

FROM LEGACY ARCHIVE TO AI-READY RESOURCE: A PETROPHYSICAL AND PETROGRAPHIC PROPERTY DATABASE FOR CARBONATES

Ralf J. Weger, Christopher C. Mader¹, Cédric M. John², and Gregor P. Eberli

¹⁾ Institute for Data Science and Computing (IDSC), University of Miami

²⁾ College of Earth, Ocean, and Atmospheric Sciences, Oregon State University

KEY POINTS

- In its current state, the CSL database links >2,100 samples to ~18,000 velocity records, digital image analysis (DIA) derived pore shapes, composition, and — newly — Tonietto's diagenetic parameters.
- When combined with other available data archives (~2,700 AutoLab samples, ~2,900 DIA thin sections, ~1,500 shared), this can become an AI-ready resource without the need for any new measurements; we invite the associates to help shape it.
- Physical measurement data are nearly complete (porosity 95%, permeability 68%), but petrographic descriptors (Dunham, Choquette and Pray, Lucia) cover only 24–35% of samples, mostly as free text.
- Pore geometry, not porosity, controls property variability (e.g., wet P-wave velocity in the database spans 3,100 to 6,350 m/s) and AI might improve upon existing relationships.

INTRODUCTION AND OBJECTIVE

Thin-section petrography links the geologic history of a carbonate to its physical behavior. Texture, grains, cements, and pores seen under the microscope guide facies interpretation, rock typing, and the calibration of logs and seismic data. Yet carbonate thin sections are notoriously complex (Fig. 1): depositional fabrics are overprinted by dissolution, cementation, and dolomitization, producing pore systems from intragranular micropores to centimeter-scale vugs.

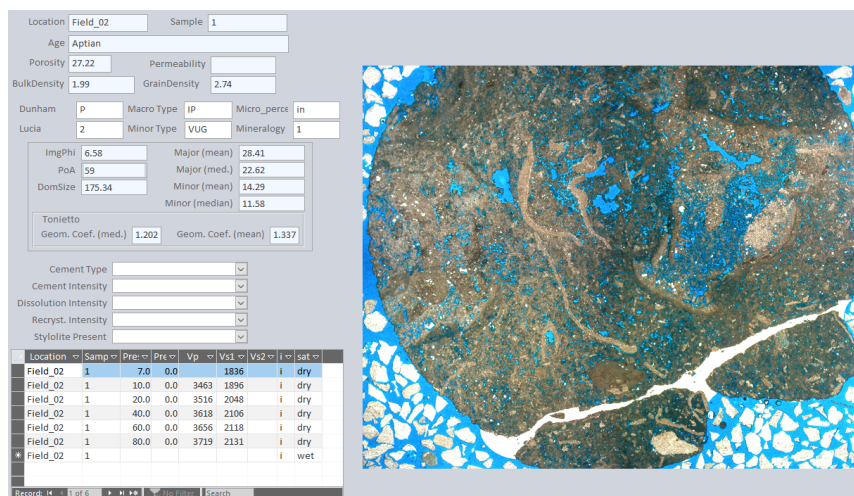


Figure 1: A sample record in the CSL database: physical properties, Dunham and Lucia classes, DIA pore geometry, Tonietto geometric coefficients and diagenesis fields (being populated), velocity runs by pressure step, and the thin-section image.

Machine learning promises faster, more reproducible description, but published carbonate image models train on a few hundred to a few thousand images with descriptive labels only. Missing is a large set of thin sections paired with measured properties of the same rock. The CSL holds one; we describe it, show why the pairing matters, and outline the effort to make it AI-ready.

WHAT WE HAVE TODAY

The core of the CSL holdings is a relational database in Microsoft Access, keyed by location and sample. A Samples table of more than 60 fields (densities, porosity, permeability, formation factor, Dunham and Lucia classes, pore descriptors) links one-to-many to velocity runs (Vp, Vs1, Vs2 by confining and pore pressure, saturation, and frequency) and to DIA pore-shape statistics, and one-to-one to geochemistry and to dated locations. Fig. 1 shows one sample record. The current export holds 2,122 samples with 18,436 velocity records from 20 locations, on average nine pressure steps per sample, mostly wet (13,936) and dry (4,169).

Tonietto parameters. We recently added the parameters of Tonietto et al. (2014), whose pore characterization combines pore geometry, connectivity, and diagenesis into a quantitative result for mapping flow units and diagenetic trends. A geometric coefficient is computed from the DIA pore shapes (1.20 median, 1.34 mean for Fig. 1), and cement type, cementation, dissolution, and recrystallization intensity, and stylolites are recorded as controlled pick-list terms, making diagenesis searchable and computable.

Completeness and vocabulary. Physical data are nearly complete (grain density 97%, porosity 95%, permeability 68% of samples), but petrographic descriptors are sparse: Dunham class 35%, macro-pore type 34%, and Lucia class 24%. Where present, they are free text, with 55 distinct Dunham strings (e.g., "P-G", "(GR", "rDol") and 45 pore-type strings. Controlled vocabularies modeled on the Tonietto fields are the first step toward closing this gap.

THIN-SECTION DESCRIPTION: ESSENTIAL BUT SUBJECTIVE

Four schemes dominate carbonate petrography: Folk (1959) names rocks by grain type and micrite versus spar; Dunham (1962), extended by Embry and Klovan (1971), classifies depositional texture by mud versus grain support; Choquette and Pray (1970) separate fabric-selective from non-fabric-selective porosity; and Lucia (1995) links rock fabric and vug type to permeability and saturation. All are indispensable, but qualitative and interpreter-dependent. Dunham does not capture pore type or size (Cantrell and Hagerty, 2003), Choquette and Pray classes show no reliable porosity-permeability relationship (Lønøy, 2006), and when petrographers described the same 24 samples, single specimens received 22 to 131 different names (Lokier and Al Junaibi, 2016). Neural networks trained on such labels (e.g., Koeshidayatullah et al., 2020; Basso et al., 2025) learn the nomenclature, inconsistencies included, rather than the physics.

WHY PORE GEOMETRY MATTERS

At equal porosity, carbonate P-wave velocity can differ by up to $\sim 2,500$ m/s within an overall range of $\sim 2,500$ – $6,000$ m/s (Anselmetti and Eberli, 1993; Fig. 2a), and permeability by up to four orders of magnitude (Fig. 2b). Archie's (1942) cementation exponent m ranges from about 1.7 to 4.1 in carbonates (Verwer et al., 2011) and up to 5.8 in travertine, against about 2 in most sandstones (Fig. 2c). Three CSL samples at 21–23% porosity illustrate this: oomoldic (0.15 mD, $m = 3.27$), vuggy (5,564 mD, $m = 3.02$), and microporous (7.94 mD, $m = 1.91$).

Quantitative pore geometry resolves much of this scatter. Digital image analysis (DIA) parameters such as dominant pore size and perimeter-over-area raise R^2 for velocity prediction from 0.54 to 0.84 (Weger et al., 2009); micro-CT descriptors raise it from 0.67 to 0.95 for permeability (Archilha et al., 2016). Appearance can mislead: two CSL dolomites with 30 and 300 μm crystals share nearly identical porosity, permeability, and m . Models must therefore learn from measured physics, not visual similarity.

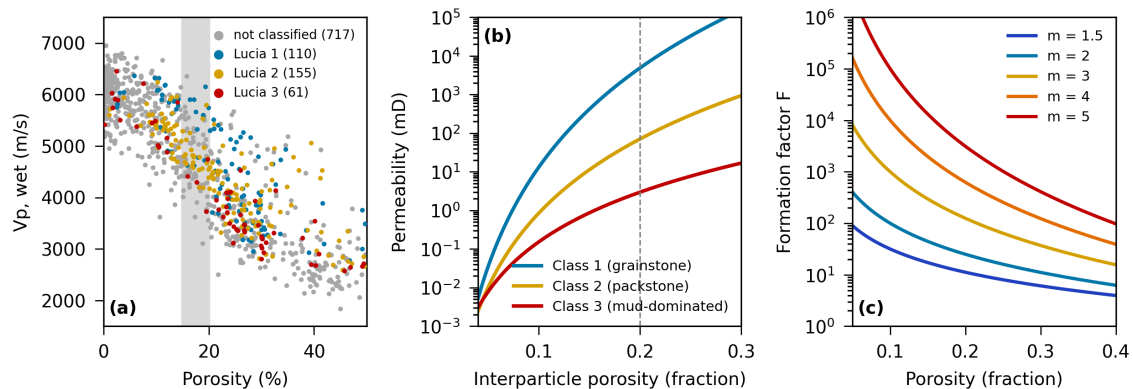


Figure 2: Pore geometry, not porosity alone, controls physical properties. (a) Wet P-wave velocity versus porosity for 1,043 samples at ~ 20 MPa; shaded: 15–20% porosity. (b) Rock-fabric permeability classes of Lucia (1995); at 20% porosity (dashed) they differ by more than three orders of magnitude. (c) Formation factor $F = \phi^{-m}$; carbonates and travertine often span $m \approx 1.7$ – 5.8 .

THE LINKED INSTRUMENT AND IMAGE ARCHIVES

Rock-Physics Archive

Beyond the relational core, the NER AutoLab 1000 systems have measured $\sim 2,756$ unique samples since 1998 across some 70 projects, producing 5,053 primary data files (13.8 GB) and an estimated 28,646 pressure steps. Velocity dominates the record, with $\sim 2,300$ wet and $\sim 1,100$ dry velocity files, followed by 1,485 resistivity files and 500 permeability files, the latter always co-measured with velocity or resistivity. Because each run steps through confining and pore pressure, every pressure step is a separate label rather than a single value per plug. The largest projects are VM16 (282 samples), IODP Expedition 359 (208), the Neuquén Basin (195), BethVandenBerg (145), and CARBO_DM (142). The files are internally consistent and documented at the run level, but stored as per-project file trees rather than database records, and the link to legacy sample identifiers and to image analysis is often incomplete.

Thin-Section Archive

About ~2,900 thin sections have been processed with full DIA, spread across four archive branches: personal working folders (~1,050), images organized by location (~1,000), contract projects (~750), and the CARBO project (~120). The largest collections come from PDO Fahud (~390), Mallorca (~280), the Boca Chica core and the Dolomites (~180 each), Qatar (~130), and Buhasa (~120). Every processed sample includes stitched PPL and XPL mosaics in three stage orientations, a full-resolution image with scale bar, region-of-interest files, black-and-white pore-segmentation masks, ImageJ particle-analysis results, MATLAB pore-geometry statistics, and a shape-summary spreadsheet and report. These counts are conservative: several directory listings were truncated during the review.

Same Rocks

The archives describe the same physical samples, drawn from carbonate systems worldwide: the Bahamas, Mallorca, Maiella and the Apennines, Oman (Fahud, Buhasa), Qatar, Malampaya, the Neuquén Basin, and ODP Leg 194 and IODP Expedition 359. About 1,500 samples carry both a segmented thin-section image and at least one acoustic measurement. This intersection is the highest-value subset, the natural starting point for petrographic enrichment, and the core of any training set; no published carbonate AI data set pairs images with measured velocity, permeability, and resistivity at comparable scale.

FROM LEGACY ARCHIVE TO AI-READY RESOURCE

No new measurements are needed; the value lies in making what exists usable, and that is painstaking work. Holdings are spread across instrument file trees, DIA folders, project archives, spreadsheets, and theses. The same specimen may appear under a project label, a plug number, and a thin-section number that do not resolve to one another automatically; geologic descriptors are free text; and measurement conditions are sometimes recorded only in a notebook. We plan a three-year effort in four steps proposed to the NSF AI Datasets program (NSF 26-512).

Sort and Match

Inventory all 5,053 AutoLab files, the ~3,000 DIA folders, and the relational database, then build an identifier crosswalk that preserves every historical label

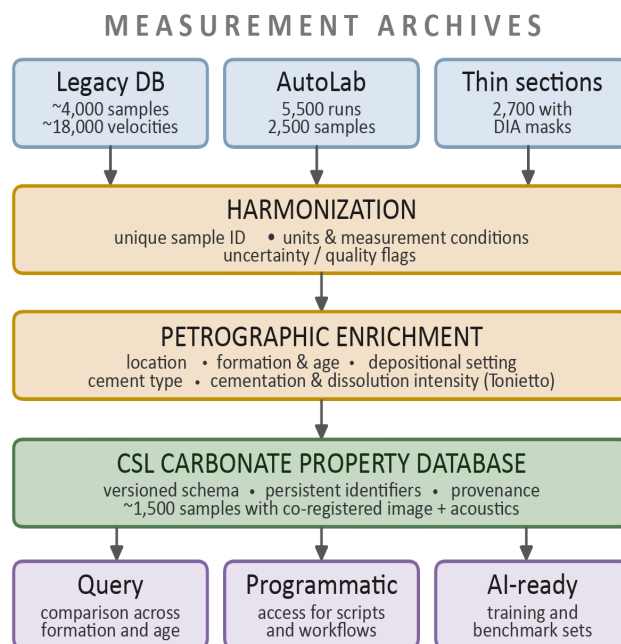


Figure 3: Planned architecture of the CSL carbonate property database. The measurement archives are reconciled through a harmonization layer, enriched with petrographic and geologic descriptors.

while assigning one persistent key to each physical specimen. The crosswalk resolves which plugs, thin sections, and velocity runs belong together and yields the first defensible counts of samples by locality, formation, and age.

Harmonize

Each value receives its units, measurement conditions (confining and pore pressure, increasing or decreasing pressure path, saturating fluid, frequency), plug orientation, and a quality flag. PPL and XPL mosaics are registered with their regions of interest and segmentation masks. The 55 Dunham strings and 45 pore-type strings are mapped to controlled vocabularies; location, formation, age, and depositional setting become structured fields, and diagenesis is scored with the Tonietto pick lists.

Validate

Duplicates, gaps, and outliers are flagged rather than silently discarded, and values whose conditions cannot be recovered are retained with a flag. Automated cross-checks compare image porosity with measured porosity and test velocities against physical bounds. Texture and pore-type labels proposed by AI-assisted segmentation and classification are confirmed by experts, starting with the ~1,500-sample image-plus-acoustics subset.

Release

With IDSC we plan a schema-versioned relational core, an object store for images and raw instrument files, persistent identifiers at the sample and measurement level, and a programmatic interface, so that a query can be issued from a script as readily as from a form. Provenance is recorded per record: who measured it, when, on which instrument, in which project. Machine-learning splits are drawn by location and formation, not by sample, so that benchmark scores reflect performance on genuinely new rocks; documented access rules and a contribution process allow new data to be added without compromising existing holdings.

WHAT AN AI-READY ARCHIVE ENABLES

The archive will support five families of tasks: texture classification (Dunham, Folk); pore typing (Choquette and Pray, Lucia); quantitative pore-geometry metrics from segmentation; prediction of porosity, permeability, velocity, and m with uncertainty, tested on laboratory measurements; and retrieval of the nearest measured analogs for any new thin section. Because images are tied to measured physics, models can be judged by how well they predict the rock's response, not only by how well they reproduce a label.

For the associates, the immediate value is comparative and works in both directions. From the geologic side, "what porosity, permeability, and velocity should I expect in a grainstone of this age and setting?" becomes a query rather than a literature search, which helps estimate properties of little-known strata in frontier regions. From the geophysical side, a velocity-porosity outlier can be traced to the pore types and diagenesis that explain it. Beyond the consortium, the archive would serve rock physics, reservoir characterization, hydrogeology of carbonate aquifers such as the Floridan, and carbon-storage and geothermal

assessment, and would give machine-learning research a physics-labeled benchmark that carbonates presently lack.

THREE-YEAR PLAN AND INVITATION

- **Year 1:** inventory, QC, and matching images to laboratory data; metadata schema and provenance model.
- **Years 1–2:** harmonization, AI-assisted segmentation, and expert-validated labels.
- **Years 2–3:** versioned ML-ready release with baseline models for classification and for ϕ , k , V_p , and m . Documentation, access rules, workshops, and training.

REFERENCES

- Anselmetti, F. S., and Eberli, G. P., 1993, Controls on sonic velocity in carbonates: *Pure and Applied Geophysics*, v. 141, no. 2, p. 287–323.
- Archie, G. E., 1942, The electrical resistivity log as an aid in determining some reservoir characteristics: *Transactions of the AIME*, v. 146, p. 54–62.
- Archilha, N. L., Missagia, R. M., Hollis, C., Ceia, M. A., McDonald, S. A., Lima Neto, I., Eastwood, D. S., and Lee, P., 2016, Permeability and acoustic velocity controlling factors determined from X-ray tomography images of carbonate rocks: *AAPG Bulletin*, v. 100, no. 8, p. 1289–1309.
- Basso, M., Souza, J. P. d. P., Chinelatto, G. F., *et al.*, 2025, Petrographic image classification of complex carbonate rocks from the Brazilian pre-salt using convolutional neural networks: *Scientific Reports*, v. 15, 30700.
- Cantrell, D. L., and Hagerty, R. M., 2003, Reservoir rock classification, Arab-D reservoir, Ghawar field, Saudi Arabia: *GeoArabia*, v. 8, no. 3, p. 435–462.
- Choquette, P. W., and Pray, L. C., 1970, Geologic nomenclature and classification of porosity in sedimentary carbonates: *AAPG Bulletin*, v. 54, no. 2, p. 207–250.
- Dunham, R. J., 1962, Classification of carbonate rocks according to depositional texture, in Ham, W. E., ed., *Classification of Carbonate Rocks*: AAPG Memoir 1, p. 108–121.
- Embry, A. F., and Klovan, J. E., 1971, A Late Devonian reef tract on northeastern Banks Island, N.W.T.: *Bulletin of Canadian Petroleum Geology*, v. 19, no. 4, p. 730–781.
- Folk, R. L., 1959, Practical petrographic classification of limestones: *AAPG Bulletin*, 43, 1, p. 1–38.
- Koeshidayatullah, A., Morsilli, M., Lehrmann, D. J., Al-Ramadan, K., and Payne, J. L., 2020, Fully automated carbonate petrography using deep convolutional neural networks: *Marine and Petroleum Geology*, v. 122, 104687.
- Lokier, S. W., and Al Junaibi, M., 2016, The petrographic description of carbonate facies: are we all speaking the same language?: *Sedimentology*, v. 63, p. 1843–1885.
- Lønøy, A., 2006, Making sense of carbonate pore systems: *AAPG Bulletin*, 90, 9, p. 1381–1405.
- Lucia, F. J., 1995, Rock-fabric/petrophysical classification of carbonate pore space for reservoir characterization: *AAPG Bulletin*, v. 79, no. 9, p. 1275–1300.
- Verwer, K., Eberli, G. P., and Weger, R. J., 2011, Effect of pore structure on electrical resistivity in carbonates: *AAPG Bulletin*, v. 95, no. 2, p. 175–190.
- Weger, R. J., Eberli, G. P., Baechle, G. T., Massaferro, J. L., and Sun, Y.-F., 2009, Quantification of pore structure and its effect on sonic velocity and permeability in carbonates: *AAPG Bulletin*, v. 93, no. 10, p. 1297–1317.
- Tonietto, S.N., Smoot, M.Z. and Pope, M., 2014. PS Pore Type Characterization and Classification in Carbonate Reservoirs, AAPG Annual Convention and Exhibition. Article, Houston, Texas.