



Australian Research Data Commons

Bridging the data discovery gap: User-centric recommendations for research data repositories

17 April, 2026, ESIP Discovery Cluster Meeting

Mingfang Wu
Australian Research Data Commons



ARDC is enabled by NCRIS

- Introduce the Research Data Alliance Data Discovery Paradigms IG
- Present “Bridging the data discovery gap: User-centric recommendations for research data repositories”
- Discuss AI-driven paradigm shift in research data discovery

To be Findable:

F1. (meta)data are assigned a globally unique and eternally persistent identifier.

F2. data are described with rich metadata.

F3. (meta)data are registered or indexed in a searchable resource.

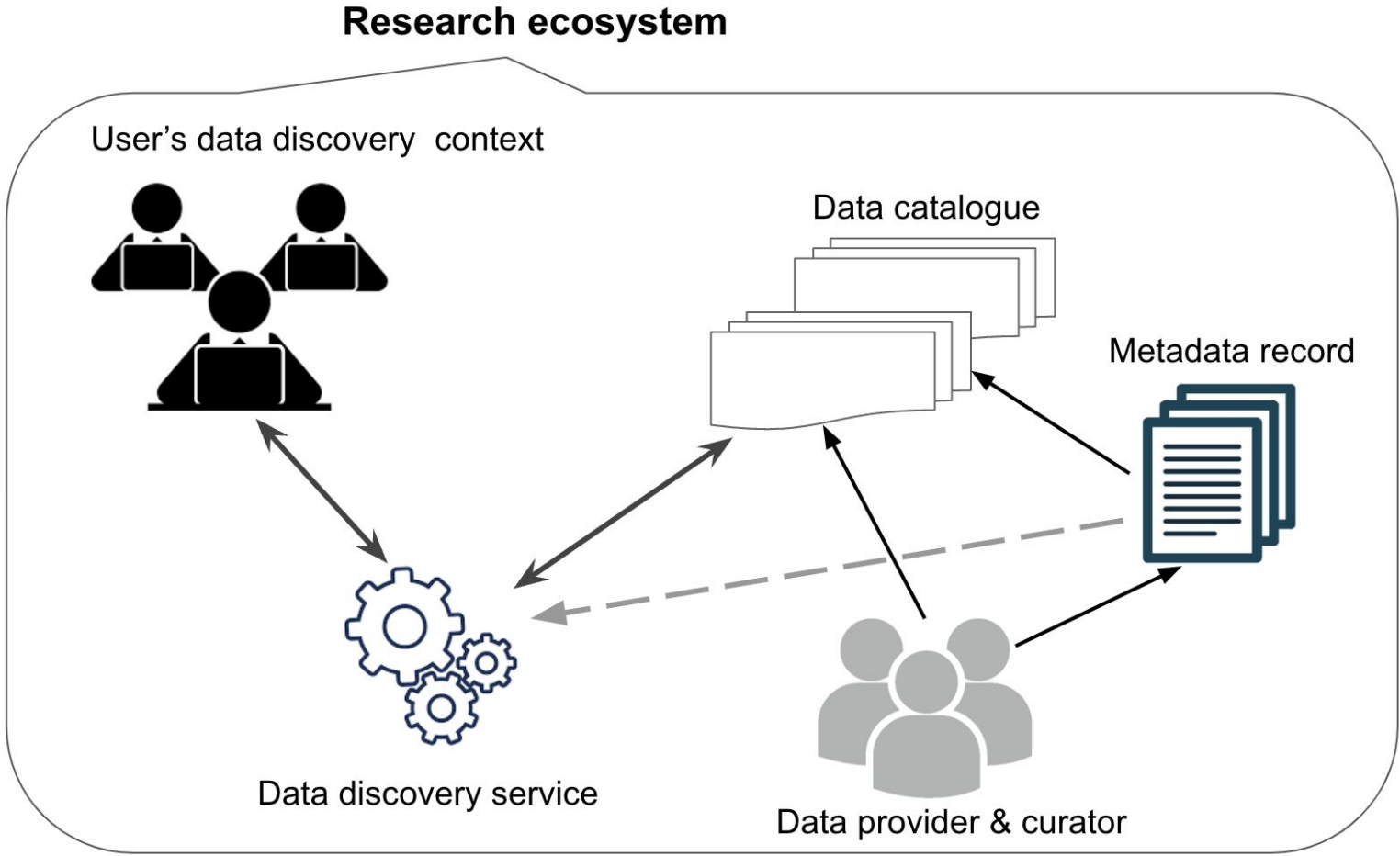
F4. metadata specify the data identifier.

RDA Data Discovery Paradigms Interest Group: Motivation

Findability ≠ Discoverability
Repository view vs User centric

Principle	Focus	Limitation
F1 (Metadata is assigned a unique identifier)	Machine identification of datasets.	Does not ensure the metadata <i>content</i> is relevant or complete for a user's specific query/need .
F2 (Data is described with rich metadata)	Repository obligation to provide comprehensive metadata.	Does not account for how a user actually searches (e.g., using domain-specific jargon vs. standard terms) or the quality of the discovery interface .
I1 (Use a formal, accessible, shared, and broadly applicable language)	Technical interoperability via standards (e.g., JSON-LD).	Data may be syntactically interoperable but remain semantically irrelevant or difficult to find without a tailored search experience .

User-centric data discovery



Data Discovery Paradigms Interest Group Charter

Objectives:

- Provide a forum where representatives across the spectrum of stakeholders and roles can **explore shared issues** of how to **improve data discovery**.
- Produce **actionable recommendations** for data producers, data repositories/catalogues, users/data seekers and data discovery service providers.

Group task forces and outputs

Kickoff: 2016 (P8)

Progress: P9 — P25

Current: (2026)

Task Forces

- Use cases, prototyping tools and test collections
- Guidelines for making data findable
- Relevancy ranking
- Metadata enrichment for discovery
- Data discovery on granularity (WG)
- Using schema.org for research data discovery (WG)
- User study of data discovery

Key Outputs

Data Discovery Paradigms: User Requirements and Recommendations

Data Science Journal (2019) • DOI: 10.5334/dsj-2019-003

79 Use Cases Published to Zenodo

(2017) • DOI: 10.5281/zenodo.1050976

Eleven Quick Tips for Finding Research Data

PLoS Comput Biol (2018) • DOI: 10.1371/journal.pcbi.1006038

A Survey of Current Practices in Data Search Services

(2018) • DOI: 10.17632/7j43z6n22z.1

Ten Principles to Improve Dataset Discoverability

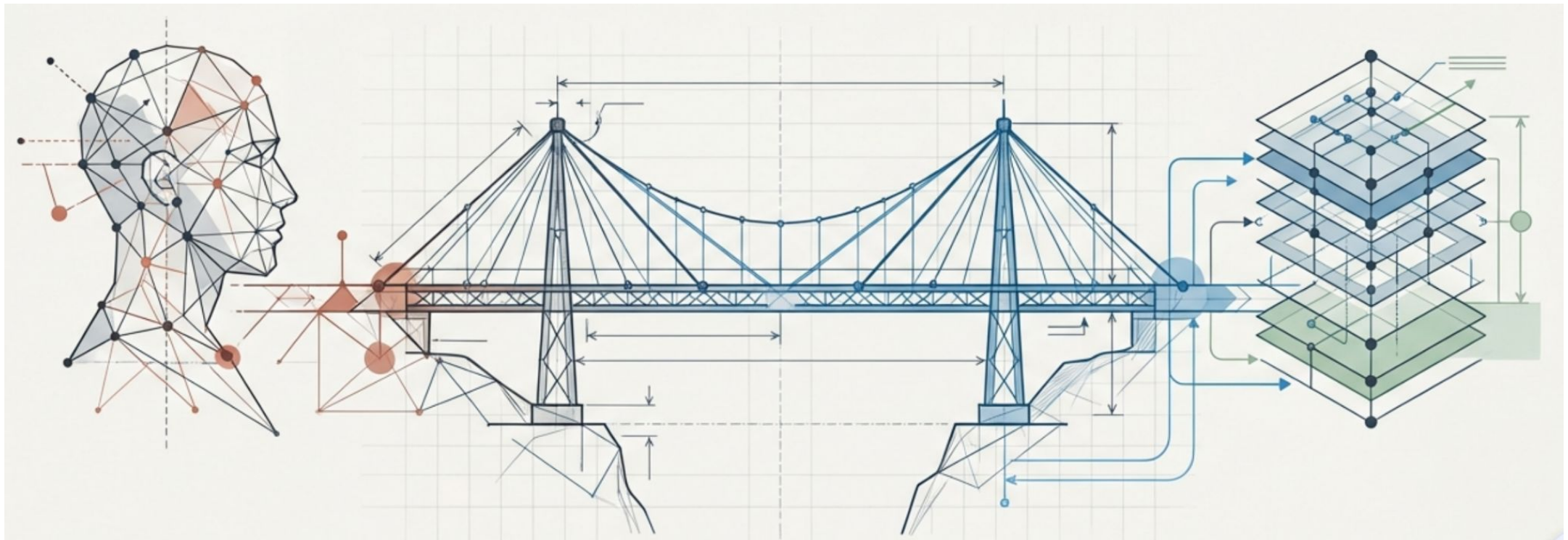
Research Data Alliance (2024) • DOI: 10.15497/rda/00120

Bridging the Data Discovery Gap: User-Centric Recommendations

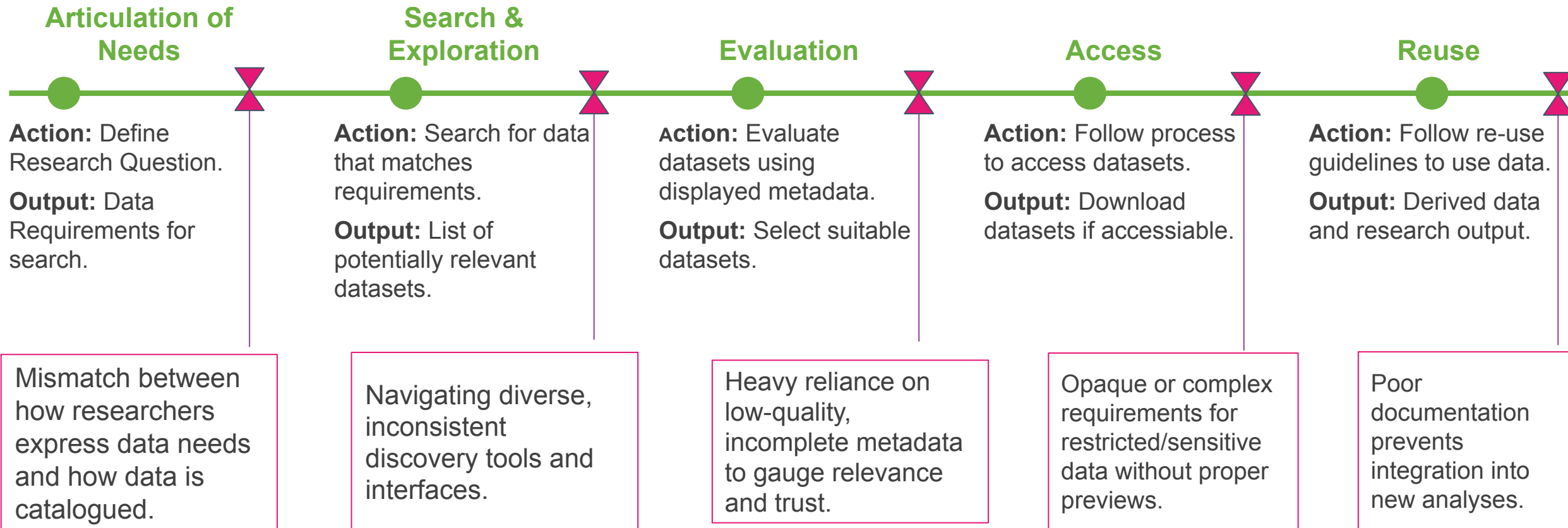
Data Science Journal (2026) • DOI: 10.5334/dsj-2026-006

Bridging the Data Discovery Gap: User-Centric Recommendations for Research Data Repositories

Wu, M., Löffler, F., Mathiak, B., Psomopoulos, F., Schindler, U., Aryani, A., Sempere, J.B., Culina, A., Czerniak, A., Erdmann, C., Gregory, K., Juty, N., Lister, A., Liu, Y.-H. and Pearman-Kanza, S. (2026) 'Bridging the Data Discovery Gap: User-Centric Recommendations for Research Data Repositories', *Data Science Journal*, 25(1), p. 6. Available at: <https://doi.org/10.5334/dsj-2026-006>.



User research data discovery process



Four Recommendations on user centric data discovery

Recommendation 1

User-centric design

Understand user's needs and search behaviours to align services with research workflows.

Recommendation 2

Leveraging external research ecosystems

Expand data discovery by integrating with broader digital research infrastructures and global ecosystems.

Recommendation 3

High-quality metadata

Ensure metadata is structured, interoperable, and discoverable to support automated and manual search processes.

Recommendation 4

Flexible search and user-friendly interfaces

Support flexible search modalities that cater to both expert and novice users through intuitive interface design.

Recommendation 1: Under-centric design: Understand user's needs and search behaviours

Effective data discovery services are built upon understanding their users by addressing key questions, e.g.

- **User profiles:** Who are the intended users? What is their familiarity with the discovery system and their knowledge of the repository's subject domains?
- **Search motivations:** Why do users seek data? What are their specific data requirements, what triggers their search within the service, and what are their intended uses for the sought data?
- **Discovery strategies:** How do users typically find data? Do they utilise interactive portal search features, follow literature, browse, or use APIs?

Recommendation 1: Under-centric design: Understand user's needs and search behaviours - Comparison of user study methods

Survey

Description

Collect insights on user needs, requirements, and satisfaction.

Pros

- Cost-effective
- Scalable
- Large user reach
- Quant/Qual data

Cons

- Self-reported bias
- Lacks depth
- Response bias

Tools

SurveyMonkey, Qualtrics, Google Forms

Interview

Description

In-depth conversations with individuals or focus groups.

Pros

- In-depth insights
- Allow for clarification
- Exploratory study

Cons

- Time-consuming
- Resource-heavy
- Skilled needed

Tools

Self-developed scripts, AI transcription tools

Log Analysis

Description

Analysing search patterns and system interactions.

Pros

- Actual behavior
- Objective data
- Specific patterns

Cons

- No "why" context
- Technical expertise
- Lack explanation

Tools

Google Analytics, Matomo, Self-coding

A/B Testing

Description

Comparing two design alternatives for impact.

Pros

- Empirical evidence
- Direct comparison
- Can be large user reach

Cons

- Complex design
- Validity risks
- Limited "why"

Tools

Crazy Egg, Hotjar, Controlled testing

Observation

Description

Watching users in natural or controlled environment.

Pros

- Direct insights
- Reveal unexpected behaviors & pain point

Cons

- Very intensive
- Observer bias
- Generalisation

Tools

Lookback, Silverback, recording

Recommendation 1: Under-centric design: Understand user's needs and search behaviours - Adopt appropriate user study methods

Conceptualisation & design

- Survey
- Interview
- Focus groups

Ideal for gathering user requirements and system architecture design.

New service development & first online version

- Interview
- A/B testing
- Observation

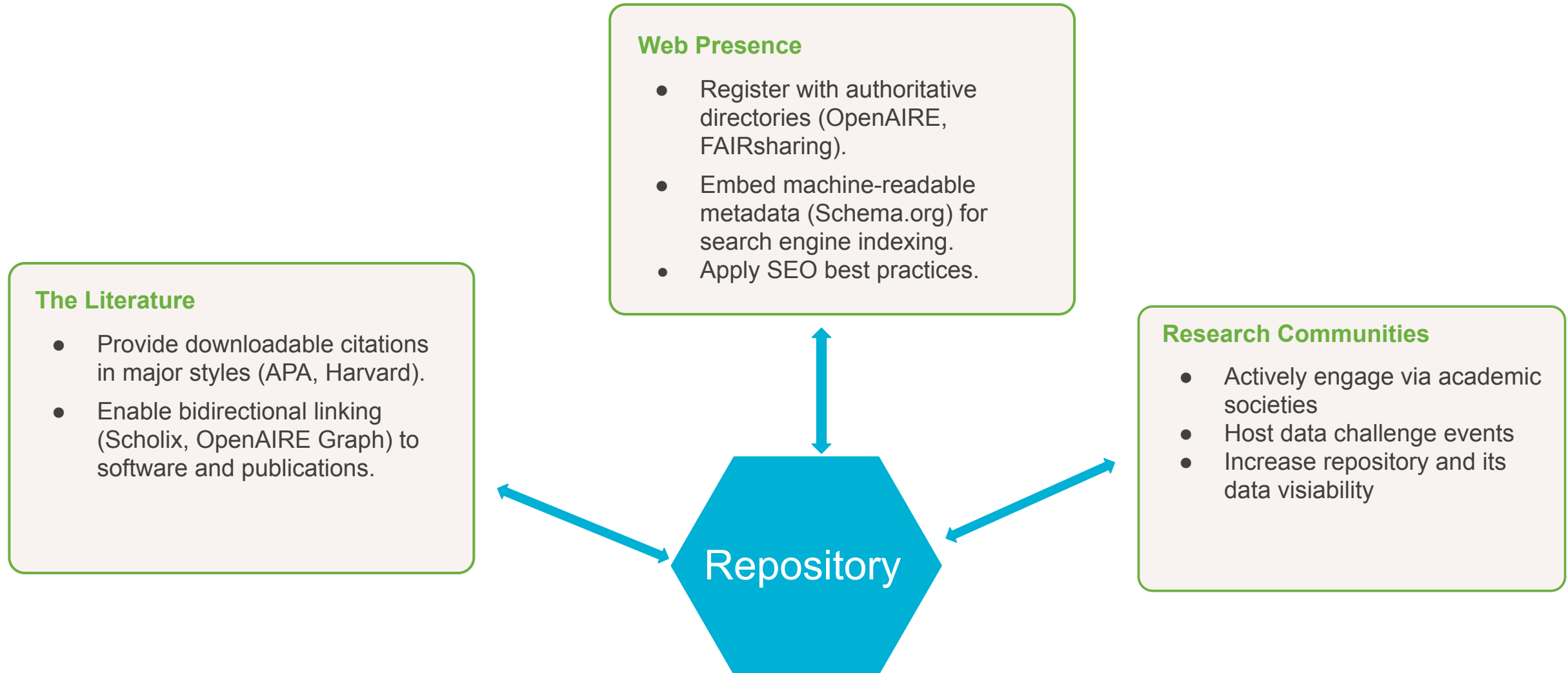
Used for gathering iterative input and testing MVP's usability and effectiveness.

Service improvement & evolution

- Continuous A/B testing
- Observation & Survey
- Interaction log analysis
- Survey & Interview

Crucial for evidence-based refinement of features as the service matures.

Recommendation 2: Leveraging external ecosystems for data discovery



Recommendation 3: High-quality, structured, interoperable, and discoverable metadata

Enhance Coverage & Interoperability

Balanced Collections: Prioritize metadata interoperability across generalist and discipline-specific repositories.

Strategic Partnerships: Regularly exchange metadata for harvesting via APIs.

Standardized Schema and Vocabularies: Use CDIF, Schema.org, DCAT, and DDI-CDI for structure and discoverability; SKOS, XKOS, and OWL for semantics; utilize FAIRsharing for standard identification.

Support Flexible Searching & Discovery

Broad & Specific Terms: Support diverse needs for both human exploratory discovery and automated agent retrieval.

Multiple Classifications: Leverage general research classification alongside domain-specific vocabularies (e.g., MeSH).

Vocabulary Mapping: Connect disparate terminologies to enable richer search filters and query expansion.

Implement Metadata Quality Assurance

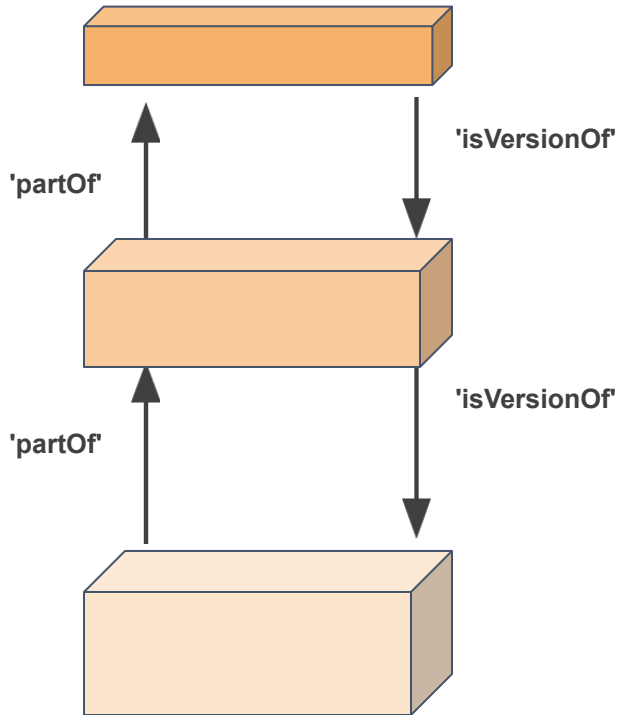
Enforce core metadata attributes: Define a mandatory core set of user-critical, domain-specific metadata attributes.

Support metadata quality assessment: Offer proactive support and tools for data providers to ensure good metadata generation.

Validate quality and FAIRness using assessment tools complemented by user-driven evaluations.

Recommendation 3: High-quality, structured, interoperable, and discoverable metadata - Structuring granularity

The Granularity Stack



Top Layer: Collection-Level (The Box)

Example: National election survey.

Function: Summarizes sub-level datasets. Enables broad exploratory navigation and topic discovery.

Middle Layer: Item-Level (The Contents)

Example: Specific survey result files or bundled publications.

Function: The specific asset retrieved. Enriched by collection-level key terms to boost search engine ranking.

Bottom Layer: Variable-Level (The Ingredients)

Example: Specific data points/variables (age group, suburb, water salinity).

Function: Enables granular reuse, semantic search, and AI extraction.

Recommendation 4: Support Flexible Search and User-Friendly Interface

Versatile Search Options

Provide broad and narrow search pathways, including:

- Simple keyword & complex Boolean queries.
- Structured advanced queries & filters.
- Map-based interfaces for geospatial data.
- Spatio-temporal subsets (e.g., OGC API EDR).

Query Refinement & Semantic Search

Enhance discovery likelihood through:

- Expansion with synonyms via terminology services.
- Integration of domain-specific tools for multidisciplinary datasets.

AI & LLM Integration

Future-proof discovery using:

- Natural language interactions for intuitive search.
- Retrieval Augmented Generation (RAG) for contextual summaries.
- Bridging the gap between research questions and raw data retrieval.

Data Discovery Paradigms: User Requirements and Recommendations for Data Repositories. Data Science Journal, 18(1), 2019. DOI: <https://doi.org/10.5334/dsj-2019-003>

Recommendation 4: Support flexible search and user-friendly interface

The Power of Semantic Search Expansion

Standard Keyword Search

User Query: Bee

A diagram illustrating standard keyword search. A white box on the left contains the text "User Query: Bee". A red dashed line extends from the right side of this box to a thick, vertical brown bar on the right, representing a search engine's index or database.

System Match: Exact text matches only.

Result: **Misses 80%** of relevant scientific datasets. User assumes data doesn't exist.

Semantic Terminology Service

User Query: Bee

A diagram illustrating a semantic terminology service. A white box on the left contains the text "User Query: Bee". A blue curly bracket extends from the right side of this box to a vertical stack of four rounded rectangular boxes on the right. These boxes contain the following text from top to bottom: "Apis", "Honeybee", "European Honey Bee", and "Merops apiaster (predator context)".

Apis

Honeybee

European Honey Bee

Merops apiaster
(predator context)

System Action: Hits a SKOS terminology service (e.g., GFBio).

Result: **Comprehensive discovery** bypassing user's lack of precise domain vocabulary.

Recommendation 4: Support flexible search and user-friendly interface

The Cluttered Search/Presentation



⊗ **Impact:** Cognitive overload. User wastes time mentally tracking duplicates and deciphering the authoritative version.

The Grouped Search



✓ **Impact:** "Overview first, details on demand." Instant clarity on relevance, format, and licensing.

From findability to discoverability

Shift the Mindset

Moving from publishing data to empowering human search behavior.

Invest in the Bridge

High-quality, semantic metadata and an effective discovery system and interface that seamlessly connect researcher to it.

Iterate Continuously

Data discovery systems are never finished; they evolve through constant, user-centric performance assessment.

By adopting user-centered design, repositories can move beyond simply making data available to **genuinely making it discoverable**

Data discovery paradigm shift: Traditional paradigm

The Traditional "Library" Model: Search and Sift

Articulation

Action: Define Research Question and requirements.

Search

Action: Search for data that matches requirements.

Evaluation

Action: Evaluate datasets using metadata.

Access

Action: Follow criteria to access datasets.

Reuse

Action: Follow re-use guidelines to use/cite.

Workflow Friction: Mismatch in cataloging • Inconsistent discovery tools • Reliance on shallow metadata • Opaque access requirements • Poor documentation

The Traditional "Library" Model (Repositories & Google Dataset Search)

- Researchers **formulate keyword queries**
- Systems return a **ranked list of datasets or links**
- Users must manually **click, read, and assess** results

AI-Powered discovery tools: “Find atmospheric datasets with greenhouse gas concentrations and emission estimates”

Here are representative atmospheric datasets that match your query:

👉 “Find atmospheric datasets with greenhouse gas concentrations and emission estimates”

I've grouped them the way an AI-assisted system would—by type and use case.

🌐 1. Emission inventory datasets (estimates)

◆ EDGAR – Emissions Database for Global Atmospheric Research

- Global greenhouse gas emissions (CO₂, CH₄, N₂O, F-gases)
- Coverage: 1970–present, by country, sector, and grid

• Includes:

- annual time series
- spatial emission maps

- Widely used for climate modelling and policy analysis

👉 Provides detailed emission estimates by sector and geography Edgar

◆ VERIFY / EU GHG emissions repository

• Combines:

- bottom-up inventories (e.g. agriculture, energy)
- top-down estimates (from satellite and atmospheric observations)

- Includes NetCDF datasets for modelling

👉 Integrates multiple methods to estimate emissions EU CAP Network

📄 2. Atmospheric concentration datasets (observations)

◆ NOAA atmospheric datasets

- Ground-based and campaign measurements

• Includes:

- atmospheric composition
- greenhouse gas concentrations
- modelling datasets

- Global observational networks

👉 Covers both observations and modelling of atmospheric composition CSL

◆ Global atmospheric GHG concentration dataset

- Long-term record (1860–2021)

• Variables:

- CO₂, CH₄, N₂O concentrations

- Derived from global monitoring networks

👉 Useful for long-term climate trend analysis Hourise

◆ WMO / global monitoring datasets

- Global averages of atmospheric GHG concentrations

- Based on distributed measurement stations worldwide

- Includes high-precision concentration trends

👉 Built from multiple global observation networks European Environment Agency

🛰️ 3. Satellite-based datasets (concentration + emission detection)

◆ NASA EMIT (Earth Surface Mineral Dust Source Investigation)

- Detects CO₂ and CH₄ concentration plumes

- Identifies point-source emissions (e.g. oil & gas leaks)

- High spatial resolution

👉 Enables mapping of emission sources from space JPL Earth Science

🧠 How these datasets fit together

Type	Example	What it provides
Emission inventories	EDGAR, VERIFY	Estimated emissions by sector/location
Observations (ground)	NOAA, WMO	Measured atmospheric concentrations
Satellite data	NASA EMIT	Spatial detection of emission sources

🎯 Key insight (AI-style interpretation)

To answer your query effectively, you typically need **both**:

- **Concentration datasets** → what is in the atmosphere

- **Emission datasets** → where it comes from

👉 Combining them enables:

- source attribution

- validation of emission inventories

- climate modelling and policy analysis



Key differences in traditional and AI-driven discovery

FEATURE	Traditional Discovery	AI-Driven Discovery
Query style	Keywords	Natural language / task-based
Output	Ranked list of datasets	Synthesised answer or recommendation
Metadata use	Explicit filtering	Implicit semantic interpretation
User role	Active evaluator	Guided / assisted decision-maker
Workflow	Iterative querying & browsing	Conversational refinement
Cognitive load	High	Reduced (but shifted to trust)

Role of data repositories in the new AI discovery paradigm

Challenges in AI discovery

- Disintermediation by AI (users bypass portals)
- Shift from search to reasoning
- Inconsistent / shallow metadata
- Lack of relationships between entities
- Complex access & governance (esp. sensitive data)
- Reduced transparency in AI results

What repositories should do

- Position as **trusted metadata providers** powering AI systems
- Provide **structured, machine-actionable metadata** (not just text)
- Improve **metadata quality, completeness, and standardisation**
- Build **connected metadata** (knowledge graphs, PIDs)
- Make **access conditions and consent machine-readable**
- Ensure **provenance, traceability, and explainability**

Repositories remain essential as the trusted foundation enabling AI-driven data discovery.

Explore unknowns in AI data discovery paradigm ...

- What problem in data discovery are we actually trying to solve: recall, precision, usability, or trust?
- Where does traditional keyword search still outperform AI-based discovery today?
- How is the user experience different on each platform for dataset discovery?
- What does “AI discovery” really mean in practice: semantic search, agents, or something else?
- Bias and validation: How can we demonstrate that AI-driven discovery does not bias results toward certain repositories or datasets? What methods can we use to detect and measure bias in AI-driven dataset discovery?
- Prompt Engineering for Scientists: Do we need to teach researchers "Data Prompting" as a new literacy, similar to how we once taught them Boolean operators (AND/OR/NOT)?
- ...

Thank you!



Australian Research Data Commons