CHAPTER 5

Data and Sample Management

Ronald Conze*, Thomas Gorgas and Knut Behrends*

The main objective of Data and Sample Management is to acquire key information about the technical and scientific works performed during the operational phase. Works are typically performed in the field, in the lab, or at the sites where sample material is stored. Ideally, the resulting output provides a comprehensive data set that can serve as a common reference. Validation of this data set should be completed when most of the science team members start their scientific work. Therefore, data and sample management is an important service during the lifecycle of a drilling project. Dedicated planning including the definition of data management policies is a prerequisite for success.

Lifecycle

The general Data Management Lifecycle is outlined in Figs. 5.1 and 5.6 from the data and sample management point of view. Starting with the first proposal (see: Chapter 11 on Proposal Writing), principal investigators should describe in detail which financial means and resources will be needed during the operational phase for the data and sample management. The proposal should include budgets for hard- and software, transport of devices such as the core scanner, or the sample material from the site to the lab and/or repository, and the travel costs for a training course or workshop that is focused on the planned operational data and sample management in the field, the labs, and storage of sample material of that specific project. Beginning with the first shift onsite, the acquisition of the primary or basic data commences. Regularly, staff should upload these field data to an official Project Web Site (e.g., http://cosc.icdp-online.org/), and store raw and processed data in an archive for secure long-term preservation.

As long as the fieldwork is going on, each shift will collect data in a way which is almost always unique and project-specific. In many cases, after the final hole has been drilled, a certain period of lab work happens next. This phase is also part of the primary data acquisition. Toward the end of the operational phase, the sample material should be ready for sampling and distribution to remotely operating members of the science team. An important final document, the "Operational Report" comprises all this information and serves as the common reference for all follow-up activities, such as scientific and engineering analyses, that usually are published in scientific journals. The Operational Report is a public document. Typically, during the operational phase, physical sample material and online data can exclusively be accessed only by registered science team members (secure access). However, after a project-based, pre-defined moratorium time period, eventually most locked-up publications and the rest of the sample material become publicly accessible. This kind of data management Lifecycle (Fig. 5.1) is repeating itself in a similar way for each new drilling project, independent of the scope of the project.

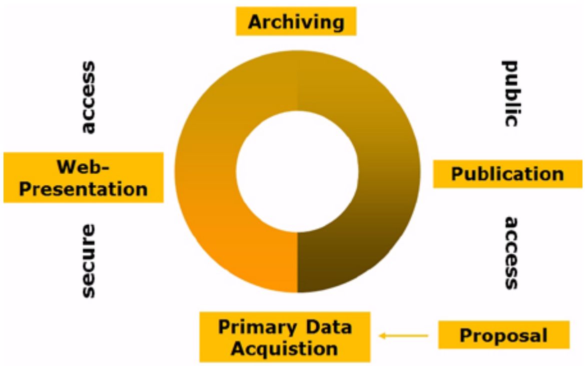

Fig. 5. 1: Lifecycle of data management in drilling

The Science Team

The Science Team is of central importance for the data and sample management because here the producers and consumers are the same people. The Principal Investigators (PIs) and cooperating PIs (CO-PIs) are naturally the major stakeholders of the project. They set priorities for incoming sample requests and proposals. They have to work out the specific plans and budgets needed for data and sample management. If needed, they also designate Chief Scientists for the different subprojects, and they finally assign scientists, students, technicians and volunteers to the science team. It is imperative to identify and prepare certain individuals to key responsibilities early in the project. This also holds for the data and sample management tasks as they arise during the project (e.g. through aforementioned training courses, workshops, etc. – see also: Chapters 10 & 11 on Outreach, Education and Training, and Proposal Writing). In most cases, the total number of science team members is larger than the group which is doing the field work and lab work during the on-site operational phase (Fig. 5.2). Therefore it is important to set up a-priori policies (see also: Chapter 12 on Project Funding and Policies) among all key players of the project to avoid conflicts of interest that may create data and sample management issues later on.

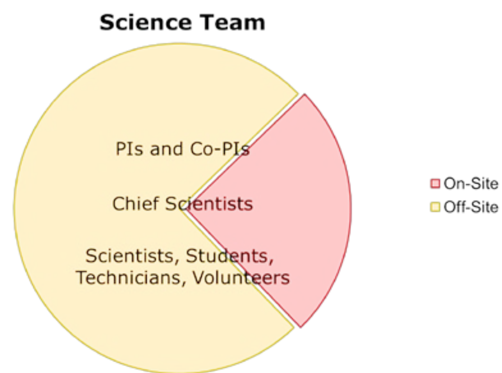

Fig. 5.2: Typical composition of a science team

In addition to the science team, a number of service companies, sub-contractors, and other project-aids (sometimes in form of volunteers) are involved throughout the various project segments. They often contribute to data acquisition in different ways, and thus are an important integral part of the science plan. These topics and fine details of a project have to be negotiated carefully beforehand.

Policies

Sound and reasonable policies for a proper project management are required throughout every successful ICDP project (see also: Chapter 12 on Project Funding and Policies). The content of these policies should already be discussed and confirmed during the proposal development phase. Each science team member should commit to these rules and guidelines before the planned start of the operational work. The main topics are:

- Moratorium periods and milestones along the timeline
- Science Team – Selection of participating scientists, responsibilities, duties and privileges
- Data acquisition and sharing
- Scheduling and distribution of reports
- Sampling strategy and sample distribution
- Publication guidelines along the timeline
- Public outreach issues and internal confidentiality agreements

Information system for Scientific Drilling

Any information system for scientific drilling projects can be divided into three levels according to the main purposes of the data and sample management as shown in Fig. 5.3:

- Data acquisition
- Data dissemination
- Data publication

For data acquisition purposes, ICDP provides the Drilling Information System (DIS) which can easily be adapted to the individual requirements of any specific project. The task of running and maintaining the project-specific DIS is usually established and located on-site, nearby the drilling operations, in field laboratories, shore bases, buildings of institutes and/or storage places for sample material. DIS is designed for the use in a small 'closed shop' environment (a small, private, local area network) and optionally communicates project data via the Web-based eXtended DIS interface (X-DIS).

The DIS-Administrator is able to define which data are shown, which forms, reports, or data views can be selected, and who is authorized to edit (insert, overwrite, delete) which subset of data via X-DIS. One advantage is that the DIS-administrator can perform certain maintenance features remotely, another benefit is that certain project members, e.g. principal investigators or chief scientists can use it for remote cross-checking and quality control (QC) purposes.

For data dissemination it is recommended to use modern Web based transfer mechanisms and online media. This online interface platform acts as user interface for the public in order to fulfill outreach purposes and as user interface for the science team members to access internal project data and information.

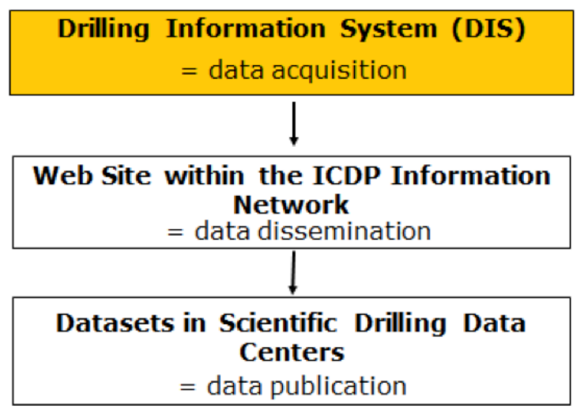

Fig. 5.3: Three level architecture of an information system for scientific drilling projects

For data publications, it is recommended to use established data-sharing services from institutional or commercial data centers, sometimes labeled as "World Data Centers". Many publishers allow adding Supplementary Materials to online versions of already published papers.

The Drilling Information System DIS

ICDP provides the Drilling Information System DIS for operational support and on-site data management. It is designed as a software toolbox to build and maintain customized DIS instances for any distinct drilling project. The software is based on a project-specific and internally consistent database, which integrates different types of information (various measurements and data sets). The graphical user interface of DIS utilizes specific but still customizable data-input forms, and templates for both tabular data-views and printable reports. The main purpose of the DIS is data acquisition for the documentation and administration of:

- basic – initial – primary data
- initial measurements and reports
- sample requests, sample curation and sample distribution

in order to establish a common data set and reference for all science team members.

The DIS is designed to be used on-site 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. Toward the end of the on-site drilling phase, the collected data should go through a depth matching process to synchronize different depth regimes and to integrate downhole logging data. Finally, the built-in templates of DIS can be used as source for the Operational Report.

However, the DIS will never be an active online real time monitoring system, or an active measuring or logging system. The DIS does not include any applications performing sophisticated exploratory data analysis for interpreting or evaluating data. These software design-decisions have been made on purpose. Experience shows that researchers prefer their own toolsets for analyzing and visualizing science data anyway.

The concept of DIS defines data-acquisition workflows that focus on certain automated data-consistency checks and human quality controls. 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 within relational tables within the DIS project database. The data contents of the measurements can easily be transferred into external data-processing applications and spreadsheets.

Technical setup and scalability

The DIS can be installed as a single standalone, even mobile system, or in context of a local area network depending on the environmental options on-site (Fig. 5.4). The central part is always a dedicated personal computer acting as DIS server which contains the data base system and the DIS user interface. If data acquisition facilities are being distributed across a larger area, such as a large field site, or a fleet of research vessels, the DIS server can be cloned into several instances. These can be kept in synchrony by means of a built-in mechanism known as data base replication.

Any number of DIS client computers can connect to the dedicated DIS server. They can be added using wired or wireless network connections. A DIS client does not store any data locally, but instead has only the user interface for data input installed. Other external devices such as core scanners or core loggers can be also part of that network. The simplest interface is a shared file system of the used network. If the device allows, a DIS interface can be added.

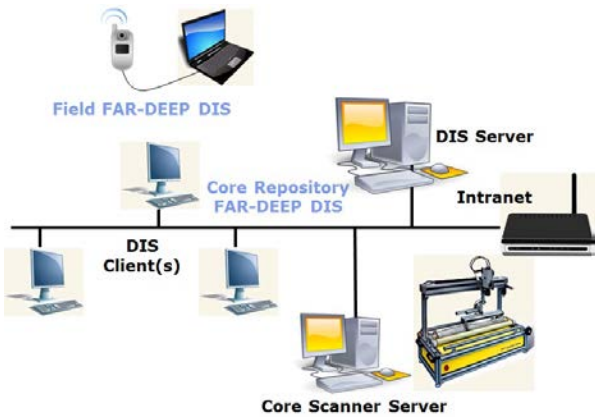

Fig. 5.4: Set up of a typical field or core repository DIS, showcasing the DIS of the FAR-DEEP

If the DIS system at the field site can be connected to the internet, it is possible to upload daily updates and progress reports to the dedicated project website and/or archive servers. It is also useful for remotely supporting the DIS operator and the DIS system itself. Under certain circumstances it might be required to configure and enable the eXtended DIS interface, which allows a secure remote access to the operational DIS on-site. For more details and valid versions of the DIS operational procedures please **click here** (www.icdp-online.org – Support – Data Management –Technical Requirements).

Standards and naming conventions

For describing the most important features and attributes of wellsite data and geological field data, ICDP uses similar terms and naming conventions as IODP has introduced. In both data-models, the terms are arranged in a relational hierarchy:

- *Expedition* is the operational phase of a scientific drilling project
- One or more *Sites* can be visited during an Expedition
- One or more *Holes* can be drilled on a Site

The data model of the German data center PANGAEA also includes an extension:
One or more *Events* can take place on a Site, e.g. the event 'Drilling a Hole' (or 'Water Sampling')

ICDP / IODP	ODP	PANGAEA
Program		
Expedition	Leg / Cruise	Project / Campaign
Site	Site	Site
Hole	Hole	Event

Table 5.1: Conceptual schemes and terms of IODP, ICDP, and PANGAEA

As shown below, the used naming conventions can be quite different. To over-come these differences, the DIS allows unconstrained/free-formatted naming sche-mes as long as these are used consistently throughout a single expedition or project.

	Expedition	Site	Hole
IODP	312	M0005	A
ICDP	5023	1	A
LacCore	GLAD5-BOS04	1	A

Table 5.2: Conceptual schemes and terms of IODP, ICDP, and PANGAEA to define expedition, site and hole for each project

Main tasks and personnel requirements

The Chief Geologist is responsible for maintaining the integrity of the pre-defined science plan and sampling plan. Accordingly the chief geologist selects supervisors and helpers (program-aids called within IODP) who perform the daily work of data and sample acquisition and management (Fig. 5.5). For installing and operating DIS, a staff member with a certain expertise and skill set should act as IT expert for the administration and maintenance of the system. Additional responsibilities of this IT manager role are:

- guide and educate additional data entry staff,
- take care of shift plans of the data entry teams,
- oversee consistency and quality/security of the data acquisition (Fig. 5.5),
- interface as relay for distributing reports.

Rule: Data and sample management is not a just technical issue.

Expedition

It is good practice to consider Data and Sample Management as long-term tasks that start with the field work and extend throughout the entire period of the expedition and project duration. Therefore data management deserves a high-level of attention from both the project management and entire science team.

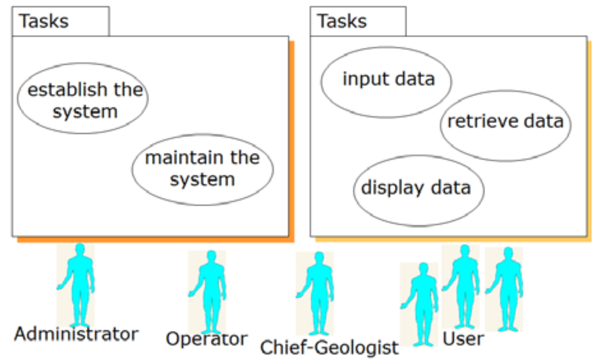

Fig. 5.5: Tasks and personnel for data and sample management

In a typical timeline of a scientific drilling project, a training course or workshop is the first major event to kick-off data- and sample management activities. Such a training and/or workshop should be conducted within a six-month period prior to starting drilling operations. Generally, the crucial phase of the expedition starts with rigging up for the first hole and ends with the completion of the operational and/or operational report. During the drilling phase, the initial project data are collected as they relate to a multitude of drilling parameters and the intrinsic details of the drilling operations, the recovery of the material extracted from the hole, its sampling, creation of descriptions and documentation, downhole logging data, and so forth. However, in many cases not all of these tasks can be executed and performed on-site due to harsh environmental conditions and/or a lack of available space. Consequently, the expedition is then divided into two phases of drilling operations and lab work. This often takes place with a significant lag time due to the transfer of all the sample material from the sites to the target lab (Fig. 5.6).

Rule: Avoid any task which is not directly important for decisions regarding field operations, and which can be taken care of better in the lab than in the field.

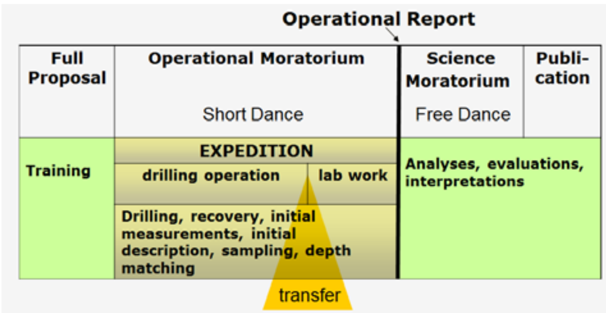

Fig. 5.6: Typical timeline of a scientific drilling project

Expedition DIS

The Expedition DIS is designed for use in the field or on board (onsite/offshore) and in the lab (onshore). Due to its relatively simple technical setup and scalability it is easy to handle. The basic architecture of a typical data acquisition and workflow model is shown in Fig. 5.7.

Ideally the grant proposal and the science plan contain the outline of a data management workflow. This only exists in conceptual form on paper. The DIS operator must convert the predefined, conceptual workflow into an individually designed ExpeditionDIS of the project.

This should happen before drilling starts. The ExpeditionDIS can be customized according to the actual environmental situation and requirements. This customization can be a complex, unfamiliar task to most people on the science team. The ICDP OSG offers training and support before and during the field operations as well as remote support after the initial set-up in the field.

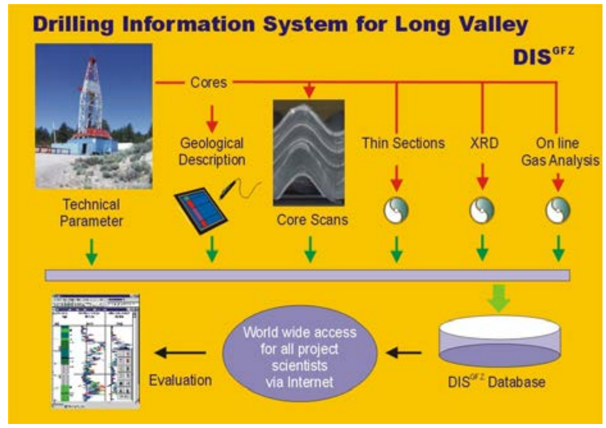

Fig. 5.7: Scheme of data management for the Long Valley project in ICDP

Many drilling projects limit the onsite workflow to capturing the technical parameters of the drilling operations and producing corresponding 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 have to be tracked. To this end, persistent identifiers can be used. Recently, the capability to tag samples with "International Geo Sample Number" (IGSN) identifiers has been added to the ExpeditionDIS system. IGSNs are worldwide unique IDs that can be used as digital link to almost all information related to the object. These samples may be treated for onsite thin section preparation and analyses or even XRD measurements. Preferably the basic data acquisition can be entirely done onsite, as demonstrated by the Chinese Continental Scientific Drilling Project near Donghai.

The scope of subsequent analyses encompasses mainly descriptive methods and automated measurement procedures. Each method and each step along the workflow can produce various data formats in different units and scales of resolution. The DIS allows for configuring specific scripts ("data pumps") to harmonize these data using a common naming convention and standards for date and time, depth scales and units. As soon as the data are stored in DIS tables the data can be copied to the project specific Web sites and/or further processed for reports.

Sample strategy

The statement below is fully applicable to scientific drilling although it is derived from planetary and space science (Allen Carlton et al. 2013, Curating NASA's extraterrestrial samples. EOS 94(29):16.7.2013):

"Through nearly a half century of work on analyzing and curating samples from places beyond Earth, a few key messages stand out.

First and foremost, the main point of any sample return mission is laboratory analysis. Everything must be designed, built, and operated to get the highest-quality samples to the best laboratories.

Further, curation starts with mission design. Samples will never be any cleaner than the tools and containers used to collect, transport, and store them. Scientists and engineers must be prepared in case missions do not go according to plan. Really bad things can, and do, happen to missions and to samples. Careful planning and dedicated people can sometimes save the day, recover the samples, and preserve the science of the mission.

Every sample set is unique. Laboratories and operations must respond to the diversity and special requirements of the samples.

Finally, curation means that those involved are in it for the long haul. Samples collected decades ago are yielding new discoveries that alter scientific understanding of planets, moons, and solar system history. These discoveries will inspire new generations of scientists and research questions and will drive future exploration by robots and humans. Curation is—and will remain—the critical interface between collecting samples and the research that leads to understanding other worlds."

Sample requests and sample distribution

The central rule is 'No sampling without sample requests'. In order to achieve this it is recommended to publish 'Calls for Sample Requests'. This can be done already before the planned drilling operations actually start. This call should be repeated, but be announced not later than the closing sampling party or science workshop toward the end of the expedition. These reasons are important:

- To inform the science team about the actual drilling targets
- To review the individual sample requests
- To detect sections which are over-sampled, or which are not requested enough
- To review and adjust the general sampling strategy
- To improve the sampling procedure

The first "call for sample requests" is especially important for samples that have to be taken on-site simultaneously with the drilling operations. Here is the chance to check whether this type of sampling is really necessary, and if yes, how it can be integrated into the on-site workflow. The second "call for sample requests" should be done when the holes and the sample materials have been initially measured and documented. In both cases, the Web based project site is very useful as interface to the science team members. Images, scans, lithological descriptions, logs along the sample material and inside the holes are basic information about the quality of recovery and geo-properties that can support the selection of appropriate sampling spots.

Rule 1: No sampling without sample requests.
Rule 2: On-site sampling is restricted to special requirements based on a consolidated science plan and accepted through an approved sample request.

Sample curation

ICDP does not maintain its own storage sites (repositories) for sample material. In general, the project has to take care of appropriate facilities and accessibility for a long-term period after the end of the project. Additional to that, ICDP is allowed to use storage facilities from IODP, LacCore and GESEP.

Repository	Program	Country	Envir.
Bremen Core Repository (BCR)	IODP	Germany	cooled
Gulf Coast Repository (GCR)	IODP	U.S.A.	cooled
Kochi Core Center (KCC)	IODP	Japan	cooled
National Lacustrine Core Facility (LacCore)	NSF, CSDCO	U.S.A.	cooled
Rutgers … Core Repsoitory	IODP	U.S.A.	cooled
National Core Repository	BGR, GESEP	Germany	not cooled

Table 5.3: Core Repositories cooperating with ICDP

ICDP is offering the CurationDIS as a Drilling Information System administrative tool for managing inventory stored in data repositories (e.g. Tab. 5.3). These repositories host and preserve sample material and conduct professional sample curation. One big advantage of the DIS work philosophy is that the content of an ExpeditionDIS can easily be transferred into a CurationDIS. Another advantage is the assignment of International Geo Sample Numbers (IGSN). Already on ExpeditionDIS level IGSNs are assigned to holes, cores and sections, mud, cuttings, gas or other material extracted from the project drill holes. IGSNs are assigned to any on-site sample; corresponding labels can be printed already in the field (Fig. 5.8). As sampling continues in a core repository, the same procedure is performed by the CurationDIS.

Fig. 5.8: IGSN encrypted in QR code on core sample

Depth matching and composite profiles

Depth Matching is an important issue of wellbore data consistency. Typically, during the operational phase of a drilling project, many different depth systems are being used:

- The Driller Depth is calculated from the length of lowered drill string
- Lag Depth (see: Glossary) is a calculated depth derived from the mud circulation and is used for any kind of mud samples, e.g. cuttings or gas

- Log Depths derive from downhole measurements. Log depths are usually continuous and most accurate

If log depths are available, it is recommended to correlate or match all other depth systems to log depth. Composite Depths are resulting from splicing selected sections retrieved from different holes. True Vertical Depth can be calculated if the trajectory of the hole is known.

These three features are supported by the DIS:

- Transfer any depth measures in meter units
- Define common reference level for all holes of a site
- Build composite or spliced profiles in case of multiple, partly overlapping holes on a site

These two features that require specialized software tools such as WellCAD and/or CORRELATOR:

- Correlate all types of depths with a selected master of the downhole logs
- Calculate true vertical depth

Spliced data profiles (including line scan images) can be generated by using, for example, the open-source tools CORRELATOR and CORELYZER to produce a composite site image overlaid by the various data sets (e.g., from logging or physical property measurements). This also extends into the task of 'Depth -&- Data Matching', which is, generally speaking, a mandatory prerequisite for the overall quality of the data set(s) obtained in the field and laboratories after the field operation has been concluded.

Operational/Expedition reports

The often short expedition period has a more standardized structure compared to the longer, subsequent period of the science moratorium. The moratorium period has essentially no predefined workflows, because it is strongly dependent on the outcome of the drilling phase, the general financial situation, participant turnover, and other factors (Fig. 5.6). Therefore, the Operational Report is an important milestone and landmark between these two phases. The Operational Report must be finalized not later than six months after a sampling party where the samples are distributed to the science team. All data sets and results produced during these scientific analyses, evaluations, simulations, and interpretations become parts of the science papers to be published toward the end of the pre-defined science moratorium.

To make the operational report more attractive

- it should be reviewed by external reviewers
- all science team members are authors and have to contribute (this makes the selection of science team members even more important)
- it should be published as digital supplement including the basic data sets under open access license of a regular journal such as the Scientific Drilling Journal (see: Chapter 10 on Education and Outreach)
- it should be a public document
- it can be a door opener for an additional post-drilling workshop where all science team members gather to discuss the gained knowledge and sample material, and plan the next steps of the subsequent scientific work

A template for a Table of Contents is shown below:

- Title page
- Publisher's notes
- Expedition participants
- Abstract
- Introduction
- Geological Setting
- Scientific Objectives
- Strategy
- Synthesis
- Site Overview
- Preliminary Scientific Assessment
- Topics according to the specific expedition (e.g. lithostratigraphy, micropaleontology, sedimentation rates, petrophysics, chemistry, microbiology, others)
- Operations
- Site Operations
- Acknowledgements
- References
- Tables
- Figures

One example is available **here**.

The project web site

In addition to providing the operational tools and procedures for the data and sample management in the field, labs, and repositories for sample material, the project is hosted on ICDP's Web site on the World Wide Web. ICDP usually creates a Web site for each ICDP project after the first grant proposal for a workshop has been approved.

Within the conceptual design of ICDP, each project receives the same initial screen space and weight within the ICDP Web site structure. Generally, each project is described on a project profile that derives from the proposals. Topics such as News, Scientists, Press & Media, Publications, Workshops, etc. are updated as required. With the project developing and according to actual project activities, the project Web site also grows. When the project is ongoing, it usually receives more attention from the general public. Accordingly, the project will be featured as an ICDP Highlight on the web site.

In order to enhance the outreach effect, a project can also maintain its own Web site. Project PIs can use their own preferred choice of modern social media. ICDP web site creates an abundance of links to the project-specific contents of these external media. More scientific project data are usually confidential and 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 for selecting samples.

Long-term monitoring equipment

Some projects are using the drilled holes for long-term measurements and observations. Typical examples are downhole seismometers as part of large scale seismic networks; or pressure/temperature sonde in conjunction with geological injection or production of fluids. The latter are often supported by surface installations such as tanks and power stations.

Generally, these sensor systems in the holes and in surface installations are measuring and transferring data to their own central control system, or they store their data locally. Data read-out happens often manually, and regularly.

It is possible to couple such sensor control systems with a data acquisition system such as the DIS. When they are indeed coupled, custom programming is needed to ensure that the different time series and other data are synchronized.

Information Technologies

Drilling technologies have been in existence at least since the 19th century. Before the computer age a huge number of explorations have been carried out, and many of these were even scientific wells – and it worked. The modern computerized techniques provide a lot of enhancements and new options to make the operations around drilling easier. The crew -&- staff members on-site should not be overloaded with technical features that distract them from performing their actual work. Therefore information technology used on-site should be as simple as possible, yet as much as necessary. Training is essential. This holds even today as almost everyone is working with computers. The use of IT in drilling projects is a special field that has to be prepared for carefully. The DIS, for instance, is technically not a big deal, but its customizable scientific workflows and its rigorous focus on data integrity requires some training sessions and possibly remote follow-up training and online support during the field operation. Nowadays this can often be accomplished with efficient video-conferencing tools.

Emerging technology trends affecting science projects such as mobile technical and social networking are becoming more important. This holds for outreach issues as never before, as well as for the intra-project coordination.

The publishing landscape is changing rapidly, pushed by open access media, Digital Object Identifiers (DOIs) for data sets and International Geo Sample Numbers (IGSN), digital supplementary materials added to publications. Big data deriving from monitoring of ultra-high resolution are demanding even more powerful technology infrastructure highly integrated Internet services, and more skillful users.

** Ronald Conze, Thomas Gorgas and Knut Behrends*
Operational Support Group ICDP, GFZ German Research Centre for Geosciences, 14473 Potsdam, Germany
r.conze@icdp-online.org