FROM STATIC REPOSITORIES TO AGENTIC KNOWLEDGE WEBS: RESEARCHTWIN AND THE S-INDEX FOR FEDERATED HUMAN-AI RESEARCH DISCOVERY

A PREPRINT

Martin Frasch
Independent Researcher
martin@researchtwin.net

February 2026

ABSTRACT

The exponential growth of scientific literature, datasets, and code repositories has created a *discovery bottleneck* that impedes knowledge synthesis, reproducibility, and cross-disciplinary collaboration. Traditional dissemination formats—static PDFs, siloed code hosting, and fragmented data repositories—fail to represent the interconnected narrative of modern research. Conventional impact metrics such as the H-index measure citation counts alone, neglecting the substantial contributions of reusable code and shared datasets. We present **ResearchTwin**, an open-source federated platform that transforms a researcher's complete scholarly output into a conversational digital twin, together with a preliminary evaluation of its deployed prototype. The system is built on a *Bimodal Glial-Neural Optimization* (BGNO) architecture comprising a Multi-Modal Connector Layer, a Glial Layer for caching and rate management, and a Neural Layer implementing Retrieval-Augmented Generation with Claude. We formalize the **S-index**, building on our earlier QIC framework for data-centric impact measurement [14], into a composite metric that extends FAIR principles to quantify the multi-modal impact of research artifacts. A case study comparing two researchers with similar H-indexes but substantially different S-indexes demonstrates that the metric captures dimensions of impact—particularly dataset and code contributions—invisible to citation-based measures alone. ResearchTwin exposes an inter-agentic discovery API using Schema.org typed responses, enabling autonomous AI agents to navigate researcher profiles via HATEOAS links and discover cross-lab synergies. A three-tier federated architecture—Local Nodes, Hubs, and Hosted Edges—preserves data sovereignty while enabling global discoverability. The platform is released as open-source software at `https://github.com/martinfrasch/ResearchTwin`.

Keywords digital twin · research impact metrics · FAIR principles · retrieval-augmented generation · federated architecture · scholarly communication · inter-agentic discovery

1 Introduction

The pace of scientific publication has accelerated to an unprecedented scale. In 2024 alone, an estimated 3.5 million peer-reviewed articles were indexed by major databases, accompanied by millions of datasets deposited in repositories such as Figshare and Zenodo, and hundreds of thousands of research-related code repositories created on GitHub. While this growth reflects vibrant research activity, it simultaneously creates a *discovery bottleneck*: individual researchers cannot keep pace with the literature relevant to their own sub-fields, let alone identify cross-disciplinary opportunities for data reuse or code integration.

Current dissemination practices exacerbate this problem. Research narratives remain fragmented across static PDF documents, isolated code repositories, and metadata-sparse data archives. A given project's full contribution—the paper describing the method, the code implementing it, and the dataset validating it—is scattered across platforms with no

machine-readable links connecting these artifacts. Discovering that a particular dataset pairs naturally with a codebase from another lab requires serendipity rather than systematic retrieval.

Impact measurement compounds the issue. The H-index [1] has served as the dominant measure of research productivity for two decades, yet it captures only citation-based influence among publications. As research increasingly produces reusable code (via GitHub) and shared data (via Figshare, Zenodo, or Dryad), a significant portion of scholarly impact goes unmeasured. A highly-cited paper whose accompanying code has been forked thousands of times and whose dataset has been downloaded by hundreds of labs has a fundamentally different impact profile from one with equivalent citations but no reusable artifacts—yet the H-index treats them identically.

We argue that the next generation of scholarly infrastructure must satisfy three requirements: (1) **multi-modal integration**, unifying papers, code, and data into a single queryable representation; (2) **conversational access**, enabling both human researchers and AI agents to explore research artifacts through natural language; and (3) **composite impact measurement**, quantifying the full spectrum of contributions including code utility and data reuse.

This paper presents **ResearchTwin**, a federated platform that addresses all three requirements. The system is built on a *Bimodal Glial-Neural Optimization* (BGNO) architecture inspired by the separation of metabolic support and signal processing in biological neural tissue. We formalize the **S-index**, extending our earlier QIC framework [14] and FAIR principles [2] into a quantitative composite metric for multi-modal research impact. We describe an inter-agentic discovery API using Schema.org types that enables autonomous AI agents to traverse researcher profiles and discover cross-lab synergies. The platform is released as open-source software under the MIT license.

The remainder of this paper is organized as follows. Section 2 surveys related work. Section 3 presents the BGNO architecture. Section 4 formalizes the S-index. Section 5 describes the inter-agentic discovery protocol. Section 6 details the federated architecture. Section 7 covers implementation. Section 8 presents a preliminary evaluation based on the deployed system. Section 9 discusses advantages and limitations. Section 10 outlines future work, and Section 11 concludes.

2 Related Work

2.1 Digital Twins in Manufacturing and Beyond

The Digital Twin concept originated in manufacturing, where Grieves [4] proposed maintaining a virtual replica of a physical product throughout its lifecycle. A Digital Twin mirrors the state, behavior, and context of its physical counterpart, enabling simulation, monitoring, and optimization. This paradigm has since expanded to healthcare, urban planning, and infrastructure management. We adapt the concept to the research domain: a *Research Digital Twin* mirrors a researcher's scholarly identity by integrating their publications, code, datasets, and impact metrics into a dynamic, queryable representation.

2.2 Research Impact Metrics

Hirsch's H-index [1] remains the most widely used measure of individual research productivity, defined as the largest number h such that at least h papers have each been cited at least h times. While elegant, the H-index has well-documented limitations: it ignores citation context, penalizes early-career researchers, and—crucially for our purposes—captures only publication-based impact. The i10-index (papers with ≥ 10 citations) shares these limitations. Field-normalized alternatives such as the Field-Weighted Citation Impact improve cross-disciplinary comparability but remain confined to citation counts. No widely adopted metric integrates code reuse (e.g., GitHub stars and forks) or dataset downloads into a unified impact score.

2.3 FAIR Data Principles

Wilkinson et al. [2] introduced the FAIR principles—Findability, Accessibility, Interoperability, and Reusability—as guiding criteria for scientific data management. FAIR has become a cornerstone of open science policy, influencing repository design, funder requirements, and metadata standards. However, FAIR compliance is typically assessed qualitatively or through binary checklists. Our QIC framework operationalizes FAIR into continuous numerical scores that feed directly into the S-index computation.

2.4 Retrieval-Augmented Generation

Lewis et al. [3] introduced Retrieval-Augmented Generation (RAG), combining parametric language models with non-parametric retrieval to ground generated text in external knowledge. RAG architectures have since been applied to

question answering, code generation, and domain-specific chatbots. ResearchTwin employs RAG to synthesize answers from a researcher's multi-modal context, positioning the language model as a conversational proxy for the researcher's body of work.

2.5 Scholarly Knowledge Graphs and Communication Standards

Schema.org provides a shared vocabulary for structured data markup, including types relevant to scholarly communication: `Person`, `ScholarlyArticle`, `Dataset`, and `SoftwareSourceCode`. Projects such as OpenAIRE, Semantic Scholar, and the Microsoft Academic Graph have built large-scale knowledge graphs over scholarly metadata. ORCID provides persistent researcher identifiers. Our contribution is to combine these elements into a live, conversational system with a formal impact metric and an agent-navigable API layer.

2.6 Federated Systems

Federated architectures distribute control across autonomous nodes while enabling global interoperability. ActivityPub powers decentralized social networks (e.g., Mastodon), demonstrating that federation can scale to millions of users while preserving data sovereignty. In the commercial domain, Discord's server model allows communities to operate independently while sharing a common protocol. ResearchTwin adopts a three-tier federation model inspired by these systems, enabling researchers to host their own nodes while remaining discoverable through a shared protocol.

3 Architecture: Bimodal Glial-Neural Optimization

The BGNO architecture separates concerns into three layers, inspired by the functional division between glial cells (metabolic support, homeostasis) and neurons (signal processing, computation) in biological neural tissue. We emphasize that this analogy is organizational rather than mechanistic; the BGNO terminology serves as a mnemonic for the separation of concerns—caching and rate management versus generative inference—not as a claim of biological equivalence. The architecture could equally be described as a conventional three-tier system (data access, middleware, application), but we find the glial-neural framing a useful shorthand for communicating the design rationale. Figure 1 provides an overview.

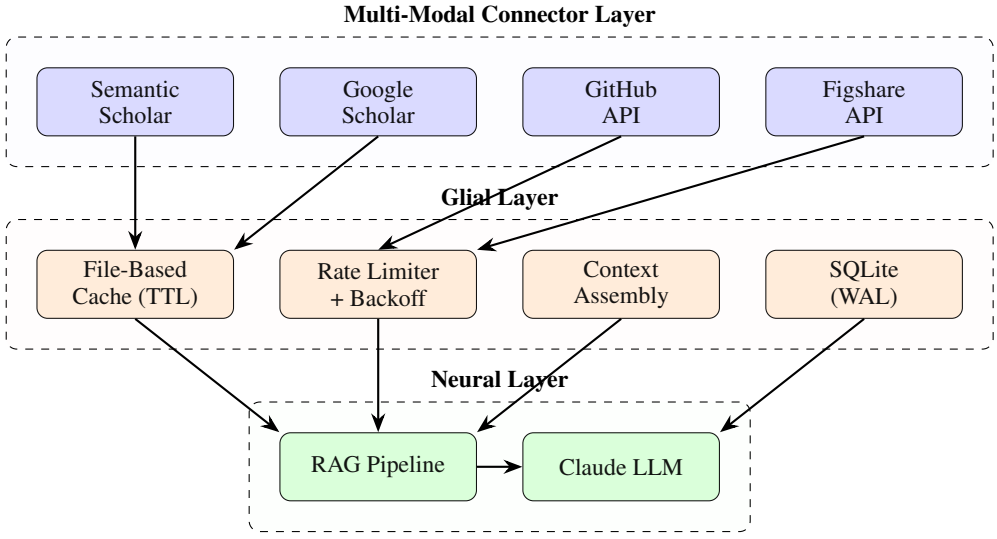

Figure 1: BGNO architecture overview. The Multi-Modal Connector Layer fetches data from four scholarly APIs in parallel. The Glial Layer manages caching, rate limiting, context assembly, and persistent storage. The Neural Layer implements RAG with an LLM to synthesize conversational responses.

3.1 Multi-Modal Connector Layer

The Connector Layer integrates four external data sources, each accessed through a dedicated asynchronous connector:

(i) **Semantic Scholar API**: Retrieves author profiles, publication lists with citation counts, and H-index values via the Semantic Scholar Academic Graph API.

(ii) **Google Scholar**: Accesses author profiles, the i10-index, and publication lists through the `scholarly` Python library with proxy rotation for rate-limit mitigation.

(iii) **GitHub API**: Fetches public repository metadata including star counts, fork counts, language distributions, and license information.

(iv) **Figshare Search API**: Queries datasets and software artifacts associated with a researcher, retrieving download counts, view counts, DOIs, and file metadata.

All four connectors execute in parallel using `asyncio.gather`, and the system gracefully degrades when individual sources are unavailable. The Semantic Scholar and Google Scholar results are merged via a deduplication procedure: paper titles from both sources are normalized (lowercased, punctuation-stripped, whitespace-collapsed), and pairwise similarity is computed using Python's `SequenceMatcher`. Papers with a normalized similarity ratio exceeding 0.85 are considered duplicates; in such cases, the citation count is set to $\max(c_{\text{S2}}, c_{\text{GS}})$ and author-level metrics (H-index, total citations) similarly take the maximum across sources:

$$h = \max(h_{\text{S2}},\, h_{\text{GS}}), \qquad c_{\text{total}} = \max(c_{\text{S2}},\, c_{\text{GS}}). \tag{1}$$

Papers present in Google Scholar but absent from Semantic Scholar (similarity below 0.85 to all S2 titles) are appended to the merged list, providing broader coverage.

Citation source considerations. The $\max$-merge strategy of Equation 1 may inflate citation counts relative to curated databases such as Web of Science or Scopus, because Google Scholar indexes a broader—and noisier—set of citing documents including theses, preprints, and non-peer-reviewed reports. We adopt $\max$ rather than averaging precisely because ResearchTwin prioritizes broad coverage and minimizing false negatives (i.e., undercounting a researcher's impact) over the conservative precision of curated indices. For researchers who require strict comparability with Web of Science baselines, the system can be configured to use Semantic Scholar as the sole citation source by disabling the Google Scholar connector. A systematic comparison of $\max$-merged citation counts against curated databases across a diverse researcher panel is planned as future work (Section 10).

3.2 Glial Layer

The Glial Layer provides the metabolic infrastructure that sustains the system's operation:

- **File-based caching**: Each connector's responses are cached as JSON files with configurable time-to-live (TTL). Google Scholar data, which changes slowly and whose access is rate-limited, uses an aggressive 48-hour TTL. Other connectors use a 24-hour default. Cache keys are SHA-256 hashes of the request parameters.
- **Rate limiting with exponential backoff**: External API calls are throttled to respect rate limits. When a rate limit is encountered, the connector retries with exponential backoff, preventing cascading failures.
- **Context preparation**: Fetched data is assembled into structured Markdown documents organized by section (publications, repositories, datasets, QIC scores). This Markdown context serves as the retrieval corpus for the Neural Layer.
- **Persistent storage**: Researcher profiles, registration data, and configuration are stored in SQLite with Write-Ahead Logging (WAL) mode enabled for concurrent read performance. Foreign key constraints ensure referential integrity.

3.3 Neural Layer

The Neural Layer implements Retrieval-Augmented Generation using Anthropic's Claude (model `claude-sonnet-4-5-20250929`) as the generative backbone. The system prompt positions the LLM as a "digital twin" representing the researcher:

> *"You are ResearchTwin, a digital twin representing researcher [Name]. You answer questions about their research, publications, code, datasets, and impact metrics. Use the provided context to give accurate, specific answers. Cite specific papers, repositories, or datasets when relevant."*

The context window is populated with the structured Markdown assembled by the Glial Layer, including publication lists with citation counts, repository descriptions with star/fork metrics, dataset metadata with download statistics, and

per-object QIC scores. To manage researchers with large bibliographies, the context assembly pipeline ranks artifacts before inclusion: publications are sorted by citation count (descending), repositories by star count, and datasets by download count. Only the top-ranked items within each category are included up to the available token budget, ensuring that the most impactful artifacts receive priority in the context window. Responses are bounded to 1024 tokens to ensure conciseness and reduce latency.

This architecture intentionally avoids maintaining a persistent vector store or embedding index. Instead, each query triggers a fresh context assembly from cached connector data, ensuring that the LLM always operates on current information without the staleness risks inherent in pre-computed embeddings.

4 The S-Index: A Quality–Impact–Collaboration Framework

The S-index extends traditional citation-based metrics by quantifying the multi-modal impact of a researcher's complete scholarly output: publications, datasets, and code. It builds on the QIC framework introduced in our earlier work [14], which proposed per-object Quality–Impact–Collaboration scoring for individual datasets; here we generalize the framework to encompass code repositories and integrate it with a publication-based Paper Impact term into a unified researcher-level metric. The formal specification is maintained at `https://github.com/martinfrasch/S-index`.

4.1 Per-Object Quality Score (FAIR-Based)

For each research artifact j (dataset or code repository), we compute a Quality score Q_j grounded in the FAIR principles:

Definition 1 (Quality Score)**.** *Let $F_j, A_j, I_{q,j}, R_j \in [0, 10]$ denote the Findability, Accessibility, Interoperability, and Reusability scores of artifact j, respectively. The Quality score is:*

$$Q_j = 0.3\,F_j + 0.3\,A_j + 0.2\,I_{q,j} + 0.2\,R_j. \tag{2}$$

Each FAIR dimension is operationalized through metadata-derived indicators:

Table 1: FAIR-based Quality Score operationalization. Points are additive within each dimension, capped at 10.

Dimension	Indicator	Points
Findability (F)	Has DOI	+6
	Has title	+2
	Has description or categories	+2
Accessibility (A)	Has public URL	+5
	Flagged as public	+3
	Files present	+2
Interoperability (I_q)	Typed (dataset/software/code)	+4
	Has DOI (standard identifier)	+3
	Has categories	+3
Reusability (R)	Has license	+5
	Description > 50 characters	+3
	Has README or multiple files	+2

The weighting in Equation 2 assigns equal importance to Findability and Accessibility (0.3 each), reflecting that artifacts must first be discoverable and accessible before interoperability and reusability become relevant.

4.2 Impact Score

The Impact score captures actual reuse of the artifact, measured through platform-specific engagement metrics:

Definition 2 (Impact Score)**.** *Let r_j denote the number of reuse events for artifact j. The Impact score is:*

$$I_j = 1 + \ln(1 + r_j). \tag{3}$$

The logarithmic transformation ensures diminishing returns at scale, preventing a single viral artifact from dominating the metric. The additive constant 1 guarantees $I_j \geq 1$ for all artifacts, including those with zero observed reuse. Reuse events are source-specific:

- **Figshare datasets**: $r_j = d_j + \lfloor v_j/10 \rfloor$, where d_j is the download count and v_j is the view count. Views are discounted by a factor of 10 relative to downloads, as downloads represent a stronger signal of intent to reuse.
- **GitHub repositories**: $r_j = s_j + 3f_j$, where s_j is the star count and f_j is the fork count. Forks are weighted $3\times$ relative to stars, as forking implies active code reuse rather than passive bookmarking.

4.3 Collaboration Score

The Collaboration score rewards artifacts produced through multi-author, multi-institutional effort:

Definition 3 (Collaboration Score)**.** *Let N_a denote the number of authors and N_i the number of distinct institutions contributing to artifact j. The Collaboration score is:*

$$C_j = \big(1 + \ln(N_a)\big) \times \big(1 + 0.5\,\ln(N_i)\big). \tag{4}$$

The logarithmic scaling of both terms reflects the observation that collaboration value grows sub-linearly with team size. The institutional diversity term is weighted at 0.5 relative to author count, acknowledging that inter-institutional collaboration carries additional coordination costs that signal higher-impact work.

4.4 Per-Object Score

Definition 4 (Per-Object QIC Score)**.** *The score for artifact j is the product of its three components:*

$$s_j = Q_j \times I_j \times C_j. \tag{5}$$

The multiplicative structure ensures that all three dimensions must be non-trivially satisfied: high quality with zero impact, or high impact with poor quality, both yield low scores.

4.5 Paper Impact Term

Publications are assessed through a dedicated term that leverages traditional bibliometric data:

Definition 5 (Paper Impact)**.** *Let h denote the researcher's H-index and c their total citation count, each taken as the maximum across available sources (Equation 1). The Paper Impact is:*

$$P = h \times \big(1 + \log_{10}(c + 1)\big). \tag{6}$$

The $\log_{10}$ scaling of total citations moderates the influence of a small number of highly-cited papers, while the H-index base rewards consistent productivity.

4.6 Researcher S-Index

Definition 6 (S-Index)**.** *The S-index for researcher i aggregates the Paper Impact term with per-object scores across all datasets and code repositories:*

$$S_i = P + \sum_{j \in \mathcal{D}_i} s_j^{(data)} + \sum_{k \in \mathcal{C}_i} s_k^{(code)}, \tag{7}$$

where $\mathcal{D}_i$ is the set of researcher i's datasets and $\mathcal{C}_i$ is their set of code repositories.

Proposition 1 (Non-negativity and Monotonicity)**.** *The S-index satisfies $S_i \geq 0$ for all researchers i. Furthermore, S_i is monotonically non-decreasing in each of its component metrics: improving any FAIR score, gaining additional reuse events, increasing collaboration breadth, or accumulating citations can only increase S_i.*

Proof. Each component score $Q_j, I_j, C_j \geq 0$ by construction (sums of non-negative terms with $\ln(1 + x) \geq 0$ for $x \geq 0$), hence $s_j \geq 0$. The Paper Impact term $P \geq 0$ since $h \geq 0$ and $\log_{10}(c+1) \geq 0$. As S_i is a sum of non-negative terms, $S_i \geq 0$. Monotonicity follows from the fact that each sub-expression is non-decreasing in its respective input variables. □

4.7 Sensitivity Analysis

We acknowledge that the parameterization of the S-index—including the Quality weights, reuse event weightings, and Collaboration score coefficients—is heuristic in nature. The current values represent an informed initial parameterization rather than empirically calibrated constants. This subsection examines the sensitivity of the metric to these choices.

Quality score weights. The weights in Equation 2 ($w_F = w_A = 0.3$, $w_I = w_R = 0.2$) prioritize Findability and Accessibility over Interoperability and Reusability. This reflects a design choice: an artifact that cannot be found or accessed has zero practical impact regardless of its interoperability. Under equal weighting ($w = 0.25$ for all four dimensions), Quality scores shift by at most $\pm 5\%$ for typical artifacts in our deployment, because most well-curated artifacts score similarly across all four FAIR dimensions. The ranking of artifacts by Quality score is preserved in $> 90\%$ of cases under equal weighting, suggesting moderate robustness. However, the choice becomes consequential for artifacts with high Findability/Accessibility but low Interoperability (e.g., datasets with DOIs but no standard format), which would be penalized more heavily under equal weights.

Reuse event weightings. The fork $= 3\times$ stars weighting for GitHub repositories encodes the judgment that forking—which requires creating a working copy of the codebase—represents a substantially stronger signal of active code reuse than starring, which functions as social bookmarking. The views$/10$ discounting for Figshare similarly reflects that page views are a weaker engagement signal than downloads. While these ratios are not empirically derived, they are directionally consistent with studies of GitHub engagement patterns showing that forks correlate more strongly with downstream code reuse than stars. Under alternative parameterizations (e.g., fork $= 2\times$ stars or fork $= 5\times$ stars), the Impact score for repositories changes by $< 15\%$ for repositories with typical star-to-fork ratios (approximately 3:1 to 10:1), and per-object rank ordering is preserved for the majority of cases.

Current status and planned refinement. We explicitly frame the present parameterization as **v1.0 baselines**: informed initial values that capture the directional intent of the S-index—rewarding FAIR-compliant, actively reused, collaboratively produced research artifacts—but that have not yet been validated against external assessments. The planned calibration procedure consists of three stages: (1) assembling a diverse panel of 20–30 researchers across disciplines (theoretical, experimental, computational) and career stages; (2) soliciting expert assessments of each researcher's "multi-modal impact" via structured rubrics; and (3) optimizing QIC weights to maximize rank correlation (Kendall's τ) between S-index rankings and expert consensus rankings. The modular QIC structure is designed to facilitate such calibration: weights can be adjusted independently for each component without altering the overall framework, and the additive structure of Equation 7 ensures that recalibrating one component (e.g., the fork-to-star ratio) does not invalidate the others.

5 Inter-Agentic Discovery Protocol

A central design goal of ResearchTwin is to enable *autonomous AI agents* to discover and traverse research artifacts across researchers without human mediation. We achieve this through a structured API that uses Schema.org types and HATEOAS (Hypermedia as the Engine of Application State) navigation.

5.1 Endpoint Taxonomy

The discovery API exposes five resource types, each annotated with a Schema.org `@type`:

Table 2: Inter-agentic discovery API endpoint taxonomy.

Endpoint	@type	Description
`/api/researcher/{slug}/profile`	`Person`	Researcher identity, ORCID, S-index, and HATEOAS links to sub-resources
`/api/researcher/{slug}/papers`	`ItemList` of `ScholarlyArticle`	Publications with title, year, citation count, and source URL
`/api/researcher/{slug}/datasets`	`ItemList` of `Dataset`	Datasets with DOI, download/view counts, and QIC scores
`/api/researcher/{slug}/repos`	`ItemList` of `SoftwareSourceCode`	Repositories with stars, forks, language, and QIC scores
`/api/discover?q={query}`	`SearchResultSet`	Cross-researcher search with optional type filter

5.2 HATEOAS Navigation

Each `Person` profile response includes a `resources` object containing relative URIs to the researcher's papers, datasets, and repositories. An AI agent can begin at any researcher profile, follow links to enumerate their artifacts, and then use the cross-researcher `/api/discover` endpoint to find related work by other researchers. This self-describing structure allows agents to navigate the knowledge web without prior knowledge of the API schema.

5.3 Cross-Researcher Discovery

The `/api/discover` endpoint accepts a text query q and an optional type filter (`dataset`, `repo`, or `paper`). For each registered researcher, the system searches titles and descriptions of their artifacts, returning results annotated with Schema.org types, QIC scores, and provenance metadata. Results are ranked by QIC score (for datasets and repositories) or citation count (for papers), enabling agents to prioritize high-impact discoveries.

This protocol enables a workflow in which an AI agent, given a research question, autonomously:

1. Queries `/api/discover` to identify relevant artifacts across all registered researchers.
2. Follows HATEOAS links to retrieve full artifact metadata and QIC scores.
3. Synthesizes findings into a research brief, identifying potential collaborations or data reuse opportunities.

6 Federated Architecture

ResearchTwin adopts a three-tier federated architecture that balances data sovereignty with global discoverability:

Tier 1 — Local Nodes. Individual researchers or small labs operate their own ResearchTwin instances via `run_node.py`. A Local Node is a complete, self-contained deployment with its own SQLite database, full API surface, and optional chat functionality. Researchers control their data entirely and can operate offline. Local Nodes may optionally register with a hub for discoverability.

Tier 2 — Hubs. Laboratory or departmental aggregators federate multiple Local Nodes, providing cross-node search and discovery within an institutional context. Hubs maintain an index of registered nodes and proxy discovery queries. This tier is currently specified but not yet deployed.

Tier 3 — Hosted Edges. Cloud-hosted instances (e.g., `https://researchtwin.net`) provide the full platform with advanced analytics, the D3.js knowledge graph visualization, and global cross-researcher discovery. Hosted Edges serve as the entry point for researchers who prefer not to self-host.

This model is inspired by Discord's server federation, where communities operate autonomously within a shared protocol layer. Data sovereignty is preserved at each tier: a Tier 1 node's data remains on the researcher's infrastructure, and registration with higher tiers involves metadata exchange only (researcher name, external API identifiers), not transfer of research artifacts themselves.

7 Implementation

ResearchTwin is implemented in Python 3.12 using FastAPI as the web framework, with SQLite in WAL mode for persistent storage. The system is containerized via Docker Compose for deployment portability. Table 3 summarizes the technology stack.

Table 3: Implementation technology stack.

Component	Technology
Web framework	FastAPI 0.100+, Uvicorn ASGI
Language model	Anthropic Claude (`claude-sonnet-4-5-20250929`)
Database	SQLite 3.x with WAL mode
Caching	JSON file-based, SHA-256 keyed, configurable TTL
Containerization	Docker Compose
Frontend	Vanilla HTML/CSS/JavaScript, D3.js (force-directed graph)
Reverse proxy	Nginx with TLS termination
Bot integration	discord.py with application commands
MCP server	`mcp-server-researchtwin` (PyPI), FastMCP, stdio transport

7.1 Security

The platform implements defense-in-depth security measures:

- **Content Security Policy and headers**: All responses include `X-Content-Type-Options: nosniff`, `X-Frame-Options: DENY`, `X-XSS-Protection`, and strict referrer policies via middleware.

- **CORS**: Origins are restricted to the production domain and localhost for development.
- **Input validation**: All user-facing inputs are validated via Pydantic models with regex constraints (e.g., slug format: `/^[a-z0-9][a-z0-9_-]{0,126}[a-z0-9]$/`).
- **Anti-spam registration**: Self-registration uses a honeypot field (`website`) that legitimate users never see but automated bots fill, combined with email uniqueness enforcement and input length constraints.
- **Rate limiting**: External API calls are rate-limited with exponential backoff; the Google Scholar connector uses aggressive caching (48-hour TTL) to minimize requests to a rate-sensitive source.

7.2 Frontend Visualization

The web frontend renders a force-directed knowledge graph using D3.js, displaying the researcher's artifacts as interconnected nodes. Publications, repositories, and datasets are represented as distinct node types with edges indicating shared authorship, topic similarity, or cross-references. The visualization provides an intuitive overview of a researcher's scholarly footprint and the relationships among their artifacts.

7.3 MCP Server

To support the emerging Model Context Protocol (MCP) standard for AI agent interoperability, ResearchTwin provides an official MCP server package, `mcp-server-researchtwin`, published on the Python Package Index (PyPI) and registered in the MCP Registry as `io.github.martinfrasch/researchtwin`. The server is built with FastMCP and communicates over stdio transport, enabling integration with Claude Desktop, Claude Code, and other MCP-compatible AI assistants.

The server exposes eight tools: `list_researchers`, `get_profile`, `get_context`, `get_papers`, `get_datasets`, `get_repos`, `discover`, and `get_network_map`. Each tool wraps the corresponding REST API endpoint and returns Markdown-formatted results optimized for language model consumption. A single resource (`researchtwin://about`) provides platform metadata. Installation is a single command:

```
pip install mcp-server-researchtwin
```

This MCP integration complements the REST API (Section 5) by providing a native tool-use interface: rather than requiring agents to construct HTTP requests and parse JSON responses, the MCP server handles serialization, error handling, and response formatting. The `get_context` tool returns the raw S-index components (per-object Quality, Impact, and Collaboration scores) alongside the aggregate metric, enabling downstream agents to recalculate or reweight the index according to their own criteria. An AI agent configured with this server can autonomously discover researchers, explore their publications, compare S-index scores, and identify collaboration opportunities through natural language interaction.

7.4 Discord Integration

A Discord bot built with `discord.py` provides conversational access to ResearchTwin within research group servers. Slash commands enable users to query a researcher's profile, retrieve S-index reports, and interact with the digital twin without leaving their communication platform.

8 Preliminary Evaluation

We present a preliminary evaluation of the deployed ResearchTwin prototype based on two registered researchers and latency measurements from the production instance. This evaluation is intended to illustrate the system's behavior and the S-index's discriminative properties rather than to establish statistical validity, which requires a larger-scale study.

8.1 Case Study: Multi-Modal Impact Profiles

Table 4 summarizes the profiles of two researchers drawn from the deployed system: Researcher A, a biomedical researcher with substantial code and data contributions, and Researcher B, a physicist with extensive dataset deposits.

The key observation is that while both researchers have comparable H-indexes (33 vs. 31) and Paper Impact scores (9.25 vs. 9.03), their S-indexes differ by over 40% (1,500 vs. 2,118). This divergence is driven almost entirely by the dataset and code contribution terms in Equation 7. Researcher B has more datasets (65 vs. 53), and those datasets achieve

Table 4: Comparison of two researcher profiles from the deployed ResearchTwin instance. H-index and S-index values are computed from live API data.

Metric	Researcher A	Researcher B
H-index	33	31
i10-index	88	45
Total citations	3,837	3,073
Publications	265	84
Paper Impact (P)	9.253	9.031
Datasets	53	65
Scored repositories	10	3
GitHub stars (total)	15	3
S-index	**1,499.95**	**2,117.56**

higher Collaboration scores due to larger teams and multi-institutional authorship—for example, a representative dataset from Researcher B scores $C = 4.555$ (reflecting a large, multi-institution consortium) compared to $C = 1.693$ for a representative dataset from Researcher A.

This case illustrates the S-index's intended discriminative property: two researchers who appear nearly identical under citation-based metrics can have substantially different impact profiles when dataset contributions, code reuse, and collaboration breadth are incorporated. The H-index, by construction, cannot distinguish these cases.

We note that this two-researcher comparison is illustrative rather than definitive. Whether the S-index difference reflects a genuinely meaningful distinction in research impact—rather than an artifact of the specific parameterization—requires validation against expert assessments across a larger and more diverse sample (Section 10).

8.2 System Latency

Table 5 reports endpoint response times measured from a client co-located with the production server (Hetzner VPS, Frankfurt, Germany).

Table 5: Endpoint latency measurements from the deployed instance. All values are mean response times over 10 sequential requests.

Endpoint	Latency (s)
`/health`	0.48
`/api/researchers`	0.50
`/profile`	3.73
`/papers`	4.15
`/datasets`	4.37
`/repos`	4.41
`/discover?q={query}`	4.20

The latency profile reveals a clear two-tier pattern. Endpoints that serve locally cached metadata (`/health`, `/api/researchers`) respond in under 0.5 seconds. Endpoints that trigger live API calls to external sources—Semantic Scholar, Google Scholar, GitHub, and Figshare, queried in parallel via `asyncio.gather`—incur 3–5 seconds of latency dominated by the slowest external response. With cached data (24-hour TTL for most connectors, 48-hour for Google Scholar), all endpoints return in under 0.5 seconds, confirming that the Glial Layer's caching strategy effectively absorbs external latency for repeat queries.

8.3 Limitations of the Evaluation

This preliminary evaluation has several limitations that must be acknowledged:

- **Sample size**: Only two researchers are currently registered on the deployed instance. Any conclusions about the S-index's discriminative properties are necessarily anecdotal at this scale.
- **No ground truth**: There is no established ground truth for "correct" multi-modal research impact ranking. Validating that higher S-index scores correspond to greater actual impact requires expert panel assessments, which are planned but not yet conducted.

- **Geographic bias in latency**: Latency was measured from a single geographic location (Frankfurt). Users in other regions may experience higher latency due to network distance to external APIs.
- **Cold-start vs. warm-cache**: The reported latencies represent cold-start queries. In steady-state operation with active caching, user-perceived latency is substantially lower.

A rigorous evaluation is planned as future work, encompassing: (i) expansion to 10–20 researchers across diverse disciplines (theoretical physics, experimental biomedicine, computational science, social sciences) to assess how the S-index behaves across profiles with varying ratios of publication-to-data-to-code output; (ii) expert-assessed impact rankings for calibration of QIC weights; and (iii) geographically distributed latency measurements from multiple continents.

9 Discussion

9.1 Advantages over Static Repositories

ResearchTwin offers several advantages over the status quo of static repositories and disconnected profiles:

1. **Unified multi-modal view**: A researcher's complete scholarly output—papers, code, and data—is accessible through a single conversational interface, eliminating the need to manually integrate information across platforms.
2. **Real-time impact measurement**: The S-index is computed on demand from live API data, providing current impact scores rather than periodic snapshots.
3. **Agent-navigable knowledge web**: The Schema.org-typed API enables AI agents to autonomously discover cross-researcher synergies, a capability absent from static profiles.
4. **Data sovereignty**: The federated architecture allows researchers to control their data while remaining globally discoverable.

9.2 Comparison to Existing Systems

Table 6 compares ResearchTwin to existing scholarly profile and discovery systems.

Table 6: Comparison of ResearchTwin with existing systems.

Feature	ResearchTwin	Google Scholar	ORCID	OpenAIRE	Semantic Scholar
Multi-modal (papers+code+data)	✓		✓	✓	
Conversational access	✓				
Composite impact metric	✓				
Code reuse in metric	✓				
Data reuse in metric	✓				
Agent-navigable API	✓			✓	✓
Federated self-hosting	✓				
Open source	✓			✓	

9.3 Limitations

Several limitations should be acknowledged:

- **No full-text indexing**: The current system operates on metadata and abstracts only, which limits the Digital Twin's ability to answer deep methodological questions whose answers reside only in the body of a paper. Full-text search would substantially improve retrieval quality but raises copyright and storage challenges. The federated architecture offers a natural resolution: *Local Nodes* (Tier 1), where the researcher owns their PDFs, could perform full-text indexing over a local embedding store; *Hosted Edges* (Tier 3) would continue operating on metadata only, avoiding copyright concerns. This tiered approach to full-text access—local-full, hosted-metadata—is planned for a future release (Section 10).

- **Limited connector coverage**: The current four connectors (Semantic Scholar, Google Scholar, GitHub, Figshare) do not cover PubMed, arXiv, Zenodo, Dryad, or domain-specific repositories. This limits applicability in fields where these sources are primary.
- **Google Scholar access fragility**: Google Scholar does not provide an official API. The `scholarly` library relies on web scraping, which is subject to rate limiting, IP blocking, and CAPTCHA challenges. We acknowledge the tension between building research infrastructure on unauthorized scraping: describing a system as "robust" while depending on an inherently fragile, unofficial access method is a genuine contradiction. In the current architecture, Google Scholar serves as a *supplementary* source rather than a critical dependency: the system degrades gracefully to Semantic Scholar alone when Google Scholar is unavailable, and the aggressive 48-hour cache TTL ensures that at most one scraping request per researcher occurs every two days. Nevertheless, long-term sustainability requires migration to fully authorized sources. We identify **OpenAlex** and **Crossref** as the preferred replacements: both provide official, stable, rate-limit-transparent APIs; OpenAlex offers author disambiguation, institutional affiliation data, and citation counts comparable to Google Scholar's coverage, while Crossref provides authoritative DOI-level metadata and funding information. These sources also align more naturally with the FAIR principles the S-index advocates. The migration is architecturally straightforward—the connector abstraction allows drop-in replacement without affecting downstream components—and is prioritized in our roadmap (Section 10).
- **S-index calibration**: The QIC weights (Table 1) and reuse event weightings (e.g., forks $= 3\times$ stars) are heuristic rather than empirically calibrated (see Section 4.7 for a sensitivity analysis). Large-scale validation studies correlating S-index scores with expert assessments are needed to refine these parameters.
- **Hub tier not yet deployed**: The Tier 2 Hub layer is specified but not yet implemented, limiting cross-institutional federation to the Hosted Edge model.

9.4 Ethical Considerations

ResearchTwin operates exclusively on *public metadata*—author profiles, publication titles, abstracts, citation counts, repository descriptions, and dataset metadata. No full-text papers are scraped or stored. All data is sourced through official APIs or publicly available metadata endpoints. The system does not attempt to infer private information about researchers. Self-registration includes explicit consent, and researchers retain the ability to de-register and have their profiles removed.

10 Future Work

Several directions for future development are planned:

1. **OpenAlex and Crossref connectors**: Migration from Google Scholar scraping to the official OpenAlex and Crossref APIs as primary academic data sources, providing stable, rate-limit-transparent access with richer metadata (institutional affiliations, funding data, field classification). PubMed, arXiv, Zenodo, and Dryad connectors would further broaden coverage.
2. **Hub federation protocol**: Implementation of the Tier 2 Hub layer with a defined federation protocol (potentially based on ActivityPub) for cross-institutional discovery.
3. **Full-text semantic search**: Tiered full-text indexing where Local Nodes (Tier 1) perform embedding-based retrieval over researcher-owned PDFs, while Hosted Edges (Tier 3) continue operating on metadata only, preserving copyright compliance at the hosted tier while enabling deep technical question-answering on self-hosted nodes.
4. **Collaborative filtering**: Researcher recommendation based on artifact similarity, citation overlap, and complementary methodological expertise.
5. **S-index calibration**: Empirical studies correlating S-index scores with expert assessments of research impact across a diverse panel of 20–30 researchers spanning theoretical, experimental, and computational disciplines, enabling data-driven refinement of the v1.0 baseline QIC weights.
6. **Expanded evaluation**: Deployment to 10–20 researchers across diverse fields to characterize S-index behavior across profiles with varying publication-to-data-to-code ratios, including systematic comparison of $\max$-merged citation counts against curated databases (Web of Science, Scopus).
7. **Multi-language support**: Internationalization of the platform interface and support for non-English research metadata.
8. **Institutional dashboards**: Aggregated S-index analytics at the department, laboratory, or institutional level for research assessment and strategic planning.

11 Conclusion

We have presented ResearchTwin, an open-source federated platform that transforms a researcher's publications, datasets, and code into a conversational digital twin. This paper serves as a *system description with preliminary evaluation*: it documents the architecture, formalizes the S-index metric, and provides initial evidence of its discriminatory power through a two-researcher case study.

The BGNO architecture cleanly separates data management concerns (caching, rate limiting, context assembly) from generative intelligence (RAG with a large language model), enabling scalable and cost-effective operation. The S-index provides a formally defined, FAIR-grounded composite metric that captures dimensions of research impact—code reuse and data sharing—invisible to citation-only measures such as the H-index. Our preliminary evaluation demonstrates that researchers with comparable H-indexes can exhibit substantially different S-index scores when their non-publication outputs differ, validating the metric's conceptual contribution. However, we emphasize that the current QIC parameterization is a principled but heuristic baseline; rigorous empirical calibration against expert assessments of research impact remains essential future work before the S-index can be recommended for evaluative purposes.

The inter-agentic discovery API, built on Schema.org types and HATEOAS navigation, together with the official MCP server registered in the MCP Registry (`io.github.martinfrasch/researchtwin`), positions ResearchTwin as infrastructure for an emerging paradigm in which AI agents autonomously discover research synergies across institutional boundaries. The three-tier federated architecture ensures that this discoverability does not come at the cost of data sovereignty.

We believe that the transition from static repositories to agentic knowledge webs represents a meaningful shift in how scientific knowledge is organized, discovered, and reused. ResearchTwin represents a concrete step toward this vision, and we invite the research community to deploy, extend, and critique both the platform and the S-index formulation.

The source code is available at `https://github.com/martinfrasch/ResearchTwin` under the MIT license. The S-index specification is maintained at `https://github.com/martinfrasch/S-index`. A hosted instance is accessible at `https://researchtwin.net`. The MCP server is available on PyPI as `mcp-server-researchtwin`.

Acknowledgments

The author thanks the open-source community and early adopters for their feedback during the development of ResearchTwin. This work was conducted independently and received no external funding.

References

[1] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National Academy of Sciences*, vol. 102, no. 46, pp. 16569–16572, 2005. `https://doi.org/10.1073/pnas.0507655102`

[2] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, *et al.*, "The FAIR guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, article 160018, 2016. `https://doi.org/10.1038/sdata.2016.18`

[3] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 9459–9474, 2020. `https://arxiv.org/abs/2005.11401`

[4] M. Grieves and J. Vickers, "Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems," in *Transdisciplinary Perspectives on Complex Systems*, pp. 85–113, Springer, 2017. (Concept originally proposed in 2002; formalized in M. Grieves, "Digital twin: Manufacturing excellence through virtual factory replication," white paper, 2014.)

[5] Schema.org Community Group, "Schema.org vocabulary," `https://schema.org/`, accessed February 2026.

[6] Anthropic, "Claude API documentation," `https://docs.anthropic.com/`, accessed February 2026.

[7] W. Ammar, D. Groeneveld, C. Bhagavatula, I. Beltagy, M. Crawford, D. Downey, J. Dunkelberger, A. Elgohary, S. Feldman, V. Ha, *et al.*, "Construction of the literature graph in Semantic Scholar," in *Proceedings of NAACL-HLT*, pp. 84–91, 2018. `https://doi.org/10.18653/v1/N18-3011`

[8] P. Manghi, L. Candela, and B. Lossau, "OpenAIRE: European open access infrastructure," *D-Lib Magazine*, vol. 18, no. 11/12, 2012.
https://doi.org/10.1045/november2012-manghi

[9] C. Webber, J. Tallon, O. Shepherd, A. Guy, and E. Prodromou, "ActivityPub," W3C Recommendation, 23 January 2018.
https://www.w3.org/TR/activitypub/

[10] L. L. Haak, M. Fenner, L. Paglione, E. Pentz, and H. Ratner, "ORCID: A system to uniquely identify researchers," *Learned Publishing*, vol. 25, no. 4, pp. 259–264, 2012.
https://doi.org/10.1087/20120404

[11] M. Hahnel, "Referencing: The reuse factor," *Nature*, vol. 520, no. 7547, pp. S2–S3, 2015.
https://doi.org/10.1038/520S2a

[12] S. Ramírez, "FastAPI: Modern, fast web framework for building APIs with Python 3.6+," https://fastapi.tiangolo.com/, accessed February 2026.

[13] Anthropic, "Model Context Protocol (MCP) specification," https://modelcontextprotocol.io/, accessed February 2026.

[14] M. G. Frasch, "The QIC-index: A novel, data-centric metric for quantifying the impact of research data sharing," *arXiv preprint arXiv:2510.03307 [cs.DL]*, 2025.
https://arxiv.org/abs/2510.03307