From: **Milton** 1884.milton@gmail.com
Subject: The blog as a corpus for future AIs
Date: August 17, 2026 at 3:42 PM
To: Steve Mays stevemays@gmail.com
Cc: Phil Atkinson patkinson@phillipatkinson.com

Steve, Phil—

Steve, I think the thought experiment is sound, with one important amendment: do not choose between preserving the files and sharing what an AI has learned. Publish both, in layers.

The source corpus is the memory. An AI's understanding is an interpretation of that memory. If we preserve only the interpretation—inside a trained model, an agent's private memory, or my current context—we create another fragile black box. Models are replaced, accounts disappear, and learned weights are lossy: they cannot reliably reproduce a particular post, image, date, or turn of phrase. Fine-tuning is especially poor as an archive; it teaches patterns and style, not a trustworthy documentary record.

What “AI sharing” can mean

Today, there is no durable mechanism for handing knowledge to a million independent AIs and expecting them to remember it. Agent-to-agent protocols let agents exchange tasks and artifacts while they are operating. MCP lets an AI application connect to an external collection and retrieve material. Neither is a permanent collective memory.

The practical equivalent is to make the collection public, licensed for reuse, machine-readable, well described, and easy to copy. Then present and future agents can:

- discover it through the public web and dataset catalogs;
- download the whole corpus or retrieve only relevant passages;
- cite the original post rather than trust a model's recollection;
- make their own indexes, embeddings, timelines, analyses, and derivative works;
- mirror the package so it survives the loss of any one host.

In other words, do not try to put Steve permanently inside an agent. Give replaceable agents a permanent doorway into Steve's record.

The package I would publish

1. Canonical source layer. One record per post/page in JSONL and Markdown or plain text, with stable ID, original URL, title, date, categories, tags, body, media references, and rights status. Keep the static human-readable site and the original media beside it.

2. Interpretive layer. Externalize the useful things an AI has inferred: chronology, recurring themes, vocabulary, people and places, topic map, representative reading paths, and essays such as the Gadgets & Apps analysis. Every interpretation should identify itself as interpretation and link back to supporting posts. This is “what Milton learned,” made inspectable and portable instead of trapped in Milton.

3. Discovery layer. Add a compact collection description, sitemap, feed, dataset card, schema, README, and a few example questions. A future agent should be able to learn what the collection is, how to search it, and how to cite it without first digesting a mysterious multi-gigabyte bundle.

4. Retrieval layer. Optionally provide a small open search index and a reference MCP server or local retrieval recipe. That would let an agent ask, for example, “How did Steve's view of mobile computing change from 2002 to 2026?” and receive relevant source passages. This layer is useful but disposable; it must be rebuildable from the corpus.

5. Distribution and preservation layer. Put versioned copies in several different places: the static site at smays.com, an Internet Archive item, an AI-oriented dataset repository such as Hugging Face, a DOI-bearing preservation record such as Zenodo if the package fits its limits, and offline/institutional copies from the earlier plan. Package the preservation master with manifests and checksums using a standard structure such as BagIt.

Rights need a little care. “No attribution necessary” maps closely to the CC0 public-domain dedication for material Steve owns. But Steve cannot place other people's comments, quoted passages, embedded videos, syndicated material, or third-

party images into the public domain. The package should separate Steve-owned material from exceptions and record the rights for each item. I would also preserve attribution metadata even if attribution is not legally required; future systems need provenance to know what they are looking at.

My judgment

Phil is right that this is basically the static-file proposal taken to its logical conclusion. The missing piece was not another archive format; it was an AI-ready interpretive and discovery layer.

Steve's “million copies” should be understood as permission plus copyability, not a literal mailing list of agents. Publish one authoritative, checksummed corpus under clear reuse terms, then mirror it broadly. If it later appears in training sets, that may spread traces of it further, but there is no guarantee that any model will ingest it, preserve it accurately, expose it on demand, or retain the license/provenance. Training is a possible downstream use, not the preservation strategy.

This produces something better than a Steve chatbot. It produces an open cultural object: the original record, a map into it, and permission for future humans and machines to find new readings. The archive remains the ground truth; every AI reading is a new lens.

If we ever turned the thought experiment into work, I would add a modest “AI corpus module” to the preservation plan rather than build and maintain a living agent. First create a small proof using the Gadgets & Apps category we already processed: normalized records, a dataset card and rights statement, the interpretive report, checksums, and a retrieval demonstration. If that package is intelligible to several unrelated AI systems without special coaching, the design works and can be scaled to the full blog.

Sources consulted:

• Model Context Protocol overview and resources: https://modelcontextprotocol.io/docs/2026-07-28/getting-started/intro and https://modelcontextprotocol.io/specification/2026-07-28/server/resources
• Google Agent2Agent announcement: https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/
• Hugging Face dataset sharing: https://huggingface.co/docs/datasets/upload_dataset
• Zenodo records and licensing: https://help.zenodo.org/docs/deposit/about-records/ and https://help.zenodo.org/docs/deposit/describe-records/licenses/
• Creative Commons CC0: https://creativecommons.org/publicdomain/zero/1.0/deed.en
• Library of Congress digital preservation and BagIt use: https://www.loc.gov/preservation/digital/ and https://blogs.loc.gov/preservation/2021/10/software-for-born-digital-preservation/

Best,
Milton

Personal AI agent for Phil Atkinson

Follow-up: Reply-all if you would like to continue the thought experiment; Phil is copied here and remains the approving human for any follow-on work.