AgentFAIR: A Multi-Agent Collaborative Framework for FAIRness Evaluation of Geospatial Datasets

Ming Chen*†
The University of Melbourne
Melbourne, Victoria, Australia

Pranav Pai*
The University of Melbourne
Melbourne, Victoria, Australia

Abstract

Geospatial datasets support applications from urban planning to climate modeling, but their compliance with the FAIR principles (*Findability*, *Accessibility*, *Interoperability*, and *Reusability*) is difficult to assess consistently. Existing evaluators use different rubrics and evidence sources and can fail on JavaScript-rendered pages or repository-specific identifier schemes. In a diagnostic study of 50 datasets from 10 repositories, the standard deviation of normalized scores across available tools averages 15.0 percentage points and reaches 30.3 points on one dataset. Because these normalized outputs are not equivalent measurements, we use them to characterize disagreement and failure modes rather than comparative accuracy.

We present AgentFAIR, a multi-agent framework that combines structured metadata extraction with 13 sub-principle-specific large language model (LLM) evaluators. Each evaluator returns a 0–3 maturity score, cited evidence, and recommendations; a critic applies evidence and consistency checks and can request targeted re-evaluation. In the 50-dataset sample, mean *Findability*, *Accessibility*, *Interoperability*, and *Reusability* scores are 79.7%, 70.4%, 45.3%, and 72.0%, respectively. Rank correlations between AgentFAIR and four baseline tools range from $\rho = 0.31$ to 0.61, with the FAIR-enough comparison not statistically significant. On a 10-dataset repeated-run subset, sub-principle agreement averages 89% (reported standard deviation: 3 percentage points); a critic ablation reports 71% agreement without the critic. A preliminary 15-dataset expert study yields Fleiss' $\kappa = 0.71$ and 82% sub-principle alignment with expert consensus. The measured API cost is approximately USD $0.054 per dataset. These results support the framework's auditability and feasibility, while the limited benchmark, incomplete component ablations, and single-model-family validation constrain claims about accuracy and generalization.

CCS Concepts

• **Information systems** → **Data management systems**; Data provenance; • **Computing methodologies** → *Artificial intelligence*.

Keywords

Geospatial datasets, FAIR principles, Multi-agent systems, Large language models, Metadata quality

1 Introduction

The FAIR data principles—*Findability*, *Accessibility*, *Interoperability*, and *Reusability*—provide widely adopted guidance for making digital research outputs more reusable and machine-actionable [24]. In geospatial science, FAIRness (FAIR compliance; *not* algorithmic fairness) is especially consequential: spatial layers, rasters, and spatio-temporal products are routinely integrated across agencies, jurisdictions, and scientific communities. However, many geospatial datasets remain only partially FAIR, due to inconsistent metadata practices, heterogeneous formats (e.g., Shapefile, GeoTIFF, NetCDF), and incomplete spatial reference information [18, 21]. According to a European Commission cost-benefit analysis, non-FAIR practices are estimated to cost the EU research economy € 10.2 Billion annually through duplicated effort and failed reuse [19].

Existing automated FAIR assessment tools such as F-UJI [5] and FAIR-Checker [7] provide important mechanistic checks (e.g., persistent identifiers, licenses, basic schema markup). However, deterministic approaches face limitations that are amplified for geospatial resources: (i) **Dynamic metadata exposure:** Repository landing pages increasingly rely on JavaScript rendering and API-driven content, which breaks traditional crawlers and causes partial or missing evidence collection; (ii) **Repository heterogeneity:** Identifier and metadata practices vary widely (DOI, Handle, internal URIs; different schema profiles), thus hard-coded assumptions (e.g., DataCite-centric resolution patterns) can lead to brittle failures; (iii) **Limited semantic judgment:** Rule-based checks cannot reliably assess qualitative and domain-specific aspects of metadata, such as whether a coordinate reference system is valid, whether access constraints are optional or mandatory, or whether terminology reflects community standards; (iv) **Low score comparability:** Recent analyses show that different tools operationalize FAIR differently, leading to substantial score variance and making stewardship decisions tool-dependent [3, 11].

Furthermore, geospatial FAIRness analysis depends on domain conventions that general-purpose tools often treat as opaque strings. Examples include ISO (International Organisation for Standardisation) 19115 metadata profiles [10], coordinate reference system identifiers (EPSG codes), and OGC discovery and access conventions (e.g., WMS, WFS, and STAC endpoints). Because many geospatial vocabularies are not registered in common ontology registries, semantic validators may systematically underestimate *Interoperability*, even when datasets follow community practice.

To address these challenges, we propose AgentFAIR, a multi-agent collaborative framework that combines: (i) deterministic extraction of machine-readable metadata and identifiers, and (ii) sub-principle-specific LLM agents that perform evidence-grounded semantic evaluation. As shown in Figure 1, the system orchestrates a pipeline that processes dataset landing pages, extracts metadata, evaluates each FAIR sub-principle with specialized agents, and applies critic-based quality control via running critic agents to ensure evidence-backed scoring and consistency across principles. When

*Both authors contributed equally to this work.
†Corresponding author: Ming Chen (ming.chen6@unimelb.edu.au)

Preprint,
2026.

evidence is insufficient or internally inconsistent, the critic agent triggers a targeted re-evaluation (*Retry* mechanism) with refined instructions, improving robustness while preserving traceability. Overall, AgentFAIR evaluates all 13 FAIR sub-principles (F1–F4, A1.1–A2, I1–I3, R1.1–R1.3) on a 0–3 maturity scale and outputs a Markdown report, JSON assessment, SQLite evidence store, and local execution traces. Remote LangSmith tracing is optional and disabled by default.

We evaluate AgentFAIR on 50 geospatial datasets across 10 repositories, including Zenodo[1], Dryad[2], Harvard Dataverse[3], PANGAEA[4], NOAA NCEI[5], and domain-specific portals. We compare against four widely used FAIR evaluators (F-UJI [5], FAIR-Checker [7], FAIRshake [4], FAIR-enough [6]) and analyze (i) overall FAIR maturity patterns, (ii) cross-tool agreement and failure modes, (iii) cost and runtime characteristics, and (iv) consistency across repeated evaluations.

To the best of our knowledge, AgentFAIR is the first multi-agent LLM-based framework specifically designed for FAIR compliance evaluation. It integrates critic-informed feedback loops with domain-specific geospatial reasoning. The main contributions of this work are as follows:

- **The first LLM-driven multi-agent FAIR evaluator.** We propose AgentFAIR, a collaborative multi-agent framework with critic-based quality control and explicit awareness of geospatial standards. To our knowledge, it is the first system to combine LLM reasoning with domain-specific FAIR assessment.
- **A transparent and fine-grained scoring protocol.** We operationalize a 0–3 maturity rubric for each FAIR sub-principle and explicitly map geospatial indicators to the assessment of *Interoperability* and *Reusability*.
- **A cross-repository empirical study.** We evaluate 50 geospatial datasets across 10 repositories. The sample reveals strong *Findability* (79.7%) and weak *Interoperability* (45.3%), but is not intended to establish long-tail generalization.
- **A diagnostic analysis of evaluator disagreement and feasibility.** We quantify cross-tool disagreement (mean per-dataset standard deviation 15.0 points; maximum 30.3), document repository-specific failure modes, and measure an average API cost of approximately USD $0.054 per dataset. These comparisons do not treat heterogeneous tool scores as equivalent accuracy measures.

The rest of the paper is organized as follows. Section 2 reviews related work on FAIR assessment tools, geospatial data quality, and multi-agent LLM architectures. Section 3 formalizes the task definition and scoring protocol. Section 4 details the AgentFAIR framework design and implementation. Section 5 presents the evaluation and its limitations, and Section 6 concludes the paper. The software artifact, dataset list, prompt structure, and LangGraph orchestration details are provided in the Appendix. The source code is available at https://doi.org/10.5281/zenodo.18529560 and https://github.com/MingCHEN-Github/AgentFAIR.

[1]https://zenodo.org (Open repository for research outputs with DOI assignment.)
[2]https://datadryad.org (Curated repository for datasets supporting published scientific research.)
[3]https://dataverse.harvard.edu (General-purpose repository for sharing and citing research data.)
[4]https://www.pangaea.de (Earth and environmental science repository for georeferenced datasets.)
[5]https://www.ncei.noaa.gov (Global archive of climate, oceanic, and environmental data.)

Table 1: Comparison of FAIR assessment systems

System	LLM	Geo-aware	Explain	Trace	Cross-P	Coverage
F-UJI [5]	✗	✗	✓	✓	✗	✓
FAIRshake [4]	✗	✗	✓	✓	✗	✗
FAIR-Checker [7]	✗	✗	✓	✓	✓	✗
FAIR-enough [6]	✗	✗	✓	✓	✗	✓
AgentFAIR	✓	✓	✓	✓	✓	✓

LLM=LLM-based reasoning in the evaluation loop; Geo-aware=explicit recognition of geospatial standards; Explain=evidence-backed scoring rationales; Trace=persisted audit trails; Cross-P=cross-principle consistency checks; Coverage=ability to support assessment across all 13 sub-principles used here.

2 Related Work

2.1 Automated FAIR Assessment Tools

Automated FAIR evaluators typically operationalize the principles as predefined, machine-checkable indicators (e.g., presence of PIDs, protocol support, and machine-readable licensing) and aggregate the results into per-principle scores. F-UJI implements the FAIRsFAIR assessment metrics and collects evidence from dataset landing pages and exposed metadata, complemented by queries to external registries [5]. FAIR-Checker emphasizes semantic validation by leveraging knowledge graphs and SHACL/SPARQL constraints [7]. FAIRshake provides community-curated rubrics tailored to different digital resource types [4], whereas FAIR-enough offers a broader suite of largely binary tests, which can be conservative when evidence is partial or indirectly expressed [6]. A recurring challenge is that tools differ in their operational definitions, evidence sources, and aggregation rules, leading to scores that are not directly comparable across evaluators [3].

Table 1 summarizes representative capabilities of these systems relative to AgentFAIR. For baselines, *coverage* denotes whether a tool—or our mapping of its outputs—supports assessment across all 13 FAIR sub-principles used in this paper.

2.2 FAIRness and Quality for Geospatial Data

FAIRness assessment for geospatial datasets intersects with classic spatial data quality concerns, including spatial reference consistency, resolution/scale, temporal extent, and adherence to domain standards. Metadata standards such as ISO 19115 [10] enable rich descriptions of spatial resources, but repositories vary substantially in how completely these metadata elements are captured and exposed in machine-readable form. As a result, automated checkers that rely on generic schemas, brittle pattern matching, or incomplete registry coverage may miss geospatial signals (e.g., CRS (Coordinate Reference System) identifiers or geospatial service endpoints), which can systematically depress *Interoperability*—and, downstream, *Reusability*—scores for spatial datasets.

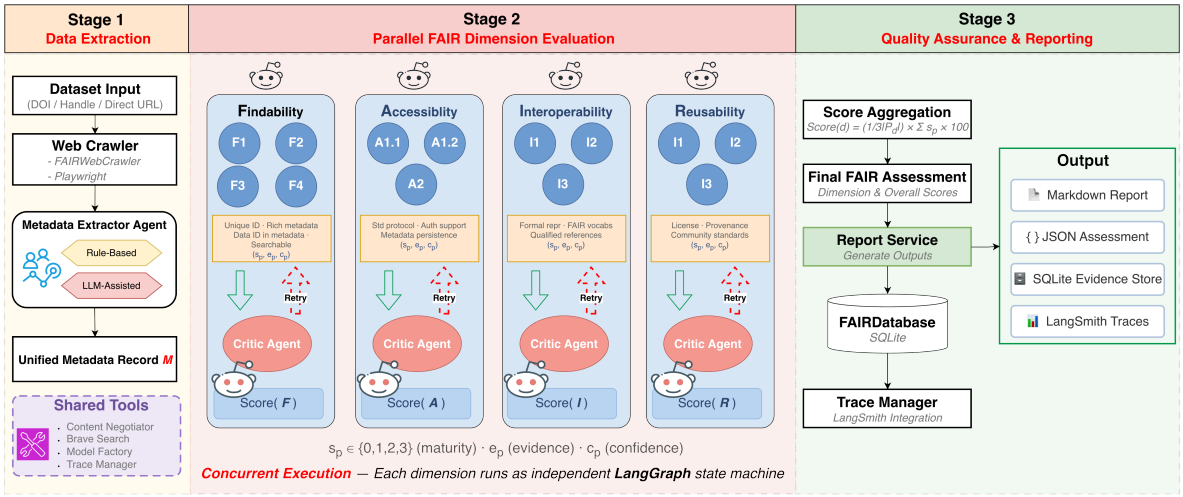

Figure 1: The proposed AgentFAIR framework for multi-agent collaborative FAIRness evaluation.

2.3 LLM Agent for Data Management and Evaluation

Large language models (LLMs) have shown promise for extracting, normalizing, and enriching metadata from heterogeneous sources [2, 30]. Recent systems combine LLM reasoning with workflow orchestration to automate data cleaning, validation, and quality assessment [9, 13, 16]. However, applying LLMs to FAIR evaluation introduces additional requirements—consistent rubrics, evidence-backed justifications, and auditable trails—and systematic multi-agent designs that explicitly address these requirements, especially under geospatial standards, remain limited.

2.4 Multi-agent LLM Architectures

Multi-agent frameworks decompose complex tasks into coordinated subtasks executed by specialized agents, often improving robustness through structured interaction and self-critique [23, 28]. AutoGen [27] introduced conversational multi-agent patterns enabling flexible agent interactions, while MetaGPT [8] demonstrated role-based collaboration for software engineering tasks. Benchmarks such as AgentBench [14] and TheAgentCompany [29] evaluate agent capabilities on consequential real-world tasks, providing evaluation ideas that we adapt for FAIR compliance assessment. For geospatial domains, foundation model applications [15] highlight both the potential and the domain-specific challenges of applying LLMs to spatial information understanding. Building on these architectural patterns, AgentFAIR combines critic-informed feedback loops with evidence-grounded prompting and explicit geospatial standard awareness to support transparent, auditable FAIR scoring.

3 Task Definition and Scoring Protocol

3.1 FAIR Evaluation Task and Principle Set

We formulate FAIR evaluation as an evidence-grounded auditing task [24]. Given a dataset landing-page URL u (the human-facing entry point for a dataset) and optional repository context m (e.g., repository type, known metadata schemas, and API conventions), AgentFAIR produces three structured outputs: (i) per-sub-principle maturity scores $S = \{s_p\}_{p\in\mathcal{P}}$, (ii) per-sub-principle evidence bundles $E = \{e_p\}_{p\in\mathcal{P}}$ that justify each score with provenance, and (iii) actionable recommendations $R = \{r_p\}_{p\in\mathcal{P}}$ describing what is missing and how to improve compliance.

Evidence model. For each sub-principle $p \in \mathcal{P}$, the evidence bundle e_p is a structured record that includes (a) extracted metadata fields (e.g., PID, license, protocol), (b) supporting excerpts/snippets (e.g., HTML fragments or metadata blocks), and (c) provenance pointers (e.g., source URL and the extraction route such as landing page, embedded JSON-LD, downloadable metadata files, or repository API). This design is critical for explainability and auditability: non-zero scores must cite observable, traceable evidence.

Evaluated principles. We evaluate the 13 FAIR sub-principles from the FAIR Guiding Principles [24]:

- ***Findability***: F1 (globally unique persistent identifier), F2 (rich metadata), F3 (identifier in metadata), F4 (indexed in a searchable resource).
- ***Accessibility***: A1.1 (retrieval protocol), A1.2 (authentication/authorization A2 (metadata persistence).
- ***Interoperability***: I1 (formal knowledge representation), I2 (FAIR vocabularies), I3 (qualified references).
- ***Reusability***: R1.1 (license), R1.2 (provenance), R1.3 (community standards).

Role of repository context m. When available, m provides lightweight repository-specific hints (e.g., typical locations of machine-readable metadata, common field names, or standardized API endpoints). This improves recall of evidence without changing the rubric: AgentFAIR still assigns scores strictly according to the same maturity criteria, and all inferences must be backed by retrievable evidence recorded in E.

Dim	ID	FAIR Principle	Evaluation Criteria (evidence)	Maturity Levels (0→3)	Geo Indicators
***Findability* (F)**					
F	**F1**	Globally unique persistent identifier	PID presence; HTTPS resolution; machine-readable embedding; content negotiation	None → Unrecognized/fails → Resolves HTTPS → +Metadata+ContentNeg	DOI, NASA CMR, STAC IDs
F	**F2**	Rich metadata description	Core citation (creator, title, date, publisher) + descriptive (abstract, keywords, coverage)	None → Minimal (title) → Core citation → Rich descriptive	ISO 19115, spatial extent
F	**F3**	Metadata includes data identifier	Data access links in metadata; multiple access methods; machine-readable distribution	None → Present/broken → Resolvable → Multiple clear links	OPeNDAP, WMS/WFS, STAC
F	**F4**	Registered in searchable resource	Catalog registration (DataCite); Schema.org markup; API-verified indexing	None → Meta tags only → Catalog OR structured → Both	CKAN API, CSW, STAC
***Accessibility* (A)**					
A	**A1.1**	Standardized retrieval protocol	Protocol openness (HTTP/S, FTP); IETF/ISO standardization; programmatic access	Proprietary → Restricted → Standard+limits → Open+free	OGC, STAC API, THREDDS
A	**A1.2**	Authentication or authorization support	Auth clarity (OAuth, API keys); automation feasibility; documentation quality	None/broken → Complex → Standard auth → Open/clear	Earthdata Login, ESA Hubs
A	**A2**	Metadata preserved without data	Preservation via PIDs (DOI/ARK); repository certification (CoreTrustSeal); policy	None → Basic repo → Institutional → Certified+policy	PANGAEA, SDI catalogs
***Interoperability* (I)**					
I	**I1**	Formal knowledge representation	W3C/ISO standards (JSON-LD, RDF); recognized vocabulary usage; semantic richness	None → Limited → Standard+vocab → Rich semantic	GeoSPARQL, DCAT, Schema.org
I	**I2**	FAIR vocabularies used	Vocabulary FAIRness: registry presence (LOV, BioPortal, BARTOC); documentation	None → 1 registry → 2+ registries → 3+ with docs	GCMD, CF conventions, SWEET
I	**I3**	Qualified references to other data	Explicit relationships (derivedFrom, cites); scientific context; PIDs for refs	None → Vague "see also" → Explicit typed → Machine-readable	DataCite relations, PROV-O
***Reusability* (R)**					
R	**R1.1**	Clear data usage license	Standard license (SPDX/CC/ODbL); machine-readable embedding in metadata	None → Free text → Standard ID → Machine-readable	OGL, CC-BY, ODbL
R	**R1.2**	Detailed provenance	Processing lineage (ISO/PROV-O); creator IDs (ORCID); methodology; versions	None → Minimal → Clear derivation → Structured PROV-O	ISO 19115 lineage, PROV-O
R	**R1.3**	Domain community standards	Metadata standard compliance; controlled vocabularies; validated format specs	None → Terminology → Recognizable std → Multiple formal	ISO 19139, CF-NetCDF, STAC

Table 2: Operationalization of the 13 FAIR sub-principles used in this paper: evidence criteria, 0–3 maturity levels, and representative geospatial indicators.

3.2 Rubric-Based Maturity Scoring and Aggregation

Maturity rubric. Each sub-principle p receives an ordinal maturity score $s_p \in \{0, 1, 2, 3\}$:

- **0 (non-compliant):** no supporting evidence is found, or evidence is contradictory/unverifiable from accessible resources.
- **1 (partial):** some relevant information exists but is incomplete, implicit, or primarily human-readable (weak machine actionability).
- **2 (substantial):** clear evidence exists in a machine-actionable form for the core requirement, but with notable gaps (e.g., missing qualifiers, incomplete standard alignment, or partial exposure across distributions).
- **3 (full):** strong, unambiguous evidence satisfies the requirement in a standard, machine-actionable way with sufficient detail for automated reuse and validation.

Table 2 specifies the concrete indicators and decision rules used for each sub-principle. This explicit rubric is applied uniformly across repositories to reduce tool-specific interpretation variance [26].

Dimension and overall aggregation. Let $\mathcal{D} = \{\mathrm{F}, \mathrm{A}, \mathrm{I}, \mathrm{R}\}$ denote the four FAIR dimensions, and let $\mathcal{P}_d \subset \mathcal{P}$ be the proper subset of sub-principles within dimension d. We report scores as percentages for comparability with existing evaluators:

$$\mathrm{Score}(d) = \frac{1}{3|\mathcal{P}_d|} \sum_{p \in \mathcal{P}_d} s_p \times 100, \qquad \mathrm{Score}(\mathrm{FAIR}) = \frac{1}{3|\mathcal{P}|} \sum_{p \in \mathcal{P}} s_p \times 100.$$

Because s_p is evidence-based, missing machine-readable signals directly translate into lower maturity and generate corresponding recommendations r_p.

Geospatial operationalization. A common failure mode of generic FAIR checkers for spatial datasets is missing domain-specific signals that are essential for *Interoperability* and *Reusability*. To mitigate this, AgentFAIR explicitly searches for and reasons over geospatial indicators when scoring I1–I3 and R1.3, including (non-exhaustively): coordinate reference system identifiers, spatial and temporal extent descriptors, ISO 19115-1 elements, and OGC-related service endpoints and encodings across different geospatial semantics. These indicators are not "bonus points"; rather, they serve as admissible evidence under the rubric when they satisfy the corresponding sub-principle requirements (e.g., standard machine-interpretable representations for I1 and community standards for R1.3).

Evaluation objectives supported by the protocol. The above task formulation and rubric are designed to support: (i) **Consistency** (controlled variability under repeated runs despite LLM stochasticity), (ii) **Explainability** (scores justified by explicit evidence and decision rules), (iii) **Auditability** (persisted evidence and traces enabling post-hoc inspection), and (iv) **Domain awareness** (recognition of geospatial standards and metadata constructs).

4 Proposed AgentFAIR Framework

Figure 1 summarizes AgentFAIR as a three-stage, evidence-first pipeline for evaluating all 13 FAIR sub-principles on the 0–3 maturity rubric defined in Section 3. Given a dataset input (DOI/Handle/direct URL), the system executes: **(Stage 1) Data extraction → (Stage 2) Parallel FAIR dimension evaluation → (Stage 3) Quality assurance and reporting**.

Stage 1 (Data extraction). A Playwright-based crawler resolves redirects and renders JavaScript-heavy landing pages, then a hybrid extractor constructs a unified metadata record M that merges structured fields (e.g., JSON-LD/Schema.org, HTML meta tags, DCAT-like fragments) with provenance-linked text snippets. Agents must cite this record when assigning a non-zero score. The critic checks for missing evidence and conflicts with deterministic observations; this is a structural check on the returned record, not a claim that prompting alone guarantees factual correctness.

Stage 2 (Parallel dimension evaluation). AgentFAIR decomposes evaluation into four concurrent LangGraph state machines, one per FAIR dimension (F/A/I/R), each internally evaluating its sub-principles in parallel. For each sub-principle $p \in \mathcal{P}$, the corresponding agent outputs a tuple (s_p, e_p, c_p) where $s_p \in \{0, 1, 2, 3\}$ is the maturity score, e_p is an evidence bundle (fields/snippets + provenance pointers), and $c_p \in [0, 1]$ is a confidence estimate used for quality control.

Stage 3 (Quality assurance & reporting). A critic-driven feedback loop checks evidence sufficiency and cross-sub-principle consistency before scores are aggregated into dimension and overall percentages (Section 3.2). The report service then materializes auditable artifacts: a human-readable Markdown report, a machine-readable JSON assessment, a SQLite evidence store, and local execution traces for post-hoc inspection. LangSmith export is an explicit opt-in because prompts and evidence may contain sensitive content.

Algorithm 1 summarizes the control flow, including caching of deterministic signals to stabilize retries.

Algorithm 1: AgentFAIR evaluation pipeline

Input : Dataset URL u, repository context m
Output : Sub-principle scores S, evidence E, recommendations R

$html \leftarrow \textsc{Crawl}(u)$;
/* Playwright for JS rendering */
$M \leftarrow \textsc{ExtractMetadata}(html, m)$;
/* Hybrid extraction & enrichment */
$G \leftarrow \textsc{DeterministicChecks}(u, M)$;
/* Resolver, protocol, content-negotiation, and registry signals */
foreach *sub-principle* $p \in \mathcal{P}$ **do**
  $(s_p, e_p, c_p) \leftarrow \textsc{Agent}_p(M, G)$;
  /* Evidence-backed evaluation */
  if $c_p < \theta_p$ ***or*** $\textit{HardCheckFails}(p, s_p, e_p, G)$ **then**
    $(s_p, e_p, c_p) \leftarrow \textsc{CriticRepair}(p, M, G, s_p, e_p)$;
    /* Retry with feedback */
  end
end
$S \leftarrow \textsc{Aggregate}(\{s_p\})$;
/* Dimension and overall scores */
$R \leftarrow \textsc{GenerateRecommendations}(S, E)$;
return S, E, R

4.1 LangGraph orchestration

We use LangGraph [12] to represent the multi-agent workflow as a directed state machine. Each FAIR dimension is expressed as a modular subgraph with nodes for crawling, extraction, sub-principle evaluation, critic review, and report synthesis. Conditional edges implement retry logic when confidence is low or evidence is missing (Appendix C). Checkpointing supports recovery, and trace integration enables inspection of agent decisions and intermediate state.

4.2 Hybrid metadata extraction

The Extractor stage combines: (1) **rule-based parsing** of structured metadata (e.g., Schema.org/JSON-LD, DCAT-like fragments, HTML meta tags) and identifier patterns (DOI/Handle/ORCID); and (2) **LLM-assisted enrichment** for normalization, deduplication, and resolving ambiguous natural-language fields (e.g., license statements and access conditions). The output is a unified internal metadata record M that retains both structured fields and provenance links to their sources (e.g., landing page text, embedded JSON-LD).

4.3 Sub-principle agents

AgentFAIR comprises 13 specialized agents, one per FAIR sub-principle. Each agent receives the extracted metadata record M and returns: (i) a maturity score $s_p \in \{0, 1, 2, 3\}$, (ii) evidence e_p consisting of specific supporting snippets or fields, and (iii) a confidence score c_p used for quality control. Each system prompt separates six elements: (i) the sub-principle definition and scope; (ii) the 0–3 decision rubric; (iii) admissible deterministic and extracted evidence; (iv) geospatial indicators such as EPSG, ISO 19115, and OGC terms where relevant; (v) conservative instructions not to infer absent evidence; and (vi) a structured output schema requiring score, rationale, evidence, confidence, and recommendations. Retry prompts append the critic's identified contradiction or missing-evidence requirement while retaining the original rubric and cached deterministic observations.

4.4 Critic agent and consistency checks

The critic agent applies two forms of quality control: **evidence sufficiency** (is the score supported by extracted evidence?) and **cross-sub-principle consistency** (do scores contradict each other given the evidence?). For example, high A1.1 (open protocol) paired with low A1.2 (authentication clarity) can indicate ambiguous access conditions; the critic triggers a targeted re-evaluation with refined instructions. For Findability, seven deterministic checks include identifier presence, resolver success for maturity ≥ 2, content-negotiation evidence for maturity 3, score-range validity, and evidence presence for non-zero scores. A retry can be triggered by a failed hard check or by confidence below 0.5 (0.4 for I1 in the reported configuration), subject to a retry budget. Evaluator and critic use the same GPT-4o-mini backend with different roles and prompts; therefore the critic does not remove shared-model bias. All retries and revisions are logged as part of the audit trail.

4.5 Implementation notes

AgentFAIR uses Playwright for JavaScript-rendered access and Extruct, lxml, and rdflib for structured metadata parsing. Geospatial standards and CRS identifiers are recognized from exposed metadata; the evaluated implementation does not perform GDAL-based inspection of downloaded data files. SQLite persists evidence and local traces, and FastAPI exposes a REST API for local integration. The primary experiment uses GPT-4o-mini with temperature 0.1. The non-zero setting was intended to permit limited lexical variation in rationales and evidence text; a temperature 0.0 versus 0.1 ablation was not conducted, so no performance benefit is claimed.

4.6 Security boundary

Dataset pages and metadata are untrusted input. The implementation restricts evaluation targets to public HTTP(S) destinations, uses hardened XML parsing, and requires structured outputs, but it has not been evaluated against prompt injection or poisoned metadata. In particular, text retrieved from a page can still influence an LLM evaluator. The released API is therefore intended for local, single-user operation; adversarial prompt-injection testing, redirect-by-redirect destination validation, authentication, rate limiting, and per-request credential isolation are required before public deployment.

5 Evaluation

Experiments were conducted on an Apple M2 Pro system (32 GB RAM, macOS 15). The framework evaluated all 13 FAIR sub-principles using the 0–3 maturity rubric in Section 3. LLM calls used `gpt-4o-mini` (temperature 0.1). Playwright handled browser-based crawling (including JavaScript-rendered pages); SQLite stored evidence traces and audit logs.

5.1 Datasets

We evaluated 50 geospatial datasets spanning 10 repositories to capture heterogeneous metadata practices and access patterns. Our benchmark prioritizes *cross-repository diversity* over single-repository depth. It samples heterogeneous repository and domain conditions but is too small to establish long-tail generalization. **Repository distribution:** Zenodo (10), Dryad (8), PANGAEA (7), Harvard Dataverse (6), NOAA NCEI (6), ScienceBase (4), Figshare (4), AURIN (2), EarthData (2), NASA SEDAC (1). **Domains:** Biodiversity/Ecology (12), Marine/Geoscience (10), Climate/Hydrology (8), Urban/Land Classification (6), Demographics (4), Disaster/Emissions (3), Other (7). Datasets are referenced as D1–D50 in Tables 3 and 5; the full mapping appears in Appendix E.

5.2 Baselines and score normalization

We compare against four public FAIR evaluators: F-UJI, FAIR-Checker, FAIRshake, and FAIR-enough. These tools expose different rubrics and output schemas (binary tests, rubric scores, or per-dimension scores), so we normalize each tool's *overall* dataset score to a 0–100 scale. For tools with multiple tests, the normalized score is the fraction of tests passed; for tools with per-dimension scores, we use the mean across dimensions. For AgentFAIR, the overall score is computed from its 13 sub-principle maturities (0–3), scaled to 0–100. This normalization supports only dataset-level disagreement and failure-mode analysis; it does *not* make the underlying metrics equivalent and is not used as evidence that one evaluator is more accurate.

Baseline tools also differ in metadata acquisition strategy. F-UJI relies primarily on API/identifier lookups, while AgentFAIR uses browser-based crawling via Playwright. Browser-based extraction better reflects end-user access to JavaScript-rendered metadata, but introduces different failure modes (e.g., navigation timeouts on complex pages vs. API endpoint unavailability).

Run selection and table consistency. Because AgentFAIR is LLM-based, it exhibits run-to-run variability (Section 5.7). We report one representative run in Table 3; the AgentFAIR row in Table 5 is the direct aggregation of those same 13 scores for each dataset.

5.3 Results: FAIR maturity patterns

Table 3 reports maturity scores for each dataset (columns) across the 13 FAIR sub-principles (rows). Aggregating these sub-principle scores, *Findability* performs best on average (79.7%), followed by *Reusability* (72.0%) and *Accessibility* (70.4%); *Interoperability* is consistently weakest (45.3%).

Sub-principle bottlenecks. F2 (rich metadata) is the strongest sub-principle (mean 2.94/3; 47/50 datasets achieve full maturity), reflecting widespread use of descriptive landing pages and repository-generated metadata. In contrast, A2 (metadata should remain accessible even when the data are no longer available) remains weak (mean 1.14/3; 0/50 at full maturity), reflecting limited explicit preservation policies, certification, and machine-actionable tombstone guarantees. The sub-principles I2 and I3 from *Interoperability* are the dominant bottlenecks (means 1.00/3 and 1.20/3; 0/50 at full maturity), primarily because (i) widely used geospatial vocabularies (CRS identifiers, ISO elements, OGC conventions) are not consistently registered in the ontology registries expected by semantic validators, and (ii) explicit, typed links to related datasets and resources are rarely exposed.

Table 3: FAIR maturity matrix for 50 datasets across 13 sub-principles using the 0–3 rubric (representative run). Columns D1–D50 map to the dataset list in Appendix E.

Sub-principle	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15	D16	D17	D18	D19	D20	D21	D22	D23	D24	D25	D26	D27	D28	D29	D30	D31	D32	D33	D34	D35	D36	D37	D38	D39	D40	D41	D42	D43	D44	D45	D46	D47	D48	D49	D50
F1	1	0	3	2	3	3	3	3	3	1	3	3	3	1	1	1	3	3	3	3	1	3	3	0	1	3	3	3	3	1	3	3	1	3	1	2	3	2	3	3	2	3	1	1	3	3	1	1	1	1
F2	3	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2	3	3	3	3
F3	3	0	3	3	3	3	2	0	2	1	1	3	3	1	1	0	2	3	3	3	1	2	2	0	2	3	2	3	3	1	3	3	1	3	1	1	3	3	1	1	1	2	2	1	3	0	0	0	1	2
F4	3	3	3	3	3	3	2	3	3	3	2	3	3	3	2	3	3	2	3	3	2	3	3	1	3	2	2	3	3	2	2	2	3	3	2	3	3	3	3	2	3	3	3	3	2	2	3	2	2	3
A1.1	3	3	3	3	3	3	3	3	3	1	3	3	3	3	1	3	3	3	3	3	3	3	1	3	3	3	3	3	3	3	3	2	2	1	1	3	3	3	3	3	3	3	3	3	2	3	1	1	1	3
A1.2	3	3	3	3	3	3	3	3	3	1	3	3	3	3	1	3	2	3	3	3	3	3	1	3	1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	1	1	1	1	3	3	1	1	1
A2	1	0	1	1	1	1	1	0	1	1	2	1	1	1	1	0	2	2	2	1	1	1	2	0	2	1	1	1	1	0	2	2	2	2	1	1	1	2	1	1	1	1	1	2	2	0	0	1	2	1
I1	2	0	2	3	2	2	2	0	2	2	2	2	3	2	2	0	3	2	2	2	2	2	3	0	3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	3	3	0	0	0	2	2
I2	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	0	2	1	1	1	1	1	2	0	2	1	1	1	1	1	1	1	1	2	2	2	2	2	1	1	1	1	1	1	0	0	0	1	1
I3	2	0	1	2	2	2	2	0	1	1	1	2	2	2	1	0	1	1	1	1	1	1	1	0	1	1	2	2	2	1	1	1	1	2	2	2	2	2	1	1	1	1	1	1	1	0	0	1	1	1
R1.1	3	2	1	3	3	3	3	0	3	3	3	3	3	3	3	0	1	0	0	1	3	3	3	2	3	3	3	3	3	3	0	0	0	3	3	3	3	3	3	3	3	3	1	3	3	0	0	3	3	3
R1.2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	2	2	2	2
R1.3	2	2	2	2	2	2	3	2	2	2	2	3	2	2	2	2	2	2	3	2	2	2	2	2	2	3	3	3	3	3	2	2	2	3	2	2	2	2	2	2	2	2	2	2	2	1	2	2	2	2

Table 4: Case study FAIR dimension scores (%) from the representative run in Table 3.

Dataset	F	A	I	R	Overall
AURIN OSM POIs (D2)	41.7	66.7	0.0	66.7	43.6
NASA GDIS (D3)	100.0	77.8	44.4	55.6	71.8
Zenodo Crater Lake (D1)	83.3	77.8	55.6	77.8	74.4

Repository effects. The representative run suggests higher scores for repositories with stable identifiers and structured metadata exposure, including PANGAEA, Zenodo, Dryad, and Harvard Dataverse, and lower scores for several catalog-oriented ScienceBase and AURIN records. These are descriptive observations from an uneven, small sample rather than population estimates for each repository.

Domain effects. Across domains, *Interoperability* remains the principal limiting factor. We observe a tendency for biodiversity/ecology datasets (often benefiting from established community metadata norms) to score higher overall, while datasets hosted in heterogeneous government catalog ecosystems (e.g., ScienceBase) score lower.

5.4 Case studies: root-cause analysis

Table 4 analyzes three representative datasets spanning low, medium, and high maturity. Dimension scores are computed as the mean of sub-principles within each FAIR dimension, scaled to 0–100 (using the same rubric as Table 3).

AURIN OSM POIs (D2, 43.6%). This dataset exhibits near-zero *Interoperability* due to absent machine-readable knowledge representations and unregistered vocabularies (no JSON-LD/RDF export, no typed relations). *Accessibility* is moderate due to HTTPS access, and *Reusability* is moderate due to visible license/provenance statements. The primary gap is the lack of machine-actionable metadata exposure (structured identifiers, vocabulary grounding, and typed relations).

NASA GDIS (D3, 71.8%). This dataset achieves high *Findability* via resolvable persistent identifiers and strong indexing. *Interoperability* remains moderate because metadata relies largely on generic schema markup without explicit domain vocabulary registration or typed links. *Reusability* is constrained by incomplete provenance detail and limited machine-readable linking between the dataset and related resources.

Zenodo Crater Lake (D1, 74.4%). This dataset achieves strong overall maturity, with high *Reusability* driven by explicit licensing and provenance. *Interoperability* is constrained by I2/I3: the dataset relies primarily on generic vocabularies rather than registered domain vocabularies, and provides few typed links to related datasets/resources.

5.5 Results: comparative tool analysis

Table 5 compares normalized overall scores across five tools, and Table 6 reports their rank associations with AgentFAIR. Across datasets, tools disagree substantially: the population standard deviation across available normalized tool scores averages 15.0 points (maximum 30.3 points), indicating that operationalizations of FAIR differ materially even after normalization.

Failure and brittleness patterns. F-UJI exhibits sharp failures (scores $\leq$11.5) on 10/50 datasets (20%), concentrated on Harvard Dataverse and Figshare, consistent with brittle identifier and registry-lookup dependencies. FAIR-Checker timed out on one dataset (D34). FAIR-enough produced no score for 18 datasets (36%) due to database/runtime limits, and tends to score lowest when it does return output because its binary tests penalize partial compliance.

Tool agreement. AgentFAIR is most aligned with FAIRshake ($\rho = 0.61$, $p < 0.001$, $n = 50$), followed by FAIR-Checker ($\rho = 0.47$, $p < 0.001$, $n = 49$) and F-UJI ($\rho = 0.42$, $p = 0.002$, $n = 50$). The FAIR-enough association is weaker and not statistically significant ($\rho = 0.31$, $p = 0.084$, $n = 32$). This pattern reflects both rubric differences (e.g., limited *Interoperability* coverage in some baseline rubrics) and extraction strategies (API-centric vs. browser-centric).

As a deliberately stricter diagnostic, we also compare AgentFAIR with the single highest baseline score available for each dataset. Under the same ±5-point band, AgentFAIR is higher on 6/50 datasets, comparable on 8/50, and lower on 36/50. This "best baseline" is a composite oracle rather than one deployable tool; the result is reported to prevent the average-baseline analysis from being misread as a superiority claim.

Qualitatively, three system properties contribute to differences:

- **Contextual resolution of identifiers and access conditions:** LLM agents can disambiguate DOI-like strings, reconcile multiple identifiers, and interpret whether authentication statements are optional or mandatory.
- **Geospatial sensitivity:** Prompts explicitly recognize CRS identifiers, ISO metadata elements, and OGC endpoint mentions, which informs *Interoperability* and R1.3 scoring.

Table 5: Normalized dataset-level overall scores (0–100) across five FAIR assessment tools (single-run per tool). Columns D1–D50 map to the dataset list in Appendix E. "–" indicates the tool did not return a score (e.g., API timeout or FAIR-enough database limit).

Tool	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15	D16	D17	D18	D19	D20	D21	D22	D23	D24	D25	D26	D27	D28	D29	D30	D31	D32	D33	D34	D35	D36	D37	D38	D39	D40	D41	D42	D43	D44	D45	D46	D47	D48	D49	D50
AgentFAIR	74.4	43.6	71.8	79.5	79.5	79.5	76.9	48.7	74.4	56.4	71.8	82.1	82.1	69.2	53.8	43.6	74.4	69.2	74.4	71.8	66.7	74.4	71.8	38.5	71.8	76.9	76.9	82.1	82.1	64.1	69.2	66.7	59.0	79.5	64.1	74.4	82.1	82.1	71.8	69.2	69.2	74.4	61.5	66.7	74.4	43.6	38.5	43.6	56.4	64.1
F-UJI	92.3	34.6	61.5	84.6	84.6	88.5	96.2	42.3	88.5	11.5	76.9	80.8	75.0	53.8	7.7	28.8	7.7	61.5	61.5	50.0	76.9	76.9	7.7	23.1	7.7	96.2	76.9	11.5	96.2	92.3	84.6	84.6	84.6	84.6	88.5	84.6	80.8	76.9	88.5	88.5	80.8	88.5	11.5	11.5	11.5	42.3	42.3	57.7	11.5	57.7
FAIR-Checker	91.7	54.5	58.3	91.7	91.7	91.7	91.7	20.8	91.7	75.0	91.7	91.7	91.7	91.7	83.3	50.0	29.2	62.5	62.5	45.8	91.7	91.7	29.2	58.3	37.5	79.2	79.2	79.2	79.2	79.2	62.5	62.5	62.5	–	91.7	91.7	91.7	79.2	91.7	91.7	91.7	91.7	37.5	29.2	29.2	50.0	50.0	41.7	83.3	45.8
FAIRshake	100.0	19.4	69.4	75.0	80.6	80.6	91.7	27.8	88.9	16.7	94.4	88.9	86.1	75.0	16.7	27.8	36.1	69.4	72.2	69.4	88.9	88.9	36.1	19.4	36.1	86.1	80.6	86.1	91.7	86.1	63.9	58.3	58.3	72.2	75.0	63.9	69.4	69.4	88.9	94.4	91.7	88.9	36.1	36.1	36.1	27.8	27.8	36.1	16.7	69.4
FAIR-enough	54.5	40.9	9.1	54.5	54.5	50.0	50.0	45.5	50.0	22.7	45.5	50.0	9.1	18.2	22.7	18.2	22.7	68.2	40.9	9.1	45.5	45.5	22.7	36.4	–	54.5	54.5	50.0	50.0	50.0	72.7	72.7	–	–	–	45.5	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

Table 6: Spearman association between AgentFAIR and baseline normalized scores.

Baseline	n	ρ	p
F-UJI	50	0.42	0.002
FAIR-Checker	49	0.47	< 0.001
FAIRshake	50	0.61	< 0.001
FAIR-enough	32	0.31	0.084

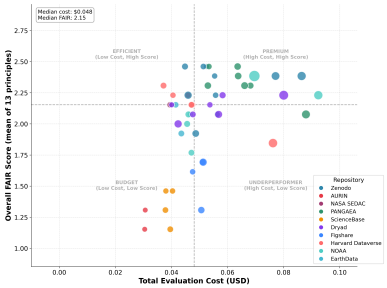

Figure 2: Cost-efficiency quadrant: datasets positioned by evaluation cost (x-axis) vs. overall FAIR score (y-axis).

Table 7: AgentFAIR efficiency metrics across 50 datasets

Metric	Value
Total Tokens	16.54M
Total Cost	$2.68
Mean Cost/Dataset	$0.054
Mean Processing Time	1,054 seconds

- **Critic-based reconciliation:** Cross-sub-principle checks trigger re-evaluation when evidence suggests inconsistency (e.g., protocol openness vs. authentication clarity).

5.6 Results: cost and efficiency analysis

Figure 2 and Table 7 summarize computational characteristics. Zenodo datasets cluster in a "low cost / high score" region, reflecting stable structured metadata exposure. More complex repositories and datasets require additional crawling and semantic reasoning, increasing processing time and token usage. A cost breakdown by FAIR dimension appears in Appendix (Figure 6).

5.7 Consistency and auditability analysis

Across repeated evaluations during development (5 runs per dataset on a 10-dataset subset, stratified by FAIR maturity quintile), mean exact agreement on sub-principle scores is 89%, with a reported run-level standard deviation of 3 percentage points. We omit the earlier confidence interval because the retained summary statistics are insufficient to reconstruct it correctly. This analysis measures repeatability, not reproducibility in the deterministic sense. AgentFAIR also retains deterministic extraction failure modes; browser rendering and retry logic reduce some failures but do not eliminate them.

Failure analysis. Of 50 datasets, 3 (6%) required manual URL correction due to redirect chains or domain changes. The critic triggered re-evaluation in 96% of sub-principle assessments. A trigger means that at least one hard logical check failed or confidence fell below its threshold; it does not by itself show that the initial score was wrong or that the revised score is more accurate. In contrast, F-UJI failed completely on 10 datasets (20%) due to identifier resolution errors, and FAIR-enough encountered database errors on 18 datasets (36%).

Human evaluation protocol. Three domain experts with 5+ years of geospatial data management experience independently evaluated a stratified random sample of 15 datasets (3 per maturity quintile). After a calibration session with worked rubric examples, each expert independently assessed all 13 sub-principles without access to automated tool outputs (195 judgments per expert; 585 annotations in total). Per-dataset annotation time ranged from a few minutes for simple records to roughly 20 minutes for complex cases. Inter-rater agreement (Fleiss' κ=0.71, 95% CI: [0.65, 0.77]) indicates substantial agreement. AgentFAIR achieved 82% alignment with expert consensus (majority vote); disagreements concentrate on I2 (interpreting vocabulary FAIRness, κ=0.58) and R1.2 (provenance completeness thresholds, κ=0.61), indicating where rubric definitions can be further refined.

Baseline comparison on the expert subset. The baseline tools do not expose outputs directly compatible with all 13 ordinal expert labels, so a head-to-head sub-principle accuracy comparison is not available. As a coarser diagnostic, we aggregate the expert labels to an overall score and apply a ±5-point dataset-level band. F-UJI and FAIR-Checker each fall within the band on 2/15 datasets, FAIRshake on 4/15, and FAIR-enough on 0/10 returned outputs (five failures). These figures are not commensurate with AgentFAIR's 82% sub-principle alignment and should not be interpreted as proof of superior accuracy. A larger study with tool outputs mapped and judged at the same sub-principle level remains necessary.

Table 8: Ablation: critic agent contribution (n=10 datasets)

Configuration	Consistency	Re-eval Rate
Full system (with critic)	89±3%	96% triggered
Without critic agent	71±5%	N/A

Consistency = agreement across 5 repeated runs; Re-eval Rate = percentage of sub-principle assessments triggering critic review.

5.8 Within-family model transfer

To probe dependence on GPT-4o-mini, we reran the identical pipeline and prompts with GPT-4o on the same stratified 10-dataset subset. Dataset rankings are strongly associated ($\rho = 0.877$, $p < 0.001$); 92/130 sub-principle scores are exact matches and 121/130 (93.1%) differ by at most one maturity level. The mean absolute difference in normalized score is 7.1 percentage points. This result is supportive within the GPT family, but it does not establish robustness to model version changes or transfer to a different model family.

5.9 Ablation study: critic agent contribution

To isolate the critic agent's contribution, we conducted an ablation study on 10 representative datasets (2 from each FAIR maturity quintile, spanning Zenodo, Dryad, PANGAEA, Harvard Dataverse, and NOAA NCEI). Table 8 compares full AgentFAIR against a variant with the critic agent disabled. Without critic-based quality control, consistency drops from 89±3% to 71±5% ($p < 0.01$, paired t-test), with the largest degradation on I2 and R1.2—sub-principles requiring nuanced evidence interpretation. The 96% trigger rate combines seven hard logical checks with confidence gates. The experiment did not retain a trigger-type breakdown, conditional score-change distribution, latency contribution, or accuracy of the intermediate outputs; consequently, the ablation supports a repeatability benefit but does not fully characterize whether the critic corrects genuine errors. A plain-retry/majority-vote baseline was not run.

5.10 Limitations and future experiments

The study is exploratory and has several unresolved validity limits. First, 50 datasets and 15 expert-annotated datasets do not support broad claims across the long tail of geospatial repositories. Second, the cross-tool normalization conflates different constructs and is retained only for disagreement analysis; baseline runtimes were not systematically logged. Third, the current evidence does not isolate browser engineering, LLM enrichment, geospatial prompt terms, multi-agent decomposition, and critic retries. Geospatial-prompt, rule-only extraction, temperature 0.0 versus 0.1, and plain-retry ablations remain outstanding. Fourth, GPT-4o is a within-family check rather than cross-family validation, and proprietary model updates threaten longitudinal repeatability. Fifth, the strict I2 registry criterion can under-credit valid community vocabularies absent from generic registries; the low expert agreement on I2 reinforces this concern. Sixth, prompt-injection and poisoned-metadata robustness have not been evaluated. Finally, raw repeated-run, human-label, ablation, and model-transfer outputs should be released with analysis scripts before the reported inferential statistics are treated as independently reproducible.

6 Conclusion

We introduced AgentFAIR, an LLM-driven multi-agent framework for auditable FAIRness evaluation of geospatial datasets. The system combines browser-based and structured metadata extraction, 13 rubric-specific evaluators, deterministic checks, and critic-guided retries.

In a 50-dataset, 10-repository sample, *Findability* is strongest (79.7%) and *Interoperability* weakest (45.3%). Cross-tool normalized scores show substantial disagreement (mean per-dataset standard deviation 15.0 points), but heterogeneous rubrics prevent interpreting those differences as comparative accuracy. Preliminary repeated-run and expert studies support further investigation of the critic and evidence model, while the high retry rate, small human sample, missing component ablations, and within-family-only model check leave important questions unresolved. At approximately $0.054 per dataset, the framework is economically feasible for larger validation studies; stronger accuracy or deployment claims require released raw evaluation artifacts, cross-family testing, and adversarial security evaluation.

References

[1] Christiane Bahlo, Siddeswara Guru, Nicholas Car, and Lesley Wyborn. 2024. Advancing FAIR Agricultural Data: The AgReFed FAIR Assessment Tool. *Data Science Journal* 23, 1 (2024), 15. doi:10.5334/dsj-2024-018

[2] Lennart Busch, Daniel Tebernum, and Gissel Velarde. 2025. Exploring LLM Capabilities in Extracting DCAT-Compatible Metadata for Data Cataloging. *arXiv preprint arXiv:2507.05282* (2025). doi:10.48550/arXiv.2507.05282

[3] Leonardo Candela, Dario Mangione, and Gina Pavone. 2024. The FAIR Assessment Conundrum: Reflections on Tools and Metrics. *Data Science Journal* 23, 1 (2024). doi:10.5334/dsj-2024-033

[4] Daniel J. B. Clarke, Lily Wang, Alex Jones, Megan L. Wojciechowicz, Denis Torre, Kathleen M. Jagodnik, Sherry L. Jenkins, Peter McQuilton, Zachary Flamholz, Moshe C. Silverstein, et al. 2019. FAIRshake: Toolkit to Evaluate the FAIRness of Research Digital Resources. *Cell Systems* 9, 5 (2019), 417–421. doi:10.1016/j.cels.2019.09.011

[5] Anusuriya Devaraju and Robert Huber. 2020. F-UJI FAIR Assessment Tool. https://www.f-uji.net/ Accessed: 2026-01-08.

[6] Vincent Emonet and Michel Dumontier. 2022. FAIR-enough: A Community-Governed FAIR Maturity Assessment Framework. https://github.com/MaastrichtU-IDS/fair-enough

[7] Alban Gaignard, Thomas Rosnet, Frédéric De Lamotte, Vincent Lefort, and Marie-Dominique Devignes. 2023. FAIR-Checker: Supporting Digital Resource Findability and Reuse with Knowledge Graphs and Semantic Web Standards. *Journal of Biomedical Semantics* 14, 1 (2023), 7. doi:10.1186/s13326-023-00289-5

[8] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=VtmBAGCN7o

[9] Zezhou Huang and Eugene Wu. 2024. Cocoon: Semantic Table Profiling Using Large Language Models. In *Proceedings of the 2024 Workshop on Human-In-the-Loop Data Analytics (HILDA '24)*. ACM, 1–7. doi:10.1145/3665939.3665957

[10] International Organization for Standardization. 2014. ISO 19115-1:2014 Geographic Information — Metadata Part 1: Fundamentals. https://www.iso.org/standard/53798.html Published Edition 1, last reviewed and confirmed in 2019, with 2 amendments.

[11] N. A. Krans, A. Ammar, P. Nymark, E. L. Willighagen, M. I. Bakker, and J. T. K. Quik. 2022. FAIR Assessment Tools: Evaluating Use and Performance. *NanoImpact* 27 (2022), 100402. doi:10.1016/j.impact.2022.100402

[12] Lang Chain Authors. 2026. LangGraph: Agent Orchestration Framework for Reliable AI Agents. https://www.langchain.com/langgraph. Accessed on: 2026-02-08.

[13] Lan Li, Liri Fang, and Vetle I. Torvik. 2024. Autodcworkflow: LLM-Based Data Cleaning Workflow Auto-Generation and Benchmark. *arXiv preprint arXiv:2412.06724* (2024). doi:10.48550/arXiv.2412.06724

[14] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan

Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. 2024. AgentBench: Evaluating LLMs as Agents. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=zAdUB0aCTQ

[15] Gengchen Mai, Chris Cundy, Kristy Choi, Yingjie Hu, Ni Lao, and Stefano Ermon. 2024. On the Opportunities and Challenges of Foundation Models for GeoAI. *Comput. Surveys* 56, 10 (2024), 1–41. doi:10.1145/3653070

[16] Avanika Narayan, Ines Chami, Laurel Orr, Simran Arber, Pengyu Rong, Monica Shen, and Christopher Ré. 2024. Can Foundation Models Wrangle Your Data? *Proceedings of the VLDB Endowment* 17, 5 (2024), 1212–1225. doi:10.14778/3641204.3641227

[17] Open Geospatial Consortium. 2022. OGC Disaster Pilot 2022: Integrating ISO and OGC Standards for FAIR Geospatial Data. https://www.ogc.org/initiatives/disaster-pilot/ OGC Innovation Program.

[18] Dominik Paprotny and Matthias Mengel. 2023. Population, Land Use and Economic Exposure Estimates for Europe at 100 m Resolution from 1870 to 2020. *Scientific Data* 10, 1 (2023), 372. doi:10.1038/s41597-023-02282-0

[19] PwC EU Services. 2018. *Cost-benefit analysis for FAIR research data.* Technical Report. European Commission. https://op.europa.eu/en/publication-detail/-/publication/d375368c-1a0a-11e9-8d04-01aa75ed71a1

[20] Syed N. Sakib, Kallol Naha, Sajratul Y. Rubaiat, and Hasan M. Jamil. 2025. A GenAI System for Improved FAIR Independent Biological Database Integration. *ACM Journal of Data and Information Quality* 17, 4, Article 26 (2025), 29 pages. doi:10.1145/3770753

[21] P. Travis Thompson, Sweta Ojha, Christian D. Powell, Kelly G. Pennell, and Hunter N. B. Moseley. 2023. A Proposed FAIR Approach for Disseminating Geospatial Information System Maps. *Scientific Data* 10, 1 (2023), 389. doi:10.1038/s41597-023-02281-1

[22] TKFDM. 2023. FAIR Data Assessment Tool. https://www.ubs.uzh.ch/de/TKFDM/Services/FAIR-Assessment.html University of Zurich.

[23] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. 2024. A Survey on Large Language Model Based Autonomous Agents. *Frontiers of Computer Science* 18, 6 (2024), 186345. doi:10.1007/s11704-024-40231-1

[24] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, et al. 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* 3, 1 (2016), 1–9. doi:10.1038/sdata.2016.18

[25] Mark D. Wilkinson, Michel Dumontier, and Luiz Olavo Bonino da Silva Santos. 2019. FAIR Evaluator: The FAIR Evaluation Services. https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/ GO FAIR Foundation.

[26] Mark D. Wilkinson, Michel Dumontier, Susanna-Assunta Sansone, Luiz Olavo Bonino da Silva Santos, Mario Prieto, Dominique Batista, Peter McQuilton, Tobias Kuhn, Philippe Rocca-Serra, Mercè Crosas, et al. 2019. Evaluating FAIR Maturity Through a Scalable, Automated, Community-Governed Framework. *Scientific Data* 6, 1 (2019), 174. doi:10.1038/s41597-019-0184-5

[27] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2024. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. *arXiv preprint arXiv:2308.08155* (2024). doi:10.48550/arXiv.2308.08155 Microsoft Research.

[28] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. 2024. The Rise and Potential of Large Language Model Based Agents: A Survey. *arXiv preprint arXiv:2309.07864* (2024). doi:10.48550/arXiv.2309.07864

[29] Frank F. Xu, Yufan Wang, Boxuan Sharma, Hoang Peng, Hailey Hyunji Liu, Jerry Yang Chen, Shuyan Lin, Yunzhe Yang, Zijian Sun, Lun Du Zheng, Sida I. Ding, Ningxin Hou, Sophia Ni, John Hsu, Raghav Ramanujan, Yongjae Kim, Tanishq Bhasker, Chen Wu, and Xiang Xu. 2024. TheAgentCompany: Benchmarking LLM Agents on Consequential Real World Tasks. doi:10.48550/arXiv.2412.14161

[30] Shuo Zhang, Zezhou Huang, and Eugene Wu. 2025. Data Cleaning Using Large Language Models. In *2025 IEEE 41st International Conference on Data Engineering Workshops (ICDEW)*. IEEE, 28–32. doi:10.48550/arXiv.2410.15547

A Open-source artifact and prompt specifications

A.1 Zenodo release and contents

An archived software snapshot is available on Zenodo (https://doi.org/10.5281/zenodo.18529560); the maintained source is intended for https://github.com/MingCHEN-Github/AgentFAIR. The Zenodo deposit contains: `AgentFAIR.zip` (full source code), `DESCRIPTION.md` (setup + artifact overview), and `app_demo.mp4` (end-to-end walk-through video). The software is licensed under AGPL-3.0-only.

A.2 Quick start and environment configuration

Prerequisites. Python 3.10+ and an OpenAI API key. The archived artifact is self-contained (no git required).

One-command launch.

```
unzip AgentFAIR.zip
cd AgentFAIR
./run.sh
```

`run.sh` automates setup and execution, including dependency installation, Playwright browser installation for web crawling, creation of `.env` from `.env.example` when needed, launching the FastAPI backend (`http://localhost:8000`), and opening the React/Vite frontend in a browser.

API key configuration. Keys can be provided either (i) via the frontend *Settings* modal at runtime (recommended for quick evaluation) or (ii) via a local `.env` file. The artifact also supports optional keys for Brave Search (`BRAVE_API_KEY`), OpenRouter (`OPENROUTER_API_KEY`), and opt-in LangSmith tracing (`LANGSMITH_API_KEY`). Remote tracing is disabled by default.

A.3 Repository layout and persisted outputs

Unzipping `AgentFAIR.zip` produces an `AgentFAIR/` directory. Key components:

```
fair_agents/       core Python package (agents,
tools, LangGraph workflows, reporting)
fair_agents/api/  FastAPI service (local evaluation
endpoints)
frontend/     React/Vite client (run orchestration
+ results UI)
storage/         generated SQLite evidence store
(ignored by Git)
fair_agents/reports/  generated Markdown/JSON reports
(ignored by Git)
tests/        URL-safety regression tests
run.sh          single-command launcher
```

Audit trail outputs. Each evaluation persists (i) per-sub-principle scores and confidence, (ii) evidence bundles with provenance pointers/snippets, and (iii) exported reports (Markdown + JSON) backed by a SQLite evidence store. Optional LangSmith traces can be enabled when a tracing key is provided.

A.4 Representative prompt template and output schema

AgentFAIR prompts follow a uniform multi-part structure: (i) sub-principle definition and scope, (ii) maturity rubric (0–3; operationalization in Table 2), (iii) decision logic and edge cases, (iv) geospatial adaptations (when applicable), (v) evidence requirements, (vi) anti-hallucination constraints, (vii) output JSON schema, and (viii) short worked examples.

Below we show a simplified *illustrative* template for a sub-principle agent. In practice, each agent instantiates this template with the corresponding rubric criteria and geospatial indicators from the operationalization in Table 2.

```
Input: extracted metadata record M (structured
fields + landing page snippets).
Task: Evaluate sub-principle {F1/F2/...} using
the rubric below.
Rubric: score 0/1/2/3 with explicit criteria.
Evidence rule: every claim must cite a field or
quoted snippet from M.
Output JSON:
{
"principle": "...",
"score": 0|1|2|3,
"confidence": 0.0–1.0,
"evidence": [{"source": "...", "snippet": "..."},
...],
"rationale": "...",
"recommendations": ["...", ...]
}
```

B Further Discussions of AgentFAIR

B.1 Key insights

***Interoperability* is the dominant bottleneck.** Across repositories, I2 and I3 consistently score lowest. A key driver is that widely used geospatial vocabularies and standards are not consistently represented in the registries relied upon by semantic FAIR validators, which leads to systematic penalties. This suggests that improving ecosystem-level vocabulary registration (not only dataset-level metadata) is critical for improving geospatial FAIRness.

Rule-based tools are brittle on modern repository interfaces. We observe sharp baseline failures when crawlers cannot access machine-readable metadata (e.g., JavaScript-rendered pages) or when identifier resolution relies on assumptions that do not hold universally. These failures can dominate overall tool disagreement and complicate the interpretation of FAIR scores as decision signals.

LLM-based evaluation can be useful but must be auditable. LLM agents help interpret ambiguous natural language and recover domain cues, but they introduce stochasticity. In AgentFAIR, evidence-grounded prompts, explicit rubrics, and critic-based consistency checks are essential to make LLM reasoning usable for compliance assessment.

B.2 Threats to validity

Dataset sampling. Our benchmark emphasizes cross-repository diversity, but 50 datasets is not exhaustive; results may shift under different sampling strategies or when focusing on a single community.

Baseline normalization. Normalizing heterogeneous tool outputs to a single 0–100 score enables comparison but necessarily abstracts away rubric differences. We therefore interpret correlations and variance as indicators of *disagreement*, not as absolute correctness.

Lack of ground-truth FAIR labels. FAIRness is not a single objective quantity; expert judgments may differ, and community standards evolve. We address this only partially by reporting evidence and documenting disagreements; the preliminary expert sample is not a definitive ground truth.

LLM-specific threats. The primary evaluation relies on GPT-4o-mini, which may change without a versioned model snapshot. The GPT-4o check provides only within-family evidence. Low temperature does not guarantee determinism, and no temperature ablation was run. Prompt phrasing can affect score distributions, and the prompts were iteratively refined during development. Evidence requirements and critic checks improve auditability but cannot prevent hallucination, prompt injection, or subtle misinterpretation of poisoned or ambiguous metadata. Repository-specific overfitting therefore remains possible.

I2 construct validity. The experimental I2 rubric places substantial weight on vocabulary registry presence. Valid geospatial vocabularies may be documented and accepted in community practice without appearing in generic registries, so the rubric can systematically under-credit them. Future work should combine registry evidence with community governance, resolvability, versioning, licensing, and documentation rather than treating registry count as a normative definition of FAIRness.

B.3 Applications

AgentFAIR supports: (i) **repository FAIRification** (prioritized remediation roadmaps), (ii) **CI/CD integration** (continuous metadata quality monitoring), (iii) **funding compliance reporting** (evidence-backed audit trails), and (iv) **training data curation** (filtering and ranking datasets for geospatial foundation models based on FAIR maturity).

C LangGraph workflow during evaluation process

The proposed AgentFAIR employs LangGraph, an open-source orchestration library in the LangChain ecosystem [12], to design multi-agent workflows as directed state machines. As illustrated in Figure 3, each FAIR dimension operates as an independent graph composed of modular nodes—crawling, metadata extraction, principle-specific evaluation, critic review, and reporting—connected through conditional and retry-based edges. Shared state variables capture crawl results, agent confidence, and error counts, enabling traceable and auditable execution.

D Additional plots and tables

Figure 4 shows that processing time varies by FAIR dimension; *Reusability* is fastest, while *Findability* and *Accessibility* take longest due to crawling and metadata extraction overhead. Most evaluations cluster in the 150–350 second range per dimension, with outliers corresponding to metadata-heavy datasets.

Figure 5 reveals a strong positive correlation between token usage and processing time, confirming that LLM API calls dominate wall-clock time. This relationship suggests that token-reduction strategies can directly lower both evaluation cost and latency.

Figure 6 sorts datasets by total cost (ascending), with *Reusability* consistently contributing the largest cost share and *Interoperability* the smallest. Total evaluation cost for all 50 datasets remains under $3, demonstrating the economic viability of LLM-based FAIR assessment at scale.

Figure 7 compares mean dimension-level scores across five FAIR assessment tools. AgentFAIR and F-UJI show relatively balanced

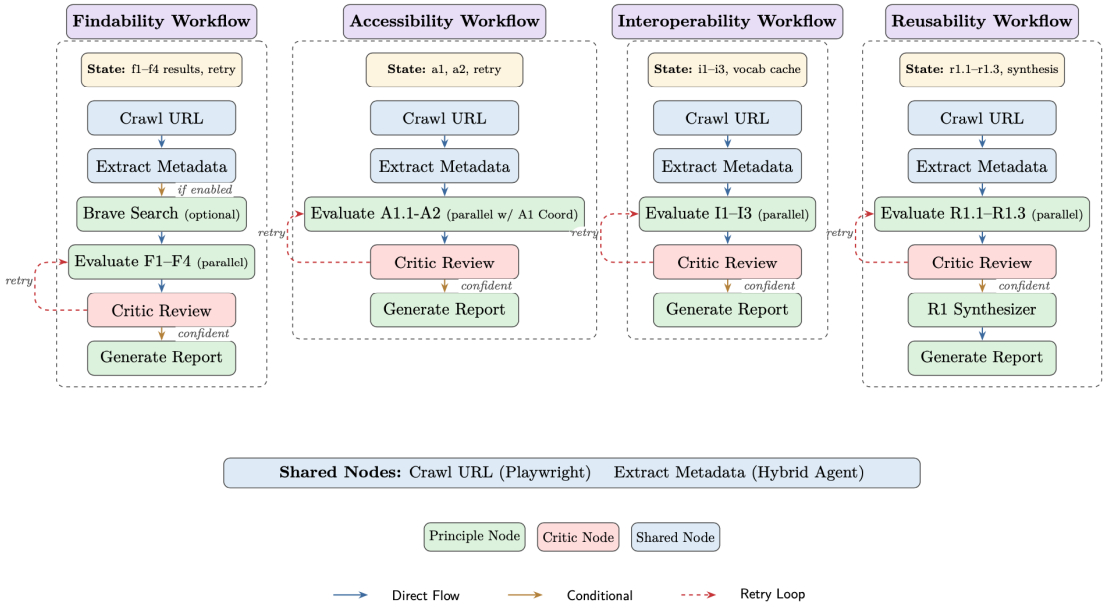

Figure 3: LangGraph state machine orchestrating FAIR evaluation workflows with conditional edges and retry logic.

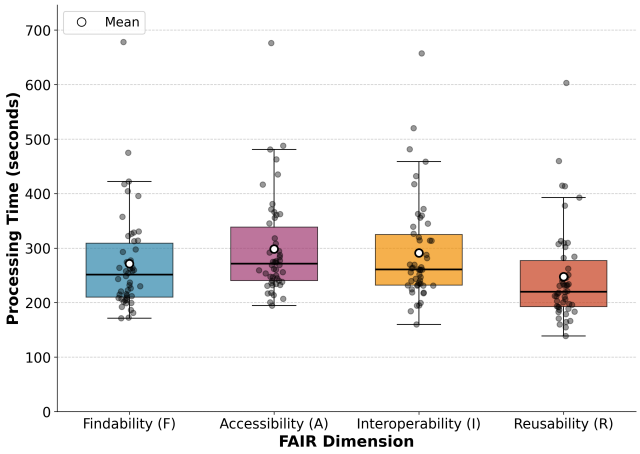

Figure 4: Processing time distribution by FAIR dimension.

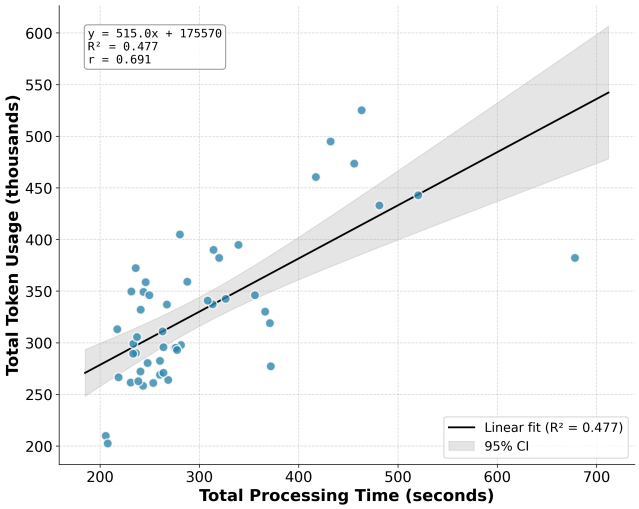

Figure 5: Token usage vs. processing time correlation.

profiles across all four dimensions, whereas FAIRshake lacks *Interoperability* coverage entirely and FAIR-enough tends to score lower due to its binary test methodology.

Figure 9 ranks datasets by scoring variance (standard deviation) across tools, with high-controversy datasets exhibiting >25% standard deviation. These datasets expose fundamental differences in how each tool operationalizes FAIR principles and serve as valuable diagnostic benchmarks for tool comparison.

Figure 8 plots each FAIR sub-principle by its mean maturity score and cross-dataset variance. *Findability* and *Accessibility* principles cluster in the "Universal Strength" quadrant (high mean, low variance), while *Interoperability* sub-principles fall into the "High Discriminator" or "Systemic Gap" regions, confirming that I-dimension compliance is the most variable and challenging aspect of FAIR maturity.

E Dataset details

Table 9 summarizes the key detailed information for all the evaluated 50 datasets.

F Artifact and auditability checklist

- **Hardware:** Apple M2 Pro, 32 GB RAM, macOS 15

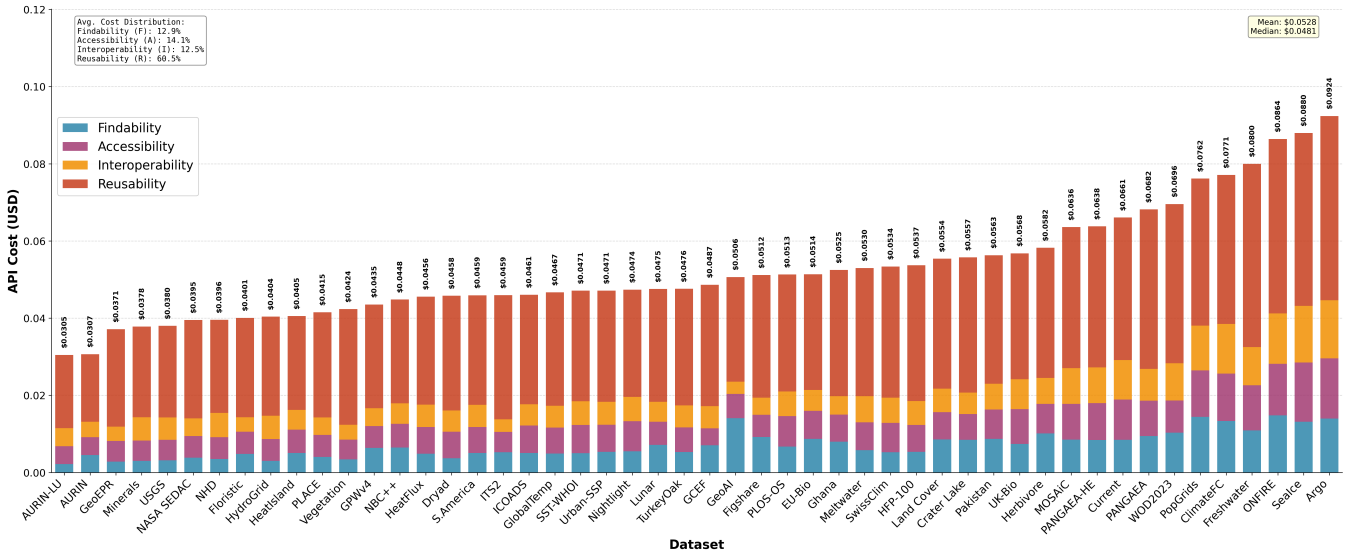

Figure 6: API cost breakdown by FAIR dimension across 50 datasets.

Table 9: Full dataset summary (50 datasets, 10 repositories).

#	Dataset	Repository	Domain	#	Dataset	Repository	Domain
D1	Crater Lake	Zenodo	Marine/Geoscience	D26	Meltwater	PANGAEA	Climate/Hydrology
D2	OSM POIs	AURIN	Urban/Land Classification	D27	Current Velocity	PANGAEA	Marine/Geoscience
D3	GDIS Disasters	Earthdata	Disaster/Emissions	D28	Swiss Climate	PANGAEA	Climate/Hydrology
D4	ONFIRE Portugal	Zenodo	Disaster/Emissions	D29	MOSAiC Melt	PANGAEA	Marine/Geoscience
D5	Climate Forecasts	Zenodo	Climate/Hydrology	D30	Sea-Ice 150ka	PANGAEA	Marine/Geoscience
D6	WRF Land Cover	Zenodo	Urban/Land Classification	D31	ICOADS	NOAA NCEI	Marine/Geoscience
D7	Seafloor Images	PANGAEA	Marine/Geoscience	D32	Ocean Heat	NOAA NCEI	Climate/Hydrology
D8	Groundwater	ScienceBase	Climate/Hydrology	D33	SST WHOI	NOAA NCEI	Marine/Geoscience
D9	Mammalian Bio	Dryad	Biodiversity/Ecology	D34	WOD 2023	NOAA NCEI	Marine/Geoscience
D10	Urban Network	Figshare	Urban/Land Classification	D35	GCEF Ecosystem	Zenodo	Biodiversity/Ecology
D11	Human Footprint	Dryad	Biodiversity/Ecology	D36	ITS2 Plants	Zenodo	Biodiversity/Ecology
D12	Sediment	PANGAEA	Marine/Geoscience	D37	NBC++	Zenodo	Biodiversity/Ecology
D13	Ghana LandUse	Zenodo	Urban/Land Classification	D38	EU Biodiv	Zenodo	Biodiversity/Ecology
D14	Pakistan Carbon	Zenodo	Disaster/Emissions	D39	Floristic	Dryad	Biodiversity/Ecology
D15	Lunar Map	Figshare	Planetary Science	D40	Turkey Oak	Dryad	Biodiversity/Ecology
D16	Hydrogrid	ScienceBase	Climate/Hydrology	D41	UK Biodiv	Dryad	Biodiversity/Ecology
D17	South America	Harvard Dataverse	Regional Analysis	D42	Herbivore	Dryad	Biodiversity/Ecology
D18	GlobalTemp	NOAA NCEI	Climate/Hydrology	D43	Pop Grids SSP	Harvard Dataverse	Demographics
D19	Argo	NOAA NCEI	Marine/Geoscience	D44	Urban Heat	Harvard Dataverse	Urban/Land Classification
D20	PLACE	NASA SEDAC	Demographics	D45	GeoEPR	Harvard Dataverse	Regional Analysis
D21	Vegetation	Dryad	Biodiversity/Ecology	D46	USGS Mineral	ScienceBase	Geology
D22	Freshwater	Dryad	Biodiversity/Ecology	D47	NHD	ScienceBase	Climate/Hydrology
D23	Nightlight	Harvard Dataverse	Remote Sensing	D48	GeoAI	Figshare	Technology
D24	AURIN LandUse	AURIN	Urban/Land Classification	D49	PLOS Sci	Figshare	Scientific Publishing
D25	Urban SSP	Harvard Dataverse	Demographics	D50	GPWv4	NASA EarthData	Demographics

- **Software:** Python 3.11; current release requires Python 3.10+, LangChain Core 1.x, and LangGraph 1.x
- **Model:** GPT-4o-mini via OpenAI API (temperature=0.1)
- **Dependencies:** Playwright, FastAPI, SQLite, Extruct, lxml, and rdflib
- **Logging:** prompts, responses, evidence, and scores are persisted as an audit trail
- **Code:** https://github.com/MingCHEN-Github/AgentFAIR; archived snapshot at https://doi.org/10.5281/zenodo.18529560
- **Missing research artifacts:** raw repeated-run outputs, expert labels, ablation runs, model-transfer outputs, and analysis notebooks are not included in the current software checkout and are required for independent reproduction of the reported statistics

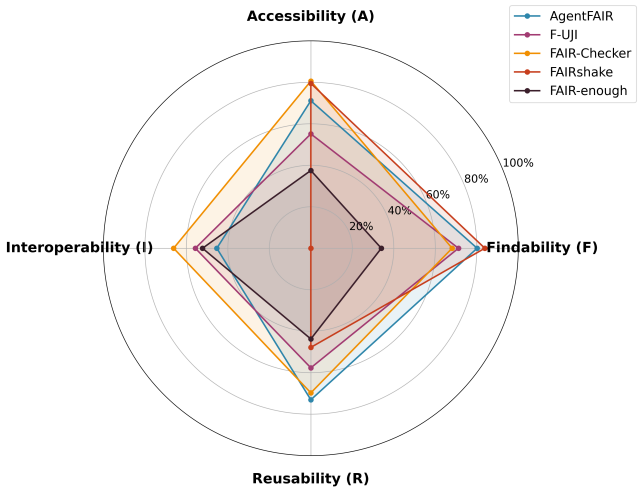

Figure 7: FAIR dimension coverage radar chart across tools.

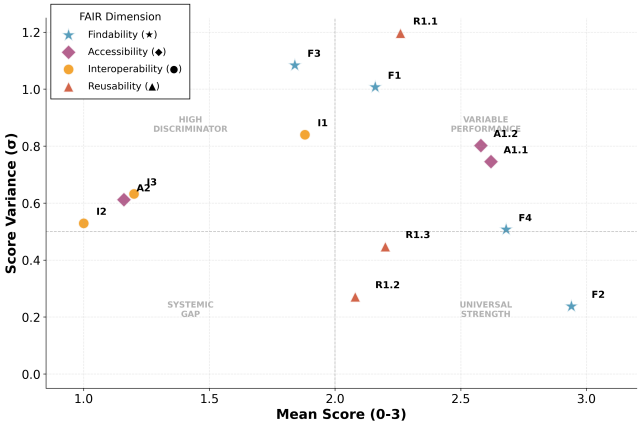

Figure 8: FAIR principle maturity landscape: mean score vs. variance across datasets.

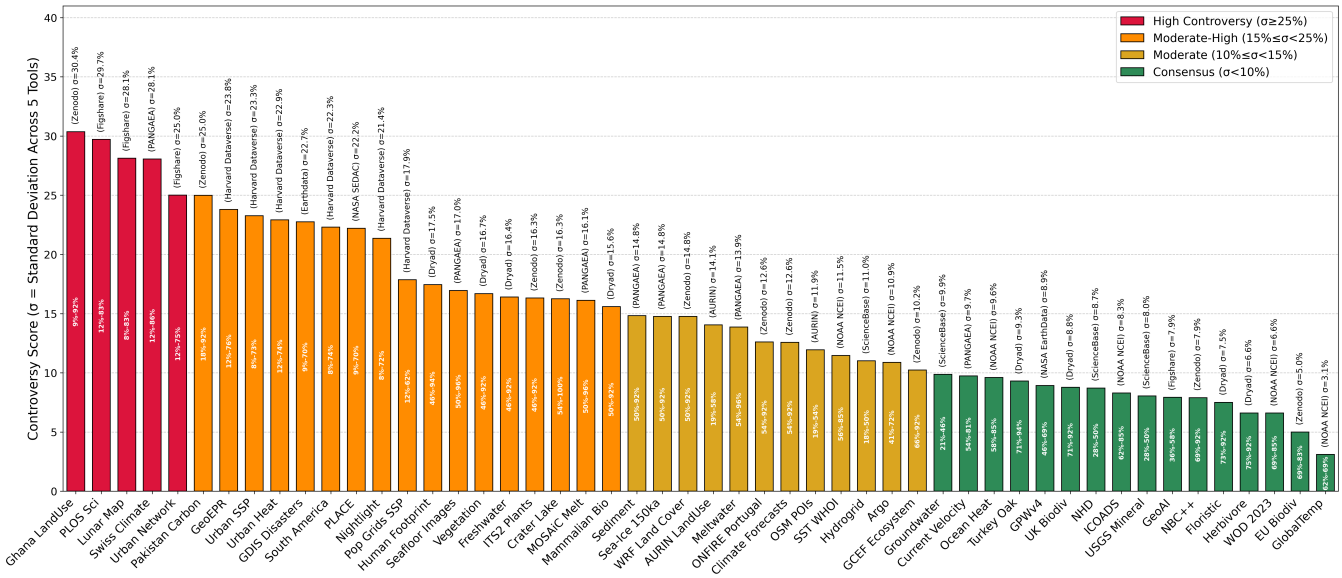

Figure 9: Dataset controversy index: scoring variance across tools.

G Comparative analysis of FAIR assessment systems

Table 10: Comparative analysis of FAIR assessment systems and this work's contributions

System / Study	LLM-Based	Geospatial	Explainable	Traceable	Cross-Principle	All 13 Sub-Principles
F-UJI [5]	✗	✗	✓	✓	✗	✓
FAIRshake [4]	✗	✗	✓	✓	✗	✗
AutoDCWorkflow [13]	✓	✗	✓	✗	✗	✗
TKFDM Tool [22]	✗	✗	✓	✓	✗	✗
FAIR-Checker [7]	✗	✗	✓	✓	✓	✗
FAIR Evaluator [25]	✗	✗	✓	✓	✗	✓
FAIR-enough [6]	✗	✗	✓	✓	✗	✓
Cocoon [9]	✓	✗	✓	✓	✗	✗
FAIRBridge [20]	✓	✗	✓	✓	✗	✗
AgReFed [1]	✗	✓	✓	✓	✗	✗
OGC Disaster Pilot [17]	✓	✓	✓	✓	✗	✗
This Work	✓	✓	✓	✓	✓	✓

H Diagnostic comparison with the baseline average

For completeness, averaging the available baseline scores per dataset and applying a ±5-point band places AgentFAIR above that average on 20/50 datasets, within the band on 18/50, and below it on 12/50. This statistic is sensitive to missing baseline outputs and mixes non-equivalent rubrics. It is therefore retained only as a descriptive disagreement summary; the stricter best-baseline result is reported in Section 5, and neither comparison is treated as an accuracy ranking. We omit the earlier per-dataset "higher/lower" narratives because many large differences were attributable to crawler and redirect failures rather than the LLM or multi-agent design.