Presentation of the External Data work package

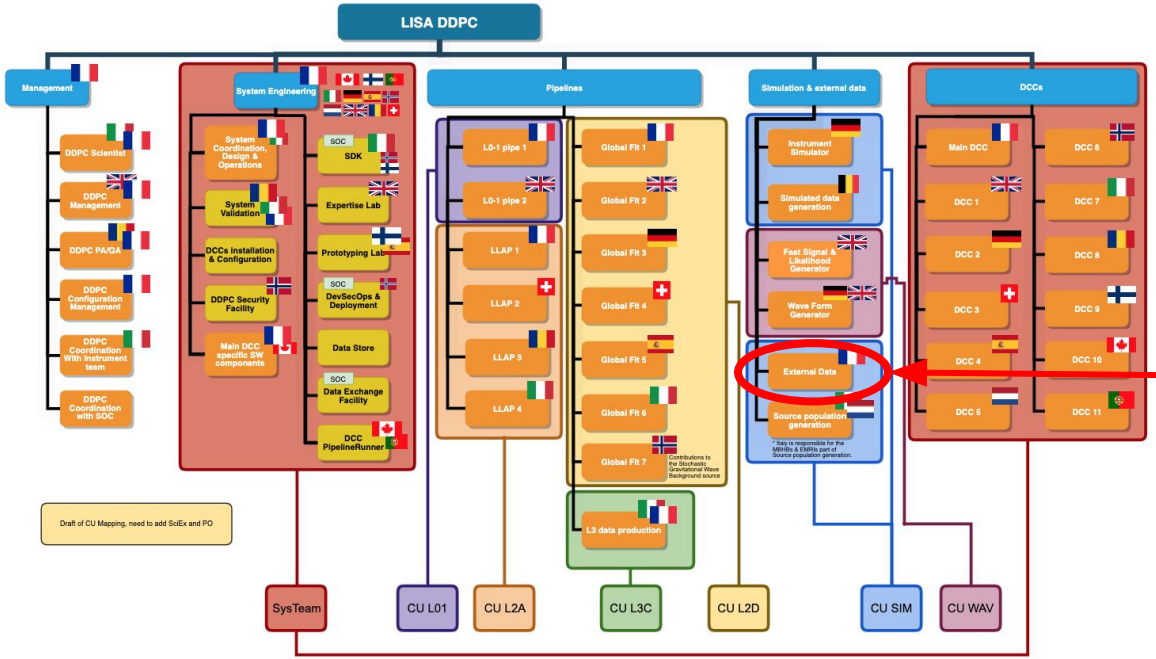
Adrian Macquet, on behalf of the External Data group

Journées LISA France, 4-5 mai 2026, Institut d'Astrophysique de Paris

irap
astrophysique & planétologie

L2IT

External Data



Work package part of the Simulations Coordination Unit within the DDPC.

General objective: collect external (astrophysical) data needed for LISA data analysis and distribute them within a standardized, DDPC-friendly framework.

➔ Ensure consistency and reproducibility of analyzes.

Responsibility: IRAP and L2IT (Toulouse)

Relatively recent project: many features remain to be precised (scope, use cases, etc).

Scope of External Data

Which data can be classified as “External Data”?

- ➔ Any non-instrumental external information that is used to produce LISA science results.
  - Verification binaries, MBHB candidates, SMBH population priors, etc
- ➔ Broad, not so well defined scope. We start with a smaller, better-delimited scope (VBs) to start developing the work package over the LISA MDCs.

External Data is a service to the LISA community:

- Collect needs from the LISA community: which external data will you use and how?
- Centralize and distribute these external data in a standardized, uniform, and DDPC-compliant way.

Goal: ensure that external data are used in a consistent and traceable way among the different analyses and projects.

Use cases

Who needs external data? For which purposes?

1. Generation of simulated data for LISA Data Challenges.
   - External Data provides list of VB parameters for Mojito Heavy (Fall 2026).
2. Priors for global fit pipelines
   - Priors on VB parameters
3. Multi-messenger follow-up?
   - MBHB candidates
4. Archival searches of stellar-mass binary black holes from ground-based detectors.

➔ New use cases should emerge as we interact with the community.

First deliverable

External data prototype: database of VBs and its API.

➔ Objective: June 2026.

Verification Binaries:

- Galactic compact binary systems (usually double WD).
- Known from electromagnetic observations (X/optical/radio).
- Predicted to emit GW detectable by LISA.
  - Orbital frequency within LISA band.
  - High enough SNR.
- ~50 sources identified so far, expect more in the coming years (e.g LSST).

➔ One of the most obvious use cases of external data, required for Mojito Heavy data generation.

Catalog of Verification Binaries

Source catalog: https://gitlab.in2p3.fr/LISA/lisa-verification-binaries (Kupfer et al. 2024, https://arxiv.org/pdf/2302.12719).

➔ 57 galactic compact binary systems.

Relevant parameters:

- f0
- sky position

Always defined, well constrained.

- masses
- distance

Always defined, but sometimes large or not statistically meaningful uncertainties.

- fdot
- inclination angle

Defined for only a few systems.

- Polarization and phase not defined.

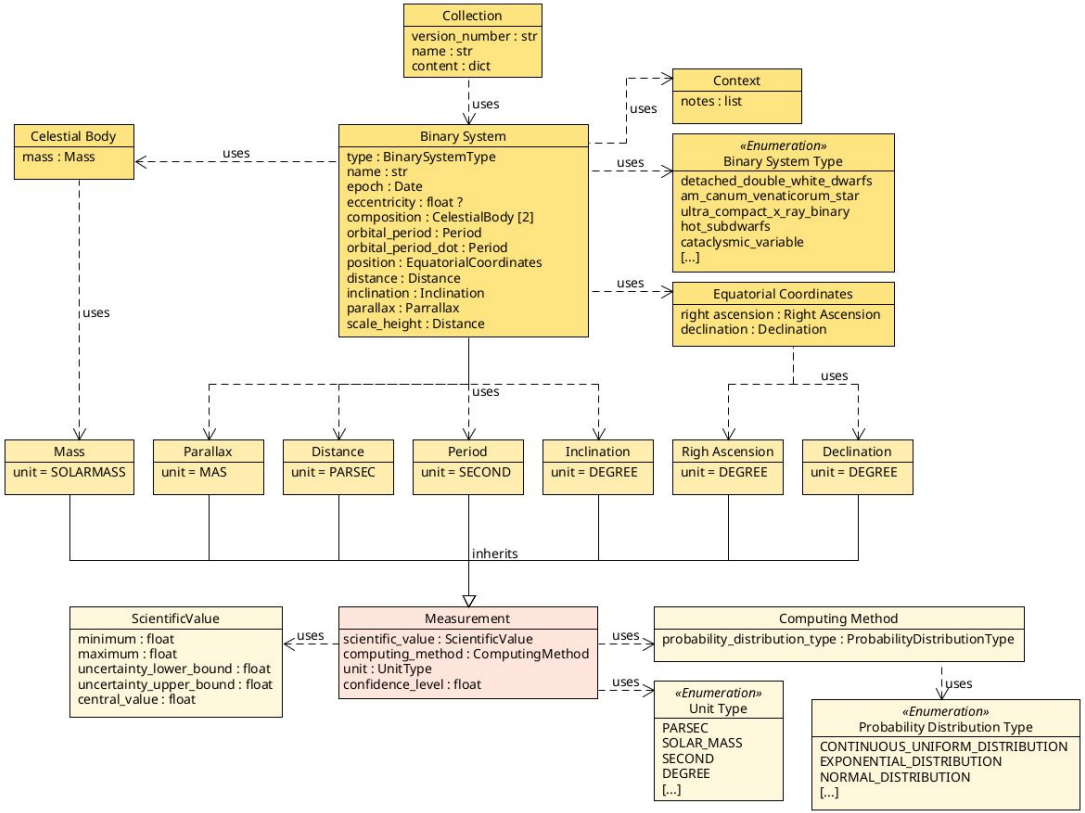
m1	delta(m1)
0.55	-
0.61	+0.017 / -0.022
0.349	+0.093 / -0.074
[0.8]	[0.1]
[0.8]	[0.1]
0.247	+/-0.0125

Parameter and their uncertainties are reported in heterogeneous formats.

➔ Come from different instruments, methods, teams.

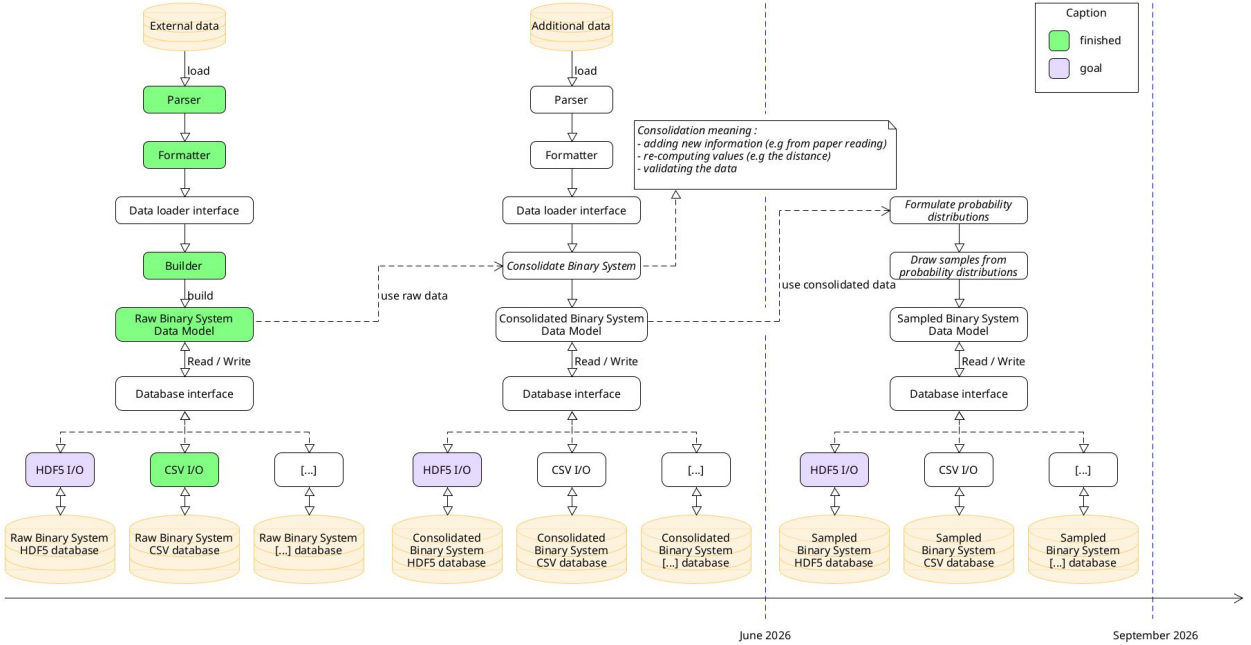
➔ Need to find a consistent and unified framework to represent these values.

Data model



- High-level data model: each VB is an instance of the *BinarySystem* class.
- Physical parameters represented with the *Measurement* class
  - *ScientificValue*: central value, uncertainty, range of authorized values.
  - Type of PDF.
  - Unit.
- *Collection* class with I/O methods to parse/write into different formats.
  - Reference format for the database is HDF5/

Implementation



- **External Data:** input VB dataset from Kupfer et al (2024)
  - ➢ Data provided by the community.
- **Additional data:** information and physical assumptions made internally
  - ➢ Necessary to standardize, handle different input formats and missing values. Well motivated and as generic as possible.
  - ➢ Clear separation between input data and scientific data.
  - ➢ Validation process.

From parameters to probability distributions

LISA data analysis is strongly Bayesian: physical quantities should be considered as random variable with a probability distribution.

- VB parameters are provided and represented in a frequentist framework: central value + upper and lower bound.
- Our use cases require to be able to convert them into probability distributions.
  - PDF to be used as priors for global fit pipelines.
  - Draw samples for Mojito Heavy generation.

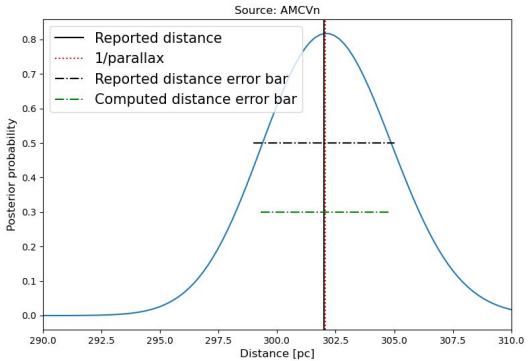
This is not always trivial: assumptions must be made on the (unknown) underlying probability distribution of parameters.

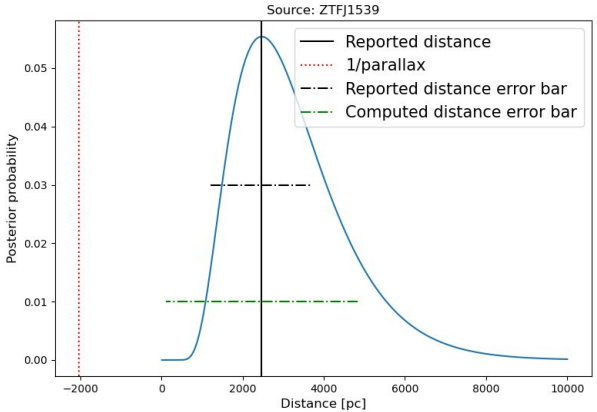
➔ Understand how each VB parameter is estimated to find the underlying PDF.
➔ When needed, make reasonable assumption on the shape of the PDF.

Distance

Bayesian probability distribution computed from Gaia parallax (Gaussian likelihood) + prior.

- ➔ Well constrained parallax : quasi-Gaussian distributions peaked at 1 / parallax (left plot)
- ➔ Uncertain parallax : proba distribution driven by uninformative prior (right plot).





- ➔ The reported parameter comes directly from a probability distribution, so it is straightforward to convert it back.

Masses

- ➔ Murkier than distance. The probability distributions associated with the reported parameters and uncertainties are not straightforward so assumptions must be made.
- ➔ Different methods to estimate the mass:
  - ◆ Radial velocities.
  - ◆ Eclipsing systems: radius + mass/radius relation. (systematics)
  - ◆ Inferred from binary evolution models.

Method used not always explicitly reported: **need to find a unified framework.**

- ➔ Use a truncated Gaussian or logNormal probability distribution compatible with the uncertainty reported.
  - ◆ Prevent negative mass values (large relative uncertainty).

Not a perfect model (mass/inclination correlation not taken into account).

- ➔ Will be improved in future versions.

m1	delta(m1)
0.55	-
0.61	+0.017 / -0.022
0.349	+0.093 / -0.074
[0.8]	[0.1]
[0.8]	[0.1]
0.247	+/-0.0125

Conclusion

Goal: provide a standardized framework to collect and distribute external data.

- First External Data prototype : database of VB and its API (python package).
  - ➢ Deadline: June 2026.
  - ➢ Use case: Mojito heavy data generation (May 2026).
- Ongoing definition of the scope of the work package and its use cases.
  - ➢ Still flexible.
  - ➢ Coordination with other LISA working groups working on similar topics.
- External Data will provide the set of Verification Binaries used in Mojito Heavy.

➔ **Service to the LISA community.**
  - ➢ What data do you need? In which format?
  - ➢ People are welcome to join this effort!

➔ Ensure consistency and reproducibility.

➔ Integration into the DDPC software environment.

Contact: *LISA-EXTERNAL-DATA-L@IN2P3.FR*