

Response provenance and the limits of normalization

Курпишев Иван Борисович / Ivan Borisovich Kurpishev

01 / Established findings

All nine JARVIS OPTIMADE response bodies have been retrieved. Their SHA-256 digests match those retained in the original snapshot. All 49 records match the snapshot under exact decimal and type-aware comparison. Inconsistent composition metadata is already present in those responses; the audited export did not introduce these differences.

Check	Result
Responses matching retained SHA-256	9 of 9
Records matching the snapshot	49 of 49
Pilot channels matched to XML	3 of 3
Original calculation files received and verified	0
New module tests	28 passed
Historical blockers / removed	37 / 0

Progress since the previous checkpoint

B3.1 detected inconsistent composition metadata. B3.2 pinned the exporter and tested its mechanics, but had no response bodies. B3.3 recovers this content against the retained hashes and adds the published XML for JVASP-114724 itself. JVASP-13829 was not substituted for the pilot.

Meaning of the hash match

Under the ordinary collision-resistance assumption for SHA-256, the recovered content is bound to the retained hash. Retrieval occurred on 20 September. The historical date, provenance of the hash commitment itself and historically executed revision remain without independent attestation. The internal provider-side cause has not been localized.

The original snapshot and full B3.2 archive remain byte-for-byte unchanged. A separate derivative is marked QUARANTINED_NOT_ADMITTED. This audit does not certify full API compliance with OPTIMADE. [S1, S2]

02 / Normalization: definition and proof

The domain is restricted to fully occupied, single-element sites with known chemical symbols and mutually consistent species_at_sites, nsites and formula. Mixed occupancy, an unknown site species or a formula conflict causes rejection. Physical correctness of the source geometry is not proved.

$E = \text{sort}(\text{unique}(\text{sites})); c(e) = \text{number of sites labelled } e; n = \sum c(e).$

$N(x): \text{elements} := E; \text{nelements} := |E|; \text{elements_ratios} := [c(e)/n \text{ for } e \text{ in } E].$

All other fields are copied. JSON ratios use floating-point numbers; exact rational $c(e)/n$ values are retained in the journal. This repairs metadata relative to the accepted site description and does not produce new chemical knowledge.

Proposition 1. Idempotence

$N(N(x)) = N(x)$. The transformation does not alter site data. Hence E , c and n are identical on the second application. Element ordering and rational-to-JSON conversion are deterministic, so the three assignments reproduce the same result.

Proposition 2. Preservation and journal-assisted reversibility

For $A = \{\text{elements}, \text{nelements}, \text{elements_ratios}\}$, projection outside A obeys $P(\text{not } A)(N(x)) = P(\text{not } A)(x)$. The journal $J(x)$ retains old and new A values and record digests. After digest validation, $R(N(x), J(x)) = x$ exactly for the parsed JSON record. Original bytes are preserved separately; recovery of original JSON formatting is not claimed.

N alone is not invertible: two inputs with identical sites and different erroneous metadata may map to the same output. The journal recovers the pinned input/output pair only. This is a conditional software proposition, not a proof of the KLT model.

Changed field	Records
elements	49
nelements	49
elements_ratios	43
Total field replacements	141

Three classes of inconsistency affect all 49 source records. In six records the ratio array already matched the final array; correcting the first two fields required no ratio replacement. The derivative passed the limited composition check, reapplication and restoration checks. Training remains closed.

03 / Pilot dossier: JVASP-114724

The provider XML for the same record and phase label has been retrieved. Three values were matched separately at exact paths. Matching published fields binds channels to the provider record but does not verify the full calculation context. Property values and coordinates are omitted from the public dossier. [S3]

#11 · DFT-OptB88vdW

/basic_info/main_relax_info/scf_indir_gap

OPTIMADE: _jarvis_optb88vdw_bandgap

#12 · DFT-TBmBJ

/basic_info/main_optics_mbj/main_optics_mbj_info/opto_indir_gap

OPTIMADE: _jarvis_mbj_bandgap

#13 · DFT-OptB88vdW-Voigt-bulk

/basic_info/main_elastic/main_elastic_info/voigt_bulk_modulus

OPTIMADE: _jarvis_bulk_modulus_kv

Embedded CONTCAR representation

The embedded CONTCAR text contains a cell and five sites. At unit scale in Cartesian coordinates, the cell and multiset of element-labelled positions match in exact decimal arithmetic. Same-element site permutations and $-0 = 0$ are accepted. No symmetry search or general cell transformation was performed. [S5]

The extracted text is labelled as a representation embedded in XML. It is neither authenticated as an original calculation file nor an independently generated KLT structure. Element labels do not establish which pseudopotentials were used.

Convergence information

The XML contains 11 points in the k-parameter series and 6 in the cutoff series, plus provider-declared convergence settings. Series shape was checked. The convergence criterion, binding of these series to each channel and complete calculation protocols remain unverified. No DFT calculation was run.

04 / Reproduction and the next checkpoint

The new collector saves HTTP bodies before parsing, including complete error responses. Receipts record URLs, status, size, SHA-256, local time and collector digest. These are HTTP payload bytes obtained with Accept-Encoding: identity, not TLS wire captures or independent timestamps.

The new module passed 28 tests: 20 synthetic and 8 against pinned research inputs. Coverage includes byte substitution, duplicate JSON keys, numeric and type precision, incomplete responses, unauthorized field changes, restoration, idempotence, separate XML channels and forbidden status promotion. The public package explicitly skips the 8 private-input checks. The historical 120 core tests were not rerun here; the core is preserved in the parent archive.

Terms of use

The Terms of Use page was captured, but no applicable licence identifier for the pinned snapshot was established. This does not assert that the provider lacks rights. The public package contains code, synthetic examples and audit metadata; response bodies, coordinates, property values and the derivative remain in the research package. [S4]

CP2B-B3.4 — evidence still needed

1. Channel-specific calculation protocols, INCAR, KPOINTS, pseudopotential identifiers or checksums, OUTCAR/vasprun or equivalent verifiable evidence. 2. Applicable terms for the particular data. 3. Independent KLT structure construction and a verifiable mapping to the material structure. 4. Only then, reconsider admission. Training and real-data forecasts remain disabled.

All 37 historical blockers remain; 0 removed. New KLT mappings: 0; source mutations: 0. science=ABSTAIN; DHC_chem=OPEN-NOT-CONFIRMED; EXT-RUN=NOT_RUN; truth_layer_promotion=0; T6-RA5=NOT_ISSUED; axiom_promotion=DISABLED. physical_reachability=NOT_ESTABLISHED; parameter_identification=NOT_PERFORMED.

S1 · [JARVIS OPTIMADE](#) · [structures](#)

S2 · [OPTIMADE 1.1.0](#) · [specification](#)

S3 · [NIST JARVIS](#) · [JVASP-114724 XML](#)

S4 · [JARVIS](#) · [Terms of Use](#)

S5 · [VASP Wiki](#) · [POSCAR](#) / [CONTCAR](#)