Keeping LEP-era knowledge usable: a DELPHI knowledge base for physicists and AI agents

A prototype for preserving the knowledge needed to reuse DELPHI data

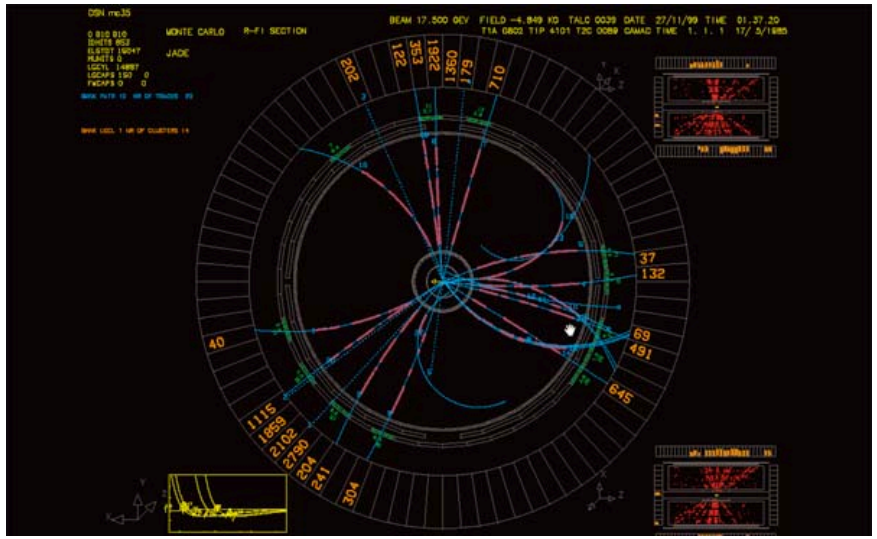
D. Liko[1] and U. Schwickerath[2]

[1] MBI · [2] CERN/IT

MARIETTA BLAU INSTITUTE FOR PARTICLE PHYSICS

Why revisit data from an older experiment?

- Advances in theory and methods can reveal new insights in archived data.
- **JADE:** new α_s measurement from PETRA data at $\sqrt{s}$ = 14–44 GeV, published in 2009—23 years after PETRA ended.
- **For LEP:** 20+ years on, experts on its data, detector, software and analyses are dwindling.



JADE event display. Image: S. Bethke, CERN Courier (2009)

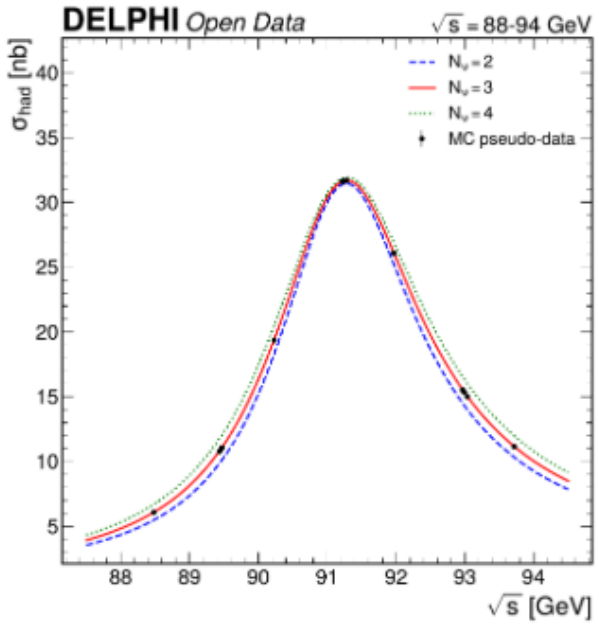
CERN Courier (2009), "Study group considers how to preserve data"; Bethke et al., EPJ C 64 (2009), 351–360.

DELPHI in the DPHEP picture

- DPHEP coordinates long-term data preservation across HEP.
- DELPHI released its full dataset, software and documentation.
- Licences: data CC0, software open source, documentation open access.
- DELPHI software ported to modern architectures; CERNLIB is open source.
- Channels: Open Data portal (data), GitLab/CVMFS (software), CDS (notes)

DELPHI data release policy, doi:10.7483/OPENDATA.DELPHI.0X8O.FBUH; cernlib.web.cern.ch

Agentic analysis: capability—and a caution



The data points are MC simulation; no real DELPHI data are shown.

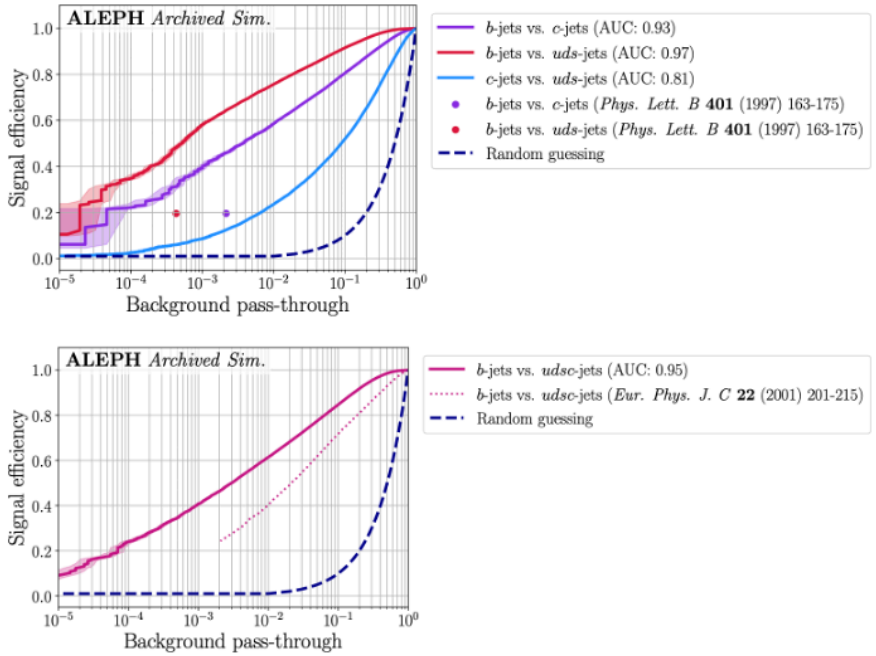
Study by Moreno et al.:
AI Agents Can Already Autonomously Perform Experimental High Energy Physics
arXiv:2603.20179 · JFC project site

DELPHI example: N_ν from Γ_{inv}. The JFC site marks it "Archived — not public."

My interpretation: incomplete knowledge misled the agent, giving wrong results.

Modern environment: Key4hep

- Key4hep/EDM4hep: FCC-ee software
- Motivation: modern tools, esp. ML
- Real-data tests and feedback for FCC
- ALEPH 1994 data: modern b-tagging gives ~2× efficiency at fixed background
- First DELPHI jet-tagging results



Defranchis et al., JHEP 07 (2026) 096, arXiv:2603.06524; Fanini et al., arXiv:2603.15493. DELPHI: "Instance-Level b-Hadron Tagging with DELPHI Open Data," ML4Jets 2026, Vienna, 14–18 Sept 2026.

Data alone are not enough for an analysis

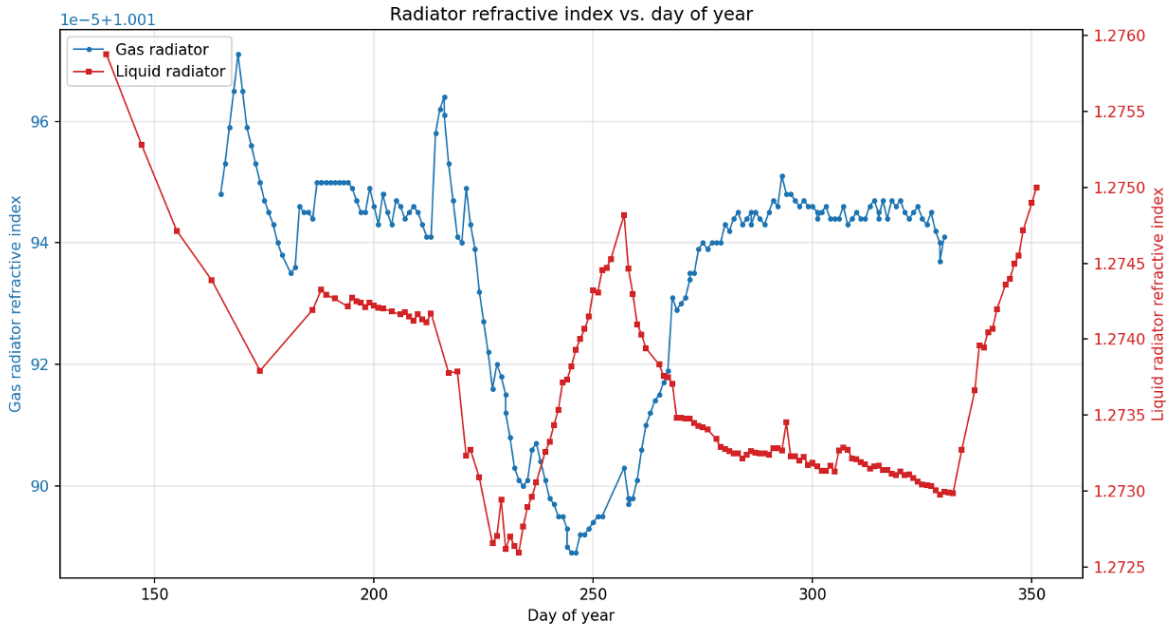
Interpreting the data requires metadata on how they were taken and processed:

- **Run conditions:** beam energy, beam parameters, beam spot
- **Detector status:** configuration, calibration, efficiencies
- **Detector changes:** upgrades, repairs, operating conditions over time
- **Software:** versions, reprocessing, matching simulation
- **Analysis choices:** selections, corrections, conventions

These details are rarely in the papers: they live in notes, logbooks and code.

Example: RICH refractive index

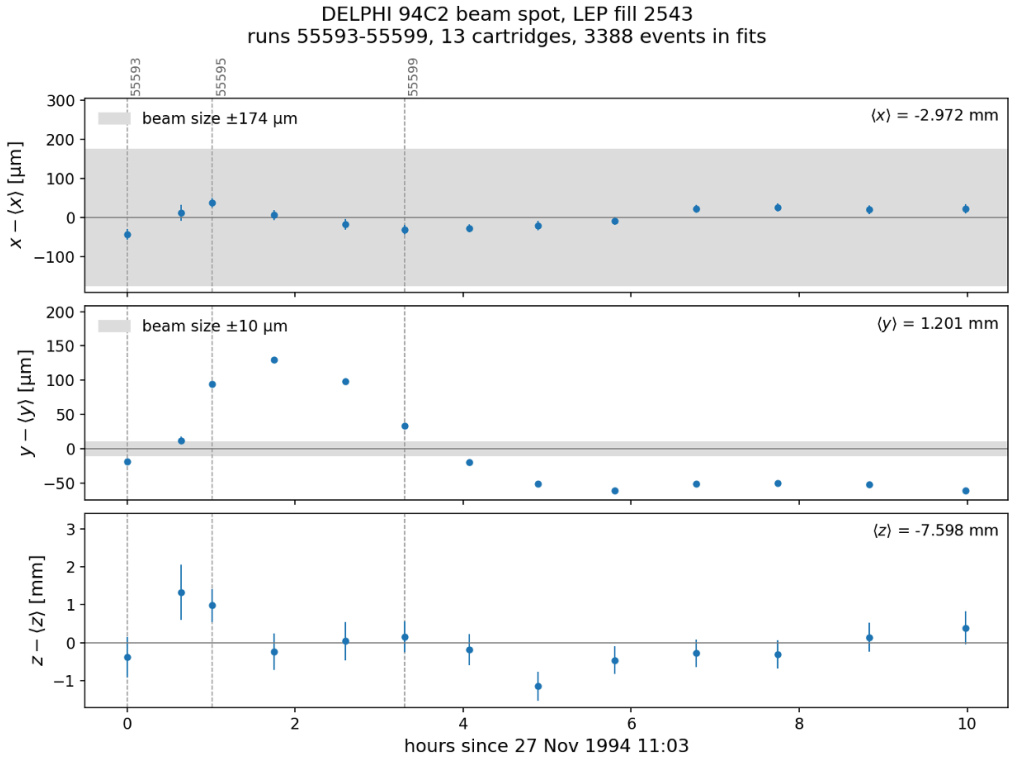
- DELPHI: first operational 4π RICH
- 1994: leaks mixed barrel gas and liquid radiators; refractive indices drifted
- 1995 notes: fluid separation and refractivity monitoring
- Demo: note → passage → plot



DELPHI-95-46-RICH-69; DELPHI-95-123-RICH-72; Fürstenau et al., NIM A 371 (1996) 263–267

Example: Beamspot

- Beam spot positions fitted in LEP fill 2543, runs 55593–55599
- No per-event record: the within-fill history is derived from tracks
- Notes explain how it was recovered



Source plot and supporting note to be confirmed

DELPHI corpus: 4,515 catalog records

Document type	Records
Internal notes	2,058
Conference contributions	1,359
Publications	549
Theses	391
Reviews	9
Other	149

What is counted	Count
Catalog records	4,515
Records with a retrieved PDF	3,277
Records without a PDF found	1,238
Retrieved PDFs	3,294
Pages in retrieved PDFs	113,106

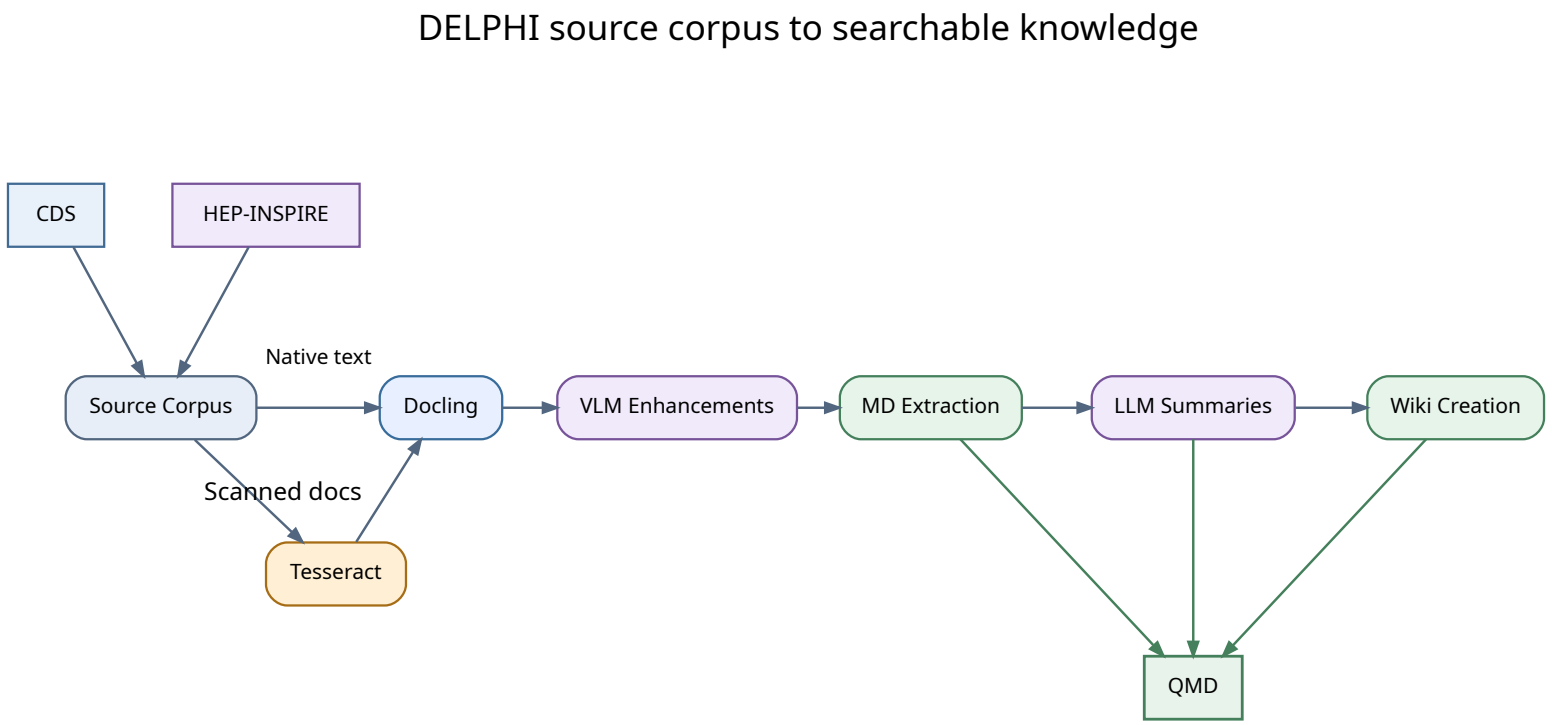
- Conversion modes: 1,914 native, 1,001 OCR-layer, 306 OCR, 73 mixed
- Internal notes capture details omitted from the papers
- Source code: the ultimate record of what was done

Aim: build a system for humans and agents

- **Traditional RAG system:** retrieve passages from the full text by keywords (BM25) and by meaning (embeddings). This is the evidence.
- **Karpathy-style LLM wiki:** an LLM compiles the corpus into linked, reviewable pages per document and topic. This is the synthesis.
  - **Global view:** e.g. knows which results are final
  - Everything traceable to its source
- **MCP service for the source code:** navigate PATCHY/Fortran routines, COMMON blocks and calls. This is how the analyses were implemented.

Prototype search: QMD, *Query Markup Documents* (github.com/tobi/qmd, by Tobi Lütke), an on-device search engine for Markdown that combines BM25 and vector search.

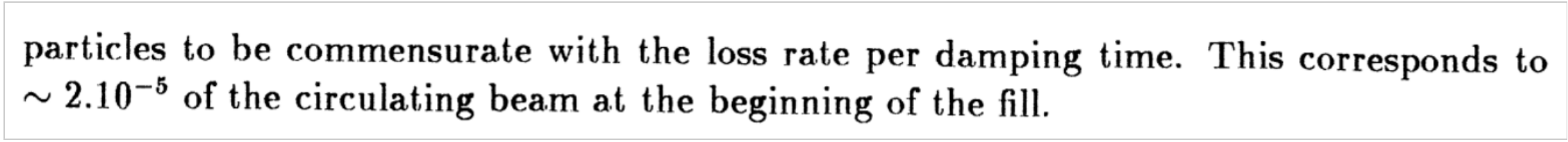
From source corpus to searchable knowledge



VLM corrections grounded in OCR text and page scan; page provenance kept.

VLM-mediated corrections of OCR snippets

Page scan


particles to be commensurate with the loss rate per damping time. This corresponds to $\sim 2.10^{-5}$ of the circulating beam at the beginning of the fill.

Tesseract OCR

```
particles to be commensurate
with the loss rate per
damping time. This
corresponds to ~ 2.10~° of
the circulating bea ng
~ 2.10~° of the circulating
beam at the beginning of the
fill.
```

VLM correction (changes marked; typeset below)

```
particles to be commensurate with the loss rate per
damping time. This corresponds to ~$\sim 2.10~° of the
circulating bea ng
~ 2.10~°2.10^{-5}$ of the circulating beam at the
beginning of the fill.
```

---

particles to be commensurate with the loss rate per damping time. This corresponds to $\sim 2.10^{-5}$ of the circulating beam at the beginning of the fill.

Qwen3.8-27B pilot, delphi-1988-069-gen-082 p. 10; 620 blocks, 615 accepted.

Document processing: challenges and lessons

- **Varying quality:** born-digital PDFs, TeX output, typewritten and scanned notes. About 40% of the PDFs come from scans (OCR-layer, OCR or mixed).
- **Text extraction:** native text, or Tesseract OCR (better than EasyOCR in our tests), followed by Docling, chosen for its handling of formulae and tables.
- **Symbols and inline formulae** come out garbled even from born-digital PDFs, so these documents also get the VLM correction step.
- **LLM artefacts:** a VLM can hallucinate, drop or rewrite content. A filter rejects edits such as invented numbers. The scan remains authoritative.
- **To be done:** compare OCR tools on degraded scans (e.g. Nougat, olmOCR).

Search with QMD: optimisation and lessons

- **Keyword search (BM25):** symbols and formulae break matching. A QMD patch maps variants (`alpha_s` , `αs` , `$\alpha_s$`) to one form via a lexicon from the text. Proved useful.
- **Semantic search:** heading-based chunks (~900 tokens); EmbeddingGemma-300M and Qwen3-Embedding-0.6B indexes built, comparison ongoing.
- **Reranker:** too slow on lxplus, so it is not used.
- **Deployment:** QMD on lxplus works well for both lexical and vector search.

Scaling an LLM wiki

Karpathy-style wiki: an agent iteratively builds linked notes from source documents.

At 4,515 documents, human curation and repeated LLM passes both become costly.

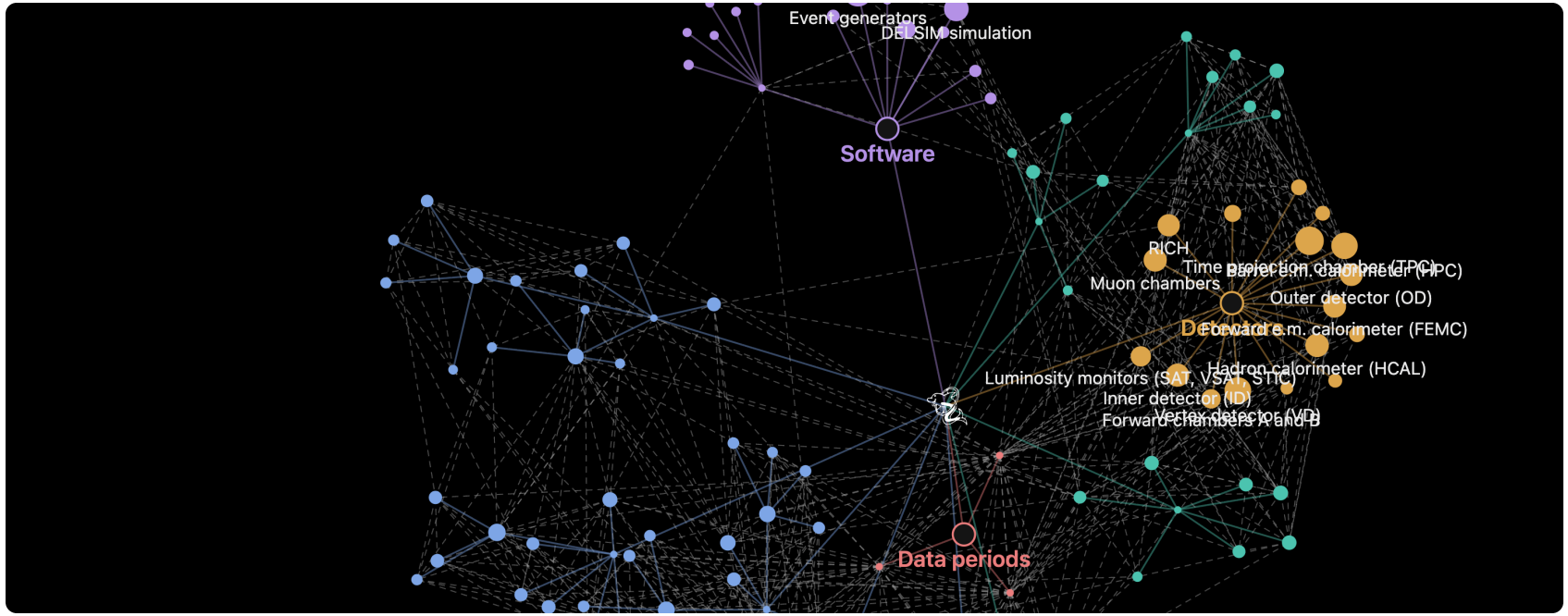
- **Skeleton:** generate document pages, metadata, and links deterministically.
- **Summaries:** Qwen3.8-27B creates document cards with a predefined typology, which also defines the wiki taxonomy. Numbers are checked for hallucinations; pages combine cards and metadata.
- **Topics:** a program collects material from cards and cited sources; the CERN-hosted `glm-5.3-flash-int4` synthesises cross-document pages in the **OpenCode** harness.
- **Checks:** automated checks at each step verify quoted numbers and provenance.

All 125 topic pages are generated. They can and should be checked by humans.

The wiki as a static website

The wiki is exported to a static website, with a link graph of documents and topics.

The wiki lives in git: changes are committed and pushed.

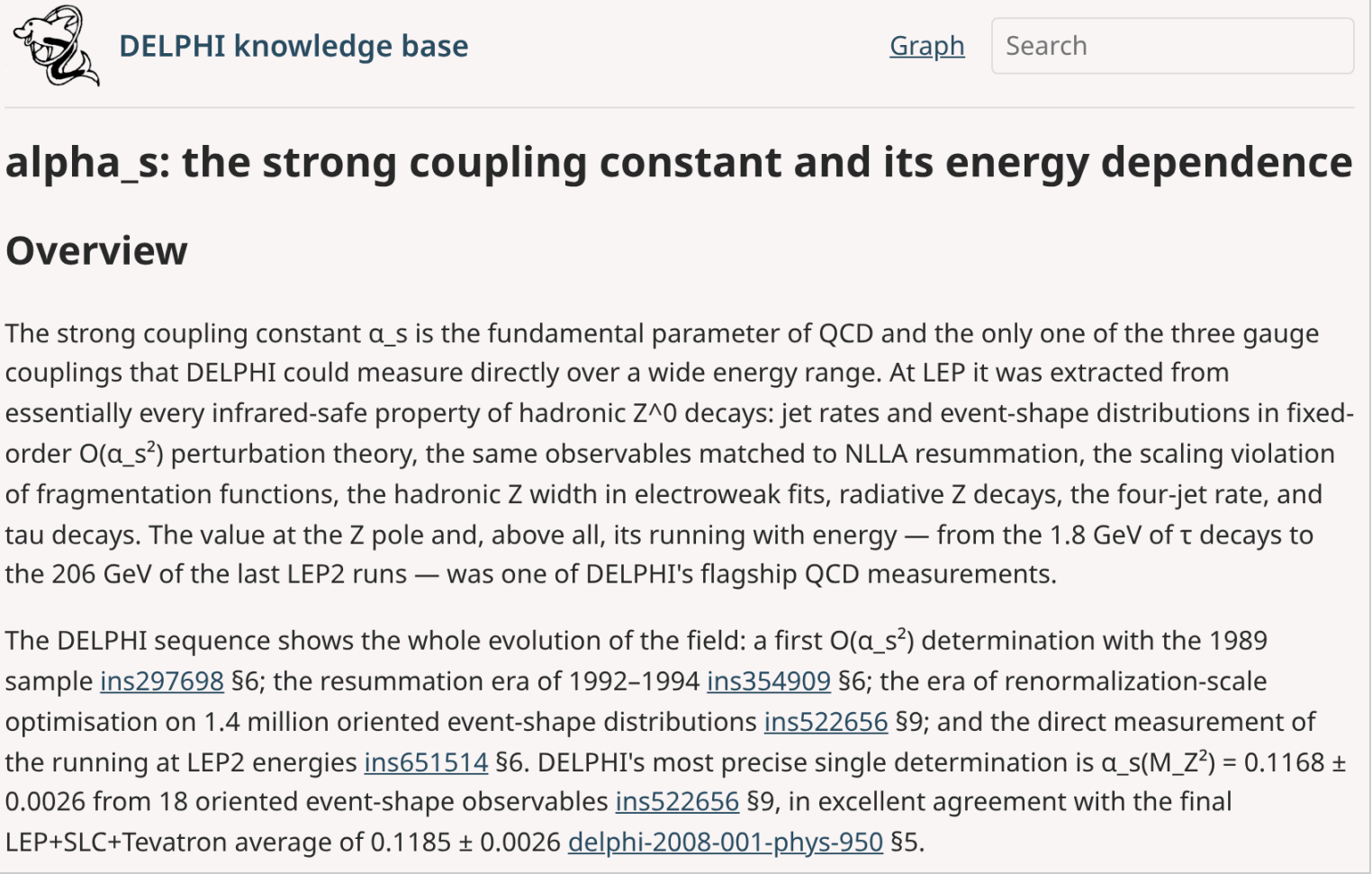


Live wiki (CERN login): **cern.ch/liko/delphi-kb**

Example: compilation of α_s results

- **Overview:** methods and history, with citations
- **Results:** 60 values with √s, years, source, status
- Errors as printed

Live: **α_s topic page**



DELPHI knowledge base

Graph Search

alpha_s: the strong coupling constant and its energy dependence

Overview

The strong coupling constant α_s is the fundamental parameter of QCD and the only one of the three gauge couplings that DELPHI could measure directly over a wide energy range. At LEP it was extracted from essentially every infrared-safe property of hadronic Z^0 decays: jet rates and event-shape distributions in fixed-order O(α_s²) perturbation theory, the same observables matched to NLLA resummation, the scaling violation of fragmentation functions, the hadronic Z width in electroweak fits, radiative Z decays, the four-jet rate, and tau decays. The value at the Z pole and, above all, its running with energy — from the 1.8 GeV of τ decays to the 206 GeV of the last LEP2 runs — was one of DELPHI's flagship QCD measurements.

The DELPHI sequence shows the whole evolution of the field: a first O(α_s²) determination with the 1989 sample ins297698 §6; the resummation era of 1992–1994 ins354909 §6; the era of renormalization-scale optimisation on 1.4 million oriented event-shape distributions ins522656 §9; and the direct measurement of the running at LEP2 energies ins651514 §6. DELPHI's most precise single determination is α_s(M_Z²) = 0.1168 ± 0.0026 from 18 oriented event-shape observables ins522656 §9, in excellent agreement with the final LEP+SLC+Tevatron average of 0.1185 ± 0.0026 delphi-2008-001-phys-950 §5.

Lessons learned: wiki generation

- **Reduced agentic capabilities** of CERN-hosted `glm-5.3-flash-int4` vs Z.AI-hosted GLM-5.3-Flash-FP8 in my setup (quantisation, effort)
- **Material collection** is therefore done by a program, not the agent.
- **Topic pages:** content similar to the Z.AI model's.
- **Claude Sonnet 5.5** showed a somewhat better grasp of the material.
- **No prompt caching** in my setup, so large contexts cost time and tokens.
- **CERN-hosted models** could address data-protection concerns. For DELPHI, with all material openly released, this is not required.

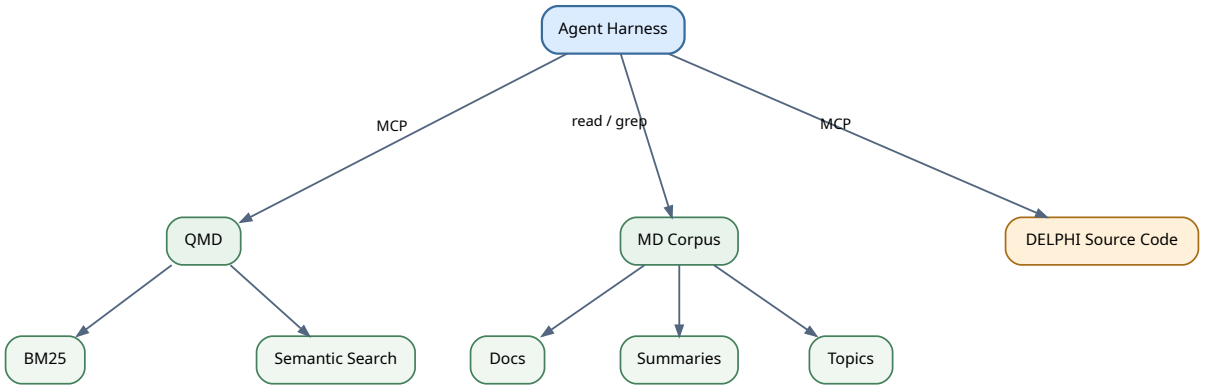
DELPHI PATCHY/Fortran MCP server

- Frontier models understand Patchy/Fortran; an index helps navigate the code.
- Python parser: routines, calls, COMMON/KEEP blocks, Patchy corrections.
- Author comments and LLM summaries give context for each element.
- A lookup table explains CERNLIB symbols used by DELPHI code.
- The MCP links results to source; inferred relationships are marked as such.

delphi-pam-mcp on GitLab

Agent harness: three routes into DELPHI knowledge

DELPHI knowledge base and Agentic Tools



- The QMD route exposes lexical and semantic search.
- Direct read/grep operates over Markdown documents, summaries, and topic pages.
- The source-code MCP exposes DELPHI legacy-code navigation.
- Claude Code is the prototype interface; other harnesses can use the tools via MCP.

Test 1: D* reconstruction

Task: from literature and code, describe how DELPHI reconstructed $D^{*+} \to D^0\pi^+$ at the Z, cited.

Result:

- 15 analyses (1990–2000) compared: channels, cuts, resolutions, yields, systematics
- 30 documents and 176 line-level citations; inferences and open questions marked
- Inconsistent numbers in the sources flagged
- Code traced to RECDSR finder and DST bank; a RECDSR bug found by reading the code

Expert assessment: good extraction and overview of selections; the code bug is real.

Test 2: evolution of the beam-spot determination

Task: a cited timeline of how DELPHI determined the beam spot, 1989–2000.

Result:

- Timeline: per fill (1990) → per run (1991) → per cartridge of ~200 hadrons (1992–95) → every 20–30 min with the LEP beam monitors (LEP 2); precision as quoted for each period
- 11 documents cited; those behind the tables re-read and verified
- Code traced from the database reader to the DST words
- Gaps stated (1989, LEP 2 precision); a non-existent fill in the prompt flagged

Expert assessment: method and evolution correct; subroutines correctly identified.

Resource consumption

Step	GPU (ml.cern.ch)	CERN LLM gateway
PDF conversion (Docling)	~7 h	–
Text correction (VLM)	~4 h	–
Document cards (Qwen3.8-27B)	~14 h (~70M tokens)	–
Search index (QMD)	~7 h	–
Topic pages (125, glm-5.3-flash-int4)	–	~12 h, ~260M tokens
Total	**~32 GPU hours**	**~260M tokens**

- 4,515 catalog records and 113k pages turned into a searchable wiki
- GPUs: RTX PRO 6000 and T4; one GPU at a time was enough
- The same topic-page work on OpenRouter would cost about $25–40.

Ongoing evaluations and plans

- **Agentic search:** determining performance metrics.
- **Agentic evaluations:** with the knowledge base as information source.
- **Ablations (planned):** what each component adds over plain Claude Code.
- **Towards a service:** the results define the required functionality; the knowledge base runs outside the user's environment.
- **Open to the community:** users such as DELPHI physicists, who may not be experts in modern tools, should be able to add their personal knowledge.