# Measuring Interdisciplinarity with Internal University Data: An Operational Guide

A. Rivero and A.I. Scaffold

2026

**Abstract**

Universities possess rich internal data — publication repositories, co-authorship records, PhD supervision logs, and grant databases — that external bibliometric services cannot access. We present an operational measurement protocol that leverages this institutional data to assess interdisciplinary research at the departmental and institutional level. The protocol combines a three-component bibliometric panel (diversity, coherence, cross-field effect) with institutional-only indicators (co-authorship diversity, co-supervision diversity, grant panel diversity). Using a mock departmental scenario, we illustrate how institutional data reveals polymathic breadth versus genuine integration — a distinction invisible to external databases. We identify the main data obstacle (citation data for cross-field effect) and propose a two-component workaround for institutions without external database subscriptions.

## Introduction

Universities increasingly seek to foster and assess interdisciplinary research as part of strategic planning and quality assurance. However, commercial bibliometric databases (Web of Science, Scopus) provide only partial views of research activity, and their subscription costs can be prohibitive. Moreover, these databases lack access to institutional processes — co-authorships, PhD co-supervisions, grant collaborations, and departmental affiliations — that reveal how knowledge integration actually occurs within the university.

This note presents an operational guide for universities to measure interdisciplinarity using primarily internal data. The core approach combines a three-component bibliometric panel — diversity ($\Delta$), coherence ($S$), and cross-field effect ($E$), developed in a companion review (Rivero, 2026) — with institutional-only indicators that capture collaborative structures. We demonstrate the protocol on a mock departmental scenario and provide implementation guidance for data extraction, quality checks, and reporting.

# Data Landscape

## What Universities Have

Most research-intensive universities maintain:

- **Publication repositories**: Institutional repositories or CRIS (Current Research Information Systems) recording all faculty publications with metadata (authors, references, keywords).
- **HR systems**: Departmental affiliations of all staff.
- **PhD records**: Supervision records, including co-supervisors and their departmental affiliations.
- **Grant databases**: Internal records of all grants, including funding source, review panel assignment, and co-investigators.

These data sources are comprehensive within the institution but unavailable to external services.

## What Universities Need (But May Lack)

- **Disciplinary classification of references**: Mapping cited works to subject categories (e.g., Web of Science categories). Some repositories include this; most do not.
- **Similarity matrix between categories**: Required for computing Rao-Stirling diversity. Can be derived from citation data or, more recently, estimated via large language models (Cantone, 2025).
- **Citation data**: Who cites each publication, and from which fields. Essential for computing the cross-field effect ($E$). Available from Web of Science, Scopus, or the open-access OpenAlex, but not from internal systems.

The main data gap is citation data for computing $E$. The protocol below addresses this with a two-component workaround.

# Measurement Protocol

## Step 1: Bibliometric Panel

For each researcher, compute:

**Diversity** ($\Delta$): The Rao-Stirling index over cited references:

$$\Delta = \sum_{i \neq j} d_{ij} \, p_i \, p_j, \qquad d_{ij} = 1 - s_{ij}$$

where $p_i$ is the proportion of references in category $i$, $s_{ij}$ is the pairwise similarity between categories, and $d_{ij}$ is the corresponding distance. This requires reference classification and a similarity matrix.

**Coherence** ($S$): The mean pairwise bibliographic coupling among publications:

$$S = \frac{1}{\binom{n}{2}} \sum_{k<l} \cos(\mathbf{r}_k, \mathbf{r}_l)$$

where $\mathbf{r}_k$ is the reference vector of publication $k$. This requires only internal repository data.

**Cross-field effect** ($E$): The fraction of citations received from outside the researcher's primary category. This requires external citation data.

## Step 2: Institutional-Only Indicators

Beyond the bibliometric panel, compute:

**Co-authorship diversity**: The fraction of publications with at least one co-author from a different department:

$$\text{CoAuth} = \frac{\text{publications with cross-departmental co-authors}}{\text{total publications}}$$

**Co-supervision diversity**: The fraction of PhD students co-supervised with faculty from other departments:

$$\text{CoSup} = \frac{\text{cross-departmental co-supervisions}}{\text{total supervisions}}$$

**Grant panel diversity**: The number of distinct funding review panels from which the researcher has secured grants. This proxies disciplinary breadth of funding capacity.

## Step 3: Interpretation

The combination of bibliometric and institutional indicators discriminates integration from polymathy:

| Profile | $\Delta$ | S | CoAuth | CoSup | Interpretation |
|---|---|---|---|---|---|
| Integrator | High | Moderate-high | High | High | Cross-disciplinary in publications AND processes |
| Polymath | High | Low | Low | Low | Broad references but no collaborative integration |

| Profile | $\Delta$ | S | CoAuth | CoSup | Interpretation |
|---|---|---|---|---|---|
| Specialist | Low | High | Low | Low | Focused, disciplinary researcher |

**Key insight:** A researcher with high $\Delta$ and high grant diversity might appear interdisciplinary from external data, but zero co-authorship and co-supervision diversity reveal polymathic breadth without integration. Institutional data provides this discriminatory power.

This distinction mirrors the input/output separation emphasized in the companion review: high diversity of inputs ($\Delta$) can indicate either multidisciplinary breadth or integrative interdisciplinarity, and only coherence/process indicators can separate the two reliably.

## Demonstration: Mock Department

A small Physics & Materials Science department with 3 researchers illustrates the protocol. We use five Web of Science-style categories and an illustrative similarity matrix:

| ID | Category |
|---|---|
| C1 | Physics, condensed matter |
| C2 | Materials science |
| C3 | Chemistry, physical |
| C4 | Optics |
| C5 | Engineering, electrical |

| | C1 | C2 | C3 | C4 | C5 |
|---|---|---|---|---|---|
| C1 | 1.00 | 0.60 | 0.40 | 0.35 | 0.30 |
| C2 | 0.60 | 1.00 | 0.50 | 0.25 | 0.40 |
| C3 | 0.40 | 0.50 | 1.00 | 0.30 | 0.20 |
| C4 | 0.35 | 0.25 | 0.30 | 1.00 | 0.45 |
| C5 | 0.30 | 0.40 | 0.20 | 0.45 | 1.00 |

Similarity values are illustrative; in practice they would be derived from inter-category citation patterns or estimated via large language models (Cantone, 2025).

All coherence values ($S$) below are illustrative; per-publication reference vectors are omitted for brevity.

**Dr. Emma** (10 years post-PhD, $\mathbf{p}_E = (0.40, 0.30, 0.25, 0.03, 0.02)$): - Biblio panel: $\Delta = 0.42$, $S = 0.55$, $E = 0.22$ - Institutional: CoAuth = 0.20, CoSup = 0.20, Grants = 2 panels - **Profile**: Integrator. Moderate bibliometric diversity reinforced by cross-departmental collaborations and co-supervisions.

**Dr. Farid** (7 years post-PhD, $\mathbf{p}_F = (0.25, 0.25, 0.25, 0.20, 0.05)$): - Biblio panel: $\Delta = 0.58$, $S = 0.05$, $E = 0.08$ - Institutional: CoAuth = 0.00, CoSup = 0.00, Grants = 4 panels - **Profile**: Polymath. High diversity and grant breadth, but zero collaborative integration. Each paper is a disconnected single-field contribution.

**Dr. Greta** (4 years post-PhD, $\mathbf{p}_G = (0.70, 0.25, 0.03, 0.01, 0.01)$): - Biblio panel: $\Delta = 0.28$, $S = 0.75$, $E = 0.08$ - Institutional: CoAuth = 0.00, CoSup = N/A, Grants = 1 panel - **Profile**: Early-career specialist. Focused research program; low $E$ expected at this stage.

Dr. Farid's case is instructive: his grant diversity (4 panels) might suggest strong interdisciplinary capacity, but the institutional indicators reveal this is breadth without depth. The university sees what external databases cannot.

# Implementation Guidance

## Data Extraction

1. **Publication data**: Export from institutional repository with full reference lists. If reference categories are not already assigned, they must be mapped to a disciplinary taxonomy (e.g., Web of Science categories). This is the main data cleaning task.
2. **Co-authorship affiliations**: Extract from HR system. Flag all publications where co-authors belong to different departments.
3. **PhD supervision records**: Extract from graduate office. Identify co-supervisions and co-supervisors' departmental affiliations.
4. **Grant records**: Extract from research office. Map each grant to the funding agency's review panel taxonomy.
5. **Citation data (if available)**: Import from OpenAlex (free), Web of Science, or Scopus. Match citing papers to categories for $E$ computation.

## Quality Checks

- **Missing reference categories**: If >25% of references lack category assignments, $\Delta$ estimates will be unreliable (Nakhoda et al., 2023). Report coverage statistics.
- **Early-career N/A values**: Researchers with <5 years post-PhD or <10 publications should have confidence intervals reported alongside point estimates.

- **Threshold calibration and ambiguity**: The profile cutoffs used in this note are illustrative. Institutions should calibrate them to local distributions and mark cases whose confidence intervals cross category boundaries as "ambiguous, needs qualitative review."
- **Departmental affiliation changes**: Researchers who have changed departments mid-career may have artificially inflated co-authorship diversity. Stratify by time period if needed.

### Reporting

Present indicators at three levels:

1. **Individual**: Provide each researcher with their profile (Δ, S, E, CoAuth, CoSup, Grants). This supports self-assessment and career planning.
2. **Departmental**: Report distributions and medians across the department. Identify outliers for strategic discussion.
3. **Institutional**: Aggregate across departments for university-wide strategic planning. Compare distributions across fields (e.g., STEM vs. Humanities).

**Do not rank researchers by a single composite score.** The indicators are multidimensional by design. Ranking collapses the information and incentivizes gaming (Rafols, 2019).

### Workaround When Citation Data Unavailable

If the institution lacks access to citation databases, compute a **two-component internal panel** (Δ, S) plus institutional indicators (CoAuth, CoSup, Grants). This still discriminates integrators from polymaths:

| Profile | Δ | S | CoAuth | CoSup | Interpretation |
|---|---|---|---|---|---|
| Integrator | High | Moderate-high | High | High | Integration via references AND processes |
| Polymath | High | Low | Low | Low | Broad but disconnected |
| Specialist | Low | High | Low | Low | Focused, disciplinary |

The loss of $E$ reduces discrimination power for assessing cross-field *impact*, but integration *capacity* is still captured.

## Limitations and Extensions

The protocol has several limitations. First, it assumes the institution has clean, structured data. Many universities' CRIS systems are incomplete or inconsistent. Data cleaning is a non-trivial prerequisite. Second, the protocol does not address quality — it characterizes the type of interdisciplinarity, not whether it is good. Quality assessment requires independent evaluation (peer review, citation percentiles, etc.). Third, the classification thresholds (e.g., "high" $\Delta \geq 0.40$) are illustrative and should be calibrated against the institution's empirical distributions before operational use. Fourth, confidence intervals can be wide for individual-level profiles when publication counts are small; operational classification should therefore use interval-aware rules rather than point estimates alone.

Extensions could include text-based indicators (semantic similarity of titles/abstracts across a researcher's portfolio), teaching-based indicators (cross-departmental teaching assignments), and temporal analysis (tracking how $\Delta$, $S$, $E$ evolve over a researcher's career).

# Conclusions

Universities possess data that external bibliometric services cannot see. Leveraging this internal data — co-authorships, PhD co-supervisions, grant collaborations — provides discriminatory power that purely bibliometric measures lack. The measurement protocol presented here combines a three-component bibliometric panel with institutional-only indicators to distinguish genuine cross-disciplinary integration from polymathic breadth. The main implementation obstacle is citation data for computing cross-field effect; a two-component workaround (diversity + coherence + institutional indicators) provides substantial discrimination power without external dependencies. Institutional deployment requires clean data, careful quality checks, and resistance to the temptation to reduce multidimensional profiles to single-number rankings.

# References


- Cantone, G. G. (2025). Estimation of disciplinary similarity with large language models. *Scientometrics*, 130(10):5345–5373.
- Nakhoda, M., Whigham, P., and Zwanenburg, S. (2023). Quantifying and addressing uncertainty in the measurement of interdisciplinarity. *Scientometrics*, 128:6107–6127.
- Rafols, I. (2019). S&T indicators in the wild: contextualization and participation for responsible metrics. *Research Evaluation*, 28(1):7–22.
- Rivero, A. (2026). Measuring interdisciplinarity: A multi-component indicator panel for research evaluation. Manuscript.