Summary: approach and main findings

Registration section 6, item 6. Entry for the *Mapping CORDIS projects to SDGs* competition. Written under `cordis-sdg-registration-v10.md`, sha256 `c5f2a26ce6049e801a7e757efc2cbe4253689987fad761f1d97cfb5b0ad68fb2`.

Every figure and hash on this page is quoted from a committed artefact, and `scripts/test_reported_figures.py` fails if any of them stops matching.

The approach

The method was registered and ratified by hash before any pipeline code was written, so that results could not move the method. Changing it afterwards takes an amendment and a new hash, in a room where each change is recorded.

The decision path is pure Python standard library, with no model in it. A project's text and its EuroSciVoc categories are matched against a vocabulary built from the SDG target texts alone, under a written extraction rule, plus a hand-justified crosswalk from research categories to targets. Matches are literal, with word-boundary lookarounds, so the same input gives the same bytes on any machine. Scores are summed per target, thresholded, and banded by how much independent evidence was found.

No project text reached the vocabulary or the crosswalk. They were built from the taxonomy and the category list before any project was read.

The evaluation was sealed before it was run. 150 projects were drawn under a recorded seed and split 50 development, 100 held out. Labels for the held-out 100 were made by a different model family from the seat that built the pipeline, uploaded before the pipeline ever saw them. The parameters were frozen, and a four-condition guard refuses to score the held-out set unless the freeze, the parameters, the hand-off and every rule file are exactly what the freeze recorded. **The held-out 100 was scored once.**

Every row of the output carries its own provenance: the snapshot hash, the pipeline commit, the matched phrases with the field each was found in, which of the five source fields were present, and the confidence band.

The findings

Against the held-out 100, scored once under the freeze:

	TP	FP	FN	precision	recall	F1
target level	21	26	74	**0.447**	**0.221**	0.296
goal level, secondary	29	42	50	0.408	0.367	0.387

The reference is 95 pairs over 100 projects, 31 of which carry no target.

The main finding is not the precision. It is that recall is bounded by the method, and the bound is measured rather than estimated. Relaxing every threshold until any evidence at all counts gives a ceiling of **0.442**, because **53 of 95 reference pairs have no evidence at any threshold**: the link the labeller made is not present in the project's words at all. **No tuning reaches those pairs.** What the ceiling measures is how much of the project-to-SDG relation is inference rather than vocabulary, and the answer is most of it.

The failure is silence rather than noise. Of the projects where the pipeline and the reference disagree, most fail because the pipeline found nothing at all: `64 disagreeing, 46 recall-only`.

On the development 50, two independent labellers agreed at target level at `Cohen's kappa 0.677`. There is no equivalent figure for the held-out 100, because it has one labeller and one cannot disagree with himself.

Over the whole snapshot, 23,451 Horizon Europe projects:

rows in the mapping table	**28,833**
with at least one target	9,394
goal level only	6,882
nothing at all	**7,175**
bands, of 14,776 target assignments	9,228 low, 2,923 medium, 2,625 high

Nearly a third of the corpus receives nothing, and most of what is assigned sits in the weakest band. Both are reported rather than tuned away: the registration set no acceptance threshold, so a poor number is submitted with its explanation.

`docs/limitations.md` carries the limitations in full, four worked failure cases with the evidence cells that produced them, and five possible improvements with what each would not fix.

Reproducing it

registration	**`docs/cordis-sdg-registration-v10.md`**
frozen pipeline commit	`0246412b44f722567feaa2e6c1b6e695de86c923`
snapshot	`bf558b0c04db70acc9a83b5badaeae959f7f85ad8ccff472ab68d22a03dce2f3`
taxonomy	`a0c61a4f5b089863d6c8a60a83671495fb530969fa1a0e04228e3a01cc072b94`
parameters, values digest	`3808845b553dfda751bf4a3e5c87bd17f32b6ca8c195a0c54321ed25e03e09e1`
evaluation hand-off	`1603a0d7eb351ac92cf17bc488e7a1180cf40305748120076888ac30383a43cc`
reference label set	`c9fa780aae00c2c0927577cbbcb3bedbaf0f0f5cbd9b191b1bfe4355a1a25532`
frozen evaluation output	`0938fe34e41cd5d8ddccfba69238481fdf6470e54046104ad7a3535da6743709`
metrics	`babd37415581ce82ecb09c4501052ce811e834e57dad7c6735389ce321dd8a2e`
mapping table	`181d378a3078014193f3ae4f684b00b375ed62f3d24b2b34c060f6457b4a00d3`
figure counts	`03d79580081ca40bf0d1d25b5279a57880ddf09d750490553593387126d92f6f`

`README.md` carries the step-by-step block, which has been run from a clean clone by more than one seat.

Live prototype: https://cordis-sdg-prototype.vercel.app/