Annual meeting
Task 7.2

Nicolas soler - May 3 - 5

LEAPS-Innov WP7.2 (M7-M18)

"Assessment of **future needs** and **development of metrics** for data compression and reduction"

Task coordinators:

- **Nicolas Soler** (ALBA)
- **Vincent Favre-Nicolin** (ESRF)

"Information on site-specific scientific needs of experiments at LEAPS facilities will be gathered through **interviews with domain experts and scientists** taking into account the past and future needs: **bandwidth, latency, connectivity, available storage volume and compute infrastructure**.

Downstream processing and analysis needs will be identified to serve as **upper boundary limits for information loss during compression**.

Detector manufacturers will be involved in the discussion to ensure the compatibility, and future integration in production, of the resulting components."

“Research will also be conducted to **identify the techniques which generate the most data** and to develop metrics which assess the effect of data reduction on the quality of the final result. The most likely candidates are **serial crystallography and single-particle imaging**, multiple types of **tomography**, and most **high-resolution imaging** techniques.

Metrics will be developed to evaluate the data explosion problem and to assess the **impact of the data reduction and compression solutions** developed in task 7.3. The performance of the various reduction strategies will be investigated, with the goal of **achieving close to real-time performance** in data reduction.

This task will be carried out in consultation with **scientists and external industrial experts** in data reduction. Contacts have been already established with industrial experts in data compression and specifically the Blosc community, IBM and StreamHPC”

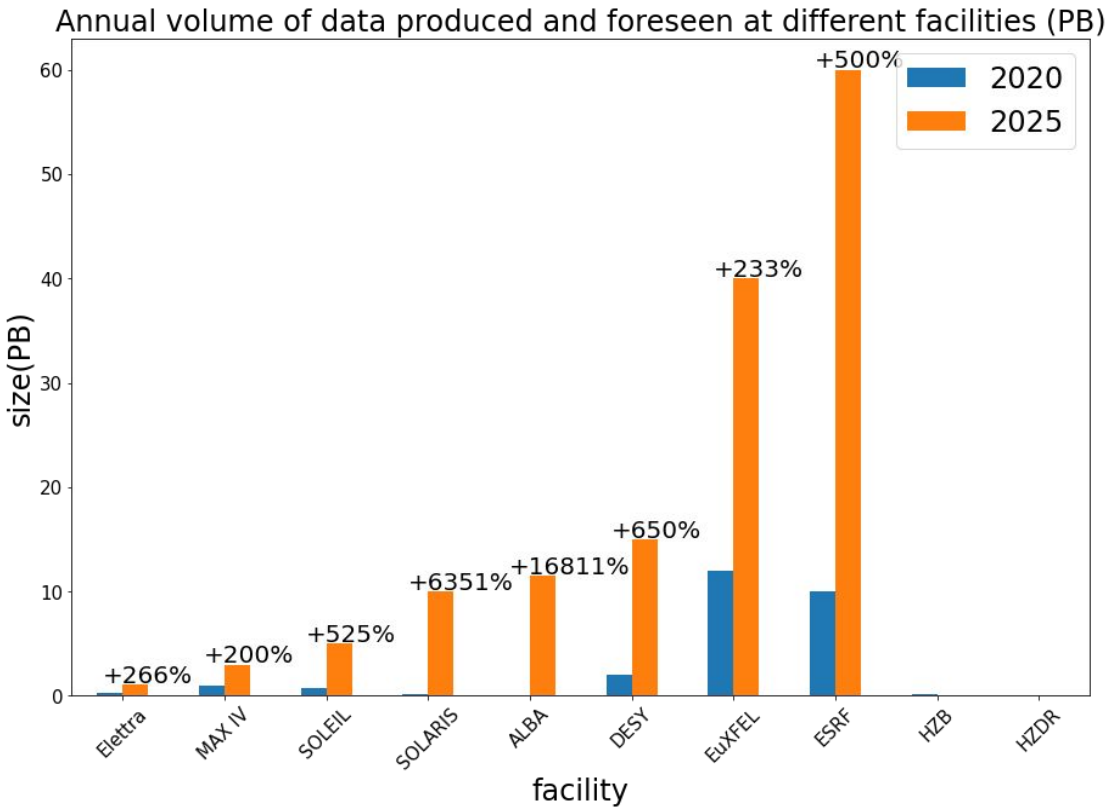
LEAPS-Innov WP7.2: main tasks

1) **Site-specific needs** (volume, bandwidth): already asked in email, we'll see the partners answers.
2) "**Upper boundary limits for information loss** during compression": **lossy compression/data reduction**. Need a specific survey among facilities for that. Or better, a poll per community (e.g. serial Xtallography, powder diffraction, tomography, etc..)
3) **Metrics**:
    a) global, technique-specific (volume reduction and compression + decompression speed)
    b) loss of information (acceptability per technique/community)

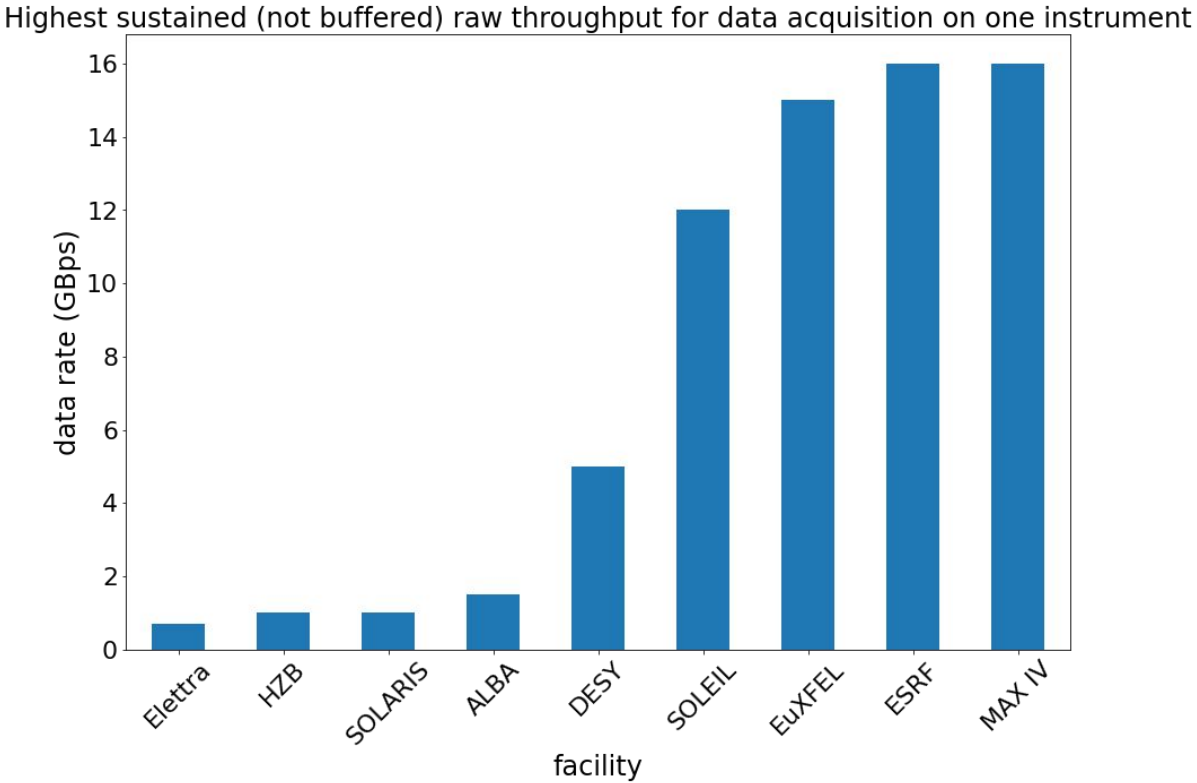
Hightlights from T7.2 report

1) Techniques and compression needs

Facility	2020	2025
Elettra	0.3	1.1
MAX IV	1.0	3.0
SOLEIL	0.8	5.0
SOLARIS	0.16	10.0
ALBA	0.07	11.5
DESY	2	15.0
EuXFEL	12	40.0
ESRF	10	60.0
HZB	0.2	NaN
HZDR	NaN	NaN



1) Techniques and compression needs



2) Data compression and reduction approaches

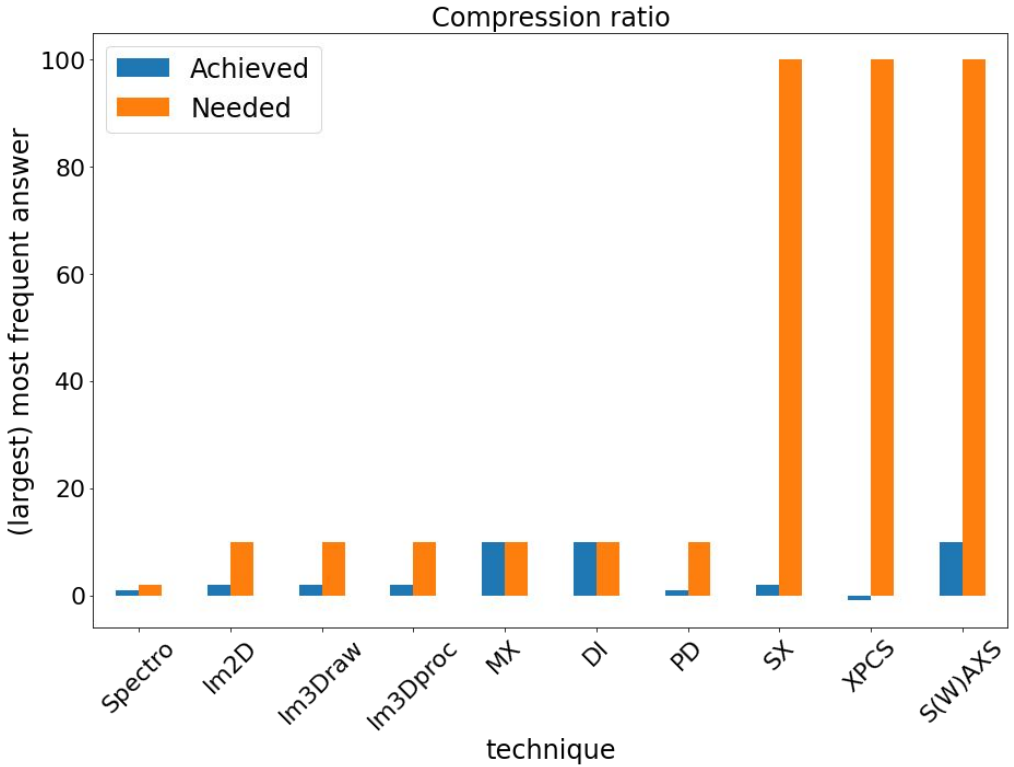


fig 3: Technique-wise compression ratio (raw volume/archived volume) achieved and needed. The bars indicate the most frequent answer (the largest compression ratio in case of equal frequencies).

2) Data compression and reduction approaches

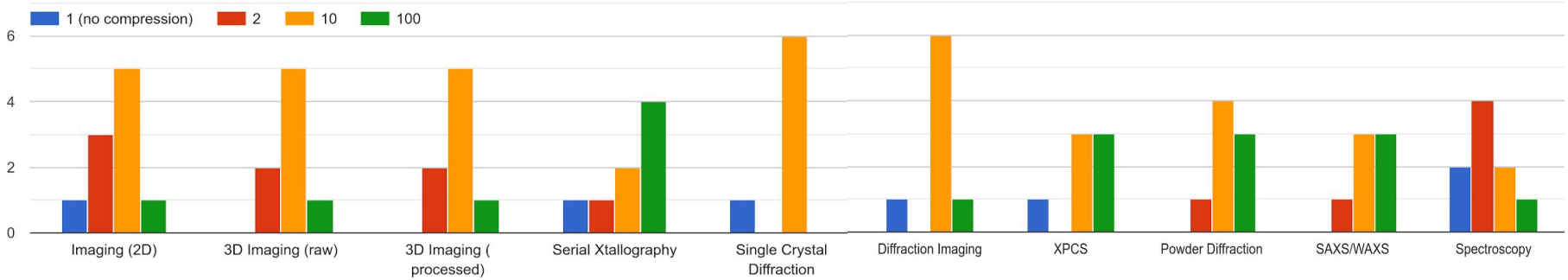
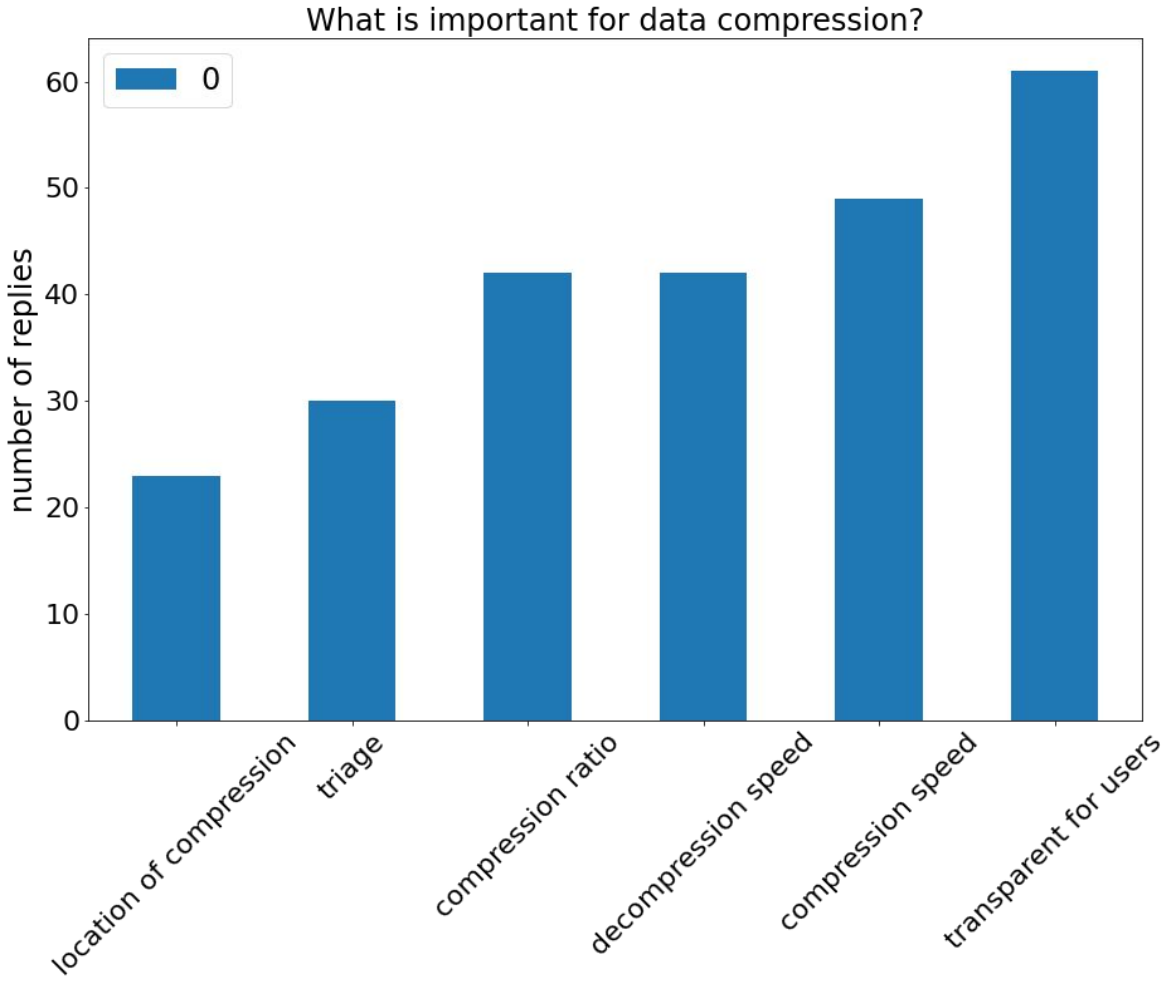


fig 4: Technique-wise compression ratio (raw volume/archived volume) needed.
(overall=including compression, reduction and triage of data)
The y-axis indicates the number of responses while the compression ratio is colour-coded.

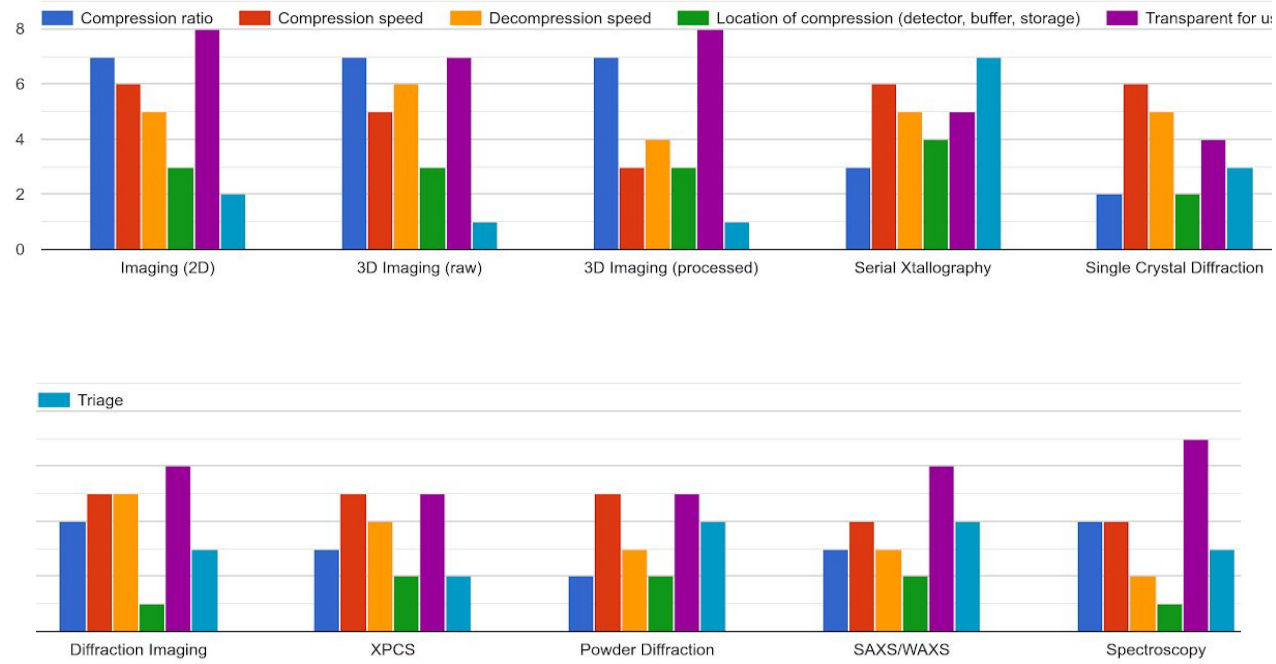
Other techniques mentioned

- cryo/material EM
- XRF, Ptychography, CDI, and FEL based techniques (FERMI)
- Single particle imaging (SPI)
- X-ray cross-correlation analysis (XCCA)
- Scanning transmission X-ray microscopy (STXM)
- SEM / FIB tomography
- Digital twins and simulation


What is important for data compression?
0
number of replies
0
10
20
30
40
50
60
location of compression
triage
compression ratio
decompression speed
compression speed
transparent for users

What is important for data compression?

fig 5: Global (top, blue barplot) and technique-wise (bottom, multicolor bar plots) distributions of factors considered important for compression.



What is important for data compression?

Technique-wise,

- tranparency for users is important everywhere
- compression ratio (blue) seems to be a concern for imaging techniques
- crystallography more concerned with compression/decompression speed (orange/yellow)
- triage (cyan), unsurprisingly is above all mentioned for XS, also other diffraction techniques such as MX, SAXS WAXS PS
- location not so crucial

Transparency for users:

This is a a challenge as the implementation of new, more efficient compression techniques (lossy or not) requires that all facility users be able to perform decompression without effort, and this implies a high level of standardisation for the compression codecs (e.g. across multiple frameworks and languages like python, matlab etc..).

Compression and decompression speed:

with a slight preference for compression speed, which can be explained by the need of making compression one of the steps of high-throughput (up to 16 GB/s) data acquisition pipelines.

What kind of compression/reduction?

Im2D

Im3Draw

Im3Dproc

SX

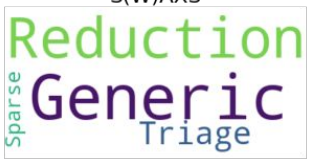
MX

DI

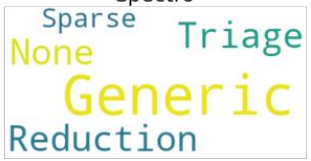
XPCS

PD

S(W)AXS



Spectro



Generic compression algorithms → frequently mentioned
(e.g. based on gzip, or other codecs which can be transparently used across many software and platforms)

Lossy → for imaging techniques (ex: JPEG already used)

Triage and reduction → diffraction techniques

Conclusions

- **general increase in data volumes and rates** already started for large facilities but also coming to smaller ones along with more high-volume techniques (imaging, CryoEM, etc) and faster detectors
- FAIR data policies being implemented → retention of at least **5-10 years** → **need to reduce the short term and long term storage** in all facilities.
- Users are also in need of compression, in order to be able to **transfer the data to their home institution**
- Compression/reduction algorithms → beyond the need for the most efficient compression (speed and compression ratio), a key issue is the **transparent nature for the end users** (their sw should be able to deal with it without users installing special libraries)--> **high level of standardisation**
- **How do you define raw data?** i.e SAXS/WAXS and powder diffraction have established/standardised 2D->1D reduction pipeline, it should not be necessary to keep raw data. What about SX? (after triage?)
- **Tomography** techniques: generate huge amounts of data difficult to compress with current techniques→ efforts in **lossy compression** should be developed.

Thanks !

Sources

Poll Answers
https://docs.google.com/forms/d/12kvEQSqPUN89cLsAiIHcrse5QHB4A1Vu7eKg2Fbc1DY/edit#responses

Jupyter notebook
https://colab.research.google.com/drive/18fRthLDsuInFt4oJlaPSa6MjaH-86B5S#scrollTo=HDdSoXgLMxq3

Answers: Excel tab:
https://docs.google.com/spreadsheets/d/18OYZ-EyDILh2tLMqQHhy6SOmdJFOG34WjGotgL4qSx0/edit#gid=765737561

Answers: Excel tab adapted to use with the notebook
https://docs.google.com/spreadsheets/d/1BB1e8Pqow0ZdsI5ET-lg5cY3BWG9nQrHlXB0sOxoCFg/edit#gid=1218228881