exequt
Where Action Meets Impact

A TECHNICAL WHITE PAPER

Keeping three million datasets findable in the language of today's research

DATMM-MeSH: the tagging model at the front door of the dataset catalog. Why the catalog speaks MeSH, what that choice demands, and how a purpose-built model came to serve both sides of every search.

ALAA SENJAB, CHIEF AI OFFICER · PREPARED BY EXEQUT · AUGUST 2026

1 in 5 datasets is findable only through its tags	**0.673** micro-F1 against expert human indexing	**30,000+** MeSH 2026 descriptors, every term since 2022 covered	**Zero** dataset text sent to outside AI services

All catalog statistics measured on the production system; all model scores from held-out, human-indexed evaluation sets.

ABSTRACT

The dataset catalog holds +3 million datasets and answers about 6.5 million search requests a month. It was built on a deliberate choice: the catalog speaks MeSH, the vocabulary behind PubMed and the indexing language biomedical researchers already trust. Every dataset entering the catalog must carry accurate MeSH tags, a dataset that does not have tags is not admitted, and every search a researcher types is translated into MeSH before it runs. That choice makes automated tagging the catalog's front door, and the quality of the tagging the quality of the catalog itself.

This paper tells the story of that requirement and of DATMM-MeSH, the compact, self-hosted model built to meet it. Trained on the full indexed biomedical literature against the current MeSH vocabulary, the model reaches 0.673 micro-F1 against professional human indexing, ahead of the leading open model at 0.608 and within five points of the National Library of Medicine's own production indexer while working from a fraction of the information. It covers 1,914 current MeSH terms the previous tagger could never produce, including every term added since 2022, and it runs entirely inside the agency's environment on ordinary CPU hardware. The same model works both sides of the catalog's search: it tags every dataset on the way in and translates every question a researcher asks.

The paper closes with the system's design principle, a stable base with independently testable add-on layers, and the road map those layers follow: substance-level tagging, calibrated confidence tiers, an annual vocabulary refresh, and, most importantly, multilingual input. The measurements are part of the deliverable: every catalog statistic in this paper can be reproduced on demand, giving the program a standing, quantified way to demonstrate the catalog's impact.

1 A catalog built around one decision

The National Library of Medicine has launched the NLM Dataset Catalog, that does for datasets what PubMed did for papers. A researcher types a question in plain language, and the catalog answers with the datasets that can support the work, drawn not from one repository but from many. That catalog already holds more than three million records (and rapidly growing). A search across the 3 million records has to come back instantly and accurately, so the hard work behind every one of those datasets stays discoverable. Behind the search box sits a metadata schema model called DATMM, which gives every dataset the same descriptive shape, and a Transform

Portal that lets repository owners bring their dataset collections in, reading each repository's own metadata and mapping it into that common DATMM schema, so that one search box can stand in front of every repository.

Early in the design, one question mattered more than any other: what language should the catalog think in? The answer was already sitting in front of every biomedical researcher. For decades, the National Library of Medicine has organized the world's health literature with MeSH, the Medical Subject Headings, a curated vocabulary of more than 30,000 precise terms [1]. PubMed is built on it, and generations of researchers have learned to trust it without ever having to think about it: search for a disease, and every relevant paper surfaces, whatever words its authors happened to choose. So the decision was to make MeSH the semantic spine of DATMM, for a reason that has nothing to do with technology. Researchers should not have to learn a new language to find data; the catalog should speak the one they already know.

That decision was written down as three working rules:

- **Every dataset must be tagged.** As datasets flow in the Dataset Catalog, each one must receive accurate MeSH terms based on whatever metadata its home repository provides: a title, a description, sometimes existing MeSH terms if the repository already uses them. Often nothing more than a paragraph of text.
- **No tags, no admission.** A dataset that ends up with no MeSH terms while ingesting is excluded from the catalog automatically, which makes tagging the admission requirement.
- **Every search translates to MeSH.** Any query a researcher types into the catalog passes through a translation layer that converts it into MeSH terms before the search begins. MeSH is the heart of the search process.

The quality of automated MeSH tagging decides which datasets exist in the catalog at all, whether any given search can reach them, and whether the newest science can even be expressed in the catalog's language. Get the tagging right and the catalog feels effortless; get it wrong and the failure is invisible: the dataset a researcher needed simply never appears, and nobody ever learns what they missed.

2 The task has a name

The industry has a name for the problem the rules created: extreme multi-label classification [11]: read a short piece of text and choose the dozen or so correct labels from a vocabulary of more than 30,000, where most labels are rare, some are brand new, and the text almost never uses the

label's own words. It is one of the recognized hard problems in machine learning, and MeSH indexing is one of the benchmark tasks the field measures itself against. It is also the same shape of problem the largest online retailers solve every day; Amazon frames its own product search as extreme multi-label classification, matching each customer query against a catalog of a hundred million products, and has open-sourced a framework called PECOS built for exactly this class of problem [12].

3 How much rides on the tags

To show how much dataset discoverability depends on MeSH, we measured the production catalog directly, reading a 37,000-dataset stratified sample against its 24.5 million tags. Three of every four tags name something the dataset never says in its own words, which means the tag is what makes the record findable under that term. A depositor types "T2D"; a clinician searches "diabetes mellitus type 2"; the MeSH tag Diabetes Mellitus, Type 2, stamped on the record when it entered the catalog, is the only bridge between them. For an estimated 582,000 datasets, one in five, every tag they have is that kind of bridge, not mentioned in their text. Without their tags, those datasets would disappear from search results entirely.

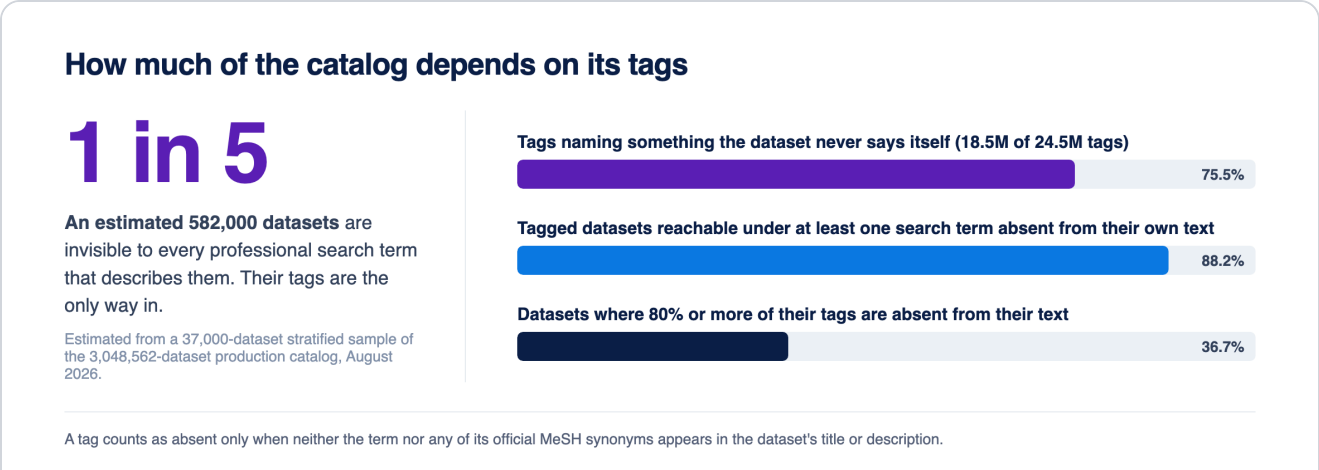


Figure 1. Findability in the production catalog, estimated from a 37,000-dataset stratified sample in August 2026. A tag counts as absent only when neither the term nor any of its official MeSH synonyms appears in the dataset's title or description.

The same dependence shows up from the opposite direction, in what searchers actually experience: when we ran eight clinical search phrases through the live search engine, four of every ten top results came back because a tag, and nothing else, matched the query. For some searches the dependence is total: every result for "anti-bacterial agents" exists because a tag bridges to datasets that say "antibiotics."

Nearly four in five of the catalog's 24.5 million tags exist because software read the dataset's text and inferred them, which makes the quality of that software the quality of the catalog's front door.

4 The tagger that time passed by

When we built the first version of the dataset catalog, we surveyed what the research world offered and chose BERTMeSH (Wellcome implementation) [2] for auto tagging, the strongest open model available for the task at the time, and a reasonable decision when it was made. Measured against the three rules stated above, though, the model was falling short, and the failures were expected to grow worse every year.

The first failure was time: BERTMeSH's vocabulary and training froze around 2021, before long COVID, the new class of weight-loss drugs, or large language models had names. Under rule three, that means the working language of the last four years of science cannot be searched, because a term the tagger cannot produce is a term no query can ever reach. Under rule two the stakes are sharper still, since a new dataset whose accurate description lives entirely in post-2021 vocabulary risks receiving no usable tags at all, and a dataset with no tags is excluded outright.

The second failure was harder to see and just as costly: every tag a model suggests carries a confidence score, the model's own estimate of how likely the term is to be correct. Those scores are what let a system act automatically: apply the certain tags, queue the doubtful ones for a person, discard the noise. As deployed, BERTMeSH's serving path compressed nearly every score into a narrow band between 0.58 and 0.73, where strong and weak suggestions look almost alike. A wrong tag at 0.58 and a certain match at 0.73 read as near-neighbors, when one belongs on the record and the other does not belong at all. This matters because rule one requires tagging millions of datasets without a person reviewing each decision, and automation like that only works when the scores can be trusted to separate the tags worth applying from the ones worth discarding; with every score crowded into the same narrow band, no threshold can draw that line.

The third failure was plain accuracy: fewer correct terms found, more incorrect ones suggested. Section 7 puts numbers to that gap, but the conclusion arrived before the measurements did, because the catalog's central mechanism was resting on a tool that could not carry it.

5 The model we could not borrow

There was one more place to look before building anything new. The National Library of Medicine already operates a world-class MeSH indexer: MTIX, the system that has tagged the biomedical literature itself since MEDLINE moved to automated indexing [4][7], which raises a fair question:

with the best tagger in the world already in the building, why build another?

The answer is that MTIX does not travel: it runs only inside NLM's internal pipeline for the literature and is not offered as a service that other systems can call. And more important than that, it was built for a different job: alongside the title and abstract it reads the metadata a published paper carries, the journal it appeared in and the year, signals that sharpen its judgment considerably. A dataset record has none of that, only a title and a description, and sometimes barely that. So the standard existed, but the tool did not fit the job and could not be borrowed for it: if the catalog wanted MTIX-class tagging on dataset metadata, the team would have to build it.

6 The model we built

Our team, decided to tackle the problem during an internal hackathon, and that produced the initial workings of DATMM-MeSH, a small, specialized model built for one job. A model like this is only as accurate as the examples it learns from, which made the choice of training data the most important decision in the build, and we chose PubMed because it is the most accurate source available for this task: more than 22 million research articles, every one carrying MeSH terms assigned under NLM's own indexing program. We train on all of it for volume, and grade only against articles a professional indexer tagged by hand, so the standard the model is measured against stays human even where the training volume is not.

The training itself follows a plain recipe. The base is a language model already fluent in biomedical text [3], so the vocabulary of medicine is native to it rather than translated. On top of that base, each of the 30,040 descriptors in the current 2026 edition of MeSH learns its own attention pattern over the text, picking out the words that signal that specific term, and a single shared scorer turns each of those readings into a confidence score. Rare and newly added descriptors, which may appear in only a handful of publications anywhere, get a dedicated training step: from each term's own MeSH record we generate a small set of synthetic training articles, half of which are deliberately written to describe the concept without ever using the term itself, so the model has to learn the idea rather than match a string. Only those descriptors' outputs are trained; everything already working is frozen, so nothing can regress. Measured on held-out data, this recovers 80% of the terms that would otherwise be unreachable, at fewer than one spurious tag per twelve datasets, and the whole pipeline can be re-run on demand, which is what lets the model track MeSH as the vocabulary grows each year. The finished model reads only a dataset's title and description, exactly what a shared dataset provides, and returns scored keywords in under tenth of a second on an ordinary CPU.

It also answers on both sides of the search box. When a repository's datasets flow in through the Transform Portal, DATMM-MeSH reads each record and stamps it with scored MeSH terms; that is rule one. When a researcher types a question, the same model translates it into MeSH before the search runs; that is rule three. The result is one model speaking one language on both sides of every match, which is what lets the catalog feel simple: the hard translation happens twice, invisibly, in under a second each time.

Because it is compact rather than a hosted service, it runs on a single machine inside the agency's own environment. No dataset text, and no researcher's query, is ever sent to an outside provider. That constraint ruled out commercial AI services from the start and shaped the design throughout.

7 Evaluation

We tested the model the hard way, against every alternative we could run, on 300 human-indexed articles that DATMM-MeSH never saw during training, while BertMeSH did see and train on. We graded only against articles indexed by professional cataloguers. We never graded against machine-generated labels, which would have flattered the result. And we ran each system at its own best operating point, meaning its own tuned settings, exactly as it runs in production.

Every system in this section is scored with the same yardstick, built from two questions. The first, called precision, asks how many of the terms a model suggested were actually right; the second, called recall, asks how many of the terms the human experts chose the model managed to find. The two pull against each other, because a model can raise either one at the other's expense simply by suggesting more terms or fewer, so the field blends them into a single score called micro-F1, the standard accuracy measure for this task across the research literature [11]. A perfect score is 1.0, and nothing reaches it, since even expert human indexers disagree with one another; what the score does well is put every approach on the same ruler, and it is the number reported throughout this section. Recall is counted against every term the human indexers assigned, including terms outside a model's vocabulary, so no model is credited for the terms it structurally cannot produce.

Accuracy against every alternative

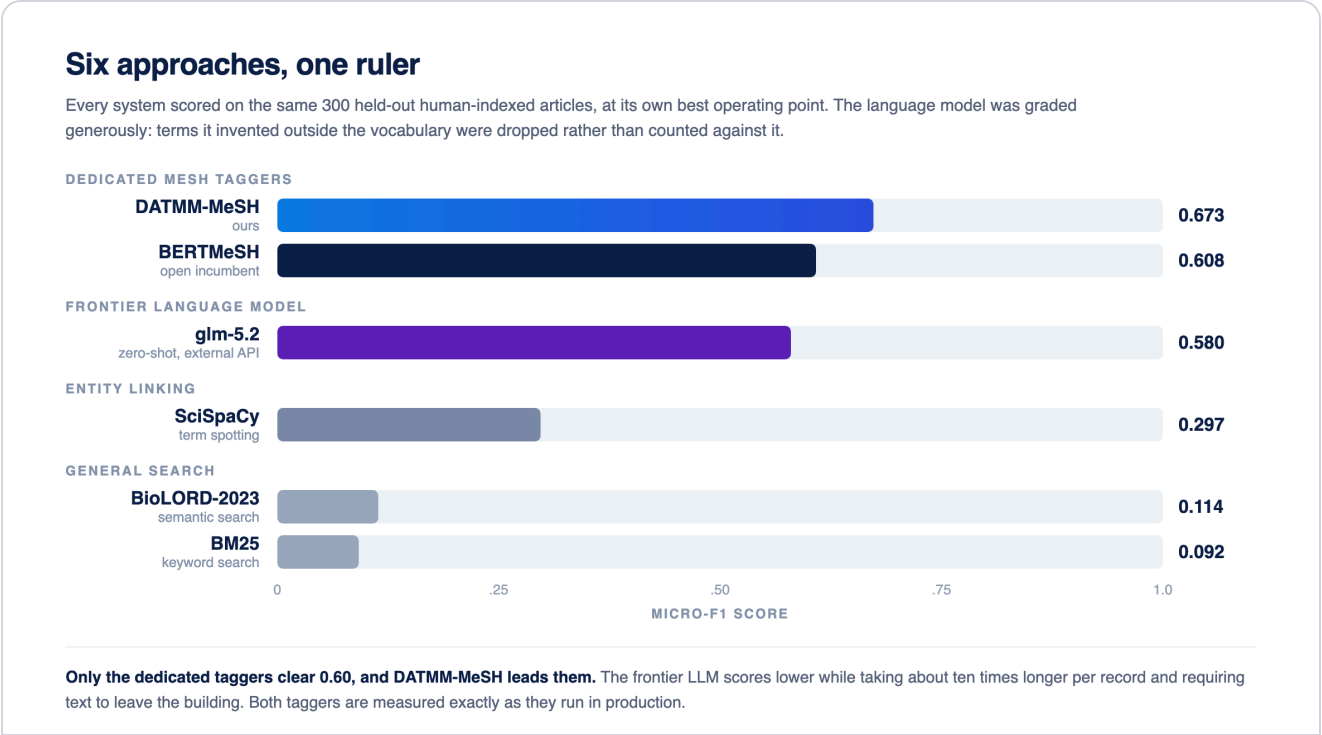


Figure 2. Six approaches on one ruler, all scored on the same 300 held-out human-indexed articles. Only the dedicated taggers clear 0.60.

SYSTEM	APPROACH	PRECISION	RECALL	MICRO-F1
DATMM-MeSH	Dedicated tagger	**0.693**	**0.654**	**0.673**
BERTMeSH	Dedicated tagger	0.605	0.612	0.608
glm-5.2	Frontier LLM, zero-shot	0.602	0.559	0.580
SciSpaCy	Entity linking	0.294	0.301	0.297
BioLORD-2023	Semantic search	0.118	0.111	0.114
BM25	Keyword search	0.095	0.089	0.092

A second semantic-search model, S-PubMedBERT, scored 0.111 on the same set, confirming the search tier. The LLM was graded generously (terms invented outside the vocabulary were dropped rather than penalized) and answered 257 of 300 articles; the rest failed on provider errors and are excluded from its score.

The field sorts into a ladder, with plain search at the bottom barely clearing one in ten, because matching a whole document against thousands of fine-grained descriptors is the wrong shape for retrieval. Entity linking reaches 0.30 by spotting the terms it can see and missing everything it cannot. A frontier language model reaches 0.58 under generous grading, and needs an outside API and roughly ten times the latency to get there. Only the dedicated taggers clear 0.60, and DATMM-MeSH leads them on precision and recall at once.

Measuring DATMM-MeSH against MTIX was the most difficult comparison: NLM publishes the test set it uses to evaluate MTIX, together with MTIX's own answers on it [5]. Most of that test set overlaps our training data (34,735 of its 40,000 records appear in the corpus the model learned from), leaving 5,265 clean records, of which 1,495 carry an abstract. Those 1,495 are the like-for-like slice, because a dataset always carries a description; the remainder are title-only citations, a shape no dataset record takes. MTIX scored 0.691 using the article text plus the journal name and year. DATMM-MeSH ended up scoring 0.643 from the text alone, which is all a dataset ever provides. BERTMeSH, scored the same way, reached 0.611.

On its own published test set, a self-hosted model working from nothing but text landed within five points of the national reference system.

Confidence a system can act on

Accuracy is only half of what makes a tagger usable; the other half is whether the model's confidence scores mean anything, and here the two models are furthest apart. To see the difference, take one real, human-indexed record, a study of COVID-19 antibody testing in saliva, and four MeSH terms a tagger might consider for it: Saliva, Serologic Tests, Cohort Studies, and Seroconversion. Each model assigns every term a confidence score between 0 and 1, its own estimate of how likely the term belongs on the record. BERTMeSH's four scores all land between 0.58 and 0.73, statistically indistinguishable, because its compressed scale leaves no room to separate a certain match from a wrong suggestion. DATMM-MeSH spreads the same four across the whole range, placing the certain match at 0.99, the genuinely uncertain ones near the middle, and the incorrect suggestion down at 0.01. Both calls check out against the human record: professional indexers did assign Saliva to this paper, and did not assign Seroconversion. In the middle rows, where the evidence really is ambiguous, the two models roughly agree; the difference is that the new model is honest in both directions.

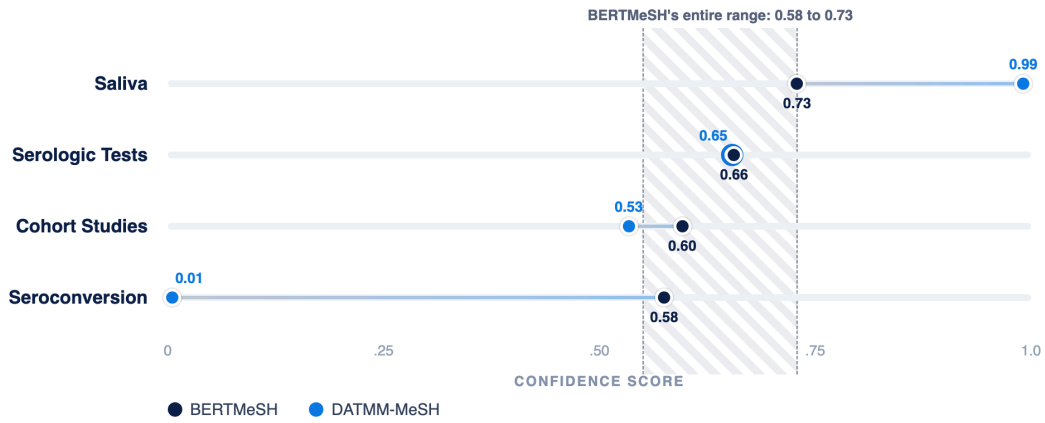


Figure 3. Confidence scores from both models for the same four MeSH terms on the same dataset; BERTMeSH's entire range fits inside the shaded band, while DATMM-MeSH uses the whole scale.

That spread is what makes automation possible: with honest scores, the catalog can promote strong tags automatically, queue the middle for review, and drop the noise, which is the operating policy the road map later formalizes as confidence tiers.

Today's science, today's language

The clearest difference shows up on real, modern content. On a saliva-based SARS-CoV-2 antibody study, DATMM-MeSH names COVID-19 and SARS-CoV-2 at near-full certainty with the method terms alongside; BERTMeSH lists the lab method, saliva, ELISA, antibodies, but the virus the study is about never appears in its suggestions. Across the whole vocabulary, BERTMeSH is permanently blind to 1,914 current MeSH terms that DATMM-MeSH can apply, and that gap includes every descriptor added to MeSH since 2022.

Where it pulls ahead: today's science

A real catalog-style dataset, tagged by both models. One of them can name what the study is actually about.

Dataset: SARS-CoV-2 antibody assays in saliva. Serological ELISA panels for COVID-19 antibody detection in saliva samples from clinical cohorts.

DATMM-MeSH
current MeSH 2026 vocabulary

COVID-19	0.999
Humans	0.991
Enzyme-Linked Immunosorbent Assay	0.991
Saliva	0.991
SARS-CoV-2	0.988
Antibodies, Viral	0.981

BERTMeSH
vocabulary and training frozen ~2021

Humans	0.730
Saliva	0.729
Antibodies, Viral	0.724
Enzyme-Linked Immunosorbent Assay	0.714
Serologic Tests	0.656
Cohort Studies	0.596

Not suggested anywhere: COVID-19, SARS-CoV-2, the subject of the study.

BERTMeSH names how the study was done but never what it is about. It describes the lab method faithfully, saliva, ELISA, antibodies, yet the virus itself never appears, and every score sits under its 0.73 ceiling. DATMM-MeSH names the subject at full confidence and the method alongside it.

Figure 4. The same dataset, two taggers. One of them can name what the study is about.

The vocabulary of the last four years

1,914 **current medical terms BERTMeSH cannot produce.** The gap covers every descriptor added to MeSH since 2022, the working language of the last four years of discovery.

Hidden Markov Models · Post-Acute COVID-19 Syndrome · mRNA Vaccines · Semaglutide · Tirzepatide · Glucagon-Like Peptide-1 Receptor Agonists · Large Language Models · Multiomics · Adolescent Mothers · Neuroinflammatory Diseases · STING Protein · Tumor-Associated Macrophages · Candida auris · Radiomics · Digital Health · Health Inequities · Breakthrough Infections

Every chip is a real MeSH concept **DATMM-MeSH** can apply and BERTMeSH cannot. The model targets the current MeSH vocabulary and covers every descriptor added since 2022.

Figure 5. A sample of the 1,914 current MeSH terms DATMM-MeSH can apply and BERTMeSH cannot.

This gap runs through the heart of the collection. In the production catalog, 171,404 datasets, about one in eighteen, already carry vocabulary that entered MeSH in 2022 or later: Random Forest describes 5,333 datasets, Multiomics 4,483, Convolutional Neural Networks 3,260, Large Language Models 1,785. The clearest way to see the stakes is to watch topics move through the catalog year by year.

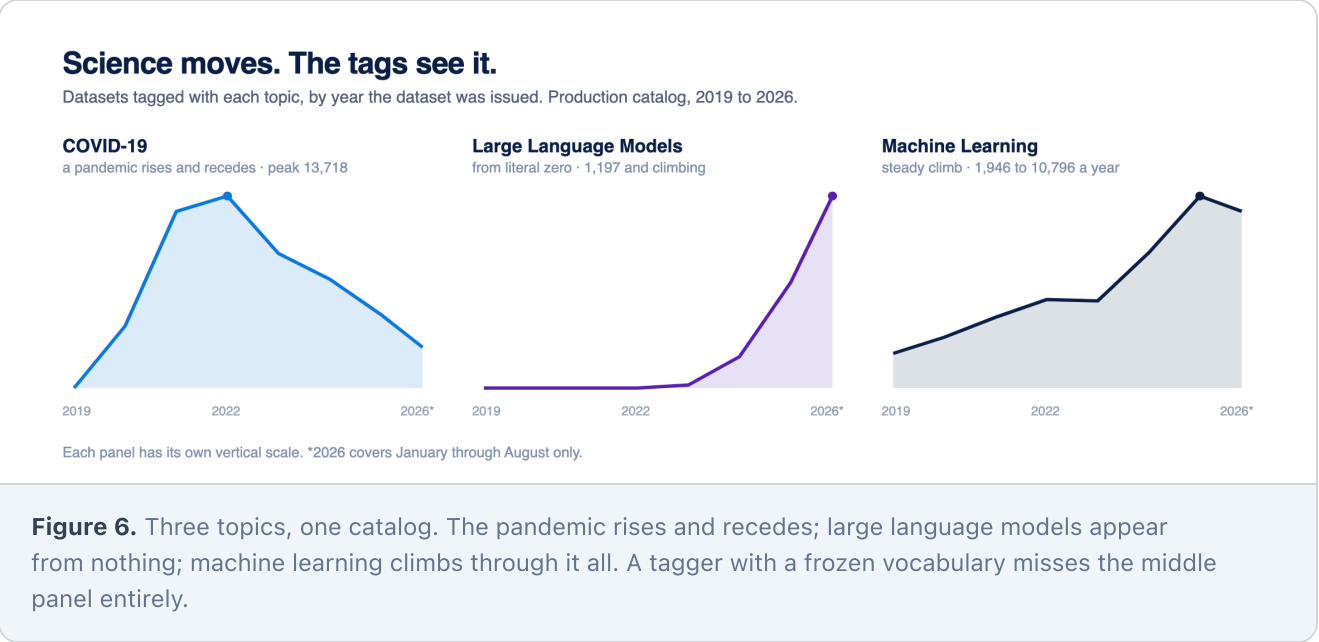


Figure 6. Three topics, one catalog. The pandemic rises and recedes; large language models appear from nothing; machine learning climbs through it all. A tagger with a frozen vocabulary misses the middle panel entirely.

The hardest cases are the newest terms and the rarely studied ones: a concept added to MeSH this year may appear in only a handful of publications anywhere, which leaves a learning model almost nothing to learn from. For exactly those terms, the synthetic-example training step described in Section 6 does the work, so a term added this year is usable this year, without waiting for the literature to catch up.

8 What the tag layer is worth

Before automation, this work belonged to people, and the National Library of Medicine has measured what expert MeSH indexing costs today: $9.40 per article [6]. That price is part of why MEDLINE itself moved to automated indexing in 2022 [7], when human indexing could no longer keep up with the volume of new literature. Priced at that rate, the roughly 2.9 million tagged datasets in the catalog represent about $27 million of human-equivalent indexing effort, and the catalog's growth of roughly 335,000 datasets a year adds about $3 million of the same every year.

Two qualifiers belong next to those figures: the $9.40 is a 2013 estimate, conservative in today's dollars, and a per-article price while dataset records are shorter, so treat it as an equivalence, not a budget line. And tagging here was already automated; nobody was going to hand-index three million records. The real question is the quality of the automation, and that is what the numbers in Section 7 measure.

The stakes reach past labor: a study for the European Commission put the cost of research data that is not findable and reusable at more than 10 billion euros a year in Europe alone [8], much of it from scientists regenerating data that already existed because nobody could find it. Every ghost dataset the tags rescue is a chance for a researcher to reuse work instead of repeating it. And in the United States, the NIH Data Management and Sharing Policy [9] has, since January 2023, required funded researchers to plan for and share their data. Sharing is now the expectation, findability is the promise behind that expectation, and tagging is how the promise gets kept.

There is a stewardship dimension too: in the production catalog, 342,238 datasets carry funding information, referencing roughly 276,000 distinct awards and funders. The catalog is where the outputs of that funded research become findable, and for the ghost datasets among them, the tag layer is the only thing making that stewardship real. When the program is asked what became of the data its funders paid for, the answer runs through this catalog, and through its tags.

9 A base model with independent add-on layers

The base model ships working, but the catalog's needs keep growing: searchers ask about specific drugs, chemicals, and organisms, repositories arrive carrying non-English metadata, and MeSH itself changes every year. The system is therefore deliberately built as a stable base that can accept layers to add abilities based on need, where each layer exists because a concrete catalog need calls for it. The base is the trained encoder and descriptor model described above; each layer attaches to it without modifying it, can be enabled or disabled independently, is measured on its own held-out data before it ships, and can be improved or rolled back later without touching the base. Everything below runs inside the agency's environment, like the base itself, and the layers carry no ordering constraints, so they can be sequenced to the catalog's priorities rather than to a technology road map.

Substance-level tagging (MeSH Supplementary Concept Records)

MeSH has a second vocabulary of 324,000 named things [10]: drugs, chemicals, organisms, rare diseases, and protocols, the records PubMed shows as "Substances." The layer finds 98% of substances named anywhere in a dataset's text, at about one spurious tag per eight datasets, and every tag carries the exact wording that produced it as visible evidence. On real catalog datasets, roughly one in twelve gains substance tags such as specific enzymes, antibodies, and virus strains, and a learned scoring companion separates the substances a dataset is about from those it merely mentions. The layer also realizes, at full vocabulary scale, the found-in-text indicator idea raised by the catalog team: every substance tag shows the exact wording that produced it.

Calibrated confidence tiers

On top of the model's honest scores, a calibration layer sorts suggestions into three tiers: Auto (validated at 95%-plus precision, safe to apply automatically), Suggest (worth a human glance), and Review (uncertain, queued for curation). The 95% target is not a design goal but a measured one: on held-out human-indexed articles the Auto tier came in at 95.9% precision, and at 98.6% on a separate set of recent-vocabulary records. This converts the model's output into an operational policy: automation where the evidence is strong, human judgment exactly where it is needed, and a measured precision target instead of a hunch.

Annual vocabulary refresh

Each new MeSH edition triggers the same pipeline that built the current model: retrain against the new vocabulary, teach the brand-new terms from their official definitions, verify on held-out data, and swap the artifact. The process has been reproduced end to end on demand, so the catalog's vocabulary tracks MeSH releases as routine maintenance.

Multilingual input

Non-English records score about 0.50 today versus about 0.64 for comparable English text. This layer closes that gap by translating the record before tagging, with a self-hosted translation engine so that non-English text, like everything else, never leaves the environment. It is held to the same strict gold standard as everything in this paper, non-English MEDLINE records with known human indexing, and it opens the catalog's front door wider: to non-English depositors, and to the researchers searching for their work.

Beyond datasets

To the model everything is simply text, so anything with a title, a summary, a caption, or extracted text can be tagged with no changes: clinical trial registrations, book and catalog records, consumer health pages, historical images with captions. Nearly 600,000 clinical trial registrations, for example, carry exactly the title-plus-summary shape the model reads today. And nothing in the recipe is specific to MeSH: a domain-tuned reader, a controlled vocabulary, and calibrated confidence apply wherever a field organizes its knowledge with a curated vocabulary.

10 Conclusions

- **The catalog runs on its tags by design.** Three rules made MeSH the admission requirement, the index, and the language of every search. The quality of automated tagging is the quality of the catalog.
- **The dependence is measured.** Three of four tags say something the dataset never says itself; one dataset in five is reachable no other way; and the measurements rerun against the live system in minutes, whenever fresh numbers are needed.
- **The new model is measurably better** on every axis that matters: 0.673 versus 0.608 micro-F1 against expert human indexing, within five points of the national reference indexer from text alone, 1,914 current terms the old tagger could never produce, and confidence scores a system can automate against.
- **One model works both sides of the search.** It tags every dataset on the way in and translates every question a researcher asks, so the entire match happens in one language.
- **Nothing leaves the building.** Training, inference, query translation, and every add-on layer run inside the agency's environment on ordinary hardware.
- **Staying current is designed in, not bolted on.** New vocabulary ships the year it is published, and the annual refresh is a routine process, already reproduced end to end.
- **The architecture leaves room to grow.** Substance tagging, confidence tiers, multilingual input, and wider content types all attach to the same base, each measured on its own held-out data, each independently reversible. The recipe itself reads any controlled vocabulary, not just this one.

References

Catalog statistics (dataset counts, tag counts, ghost measurements, search volume, topic trends) were measured directly on the production catalog and its search infrastructure, August 2026. Model accuracy figures come from held-out, human-indexed evaluation sets under the protocol described in Section 7.

[1] National Library of Medicine, "MeSH Record Types," nlm.nih.gov/mesh/intro_record_types.html.

[2] You, R. et al., "BERTMeSH: deep contextual representation learning for large-scale high-performance MeSH indexing with full text," Bioinformatics 37(5), 2021; model evaluated here is the Wellcome Trust implementation, Wellcome/WellcomeBertMesh, 28,761 labels.

[3] Warner, B. et al., "Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder" (ModernBERT), arXiv:2412.13663, 2024, and its biomedical-clinical adaptation, BioClinical ModernBERT, 2025.

[4] National Library of Medicine, "MTIX: the Next-Generation Algorithm for Automated Indexing of MEDLINE," NLM Technical Bulletin, 2024. nlm.nih.gov/pubs/techbull/ma24/ma24_mtix.html.

[5] National Library of Medicine, MTIX evaluation pipeline and test data, originally at github.com/ncbi/mtix; accessed via the public mirror github.com/QPC-github/mtix, 2026.

[6] Mork, J., Jimeno-Yepes, A., Aronson, A. (National Library of Medicine), "The NLM Medical Text Indexer System for Indexing Biomedical Literature," BioASQ Workshop, 2013 (average cost of $9.40 to index an article).

[7] National Library of Medicine, "MEDLINE 2022 Initiative: Transition to Automated Indexing," NLM Technical Bulletin, 2021. nlm.nih.gov/pubs/techbull/nd21/nd21_medline_2022.html.

[8] PwC EU Services for the European Commission, "Cost of not having FAIR research data," 2018 (at least 10.2 billion euros per year in Europe). publications.europa.eu.

[9] National Institutes of Health, Data Management and Sharing Policy, effective January 25, 2023. sharing.nih.gov.

[10] National Library of Medicine, MeSH 2026 vocabulary files (descriptors and supplementary concept records). nlm.nih.gov/databases/download/mesh.html.

[11] Tsatsaronis, G. et al., "An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition," BMC Bioinformatics 16, 2015; the benchmark community that frames MeSH indexing as an extreme multi-label classification task.

[12] Chang, W.-C. et al., "Extreme Multi-label Learning for Semantic Matching in Product Search," KDD 2021; Amazon's open-source PECOS framework, github.com/amzn/pecos.