| Deliverable Number: 4.2	Due date: October 2022
Deliverable Title: Updated Data and software inventory **WP number:** 4 **Leader Beneficiary:** HZDR **Deliverable type:** ORDP **Dissemination level:** Public	**Reporting period:** RP2 **Issue date:** 20 April 2023 **Authors:** Nicolas Bruyant, Hans Engelkamp **Reviewers:** ISABEL Coordination Board

ISABEL

Improving the sustainability of the European Magnetic Field Laboratory

---

DELIVERABLE 4.2

Updated Data and software inventory

Open research data pilot
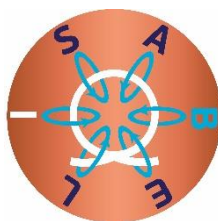
Start date of the project: 1st November 2020

Duration: 48 months

Project Coordinator: Geert Rikken – CNRS LNCMI (P1 - CNRS)

Contact: isabel@lncmi.cnrs.fr

Version	Modifications	Date	Authors
1.0	First draft	09/12/2021	Hans Engelkamp
1.1	Updates	20/04/2022	Nicolas Bruyant
1.2	Updates	28/04/2022	Eva Bezgousko, Marc Uhlarz, Thomas Hermannsdörfer
2.0	Final version	17/05/22	Eva Bezgousko, Nicolas Bruyant
2.1	Updates	03/04/2023	Nicolas Bruyant, Hans Engelkamp
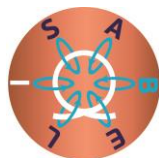
| 2.2	Corrections & final version	20.04.2023	ISABEL Coordination Board

DOCUMENT ABSTRACT

This deliverable 4.2 "Updated Data and software tools inventories", is part of Work Package 4 "Development of Data Management Plan".

Two surveys and interviews have been handled separately. In July 2021, a survey of the software used for data analysis, was sent among the local contacts of EMFL, i.e. experienced scientists at the EMFL facilities dedicated to user support. In addition, a survey was conducted among the staff in charge of the instruments on the topic of data structures and tools.

The aim of this deliverable is to report potential updates and changes since the first round of surveys. However, as it is explained, no major change justifies a modification in the first version of the deliverable.

TABLE OF CONTENTS
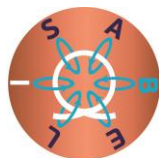
1. Software Inventory

Objective

In typical high-field experiments, the users perform at least a preliminary analysis of their data during their stay at the facility. This is made possible by software tools available at the facilities and with dedicated support from the local contacts. For various experiments, special analysis software has been developed by expert staff of the EMFL facilities. In order to generate publishable results out of the raw data, user-friendly data-analysis tools are required. So far, the users are in close contact with their local contact if help is needed for further detailed analysis or re-examining results with modified parameters.

In this task, an inventory / repository will be made on the software tools that are available and which of those are necessary for proper data analysis. It will be examined which of the software tools (including scripts for standard software packages, web-based analysis tools as well as executable customized programs) shall be and can be made available for the user community or even for the general public. Using an integrated software development repository (for instance GitLab or GitHub) a common development and improvement will be enabled. This will be a continuous process since data-analysis software evolve with time by regular updates and new programs and scripts are being developed according to the needs of in-house and external users as well as triggered by establishing novel experimental techniques. As a result of the project, analysis software can be published on decision of the partners involved.
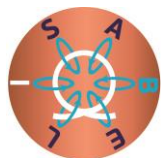
Survey

A survey has been conducted among the local contacts of EMFL concerning the typical data analysis software that is used. For this, a web-based questionnaire has been made, in order to collect information about the analysis software used by each facility for each experimental technique.

Some of the questions are:

- Is this software intended specifically for this experiment?
- Is the software home written, and if so, in which language?
- Is there some form of version management?

The full questionnaire is shown in Figure 1.

Figure 1. screen shot of the software inventory survey

Results

We received 22 responses, quite evenly spread across the facilities. 70% of the software is home written, usually by the researchers themselves. Typical programming environments are Python (32%), Labview (27%) and some C++ (9%). Among the non-home written solutions, most often the Origin software is used. Only two respondents use git for version management. 50% of the home-written programs is considered to be multi-purpose, i.e. can be used for different measurements. This suggests that this software might also be interesting for distribution among other local contacts, facilities, users and even a general public. For the other half, certainly more effort is needed.

Update

After 18 months, the survey has been repeated. This did not lead to any new insights.

2. Research-data inventory

Objective

The datasets generated by high-field experiments consist of raw data and metadata of the physical measurements. These data are formatted according to each experiment recoding system and often need to be associated to metadata describing sample and experimental parameters. One objective of this inventory is to describe them as precisely as possible to identify common points, differences and improvement in all the experiments. This inventory is not limited to raw data, but also to metadata and machine data.
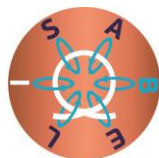
Survey

A second survey has been conducted among the local contacts of EMFL concerning the typical data format that is used. For this, a web-based questionnaire has been made, in order to collect information about the analysis software used by each facility for each experimental technique. The questions are summarized in table 1.

Topic	Question	Majority Answer
Experiment identification	Who is answering to the survey? Please indicate the lab and/or the team.	N/A
	Please indicate the experiment concerned or "whole lab" for general answers.	N/A
Data recording	Which raw data are recorded?	Experiment raw data only
	Who records the data? If raw data are stored in multiple parts please specify by whom.	Local Contact
Raw data dimensionality	What is the data dimensionality?	1D
Metadata	What is saved ?	Sample and temperature
	How is it saved ?	Manually
	Where is it saved ?	Lab Book
Format	What is the file format and/or structure?	Ascii
	If possible please provide a documentation of the format used	N/A
	Order of magnitude of file size/year	10 GB/year
Improvements	Do you plan to improve or change your data format?	No

Table 1. Questions of data inventory survey

Results

19 people in charge of experiments answered to the survey equally distributed among DC and pulsed field facilities. Most of the raw datasets are 1D recordings over time with a measured field and measured quantity columns. Higher dimensionality data are related to neutrons and optical spectroscopy experiments. Concerning the metadata, most of them are manually stored in lab-books. Two third of the data are stored in ASCII format, and the remaining in binary or specialized format. Machine data are stored separately and are not always accessible to users. The average volume of the data taken in an experiment at the EMFL sites is about 1GB/year. Finally, considering the answers, there is no specific need to change the raw data format. Concerning the metadata, it may be possible to move to an automated collection scheme if local contacts get support from data experts.

In Annex 1, survey results are detailed for each category with statistics.

Survey follow-up

One year after this study we asked all the repliers for updates on their data formats, but all of them replied that their practice did not change in the meantime.

The only difference is that some people asked for information about EMFL data management plan. A very positive outcome is that, if a new data format is needed in the future, researchers will be aware of data formats good practices.
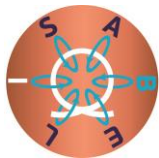
ANNEX 1
Research data inventory : Statistical results of the survey

Résultats

Questionnaire 667425

Nombre d'enregistrement(s) pour cette requête :	21
Nombre total d'enregistrements pour ce questionnaire :	21
Pourcentage du total :	100.00%

Résumé pour Lab

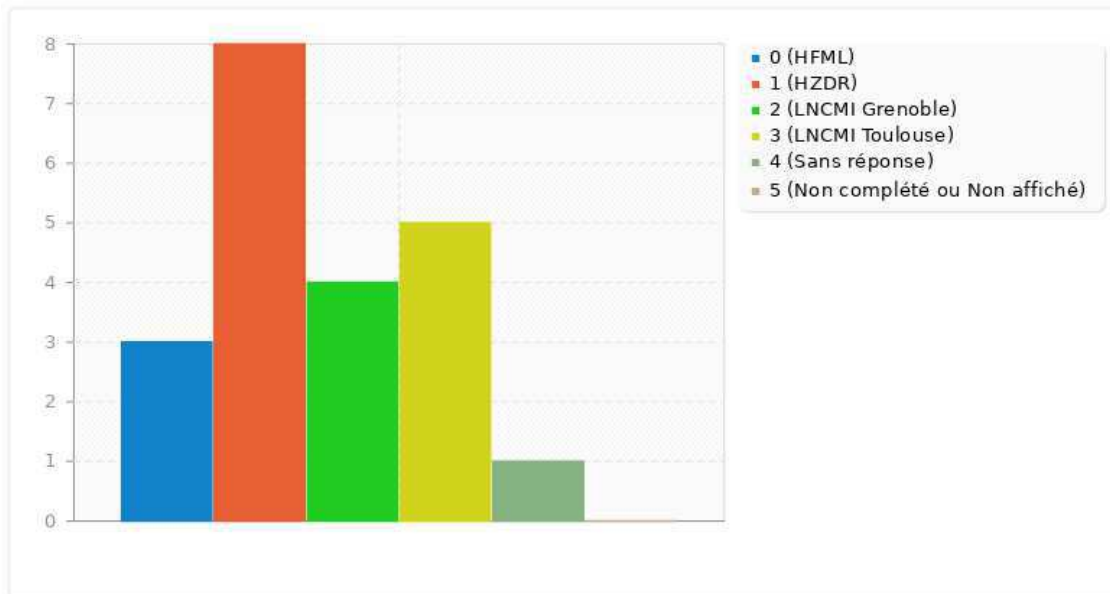## Who is answering to the survey ? Please indicate the lab and/or the team.

Réponse	Décompte	Pourcentage
HFML (A4)	3	14.29%
HZDR (A3)	8	38.10%
LNCMI Grenoble (A2)	4	19.05%
LNCMI Toulouse (A1)	5	23.81%
Sans réponse	1	4.76%
Non complété ou Non affiché	0	0.00%

Résumé pour Lab

Résumé pour Lab

Who is answering to the survey ? Please indicate the lab and/or the team.

Résumé pour Team

Please indicate the experiment concerned or "whole lab" for general answers.

Réponse	Décompte	Pourcentage
Réponse	19	90.48%
Sans réponse	2	9.52%
Non complété ou Non affiché	0	0.00%

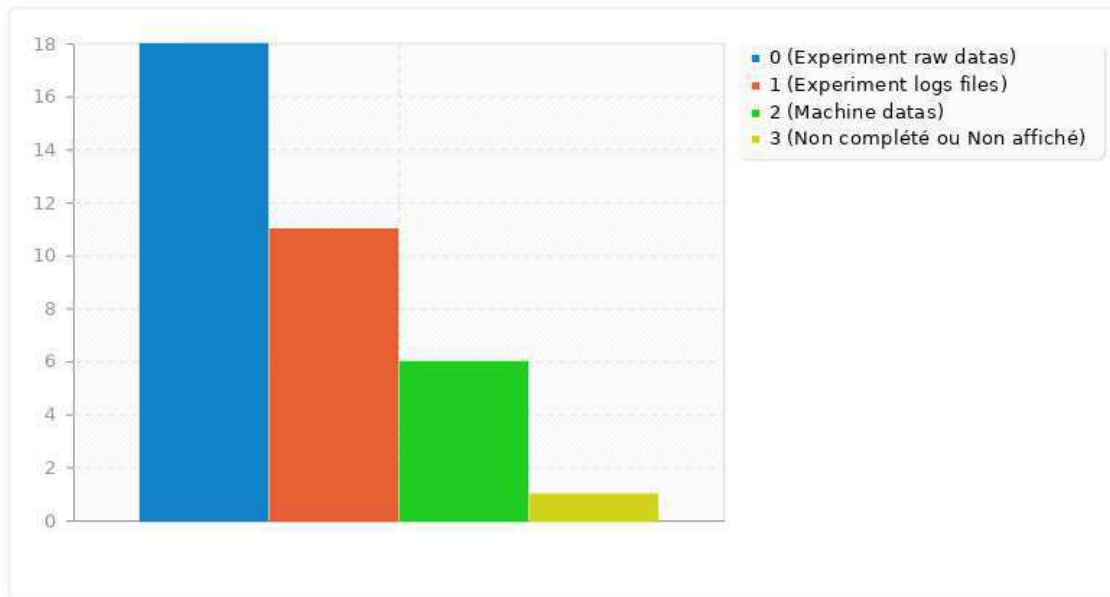| Identifiant (ID)	Réponse
2	Neutron and Xrays
3	Whole lab
4	whole lab
5	Whole lab
6	High-field ESR
8	whole lab
9	FT-IR
10	optical spectroscopy
11	ultrasound
12	whole lab
13	Magnetization & transport SCNS
14	Infrared magneto-spectroscopy
15	Magnetostriction, Ultrasound, Magnetooptics
16	Pulsed-field magnetization, electric polarization
17	Nuclear magnetic resonance in pulsed magnetic fields
18	FFC team
19	Nano team
20	NMR experiment
21	Magnetocaloric

Résumé pour G1Q00002

Which raw data are recorded ?

Experiment raw datas	18	85.71%
Experiment logs files	11	52.38%
Machine datas	6	28.57%
Non complété ou Non affiché	1	4.76%

Identifiant (ID)	Réponse
5	All the raw data are stored, so one reprocess any experiment conducted at the lab
6	Saved on the local PC as well as on the server
9	raw interferograms
12	.dat files
13	Text files with all measured data recorded in 0.1 ... 1 s intervals
16	These are data sets as a function of time to be used for further data treatments. They contain experimental results of the users.
17	Various time traces and information on RF pulse conditions are saved in a binary file format on local computer and on network drive.
21	Raw oscilloscope data
4	Temperature data for Oxford Dilution Refrigerator
13	Measuring parameters and/or comments stored in table or text file + paper logbook
16	Details regarding the various parameter settings are saved.
17	Time traces of various signals related to the pulsed magnetic fields are stored on network drives.
18	Temperature vs Time
21	Oscilloscope settings
13	centrally stored on installation computer
14	Data stored in the standard format used by Bruker Optics (OPUS format, including also experiment/configuration description)

Résumé pour G1Q00002

Which raw data are recorded ?
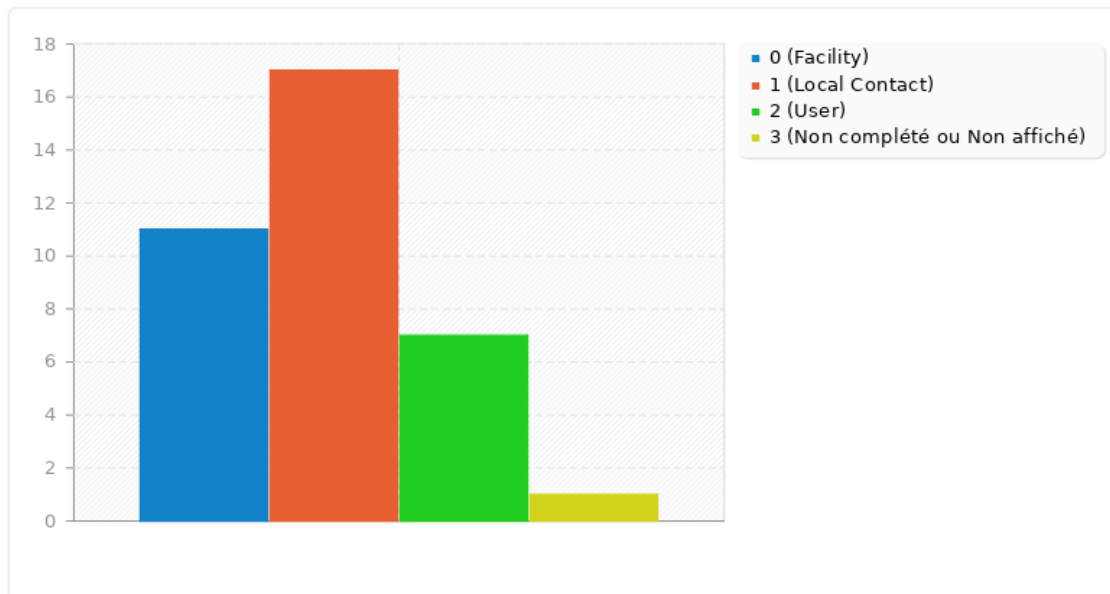
Résumé pour G1Q00001

Who records the datas ? If raw data are stored in multiple parts please specify by whom.

Facility		11	52.38%
Local Contact		17	80.95%
User		7	33.33%
Non complété ou Non affiché		1	4.76%

Identifiant (ID)	Réponse
2	ILL record all datas, ESRF is progressing toward it
4	Field data
5	The lab keeps raw data
18	Machine datas
21	Stored automatically on the file server
2	Record a copy of magnetic field datas
4	Experimental data
5	Local contact typically produces and stores processed data based on the raw data obtained,like M(H) or Rxx(H)
10	I keep all the data saved by the user to keep traces - to be able to dig inside - to be able to send them again to user later (I once was asked 7 years after an experiment for the data)
12	for back-up
13	Data recorded on local HFML computer, stored and backup at HMFL
14	All collected data, from in-house research and from users are stored and backed-up locally in the infrared lab.
16	The local contact initiates the data recording in a computer located directly at the facility site.
17	The local contact records the data of the specific experiment and for the pulsed magnetic fields. Data are stored on local computers and on network drives.
18	Raw datas
4	Experimental data
5	Users are supposed to keep the meta-data log, i.e. what was measured when
10	User selects during the experiment if he/she wants to record data or delete trial/test files
12	for use
13	dito, may also use own data acquisition system

Résumé pour G1Q00001

Who records the datas ? If raw data are stored in multiple parts please specify by whom.
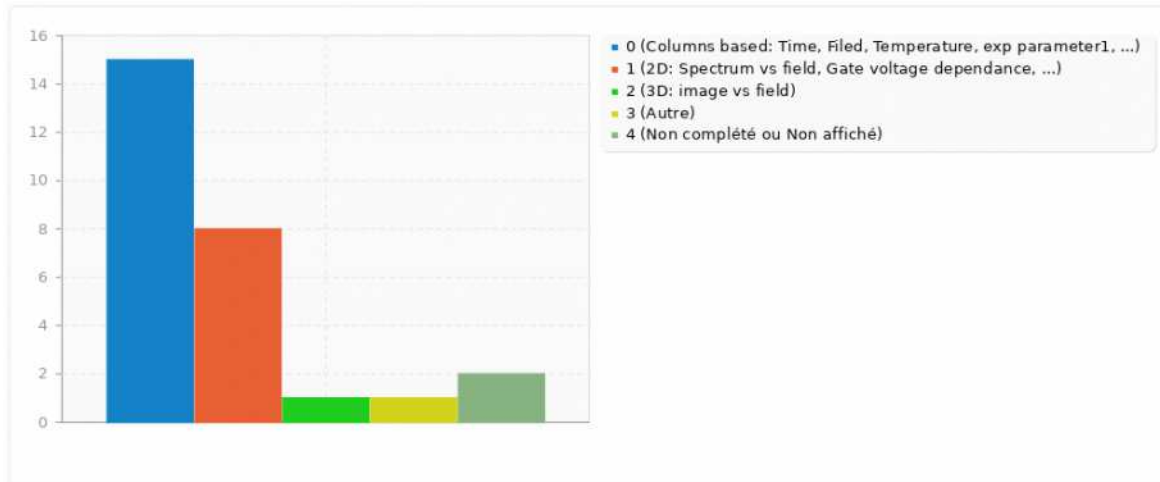
Résumé pour G2Q00001

What is the data dimensionality ?

Columns based: Time, Filed, Temperature, exp parameter1, ...	**15**	**71.43%**
2D: Spectrum vs field, Gate voltage dependance, ...	8	38.10%
3D: image vs field	1	4.76%
Autre	1	4.76%
Non complété ou Non affiché	2	9.52%

Identifiant (ID)	Réponse
5	Main way of keeping the data
6	Transmittance vs Field at certain Temperature
12	time, temperature, field, capacitance, volatege,...
16	The multicolumn data formats I use for magnetization and electric polarization typically have sizes between 5 and 10 MB.
17	The majority of the data is 1D. Mostly, various channels of time-resolved voltages are recorded. Some of the data is processed to a 2D type in the post-experiment analysis.
19	Vs time
5	2d data is saved for some specific experiments, like FBG magnetostriction
8	Voltage spectrum vs. time (point number) for pulsed-field ultrasound experiments
10	usually the CCD output in the form of a 2 columns file (wavelength - intensity). Magnetic field dependence measurements creates a single file with spectra one after another (length defined by the number of pixels - usually 1340 in our CCD - and the number of acquisition as a function of field, of gate voltage, of temperature, of excitatin wavelength etc ...)
19	Id vs Vg vs Field
14	Individual files with infrared spectra (recordered intensity versus wave number) with the magnetic field indicated in the name
20	Vs time + datas segmented vs time

Résumé pour G2Q00001

## What is the data dimensionality ?

Résumé pour G3Q00001

What is saved ?

Réponse	Décompte	Pourcentage
Sample identification (SQ004)	13	61.90%
Sample conditions : Temperature, Pressure, light (SQ005)	16	76.19%
Measurement parameters : current, power, wavelength (SQ006)	14	66.67%
Machine parameter : field quality, temperature stability (SQ007)	7	33.33%
Link to Proposal informations or UID (SQ008)	10	47.62%
Non complété ou Non affiché	2	9.52%

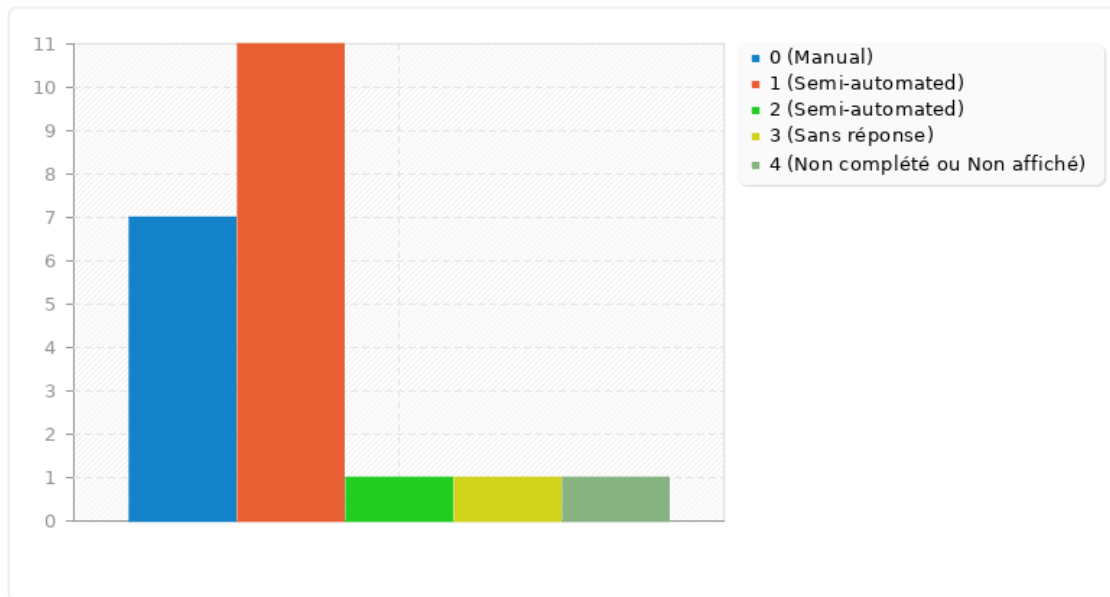# Résumé pour G3Q00001

What is saved ?
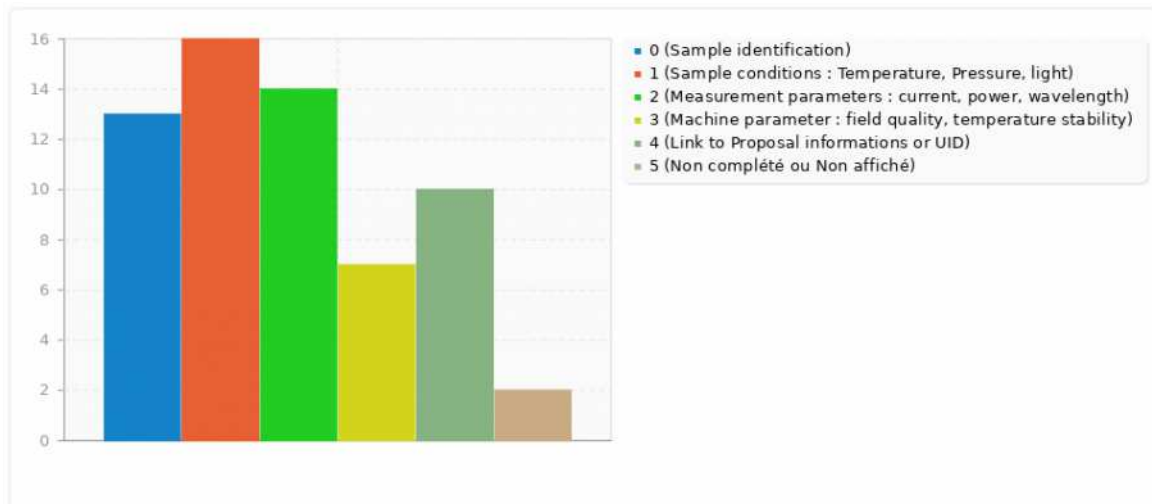
Résumé pour G3Q00002

How is it saved ?

Réponse	Décompte	Pourcentage
Manual (A1)	7	33.33%
Semi-automated (A2)	11	52.38%
Semi-automated (A3)	1	4.76%
Sans réponse	1	4.76%
Non complété ou Non affiché	1	4.76%

Résumé pour G3Q00002

How is it saved ?
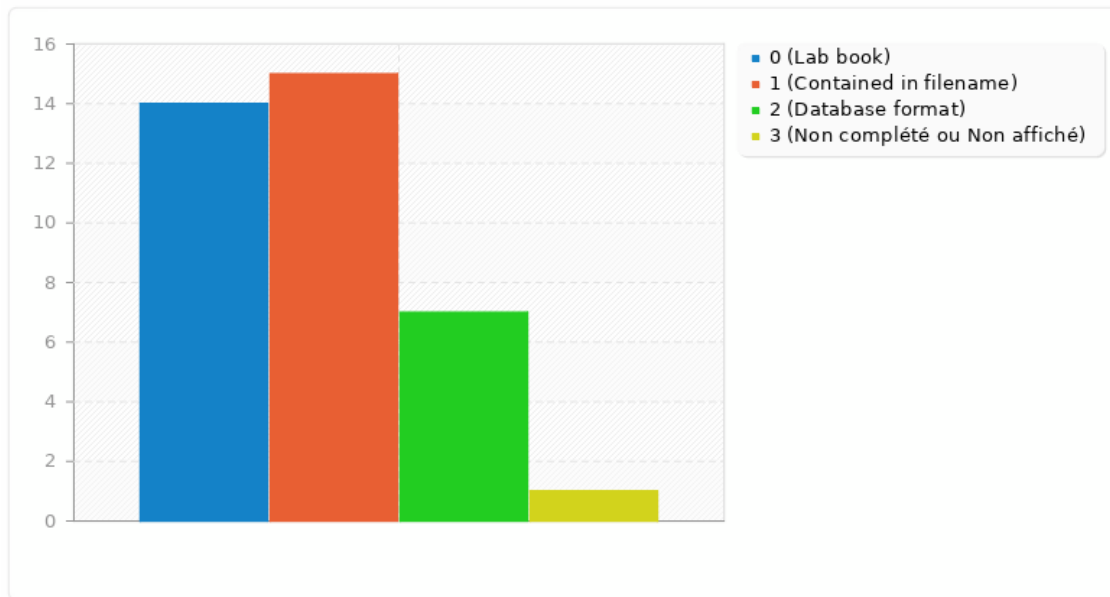
Résumé pour G3Q00003

Where is it saved ?

Lab book	14	66.67%
Contained in filename	15	71.43%
Database format	7	33.33%
Non complété ou Non affiché	1	4.76%

Identifiant (ID)	Réponse
3	Proposal ID
10	txt file for every day/week of measurements
12	hand written and/or in excel
16	In the lab book, the details of a field-pulse parameters are written in addition to the proposal number, user, and the local contact.
3	date and time in sub-sec precision and capacitor bank settings
10	metadata saved by the program of spectrometer
14	+ metadata included inside the OPUS data format
16	In addition to the above mentioned parameters, file name contains other experimental parameters such as temperature, sample orientation etc.
18	field polarity, temperature, sample id, field
4	Field data, generator log file
16	In the digital database, the lock-in settings for each of the pulse are recorded alongwith exact date and time.
17	The data is mostly saved in a binary format. Some of the data is saved in an ascii format.

Résumé pour G3Q00003

Where is it saved ?
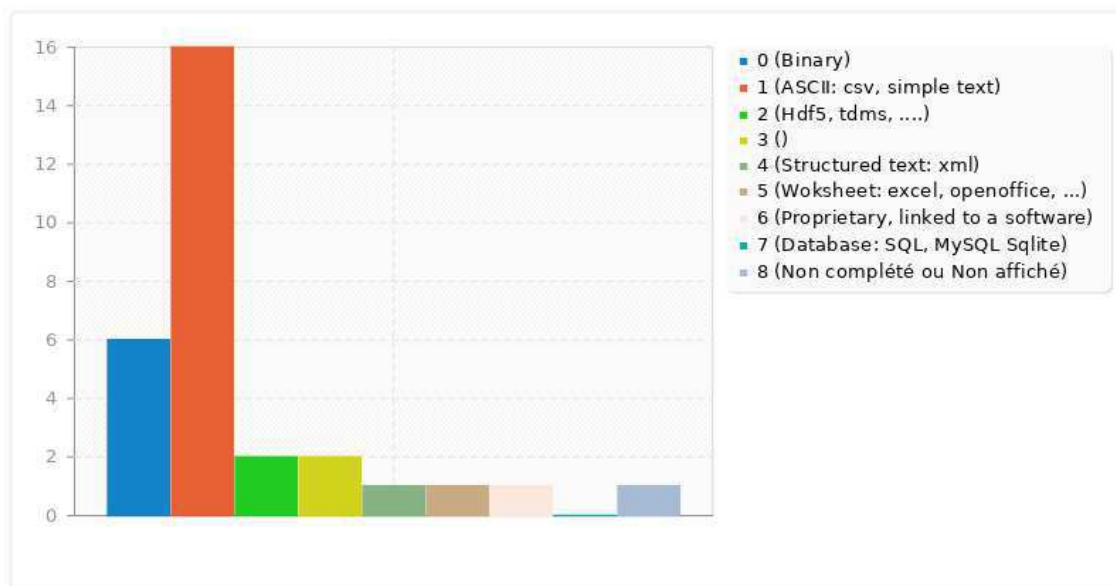
Résumé pour G4Q00001

What is the file format and/or structure ?

Binary	**6**	**28.57%**
ASCII: csv, simple text	16	76.19%
Hdf5, tdms, ...	2	9.52%
	2	9.52%
Structured text: xml	1	4.76%
Woksheet: excel, openoffice, ...	1	4.76%
Proprietary, linked to a software	1	4.76%
Database: SQL, MySQL Sqlite	0	0.00%
Non complété ou Non affiché	1	4.76%

Identifiant (ID)	Réponse
19	Raw data
16	We save the data in multicolumn ASCII format.
8	dat
8	dat
19	Metadata
19	Metadata summary
14	OPUS format , directly accessible only via Bruker's software

Résumé pour G4Q00001

What is the file format and/or structure ?

Résumé pour G4Q00002

If possible please provide a documentation of the format used

Réponse	Décompte	Pourcentage
Réponse	7	33.33%
Sans réponse	13	61.90%
Non complété ou Non affiché	1	4.76%

| Identifiant (ID)	Réponse
2	https://www.nexusformat.org/
3	- data files are stored in folders specific to magnets and pulse chambers - file names automatically contain date and time in sub-sec precision and capacitor bank settings. This allows for long term findability and readability. - file header defines all column names - data points are taken on 1 MHz sampling rate typically (or specified otherwise), i.e. row number defines time - first column: trigger signal - second column: I_shunt of magnet - third column: V_coil - fourth column: Rogowski coil at magnet leads or any other sensor - column five ... to N is experimental-technique specific - e.g. for a electrical-transport measurement: column five: pick up coil field, column six: I_sample, column seven: U_xx sample, column eight: U_xy sample
4	Column labelled ASCII depending on time
8	DAT format as saved using Python
9	Bruker Opus files as well as ascii format
20	NMR data format common with grenoble + pulsed field data format
21	1st row: channel number 2nd row: label 3rd row and further: data seperation by tabulator

Résumé pour G4Q00003

Order of magnitude of file size/year

Réponse	Décompte	Pourcentage
100MB/year (A1)	1	4.76%
1GB/year (A2)	8	38.10%
10GB/year (A3)	7	33.33%
100GB/year (A4)	3	14.29%
Sans réponse	1	4.76%
Non complété ou Non affiché	1	4.76%

Résumé pour G4Q00003

Order of magnitude of file size/year
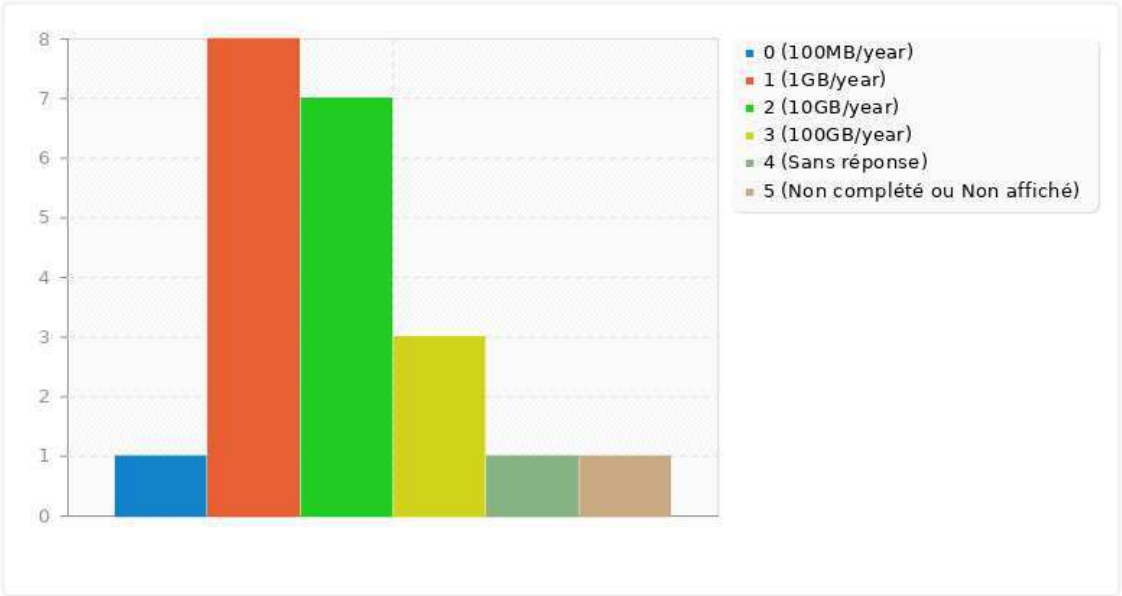
Résumé pour Improve

Do you plan to improve or change you data format ?

Réponse	Décompte	Pourcentage
Yes (A1)	0	0.00%
No (A2)	12	57.14%
Yes if i have support from ISABEL (A3)	3	14.29%
Commentaires	1	4.76%
Sans réponse	5	23.81%
Non complété ou Non affiché	1	4.76%

Identifiant (ID)	Réponse
16	The present format is very suitable for our purpose and users seems to be quite happy about it too. Therefore I do not intend to change the format.

Résumé pour Improve

Do you plan to improve or change you data format ?

- 0 (Yes)
- 1 (No)
- 2 (Yes if i have support from ISABEL)
- 3 (Commentaires)
- 4 (Sans réponse)
- 5 (Non complété ou Non affiché)