What each funder funds: specialization, complementarity, and the surprising temporal stability of funder field-portfolios in a reconciled NIH/NSF/EC grant→output graph

Author: Bucket Foundation · research-atlas working group **Version:** 1.0 (preprint draft) · **Date:** 2026-06-24 **Corpus:** research-atlas v0.1.0 (2026-06-22 build) — 1,670,434 grants / 75 funders / 470,269 grant→work edges **DOI:** [10.5281/zenodo.20774322](https://doi.org/10.5281/zenodo.20774322) (concept; this study = research-atlas v0.4.0, DOI minted on next release) **Reproducibility:** every number in this paper is emitted by `analysis/specialization/run.py` into `analysis/specialization/results.json` and pinned by `tests/test_funder_specialization.py`, which re-derives each constant from the live DuckDB and fails if the prose and the data diverge. `results.json` is the authoritative source for every statistic quoted below.

Abstract

Paper 01 in this series characterized *who* gets research grants (institutional concentration), *who shares* the resulting papers (co-funding), and *how much output accompanies a dollar* (the funding→output rate). It did not ask what is, structurally, a prior question: **what does each funder actually fund, and how distinctively?** Here we answer that on the same reconciled graph by mapping every funder's linked output to the 26 OpenAlex top-level fields — purely on *distinct-work counts*, with no dollar column anywhere — and measuring three things. **(1) A specialization gradient.** Across the 23 funders with enough linked output to estimate a portfolio, the field-concentration of that portfolio (an HHI over the 26 fields) spans a clean 5× range, from the two pure generalists — the EC (HHI **0.092**) and NSF (**0.095**), each spreading across all fields with no single field above 16% — to the hyper-specialist NHGRI (HHI **0.466**), which puts **66%** of its linked output in a single field (Biochemistry, Genetics & Molecular Biology). The NIH Institutes occupy the specialist end, each dominated by Medicine or its disease-aligned field. **(2) Complementarity, recovered from data.** Cosine similarity of funder field-share vectors recovers the agency map with no labels: the NIH Institutes form a tight cluster (internal mean cosine **0.819**), NSF is the maximal outlier (mean cosine to the NIH cluster **0.385**; the single most distinct pair is NIDCD↔NSF at **0.217**), and the EC sits in between as a partial bridge (**0.621**). **(3) Specialization is a stable fingerprint.** Comparing each funder's portfolio HHI in an early window (2016–2019) versus a late window (2021–2024), the rank order is almost perfectly preserved (Spearman ρ = **0.973**) and the mean absolute change in HHI is only **0.011** — funders do not meaningfully re-specialize over the window. The one *aggregate* compositional move we detect — Physical Sciences' share of funded output rising **+3.83pp** (95% CI [+2.87, +4.85]) from 2016 to 2024 — **vanishes under mix control**: within NSF alone the Physical-Sciences share *fell* 0.88pp, and the funder-family mix barely moved, so the aggregate wobble is a composition/coverage-endpoint artifact, not a secular shift in what funders buy. We state the scope limit plainly throughout — field assignment requires output edges, which exist for NIH/NSF/EC only, so every funder here is an NIH IC, NSF, or an EC body — and we release all code and a Zenodo-ready metadata record.

---

1. Introduction

The first paper on the research-atlas graph (`docs/papers/01-funding-landscape/`) was about the *distribution* of funding: how unequally grants are spread across institutions and countries, which funders co-fund the same

works, and how the volume of linked output per dollar varies. The second (`docs/papers/02-paper-ranking/`) left the funding graph entirely and studied paper-recommendation methods on the OpenAlex citation graph. Neither asked the question a science-policy reader asks first when they look at a set of funders side by side: **do these agencies fund the same science, or different science, and does that division of labor hold still over time?**

This is the question of *portfolio structure*. A funding system can be highly concentrated at the institution level (paper 01's finding) while being either **redundant** (every funder chasing the same hot fields) or **complementary** (funders covering disjoint parts of the research frontier). The two have opposite policy implications, and they are invisible to a concentration analysis because concentration measures *who* receives, not *what* is received. Likewise, a funder can hold a fixed mandate on paper while its *realized* portfolio drifts year to year as fashions change — or it can be a stable fingerprint. These are empirical questions, and the reconciled grant→output graph can answer them.

We answer them with three measurements, all built on *distinct-work counts* so that none of the dollar-attribution noise documented for paper 01 can touch the result:

1. 1. **A specialization gradient** — how field-concentrated each funder's realized

portfolio is, via a Herfindahl–Hirschman Index (HHI) over the 26 OpenAlex top-level fields.

1. 2. **Complementarity** — the cosine similarity between funders' field-share

vectors, which recovers, with no labels, whether funders overlap or divide the space.

1. 3. **Temporal stability** — whether each funder's specialization is a fixed

fingerprint or drifts, by comparing portfolio HHI across an early and a late window, and whether the *aggregate* field composition of funded output shifts over 2016–2024 once the obvious confounds (funder-mix change, coverage endpoints) are controlled.

The result is a coherent structural picture — a stable, complementary division of scientific labor among public funders — and an honest null on the temporal side that strengthens rather than weakens the cross-sectional story.

---

2. Data and methods

2.1 The corpus and the field mapping

We use research-atlas v0.1.0 (the 2026-06-22 build): 1,670,434 grants from 75 funders, with 470,269 `grant_work` acknowledgement edges linking grants to the 278,839 works that cite them, and an OpenAlex topic taxonomy (topic→subfield→field→domain). A *grant* carries no field of its own; it acquires one only through the works that acknowledge it. We therefore map each funder to fields by the chain **funder → grant → work → topic → field**, rolling each work's OpenAlex topics up to the 26 top-level **fields** (and the 4 **domains**) by walking `parent_atlas_id` recursively. Every quantity is a count of **distinct works**, deduplicated per (funder, field).

2.2 The scope limit (stated first, repeated throughout)

`grant_work` edges were ingested for **NIH, NSF and the EC only** (UKRI and the philanthropies have no output edges yet). Field assignment therefore exists only for those funders, so **every funder in this study is an NIH**

Institute/Center, NSF, or an EC body (EC or ERC). This is the single most important limitation and we repeat it wherever a result depends on it. Within that scope the analysis is complete and the counts are robust.

2.3 Robustness discipline (the honesty guardrail)

The graph has two documented noise sources (docs/GRAPH.md §2.5): recipient fuzzy-match noise and shared-grant dollar double-counting. **Both affect only per-organization/funder dollar sums.** This study uses *no dollar column at all* — every statistic is a distinct-work count — so neither noise source can reach any number reported here. We restrict field shares to works published in **2016–2024**, the window with near-complete OpenAlex coverage in the corpus (2015 ramps in, 2025 is partial; the same window as paper 01). A funder is admitted to the specialization cross-section only if it has **≥ 1,000 linked works** in the window (≥ 500 per half-window for the stability comparison), so each portfolio is estimable. This yields **23 funders** out of the 33 with any output edges.

2.4 Statistics

We measure portfolio concentration with the **Herfindahl–Hirschman Index** *\text{HHI}=\sum_f s_f^2* on a funder's field shares *s_f* (1/26 ≈ 0.038 if a funder spread perfectly evenly across all fields; 1.0 if it funded a single field). Funder **complementarity** is the cosine similarity between funders' 26-dimensional field-share vectors. **Stability** is the Spearman rank correlation and mean absolute HHI change between the early (2016–2019) and late (2021–2024) windows. The **aggregate composition shift** is the endpoint difference in each domain's share, with a **2,000-sample percentile bootstrap 95% CI** obtained by resampling the per-work domain assignments in the two endpoint years. All queries are read-only and live in `analysis/specialization/funder_specialization.py`.

---

3. Results

3.1 Funders span a clean 5× specialization gradient

Ranking the 23 funders by the HHI of their field portfolio produces a clean, interpretable gradient (**Figure 1**). At the **generalist** end sit the two all-field agencies: the **EC** (HHI **0.092**) and **NSF** (HHI **0.095**), each spreading across all 26 fields with its single largest field — Engineering for both — holding only **13.7%** (EC) and **15.6%** (NSF) of its linked output. The ERC, the EC's investigator-led arm, is the next-most-general (HHI 0.114, top field Physics & Astronomy at 24%). At the **specialist** end sit the NIH Institutes, each one dominated by a single field: most by **Medicine** (NIDDK 45%, NHLBI 48%, NCI 42%), the basic-science institutes by **Biochemistry, Genetics & Molecular Biology** (NIGMS 49%), and the extreme case, the **National Human Genome Research Institute (NHGRI)**, with HHI **0.466** and **66%** of its linked output in a single field (Biochemistry/Genetics) — exactly what a genome institute should look like. The 5× span from 0.092 to 0.466 is not noise; it is the realized mission of each agency, recovered from acknowledgement data alone.

Figure 1. Funder specialization gradient: portfolio HHI over the 26 OpenAlex fields, generalist (top) to specialist (bottom). Navy = US (NSF + NIH ICs), red = EC/supranational. Each bar is labelled with that funder's dominant field and its share. Distinct linked works, 2016–2024.

3.2 Complementarity: the agency map falls out of the data

Computing the cosine similarity between every pair of funders' field-share vectors recovers the structure of the public funding system **with no labels** (**Figure 2**). Three facts:

- **- The NIH Institutes are a tight cluster.** Their internal mean pairwise cosine

is **0.819** — the disease-aligned institutes (NCI, NHLBI, NIDDK, NIA, NIAID, NIAMS, …) all load heavily on Medicine and neighbouring biomedical fields, so they look alike. The single most similar pair is **NCI↔NIDDK (0.995)**.

• - **NSF is the maximal outlier.** Its mean cosine to the NIH cluster is **0.385**,

and the single most distinct pair in the entire matrix is **NIDCD↔NSF (0.217)** — a hearing/communication-disorders institute versus an all-field physical-and- engineering-sciences agency share almost no fields. NSF is *complementary* to the NIH system, not redundant with it.

• - **The EC is a partial bridge.** Its mean cosine to the NIH cluster is **0.621**

— higher than NSF because the EC's broad biomedical portfolio overlaps the NIH fields, lower than the NIH-internal cohesion because the EC also funds the physical and engineering sciences NSF covers.

This is the **division of labor** of the public research economy, drawn directly from what each funder's grants produce: a dense biomedical cluster (NIH), a disjoint physical/engineering generalist (NSF), and a broad European bridge (EC/ ERC).

Figure 2. Funder portfolio similarity: cosine between 26-field share vectors, ordered with the generalists (NSF, EC, ERC) first and the NIH Institutes after. The bright NIH×NIH block (internal mean cosine 0.819) and the dark NSF column (mean 0.385 to NIH) are the complementarity structure.

3.3 Specialization is a stable fingerprint

A funder's place on the gradient is **not a transient**. Comparing each funder's portfolio HHI in an early window (2016–2019) to a late window (2021–2024) (**Figure 3**), the rank order is almost perfectly preserved — **Spearman ρ = 0.973** (Pearson 0.990) — and the **mean absolute change in HHI is only 0.011**, with the largest single move (NIGMS, −0.039) still small relative to the 0.37-wide gradient. The generalists stay general (NSF 0.097→0.095, EC 0.093→0.092), the specialists stay specialized (NHGRI 0.480→0.460), and nobody crosses the cluster boundary. Realized funder specialization behaves like an institutional **fingerprint**: set by mandate and infrastructure, essentially fixed over the observation window.

*Figure 3. Specialization stability: each funder's portfolio HHI in 2016–2019 (x) vs 2021–2024 (y). Points hug the y = x line (Spearman ρ = 0.973; mean

ΔHHI	= 0.011) — funders do not re-specialize.*

3.4 The one aggregate shift is a composition artifact, not a trend

Given how stable individual portfolios are, what about the *aggregate* field composition of all funded output? Across 2016–2024 the four-domain split moves modestly, and only one move clears its bootstrap CI cleanly: **Physical Sciences' share rises +3.83pp (95% CI [+2.87, +4.85])**, while Life Sciences falls −2.81pp ([−3.72, −1.89]) and Health and Social Sciences are roughly flat. Taken at face value this looks like a swing toward the physical sciences.

It is not a real swing in what funders buy, and the graph lets us prove it (**Figure 4**). Two controls:

• - **Funder-family mix barely moves.** The NSF/EC family's share of linked works

goes 63.6% → 65.3% over the window — a 1.7pp drift, far too small to drive a 3.8pp domain swing.

• - **Within NSF, the trend reverses.** Holding the funder fixed (NSF only,

mix-controlled), the Physical-Sciences share of NSF's own output *falls* 0.88pp over the same window. The agency most responsible for physical-science output did not tilt toward it.

The aggregate +3.83pp is therefore an artifact of the partial-coverage **endpoint year** (2024 has the fewest linked works in the window and an idiosyncratic funder mix), not a secular change in scientific priorities. The honest conclusion is the *null*: over 2016–2024 the composition of funded output — like the individual funder portfolios in §3.3 — is **broadly stable**, and the apparent shift dissolves under the simplest mix control.

Figure 4. Left: aggregate domain composition of funded output by year — Physical Sciences (navy) appears to rise at the 2024 endpoint. Right: within NSF alone (mix-controlled) the Physical-Sciences share is flat-to-falling, showing the aggregate move is a composition/coverage artifact, not a within-funder trend.

---

4. Discussion

Three structural facts emerge, each robust to the graph's known noise because the study touches no dollar column.

Public funders divide the scientific space; they do not duplicate it. The specialization gradient (§3.1) and the complementarity structure (§3.2) together show a system organized as a *division of labor*: a dense biomedical NIH cluster of disease-and-organ-aligned institutes, a complementary physical-and-engineering generalist (NSF), and a broad European bridge (EC/ERC). The fact that the agency map can be recovered with **no labels**, purely from the field composition of each funder's acknowledged output, is a validation of the reconciliation — the ROR/ORCID/topic merge produced field vectors faithful enough to rebuild the known structure of the funding system.

Specialization is durable, not faddish. The fingerprint stability (§3.3, Spearman 0.973) says realized funder portfolios are set by mandate and standing infrastructure, not by year-to-year fashion. This matters for science policy: funders are not chasing trends, so a topic's rise or fall in the literature is mediated by *which* funder's stable portfolio it falls into, not by funders collectively pivoting. It also bounds the interpretation of paper 01's field-dynamics result: the COVID/AI risers it found are riding through stable funder portfolios, not the product of funders re-specializing.

Aggregate "shifts" demand mix control. The §3.4 result is a cautionary methodological point as much as a finding. An analyst looking only at the aggregate domain shares would report a real-looking, CI-clearing +3.8pp swing toward the physical sciences over 2016–2024. The reconciled graph — by letting us hold the funder fixed — shows that within the responsible agency the trend is flat to slightly negative, and the aggregate move is a coverage-endpoint composition artifact. Compositional time series over a heterogeneous, unevenly-covered corpus must be decomposed before they are interpreted.

---

5. Limitations

We state these plainly; none is hidden.

1. 1. **Field assignment is output-mediated and funder-bounded.** A grant has a field

only through the works that cite it, and `grant_work` edges exist for NIH, NSF and the EC only. Every funder analyzed is an NIH IC, NSF, or an EC body; UKRI and the philanthropies cannot be placed on this gradient until their output edges are ingested. This is the paper's most important caveat (§2.2).

1. 2. **Acknowledgement linkage is biased by output type.** Translational/clinical/

infrastructure funding under-links to papers (the §3.4 mechanism in paper 01), so a funder's *realized* field portfolio is the portfolio of its paper-producing work, which may differ from its full mandate. The specialization gradient is therefore a gradient of *publication-visible* specialization.

1. 3. **The 26-field taxonomy is coarse.** HHI over 26 fields captures macro

specialization (genome institute vs all-field agency) but not within-field specialization (e.g. two cancer programs in different subfields read as identical). A subfield-level repeat is a natural extension.

1. 4. **Stability is measured over one decade.** Spearman 0.973 holds over 2016–2024;

it does not speak to longer-horizon mandate changes (e.g. the creation of a new institute), which are out of the window.

1. 5. **No causal claims.** This is a structural/descriptive study of realized

portfolios; nothing here identifies why a funder specializes as it does.

---

6. Reproducibility statement

Every number in this paper is computed by `analysis/specialization/run.py` directly from `research_atlas.duckdb` and written to `analysis/specialization/results.json`; the figures are rendered by the same script into `analysis/specialization/figures/`. The constants quoted in the text are pinned by `tests/test_funder_specialization.py`, which re-derives them from the live database and fails if the prose and the data diverge. To reproduce from the published graph:

```
pip install -e .                                # duckdb, pandas, numpy, matplotlib
python analysis/specialization/run.py           # results.json + figures (idempotent)
python -m pytest tests/test_funder_specialization.py -q
```

To rebuild the graph itself from source connectors, see `docs/GRAPH.md §3`.

Data availability. The graph is published as parquet under `data/processed/` (manifest: `data/MANIFEST.json`), license CC-BY-4.0 (data) / MIT (code). Sources: NIH RePORTER, NSF Awards, EC CORDIS, OpenAlex, ROR, ORCID.

Author contributions / COI. research-atlas is developed under the Bucket Foundation open-data program. No competing financial interests.

Appendix A. Headline numbers (machine-checked)

Quantity	Value	Source field in `results.json`
Funders with output edges	33	`coverage.n_funders_with_output_edges`
Funders in specialization cross-section	23	`coverage.n_funders_in_specialization`
	202,909	`coverage.distinct_linked_works_in_window`

Distinct linked works in window (2016–24)				
Generalist HHI (EC)	0.092	`specialization_gradient[0].hhi`		
Generalist HHI (NSF)	0.095	`specialization_gradient` (NSF)		
Specialist HHI (NHGRI)	0.466	`specialization_gradient[-1].hhi`		
NHGRI top-field share	66% (Biochem/ Genetics)	`specialization_gradient[-1].top_field_share`		
NIH-IC internal mean cosine	0.819	`funder_similarity.nih_ic_internal_mean_cosine`		
NSF mean cosine to NIH cluster	0.385	`funder_similarity.nsf_mean_cosine_to_nih`		
EC mean cosine to NIH cluster	0.621	`funder_similarity.ec_mean_cosine_to_nih`		
Most distinct pair (NIDCD↔NSF)	0.217	`funder_similarity.most_distinct[0]`		
Stability Spearman ρ (early vs late HHI)	0.973	`specialization_stability.spearman_rank_corr`		
Stability mean	ΔHHI		0.011	`spec`
Physical Sciences endpoint shift	+3.83pp [+2.87, +4.85]	`composition_shift.endpoint_shift`		
Within-NSF Physical Sciences shift	−0.88pp	`composition_shift.within_nsf_domain.physical_sciences_shift_pp`		