Dtangle CORDIS SDG Mapping

Limitations, uncertain cases and possible improvements

CORDIS SDG competition, data.europa.eu

Prepared by Dtangle SA

Version 1.0

30 September 2026

Table of Contents

1. How reliable are the links? Measured against a reference reading 2

2. What the method cannot see 2

3. Uncertain cases 3

4. Possible improvements, most useful first 3

1. How reliable are the links? Measured against a reference reading

60 projects drawn at random across all parts of the programme (8 ERC, 8 MSCA, 7 climate-energy-mobility, 6 food and environment, 6 digital and industry, 6 health, 5 EIC, 4 missions, 4 culture and society, 3 security, 3 other) were read in full and annotated with their goals and targets, blind to the method's output. They were never used to build or tune the method (hold-out set). A second set of 150 projects (pilot) gives the same picture. Goal level:

Links kept	**Share confirmed by the reader (precision)**	**Share of the reader's goals found (recall)**
High only (score 4.00 or more)	0.71 (pilot 0.81)	0.43 (pilot 0.40)
Medium only (2.00 to 3.99)	0.45 (pilot 0.43)	
Low only (below 2.00)	0.11 (pilot 0.12)	
Medium + High (score 2.00 or more)	0.56 (pilot 0.60)	0.76 (pilot 0.68), F1 0.65
All links	0.33 (pilot 0.34)	0.92 (pilot 0.84)

What this means. The confidence level does its job: the higher the level, the more often a careful reader agrees. High links are right about three times in four. Low links almost never match the reader: they record that a goal is *mentioned*, not that the project is *about* it. Dtangle therefore recommends the Medium + High view (score 2.00 or more) for most uses: it finds three quarters of the goals a reader assigns, and in the hold-out none of the 17 projects the reader judged goal-free receives a Medium or High link. Low links are kept, clearly marked, for recall: with them the method finds 92 % of the reader's goals. The exact target is harder than the goal (target-level F1 0.51 at score 2.00 or more).

Caveats. The reference reading was made by one annotator, an AI model (Claude) working blind to the method's output; it has not yet been checked by a human second reader. An earlier rule-based version of the method reached F1 0.72 on the same 60 projects, but linked 14,724 projects instead of 17,437 and gave no concept-level evidence. All figures are in the file Validation_reference_reading_260924.csv.

2. What the method cannot see

- **Some targets are not research topics.** Targets about finance, trade and aid to developing countries are served by EU development and trade instruments, not by Horizon Europe research grants. Few projects write about them: 8.a (Aid for Trade) matches no project, 10.b (development assistance) one, 10.c (remittance costs) 3, 17.4 (debt sustainability) and 17.5

(investment promotion) 6 each. This is why SDG 1 (763 projects) and SDG 17 (931) are the smallest goals: the low counts describe the portfolio, not a failure of the method.

- **Only what the project wrote at the start is read.** For each project the method reads the title, keywords, call topic, acronym and objective (the public summary written when the grant was signed), in English. Deliverables, periodic report summaries and publications exist in other CORDIS datasets but were not in the extract used, so a goal that appears only in a project's results is missed.
- **Projects with no link.** 6,014 projects get no link; 4,987 of them are ERC (2,492) or MSCA (2,495) fundamental-research grants whose summaries mention no societal goal.
- **Words, not meaning.** A word used in another sense can create a link: "deprivation" in *androgen deprivation therapy* (a cancer treatment) linked DE-ESCALATE to poverty (SDG 1, score 2.00); "urban air" in *Urban Air Mobility* (drones, air taxis) linked OPENSRUM to air quality (SDG 3, score 2.00). Conversely an idea written in unforeseen words is missed.
- **Writing, not impact.** A high score means a project writes about a goal, not that it will deliver on it.
- **Length.** Scores are not normalised for text length; repetition is capped at +0.50, but longer summaries still have more chances to match.
- **Weights are a judgement**, fixed before the run, not fitted. They sit in one file and every figure can be recomputed with other values.

3. Uncertain cases

Case	Links	Example
Low confidence (score below 2.00)	28,536 (59 %)	precision 0.11 on the hold-out (see above)
A single mention in a method sentence (score 0.80)	4,733	PFAS-ITOX (immunotoxicity of PFAS in children) linked to SDG 11 because "airport" appears once
Commission classification only, no text evidence	3,880	GAMES (mass extinctions in the Mesozoic) linked to SDG 11 because its budget counts 100 % towards the EU clean-air tracker
Near the Medium threshold (score 1.80 to 2.20)	4,929	the level could flip with a small change of weights
Known false positives	2	DE-ESCALATE and OPENSRUM (above)

A sample of 77 of these cases (the 2 false positives and 25 of each other kind, fixed random seed), with their explanation and evidence, is in Uncertain_cases_sample_260924.csv.

4. Possible improvements, most useful first

1. Drop or down-weight the weakest Low links (single method sentence, structural signal only), the group with the lowest measured precision.
2. A human second reading of the 210 annotated projects, then fit the field weights on it instead of setting them by judgement.
3. Tighten the aliases behind the false positives and add guards learnt from that reading.

4. Read more text: report summaries, deliverables and publications, and non-English summaries.
5. A semantic path (embeddings) to catch unforeseen wording, keeping the lexical evidence as the explanation.
6. A length-adjusted score beside the raw one.