Reflections for FAIR data at a tomography beamline at SOLEIL

Andrew King
PSICHE beamline, SOLEIL

Plan of presentation

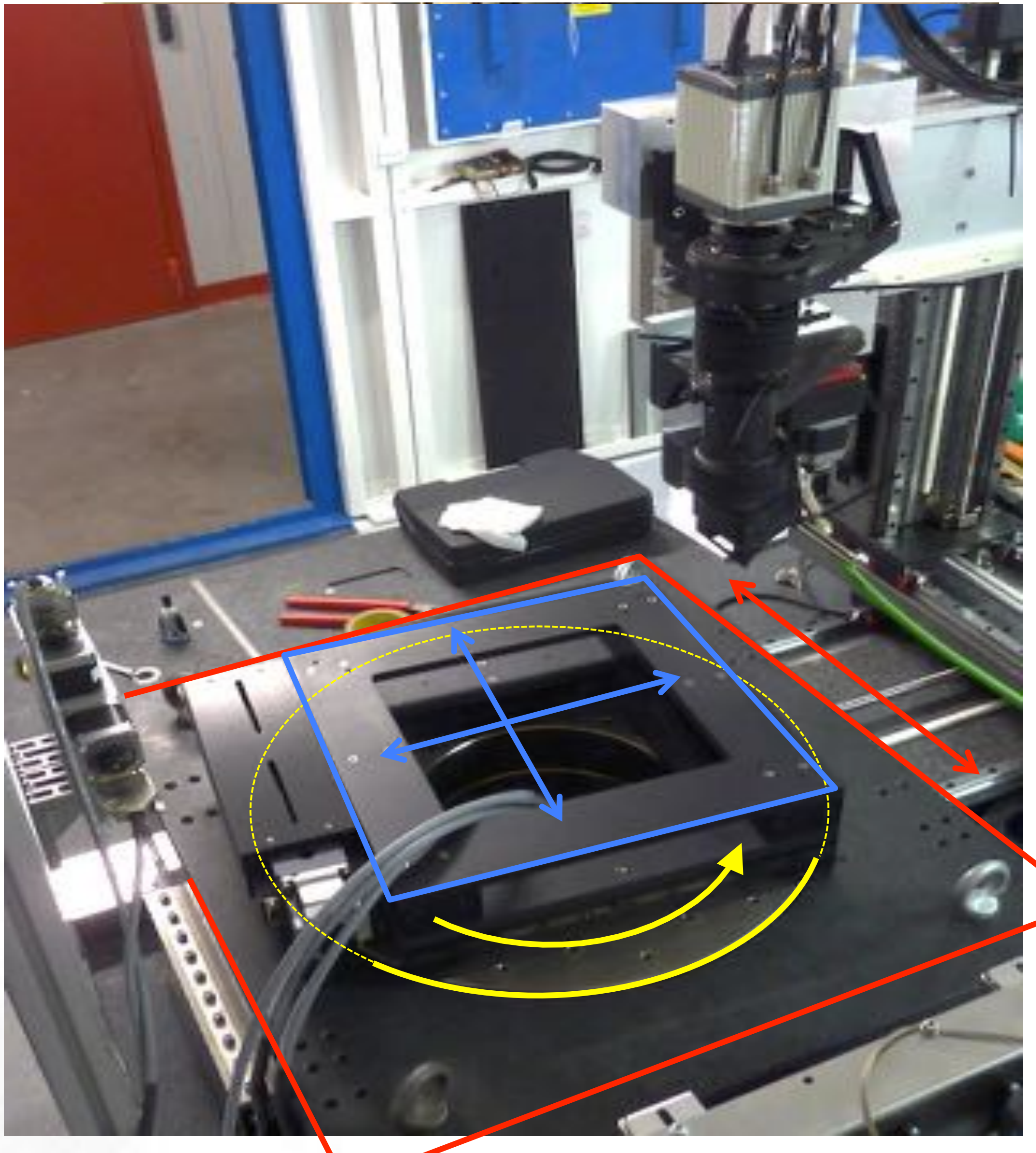
- Tomography at SOLEIL / PSICHE
- Particular challenges of FAIR data with tomography
  - Data volumes
  - Data storage and organisation
  - Data analysis
  - User supplied equipment
- Reflections and conclusions

Tomography at SOLEIL

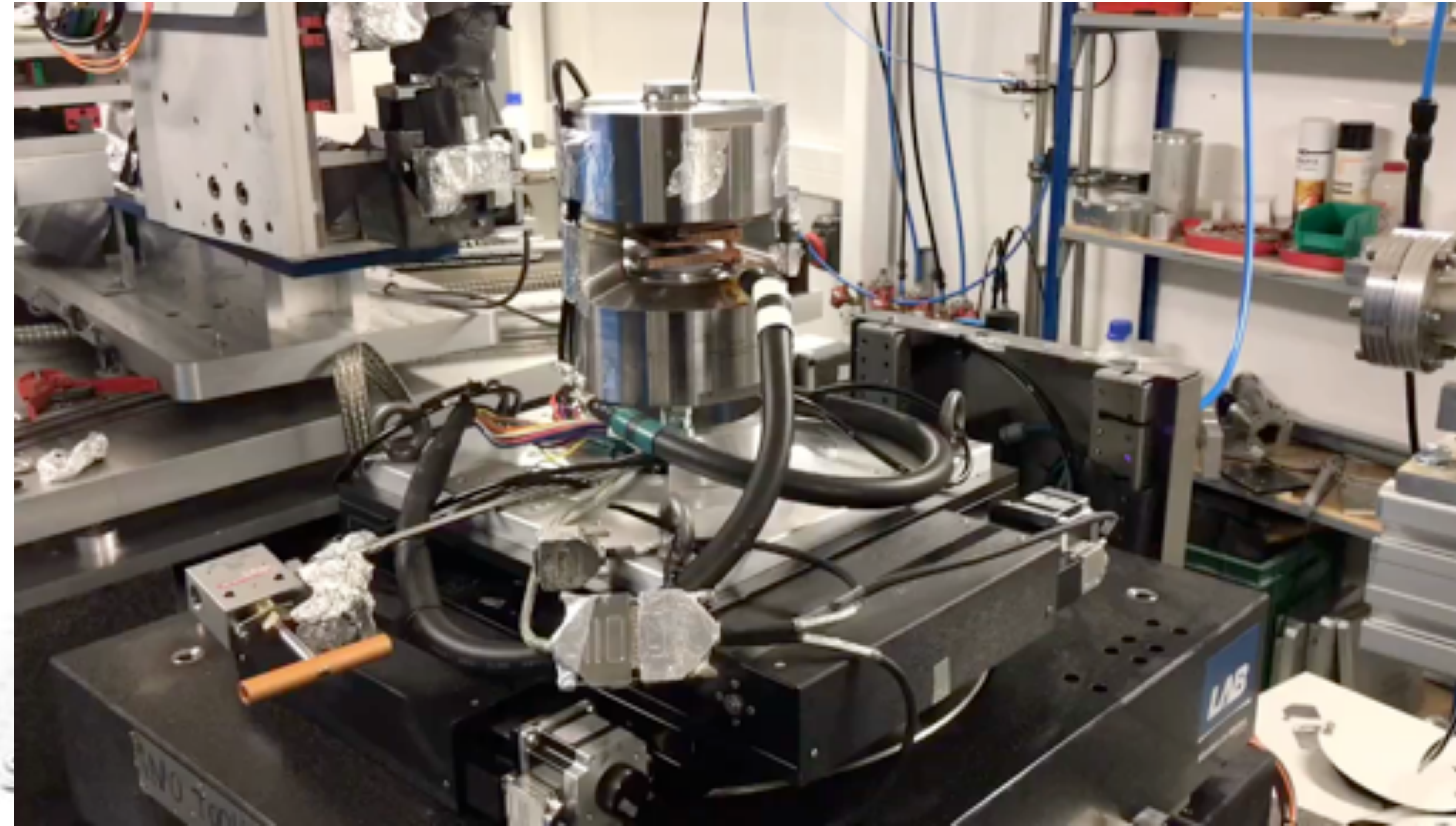
- PSICHE in user operation since 2014
- Multi purpose beamline (30-50% tomo)
  - High pressure (diffraction)
  - Tomography wide range of users
  - **Combined experiments: High pressure tomo, tomo plus diffraction, etc**
- Dedicated imaging beamline (100% tomo)
  - 200 m long
  - Two microtomography hutches
  - X-ray microscopy
- Ancient materials and cultural heritage
  - X-ray fluorescence instrument
  - Will have a tomography station (50% tomo)

The PSICHE tomograph

- Optimised for ***in-situ*** materials science
- **Fast acquisitions for time resolved experiements**
- Fully open 250 mm aperture
  - **Space for in-situ sample environments**



Example: UToPEC - Fast tomography, high P+T

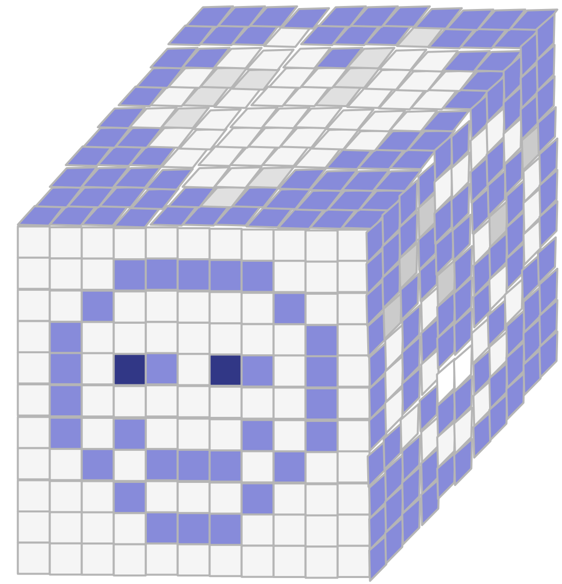


Volumes of data

- *A picture is worth a thousand words… but:*

1000 words = ~5 KB

A 5 KB tomogram would be
11 x 11 x 11 voxels
(without compression…)



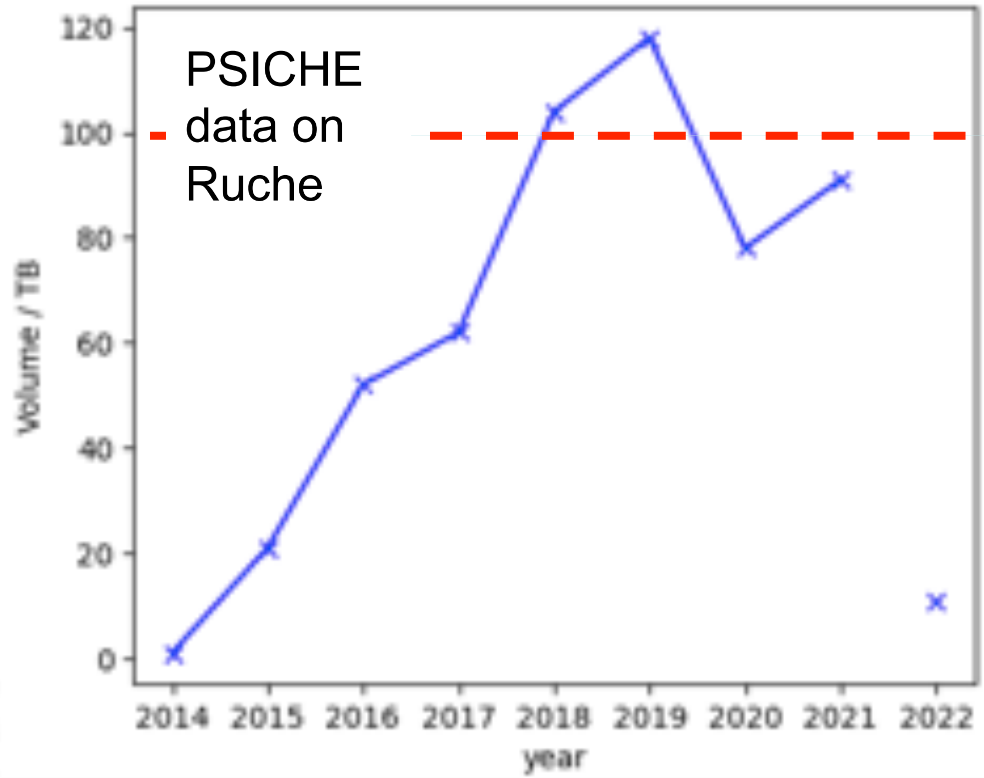
Single 2D radiograph can be 24 MB

A single 3D tomogram can be 220 GB !

Volumes of data

- Tomography produces very large datasets, and can produce them very quickly
- In-situ experiments (4D… 3D plus time) are particularly bad for this

How to actually make this volume of data available?



Organisation of workflow

- Local reconstruction solutions
- Stay at the beamline, local 10-100 gb/s Ethernet, minimal transfers across site network
- Archive only the raw data and metadata
  - Projections, which must be pre-treated using filtered backprojection to create 3D volumes.
  - Acquisition parameters from the instrument
- *What we store is not instantly reusable*

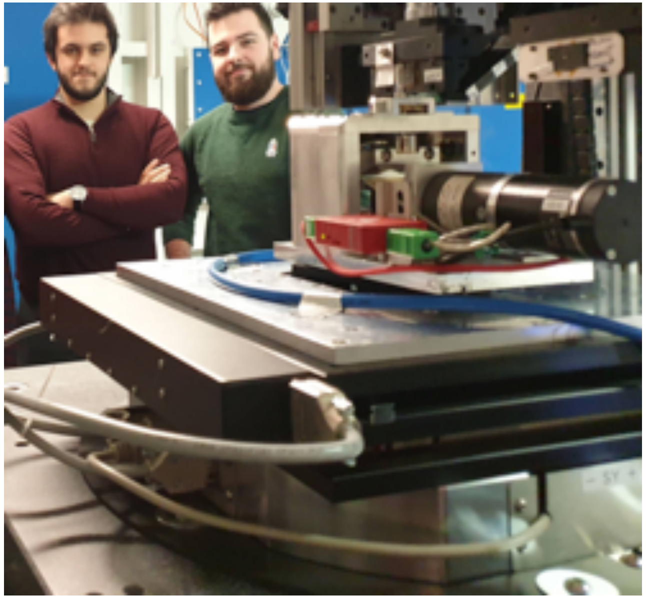
Time for user analysis

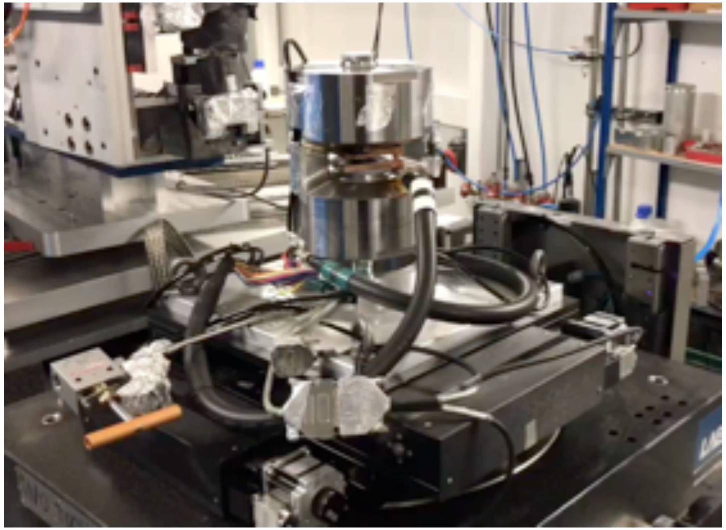
- In tomography, analysis can take a long time, especially for less experienced users.
- Also in materials science, a publication may require multiple beamtimes, certainly multiple samples, techniques, etc
- From the PSICHE publications list –
  - Legland 2022 = Girousse Feb 2018
  - Turpin 2022 = Caty Nov 2019
  - Le 2021 = Bornert Mar 2019

  -> *Typically 3-4 years to the first publication*

User metadata / in-situ equipment

- Not all metadata can be harvested from the instrument
- This is true of almost all user supplied equipment, but also some of our own equipment.
  - **Best case**: “Bulky” in-situ traction
  - Motor and load cell interfaced with SOLEIL electronics
  - **Typical case**: “UToPEC” in-situ extreme conditions
  - Pressure and temperature controlled independently.





How to get the sample metadata?

- *How many acquisitions do we have called "sample1_1, sample1_2, …" ?*
- A sample name is relatively easy to extract because it is helpful to the user.
- The history of the sample is more difficult – how to incentivize a user to fill in a sample passport? And how much of this detail can realistically be obtained?

Titanium
"Ti_fatigue_1"

Alloyed with Al, V
How much?

Thermal and mechanical treatment
Time, temperature, rolling sequence and directions?

Machine a sample
Orientation, dimensions, method?

Mechanical surface treatment
Which, what params?

Load in fatigue
Load, cycles, R-ratio, frequency?

Image of damage
Where?

Summary

- Challenges
  - Volume of data (538 TB and counting for PSICHE…)
  - Raw versus pretreated data
  - Time for users to analyse
  - Sample metadata
  - Metadata from in-situ experiments

Conclusions

- FAIR data is an important and desirable goal
- Tomography beamlines present some specific challenges
- We will be implementing and testing the first developments from our IT group