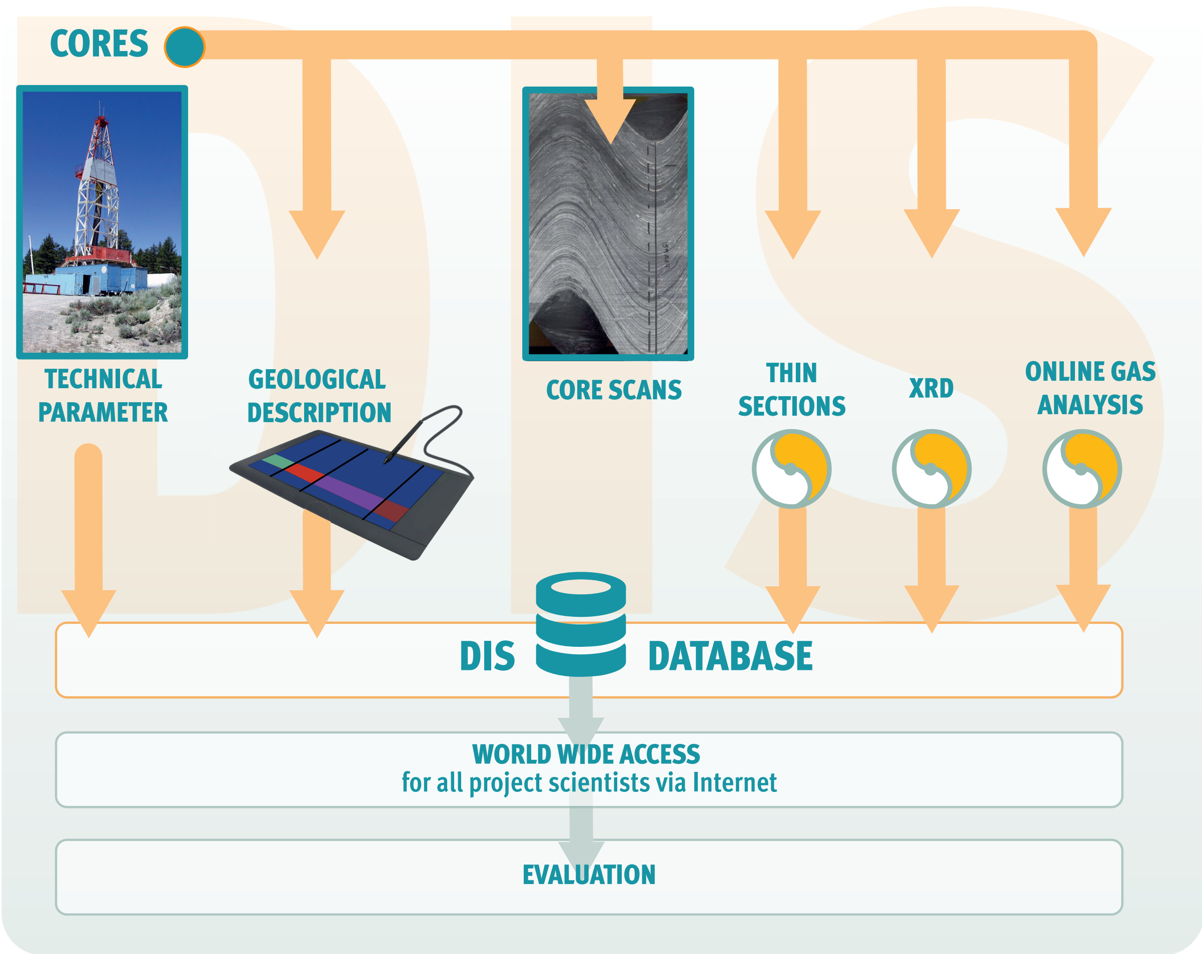

Fig. 4.3.3: Workflow example for the data management of the Long Valley project in ICDP

Consequently, the expedition is divided into two phases: drilling operations and laboratory work. This often takes place with a significant lag time due to the transfer of all the sample material from the remote sites to the target lab.

Data acquired during the laboratory work is also part of the Operational Data Set. All data gathered in the field and the laboratory will be stored in the mobile Drilling Information System (mDIS), which is established on a local server on-site, nearby the drilling operations or in field laboratories. In case of a stable internet access, the mDIS server can also be set up as a remote server, e.g., run by ICDP and the data is then accessible remotely via a web browser by authorized project members, who are not on-site and can use it for remote cross-checking and quality control (QC) purposes. Regularly, staff uploads the field data to the ICDP project web site where it is accessible to all science team members (e.g., http://cosc.icdp-online.org/).

Phase 3: Data Processing

Data acquisition is directly followed or accompanied by the data processing phase (see Chapter 4.6). Processing may include depth matching of measured core depth with the depth gathered from geophysical logging or by creating composite logs from a number of neighbouring drill holes. That processed data is the foundation of all subsequent scientific work and, thus, needs to be finalized before the sampling party.

Phase 4: Data Dissemination

For Data Dissemination, ICDP creates a web site for each ICDP project after the first approved workshop proposal at: www.icdp-online.org/projects/.
Generally, a project figure, abstract and

logo make up the cover page, while topics such as News, Scientists, Press & Media, Publications, Workshops, etc. are updated as required on the project page. With the project developing and according to actual project activities, the project web site grows; messages of the day, a photo gallery and internal data are added. When the project is ongoing, it usually receives more attention from the public. Accordingly, the project will also be featured as a Highlight on the ICDP web site as well as on the ICDP social media accounts Facebook, LinkedIn, and twitter.

To enhance outreach, project PIs can also maintain own websites and/or use their preferred choice of modern social media. The ICDP website creates an abundance of links to the project-specific contents of these external media.

New scientific data is usually confidential. Therefore, ICDP puts such data sets under secure access for registered science team members only. This protected area serves as a knowledge transfer platform within the science team and is very useful to include science team members that are not onsite and for selecting samples. If the internet connection onsite is stable or if the mDIS is hosted on a central server, it is also possible to directly grant access to the mDIS for different user groups.

After the moratorium period the Operational Data Set is published together with the Operational Report as supplement to the Science Report. PIs are encouraged to publish these reports in the Scientific Drilling Journal (see Chapter 2 and the guidelines on report writing in Supplement X), where DOIs are assigned. Afterwards, modifications on the data set are usually not possible anymore as it is the common foundation for all scientific work to follow.

Phase 5: Archiving

Data archiving is an important part of the lifecycle because it ensures that no data will be lost and available after the project ends. All raw and processed data should be stored in an archive for secure long-term preservation. At the end of a successful data management lifecycle the scientific output, e.g., publications, will stimulate discussions and ideally cumulate in further projects.

Data/Sample Training Course (mDIS)

Within the training course the data and sample management cycles and the details for data acquisition, core handling procedures and on-site workflows will be discussed and developed. Furthermore, the tool for data acquisition, the mobile Drilling Information System (mDIS), will be presented and adapted for the specific requirements of the project. At the end of the course, a final version of the mDIS should be installed on the laptop that will be exclusively available for data management at the drill site/laboratory.

For all coring ICDP drilling projects this training is mandatory and its cost should be part of the budget plan. For the training, the general regulation is that the project covers the travel costs to and from Germany (including visa). ICDP covers the accommodation in Potsdam, the daily rate for meals, all training materials, and local transport. The training will last 5 days and is usually hosted by the OSG data management group at GFZ in Potsdam, Germany. The mDIS training course should take place about six months prior to drilling.

Participants should include at least one PI and a person responsible for the data and sample management in the field and in the laboratory (if applicable). In addition, 3 - 5 members from the on-site science team should participate. It is preferential if at least the data manager and one additional person have some skills in computer handling and data acquisition. It is intended that they provide guidance and education of additional staff entering data, oversee consistency and quality/security of the data acquisition and function as relay for distributing reports. Besides the training before the drilling operation, the ICDP OSG also offers support during the field operations as well as remote support after the initial set-up in the field.

Fig. 4.3.4: The mobile Drilling Information System is platform-independent, open source and can be used on different devices due to its responsive design.

The mobile Drilling Information System

The mobile Drilling Information System is a database management application developed and provided by ICDP for capturing and curating essential data on geological samples, drilling progress, and related datasets such as images or geophysical well logs. It is easily adaptable for distinct project needs and ensures that all data gathered during the project is compiled in the Operational Data Set, which is accessible for all Science Team members during the projects' operational phases and for the whole science community at latest after the moratorium period.

The mDIS concept

The Drilling Information System is designed to be used onsite in parallel with daily operations to perform the data acquisition alongside a defined workflow. This is helpful for avoiding the excessive creation of non-synchronized and non-authorized data files.

The mDIS will be deployed and modified during the mandatory training course with the possibility for remote follow-up training and online support during the field operation.

The mDIS structure is hierarchical and reflects the ICDP naming conventions. It starts with program as highest hierarchy level, followed by expedition, site, hole, core, section, sample, etc. levels (Fig. A, see Chapter 4.5, Sample Management for details on naming conventions).

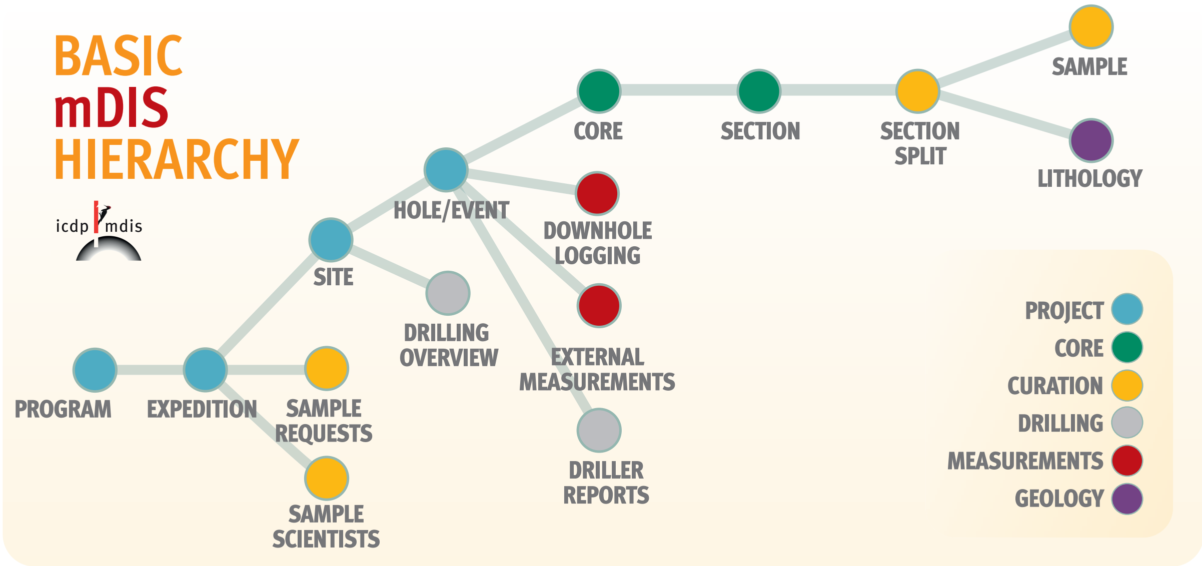

Fig. 4.3.5: Hierarchical structure of mDIS following the overall ICDP naming conventions. Colors indicate thematic groups of data (table sets).

Each hierarchy level is defined by a data table and an entry mask (form), which can be customized by trained users within the mDIS application for project specific needs in three steps:

- Create a table
- Create a form
- Use form to insert data

Data to be inserted includes all drilling and sample relevant information as the technical drilling parameters, core data, core scans, lithological description, photos and initial measurements (Fig. B).

Within the data-acquisition workflows, certain automated data-consistency checks and human quality controls ensure the data quality. Data integrity is enforced in terms of measurement units, date and time formats and naming conventions at the time of data capture, before it is safely stored in the relational tables of the mDIS project database. Furthermore, mDIS automatically assigns International Geo Sample Numbers (IGSN, Chapter 4.5) that can be exported and directly send to an allocation agent for registration.

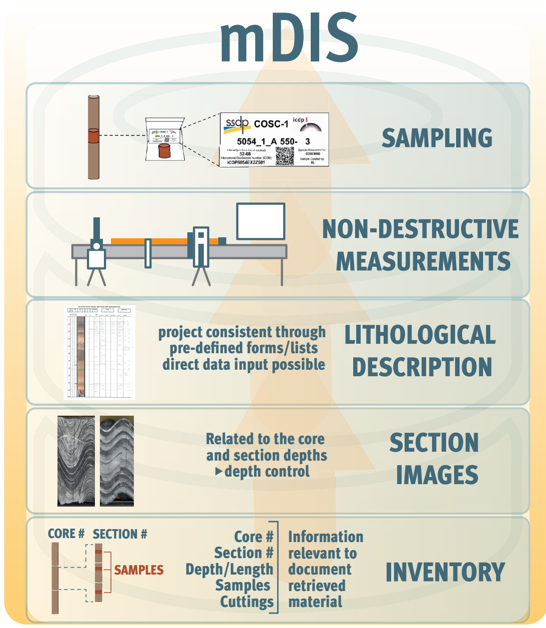

Fig. 4.3.6: Data from various sources is stored in the mobile Drilling Information System to build a complete data set available for all Science Team members during the operational project phase.

Once the data is stored it is possible to print labels for sections and samples including a QR code for easy access and findability. Additionally, the data can be printed in different kinds of reports and can be used as source for the Operational Report (Chapter 2, Operational Report) since the data can easily be transferred into external data-processing applications and spreadsheets.

mDIS set up

mDIS is platform independent, responsive and open source (Fig. 4.4.6). It is set up in a server-client environment, where a dedicated personal computer acts as mDIS server. The native mDIS runs on Linux, but it is deployed as a virtual box instance, which can be installed on Windows, Mac and Linux systems and contains the data base system and the mDIS application. As long as a device has a browser it can be used for accessing mDIS as client to view, update, edit or upload data.

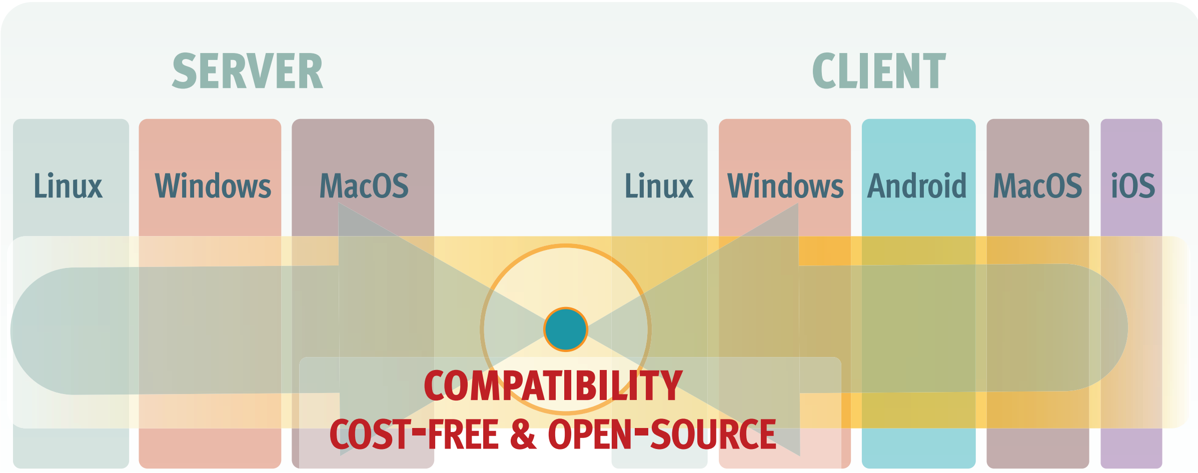

Fig. 4.3.7: The main mDIS (server) will be installed on a virtual box, which can run on Linux, Windows and MacOS. As long as a device has a web browser it can access mDIS.

Field Application

The mDIS can be used either as standalone version on a laptop, via a local network or via the internet. The standalone version is useful for data entry in remote areas without internet access, or for any small project, where all data can be inserted locally on one PC. However, if data acquisition facilities are being distributed across a larger area, such as a large field site (and separate labs), or a fleet of research vessels, it is useful to access the mDIS server via a wide area network or via the public internet. Other external devices such as core scanners or core loggers can also be included into that network. If the mDIS system at the field site can be connected to the wider internet, it is possible to upload daily updates and progress reports to the dedicated project website and/or archive servers or grant access to mDIS via a web browser for the whole Science Team. It is also useful for remotely supporting the mDIS operator and the mDIS system itself.

Technical specifications

Technically, mDIS is a relational database backed web application built on a LAMP stack (Linux, Apache, MariaDB, PHP). On the client side, mDIS is based on a Javascript framework (VueJS) that, in conjunction with the Yii PHP framework, allows for reactive two-way data binding. mDIS can be deployed in desktop environments and on servers. On mobile devices, mDIS can run as a Progressive Web App (PWA), which works almost like a native app. mDIS has REST application programming interfaces (APIs) for third-party application developers and external data providers. These APIs can be used for importing data from legacy DIS installations and other data sources such as text files and collections of core imagery.

Cindy Kunkel, Katja Heeschen*, Knut Behrends* and Ronald Conze§*
**Operational Support Group ICDP, GFZ German Research Centre for Geosciences, 14473 Potsdam, Germany*
c.kunkel@icdp-online.org; k.heeschen@icdp-online.org; k.behrends@icdp-online.org
§ Retired, formerly at OSG