Dtangle CORDIS SDG Mapping

Explainable, concept-based mapping of Horizon Europe projects to the EU SDG taxonomy: description document

CORDIS SDG competition, data.europa.eu

Prepared by Dtangle SA

Version 1.1

30 September 2026

Table of Contents

1. About Dtangle 3

Who we are 4

What we do 4

Where it is used 4

Dtangle in this project 4

Team 5

2. Summary 5

What sets the solution apart 5

Deliverables 6

3. Results 6

How a project gets linked to a goal 6

Key findings 7

1 · Goals 8

2 · Targets 11

3 · Concepts 15

4 · Evidence and coverage 19

5 · Go further 21

4. Methodology 21

At a glance 21

Data sources 22

Pipeline 22

Step 1 · The SDG → Target → Concept layer 22

Step 2 · Aliases and guards 23

Step 3 · Matching and weights 23

Step 4 · Structural signals 26

Step 5 · Scores and rankings 27

Step 6 · Funding: gross and allocated 28

Step 7 · White space analysis 30
Two BigNum codes 34
Worked example 35
Traceability 35
Reproducibility 36
5. Validation, limitations and uncertain cases 38
How reliable are the links? Measured against a reference reading 38
What the method cannot see 38
Uncertain cases 39
Possible improvements, most useful first 39
Earlier iterations and integrity checks 40
6. The project-to-SDG mapping table 41
7. The prototype: CORDIS SDG Explorer 42
8. Future work 45
1 · The question the 2 codes leave open 45
2 · A third code: the neighbourhood index 45
3 · What it cannot promise 45
4 · What it would be for 46
5 · How we would test it 46
Contact ..... **Erreur ! Signet non défini.**

1. About Dtangle

Dtangle is a Swiss deep-tech company born from CERN research. We are a spin-off with a multidisciplinary team. We develop a visual analytics technology that allows users to map complex and interconnected data without coding, and to navigate without boundaries or limits into their data in order to visualise trends, white spaces and additional insights that machines or humans alone could not understand.

Who we are

Dtangle SA was set up in late 2023 as the commercial spin-off of Collaboration Spotting X (CSX), an open-source visual analytics software developed at CERN. The name comes from “detangle”: our job is to untangle complexity and make data readable.

We balance academic excellence with commercial insight. In a market led by American and Chinese firms, we put neutrality, transparency and data privacy first, and we like to see sophisticated technology serve the greater good. Dtangle was selected for the Innosuisse Initial Coaching programme, which supports science-based innovation in Switzerland.

What we do

Offer	What it is
Visual analytics platform	A no-code tool built on CSX to navigate complex data as interactive graphs, charts and tables. Available as an on-premise enterprise version; a SaaS version is in development.
Data consulting	Bespoke support from data preparation through to advanced queries and analyses.
Data connectivity	Merging several sources into one view without altering the originals.
Maintenance and support	Keeping deployments running and users productive.

What the platform brings:

- No-code navigation of large, connected datasets.
- Seamless merging of data sources; the originals stay untouched.
- Interactive graphs, charts and tables that stay linked to each other.
- Automated, dynamic reporting.
- Domain-agnostic: the same tool works on patents, projects, companies or people.

Where it is used

- **Patents and publications analysis:** landscapes, assignees, technology clusters, white space.
- **Market research:** players, products and how they connect.
- **Intelligence and investigation:** following links across many sources.
- **Community management:** who works with whom, and on what.

Users describe it as “a bazooka for data analysis” (Head of Innovation, European institution), with a “capacity for exploring data that I haven’t seen in current BI platforms” (CEO, crowdfunding company).

Dtangle in this project

The CORDIS SDG challenge is a typical Dtangle problem: 23,451 projects, 17 goals, 169 targets, thousands of organisations and billions of euros, all connected to each other. Dtangle built the complete chain for it: a transparent, reproducible mapping of every Horizon Europe project to the EU SDG taxonomy, a web prototype that recomputes every figure, and ready-made studies that open in the Dtangle platform as live graphs of projects, goals, targets and concepts.

Team

Our members bring decades of collective expertise, both in the open-source tools started at CERN and in building an active user community around them.

This entry is submitted as a team by Dtangle SA (Geneva, Switzerland): Marc Siegrist (team leader, submitting on behalf of the team), Jean-Marie Le Goff and Paul Bagourd. Should the entry rank among the top three, the certificate should be issued under the team name Dtangle.

Name	Role
Marc Siegrist	Co-Founder
Jean-Marie Le Goff, PhD	Co-Founder
Paul Bagourd	AI Engineer

2. Summary

Dtangle mapped the 23,451 Horizon Europe projects in CORDIS to the EU SDG taxonomy: 17,437 projects (74 %) are linked to at least one goal, through 48,290 project-goal assignments and 73,893 project-target assignments. Every assignment carries a confidence score, a confidence level, a one-line explanation, the sentence that supports it and the CORDIS fields used.

What sets the solution apart

1. **Target-level precision through a concept layer:** SDG targets are too abstract to search for. Dtangle broke the 166 in-scope targets into 251 concepts that research projects actually write about (target 7.2 becomes *Solar energy*, *Wind energy*, *Bioenergy and biofuels*…), each recognised by reviewed plain-language aliases (3,437 in total) with guards against false friends.
2. **A graded, explainable confidence score:** where a concept appears decides what it is worth; a title or an aim sentence weighs more than a background sentence. Each assignment carries its score, a High / Medium / Low level, an explanation, the supporting sentence and the CORDIS fields used.
3. **Measured reliability:** against a reference reading of 60 hold-out projects, High links are confirmed 71 % of the time, Medium 45 % and Low 11 %, so every user can choose between recall and precision (section 5).
4. **Verifiable to the last digit:** the matching is a deterministic program, with no model at run time. A separate script rebuilds every published score from the evidence tables with zero mismatch, and each assignment's evidence is stored losslessly in a single integer that the prototype decodes.
5. **Answers policy questions, not only labels:** funding is shared across goals in proportion to the scores (it adds up to each grant), goal and target co-occurrence is measured with and without concepts that serve 2 goals by construction, and a white-space index shows which concepts the portfolio leaves uncovered.
6. **Built for visual exploration:** the results are delivered as a working web prototype and as ready-made studies for the Dtangle visual analytics platform, where goals, targets, concepts and projects become live, navigable graphs.

Deliverables

Deliverable asked by CORDIS	Where
Notebook: methodology, data preparation, mapping logic, results	the submission page (Abstract and Solution tabs); this document in full
Project-to-SDG mapping table	complete table online at cordis.dtangle.ch/downloads (48,290 rows), described in section 6
Visualisations of the main findings	sections 3 and 4, the figures attached to the submission, and the prototype
Limitations, uncertain cases, improvements	section 5 and the limitations document
Working web-based prototype	CORDIS SDG Explorer, described in section 7
Description document	this document

3. Results

Dtangle read all 23,451 Horizon Europe projects and asked a simple question: which of the 17 UN Sustainable Development Goals does each one work on? Here are the answers, in charts and short text. Where a chart can't show how goals hang together, we add network views from Dtangle, our visual analytics tool. The interactive versions of every chart are in the prototype (section 7).

How a project gets linked to a goal

We wrote a list of 251 concepts, from *aquaculture* to *drought* to *cancer*, each tied to one or more of the goals' targets. Then we look for them in every project: title, objective, keywords, call topic and acronym. Where a concept shows up decides how many points it earns.

Where the concept is found	Points
Title, or an objective sentence stating the project's aim	1.00
An objective sentence about expected impact	0.90
An objective sentence about method, or the keywords	0.80
An objective sentence giving context, or the call topic	0.70
A speculative sentence ("could", "in the long term"), or the acronym	0.60
Bonus when the objective keeps coming back to it (2 to 3 sentences / 4 or more)	+0.25 / +0.50

A concept counts only if it appears in a strong place (0.80 or more) or in 2 places. The **score** of a project for a goal is the sum of the points of its concepts serving that goal, plus up to 1.20 when the Commission's own classification points the same way (climate or biodiversity tracking, programme or mission, research field).

Two real projects:

- **INNOAQUA** has *aquaculture* in its title (1.00) and in the sentence stating its aim (1.00): **2.00** for SDG 2, zero hunger.

- **BOOSTER**, on drought-tolerant cereals, has *drought* in its title (1.00), its aim (1.00), 4 or more sentences (+0.50) and its keywords (0.80), so 3.30; *irrigation* in its aim adds 1.00 and its research field (water engineering) 0.30: **4.60** for SDG 6, clean water.

So the score tells how much, and how centrally, a project talks about a goal. It says nothing about the impact the project will have.

The 251 concepts are also grouped into 128 **broad concepts**. *Air pollution and health*, *chemical exposure and toxic substances* and *water and soil contamination affecting health*, for example, make up *environmental health*. At broad level, the fine concepts a project mentions are merged by broad concept: the merged concept keeps the goals of the fine concepts found and adds up their scores. Links between projects and goals, and their scores, stay the same; what changes is which concepts count as separate evidence, the concept graph and funding per concept.

Key findings

- **Three in four projects (17,437) work on at least one goal.** The other 6,014 are mostly fundamental research (ERC and MSCA grants) where we find no goal.
- **Most projects touch several goals:** 65 % of the linked projects, 2.8 goals each on average. Keep only the links where a project is really about the goal (score 2.00 or more) and that drops to 44 %; at 4.00 or more, to 20 %.
- **Health, climate and cities come first:** SDG 3 (7,918 projects), SDG 13 (5,827) and SDG 11 (5,071). Poverty (SDG 1, 763) and partnerships (SDG 17, 931) get the least attention.
- **Some goals go together.** Consumption and climate (SDG 12 × 13) meet most often: 1,375 projects with separate evidence for each (2,139 counting all). Gender and inequalities (SDG 5 × 10) pull hardest towards each other: 277 projects, each with its own evidence for both goals, 2.65 times what chance would give. Keep only the projects that are really about their goals and the top pair changes: climate and life on land (SDG 13 × 15) lead from score 2.00 up (333 projects, and 114 at 4.00 or more), each with separate evidence for both goals.
- **The money goes to a few topics:** the 20 best-funded concepts, led by cutting greenhouse gas emissions and cancer, take 42 % of the €49.9 bn we can attribute to concepts.

Three ways to read the numbers. *All links* counts every project–goal link, even a single mention in the keywords. *Score ≥ 2.00* keeps the links where the project is really about the goal, like INNOAQUA above (19,754 links, 41 % of all). *Score ≥ 4.00* keeps only the strongest 14 % (6,869 links), like BOOSTER. All the details: section 4, Scores and rankings.

1 · Goals

Projects per goal

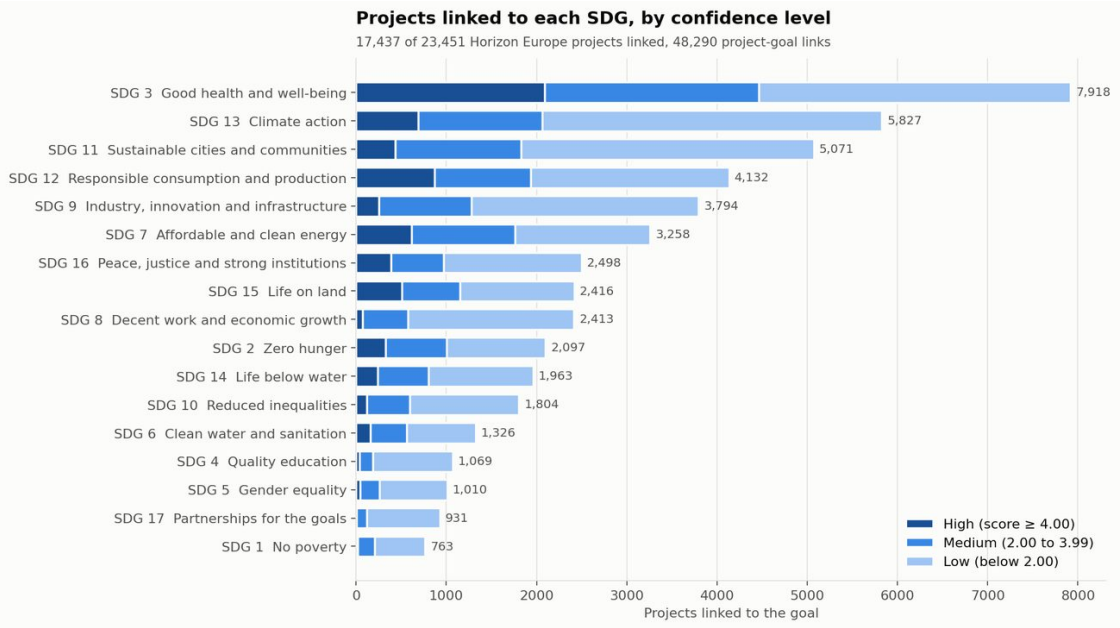


Figure 1. *Projects linked to each SDG, by confidence level.*

Goal	All links	Score ≥ 2.00	Score ≥ 4.00	Share of all projects
SDG 1 · No poverty	763	205	18	3.3 %
SDG 2 · Zero hunger	2,097	1,005	324	8.9 %
SDG 3 · Good health and well-being	7,918	4,465	2,087	33.8 %
SDG 4 · Quality education	1,069	185	40	4.6 %
SDG 5 · Gender equality	1,010	260	46	4.3 %
SDG 6 · Clean water and sanitation	1,326	561	158	5.7 %
SDG 7 · Affordable and clean energy	3,258	1,761	612	13.9 %
SDG 8 · Decent work and economic growth	2,413	572	74	10.3 %
SDG 9 · Industry, innovation and infrastructure	3,794	1,280	251	16.2 %
SDG 10 · Reduced inequalities	1,804	593	119	7.7 %
SDG 11 · Sustainable cities and communities	5,071	1,825	436	21.6 %
SDG 12 · Responsible consumption and production	4,132	1,934	869	17.6 %
SDG 13 · Climate action	5,827	2,063	690	24.8 %
SDG 14 · Life below water	1,963	804	241	8.4 %
SDG 15 · Life on land	2,416	1,153	510	10.3 %
SDG 16 · Peace, justice and strong institutions	2,498	969	387	10.7 %

Goal	All links	Score ≥ 2.00	Score ≥ 4.00	Share of all projects
SDG 17 · Partnerships for the goals	931	119	7	4.0 %

Each slice is one goal's share of all project–goal links: a project linked to 3 goals counts once in each, so the slices add up to 100 %. The legend also gives the number of projects per goal and their share of the 23,451 Horizon Europe projects; those shares add up to more than 100 %, since most projects touch several goals.

Goals that travel together

One caution before reading pairs. 41 % of the goal pairs come from a single concept that serves 2 goals at once: *air pollution* serves health (3.9) and cities (11.6), so every project on air pollution links SDG 3 and SDG 11. The project does serve both goals, but the pair was written into the concept table before any project was read: counting it shows how popular the concept is, not a combination the project chose. By default the charts below count only **independent evidence**: a project counts for a pair when one of its concepts serves the first goal and not the second, and another serves the second and not the first. The *raw* count, every project linked to both goals, is given beside it.

How the pairs are measured

Four numbers describe a pair of goals. The example is health and cities (SDG 3 × SDG 11), all links, where 17,437 projects are linked to at least one goal, 7,918 of them to SDG 3 and 5,071 to SDG 11.

Measure	What it counts	What it means	SDG 3 × 11
Shared projects (raw)	projects linked to both goals	how often the 2 goals appear in the same project, whatever the reason	2,015
Independent evidence	projects with one concept serving the first goal and not the second, and another serving the second and not the first	how often a project combines 2 separate subjects, not one concept filed under 2 goals	510
Expected count	projects of the first goal × projects of the second ÷ all projects: 7,918 × 5,071 ÷ 17,437	the number of shared projects if being linked to one goal said nothing about the other; big goals meet often by size alone	2,303
Lift	shared projects (raw) ÷ expected count: 2,015 ÷ 2,303	the chance reference: 1 means as often as chance, 2 twice as often, 0.5 half as often	0.88
Lift index	independent evidence ÷ expected count: 510 ÷ 2,303	used only to rank pairs against each other. Its numerator is a stricter count than the one behind the expected count, so it sits at or below the lift and its 1 is not the chance level	0.22

Ratios on few projects swing widely, so the lift rankings and the graph keep pairs with at least 30 shared projects (15 in the score ≥ 4.00 series; 20 and 10 for targets).

Chart: projects with independent evidence for both goals

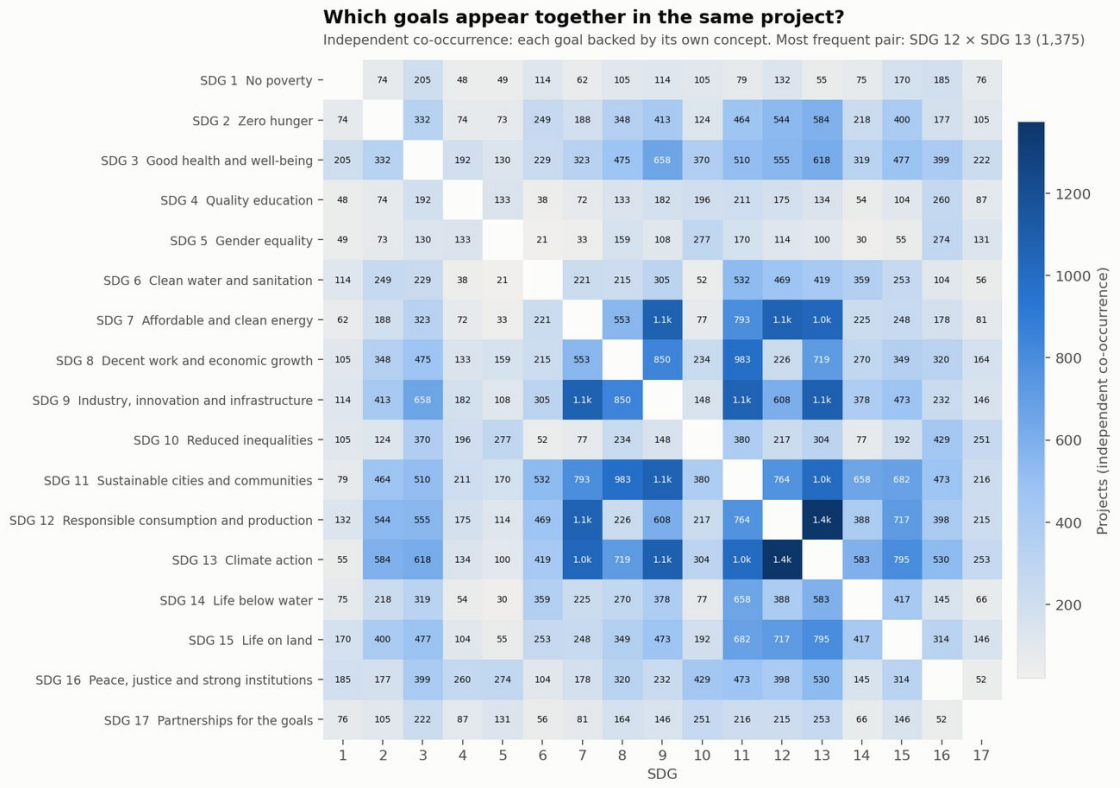


Figure 2. *Goals that appear together in the same project (independent co-occurrence).*

Most frequent pairs (independent evidence)	Projects		Strongest pull	Projects	Lift index
SDG 12 × SDG 13	1,375		SDG 10 × SDG 5	277	2.65
SDG 13 × SDG 9	1,082		SDG 10 × SDG 17	251	2.61
SDG 7 × SDG 9	1,071		SDG 17 × SDG 5	131	2.43
SDG 12 × SDG 7	1,065		SDG 14 × SDG 6	359	2.40
SDG 11 × SDG 9	1,063		SDG 4 × SDG 5	133	2.15
SDG 13 × SDG 7	1,036		SDG 1 × SDG 6	114	1.96
SDG 11 × SDG 13	1,009		SDG 16 × SDG 5	274	1.89
SDG 11 × SDG 8	983		SDG 1 × SDG 17	76	1.87

The strongest pairs

The graph can weigh its edges 2 ways, and they answer different questions. **Shared projects** is the volume: the thicker the edge, the more projects work on both goals. Big goals win here almost by default: consumption and climate (SDG 12 × 13) get the thickest edge, 1,375 projects, largely because both goals are big. **Lift** takes size out of the picture: the thicker the edge, the more often the 2 goals meet compared with what their sizes alone would give. Gender and inequalities (SDG 5 × 10) share only 277 projects, but that is 2.65 times the expected count, so they get the thickest edge there, while 12 × 13 shrinks to an average one. In short: shared projects shows where the projects are, lift shows which goals attract each other.

In the Dtangle platform

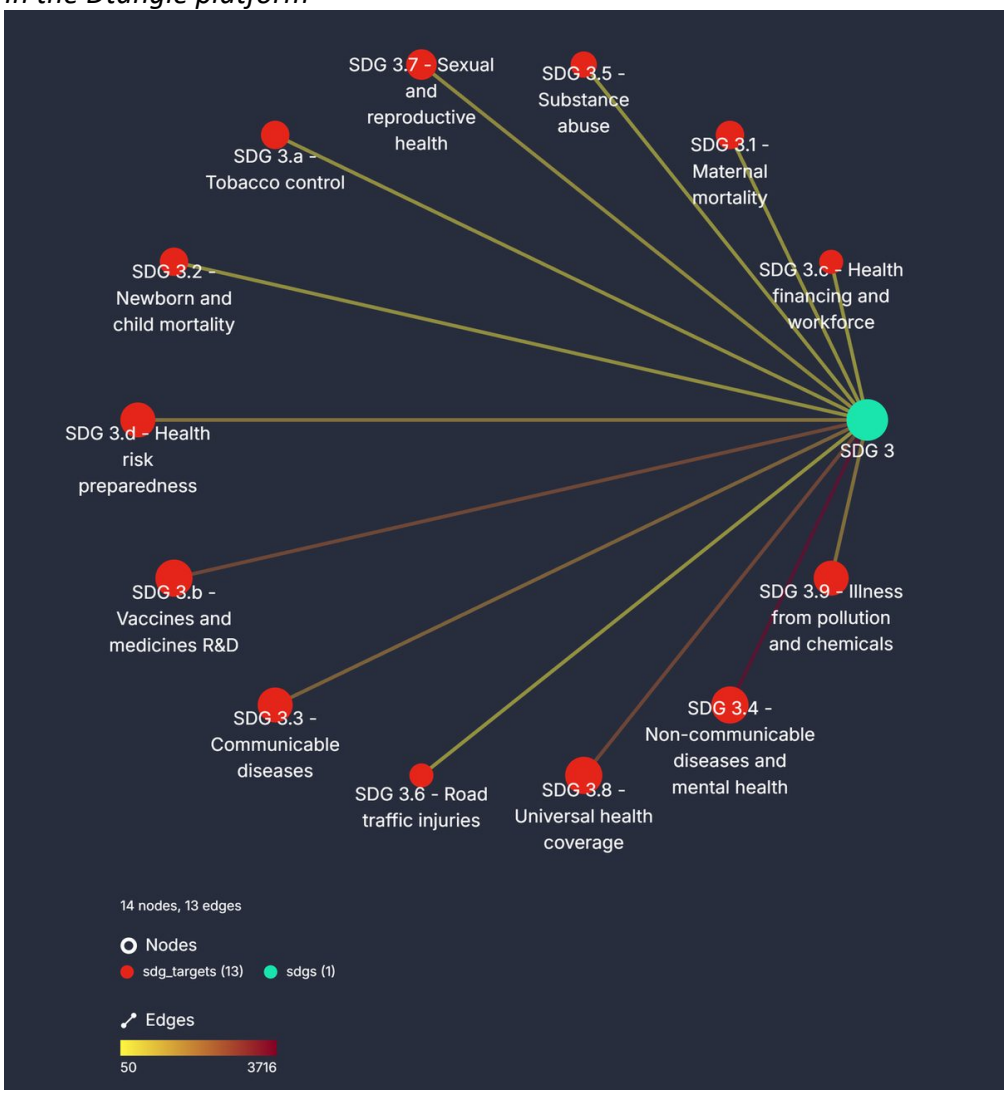


Figure 3. *SDG 3 and its 13 targets in the Dtangle platform, edges coloured by number of projects (50 to 3,716).*

2 · Targets

Projects per target

The 20 targets with the most linked projects (all links):

Target	Projects
SDG 13.1 - Climate resilience and adaptation	4,557
SDG 3.4 - Non-communicable diseases and mental health	3,726
SDG 13.2 - Climate measures in policies and planning	2,940
SDG 11.6 - Environmental impact of cities	2,715
SDG 12.2 - Sustainable use of natural resources	2,647

Target	Projects
SDG 7.2 - Renewable energy	2,336
SDG 3.8 - Universal health coverage	2,329
SDG 9.4 - Clean and resource-efficient industry	2,289
SDG 3.b - Vaccines and medicines R&D	2,240
SDG 3.9 - Illness from pollution and chemicals	2,123
SDG 12.5 - Waste reduction, recycling and reuse	1,962
SDG 7.3 - Energy efficiency	1,567
SDG 3.3 - Communicable diseases	1,461
SDG 2.4 - Sustainable food production	1,386
SDG 16.7 - Participatory decision-making	1,349
SDG 8.4 - Resource efficiency and decoupling	1,298
SDG 15.1 - Terrestrial and freshwater ecosystems	1,200
SDG 15.5 - Biodiversity loss and threatened species	1,093
SDG 3.d - Health risk preparedness	1,044
SDG 11.5 - Disaster losses	991

Targets that travel together

The same reading one level down. With independent evidence, targets linked only through one shared concept are set aside; the raw count keeps them. Pairs within the same goal are shown in the chart but left out of the "unexpected" list, where they are expected by design.

In the graph views, edge width works as for goals: **shared projects** shows the volume (targets 3.4 and 3.b, non-communicable diseases and medicines R&D, share the most projects), **lift** shows which targets attract each other beyond their size.

In the Dtangle platform

Target pair	Independent evidence	All shared projects
SDG 3.4 Non-communicable diseases and mental health × SDG 3.b Vaccines and medicines R&D	1,210	1,213
SDG 3.4 Non-communicable diseases and mental health × SDG 3.8 Universal health coverage	1,132	1,441
SDG 12.2 Sustainable use of natural resources × SDG 13.2 Climate measures in policies and planning	980	980
SDG 11.6 Environmental impact of cities × SDG 12.2 Sustainable use of natural resources	906	1,014
SDG 12.5 Waste reduction, recycling and reuse × SDG 9.4 Clean and resource-efficient industry	831	844
SDG 13.2 Climate measures in policies and planning × SDG 7.2 Renewable energy	749	959

Target pair	Independent evidence	All shared projects
SDG 13.2 Climate measures in policies and planning × SDG 9.4 Clean and resource-efficient industry	705	1,070
SDG 12.2 Sustainable use of natural resources × SDG 7.2 Renewable energy	659	848
SDG 11.6 Environmental impact of cities × SDG 9.4 Clean and resource-efficient industry	649	782
SDG 12.5 Waste reduction, recycling and reuse × SDG 13.2 Climate measures in policies and planning	635	640
SDG 12.2 Sustainable use of natural resources × SDG 12.5 Waste reduction, recycling and reuse	621	1,521
SDG 8.4 Resource efficiency and decoupling × SDG 9.4 Clean and resource-efficient industry	590	597

Start from a single target and expand it both ways: upwards to the goals its projects also serve, downwards to its concepts. Disaster resilience is a cross-cutting subject, and the graph shows it at a glance.

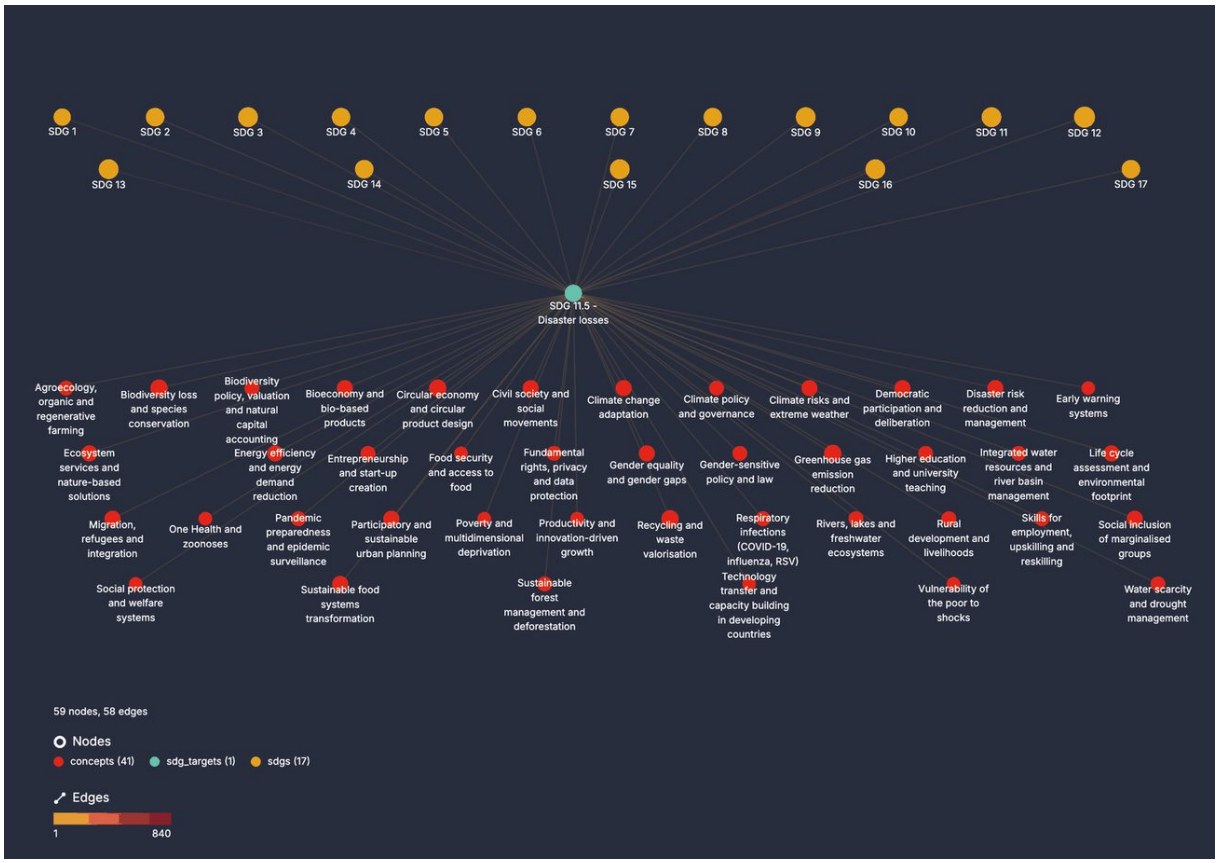


Figure 4. *SDG 11.5 (disaster losses) in the Dtangle platform, expanded in both directions: the 17 goals its projects also serve and its 41 concepts, all links (59 nodes, 58 edges).*

Raise the edge-weight filter on the whole study and only the strongest links remain: SDG 12 and 11 keep the most links to their targets; greenhouse gas emission reduction and circular economy are the concepts that still reach several targets.

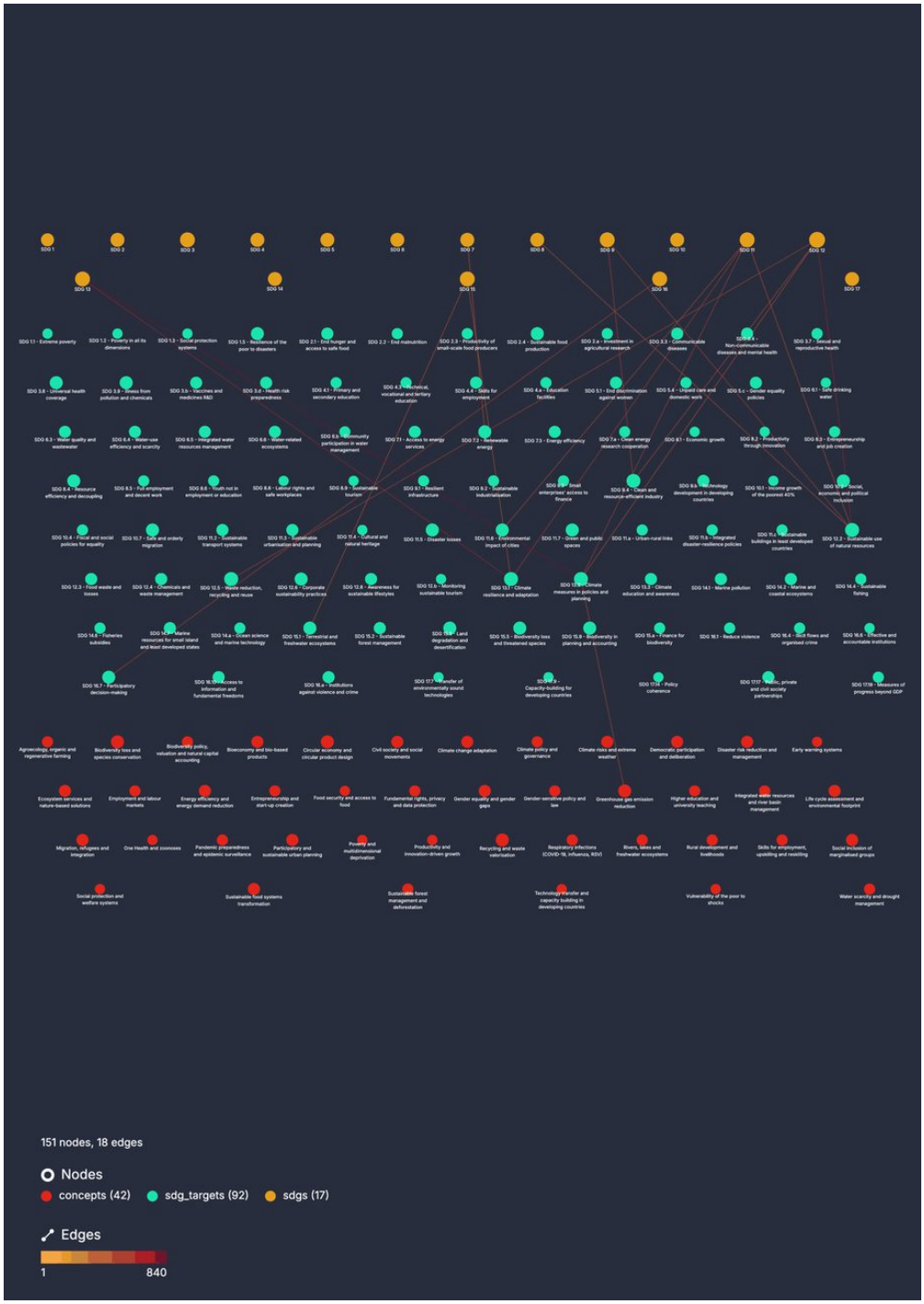


Figure 5. *The whole grid with the edge-weight filter raised: 17 goals, 92 targets and 42 concepts, only the 18 strongest links kept.*

3 · Concepts

Concepts, targets and goals

The 251 concepts (grouped into 128 broad concepts) are the vocabulary between the goals and the projects. Some goals are described by many concepts, health first; 66 targets rest on a single concept.

	Goal
Concepts	SDG 1 · No poverty
12	SDG 2 · Zero hunger
26	SDG 3 · Good health and well-being
46	SDG 4 · Quality education
15	SDG 5 · Gender equality
15	SDG 6 · Clean water and sanitation
17	SDG 7 · Affordable and clean energy
16	SDG 8 · Decent work and economic growth
18	SDG 9 · Industry, innovation and infrastructure
21	SDG 10 · Reduced inequalities
18	SDG 11 · Sustainable cities and communities
32	SDG 12 · Responsible consumption and production
34	SDG 13 · Climate action
30	SDG 14 · Life below water
21	SDG 15 · Life on land
27	SDG 16 · Peace, justice and strong institutions
23	SDG 17 · Partnerships for the goals

In the Dtangle platform, expand SDG 3 twice and the whole vocabulary of health research appears under its targets. Every edge is a set of projects you can list.

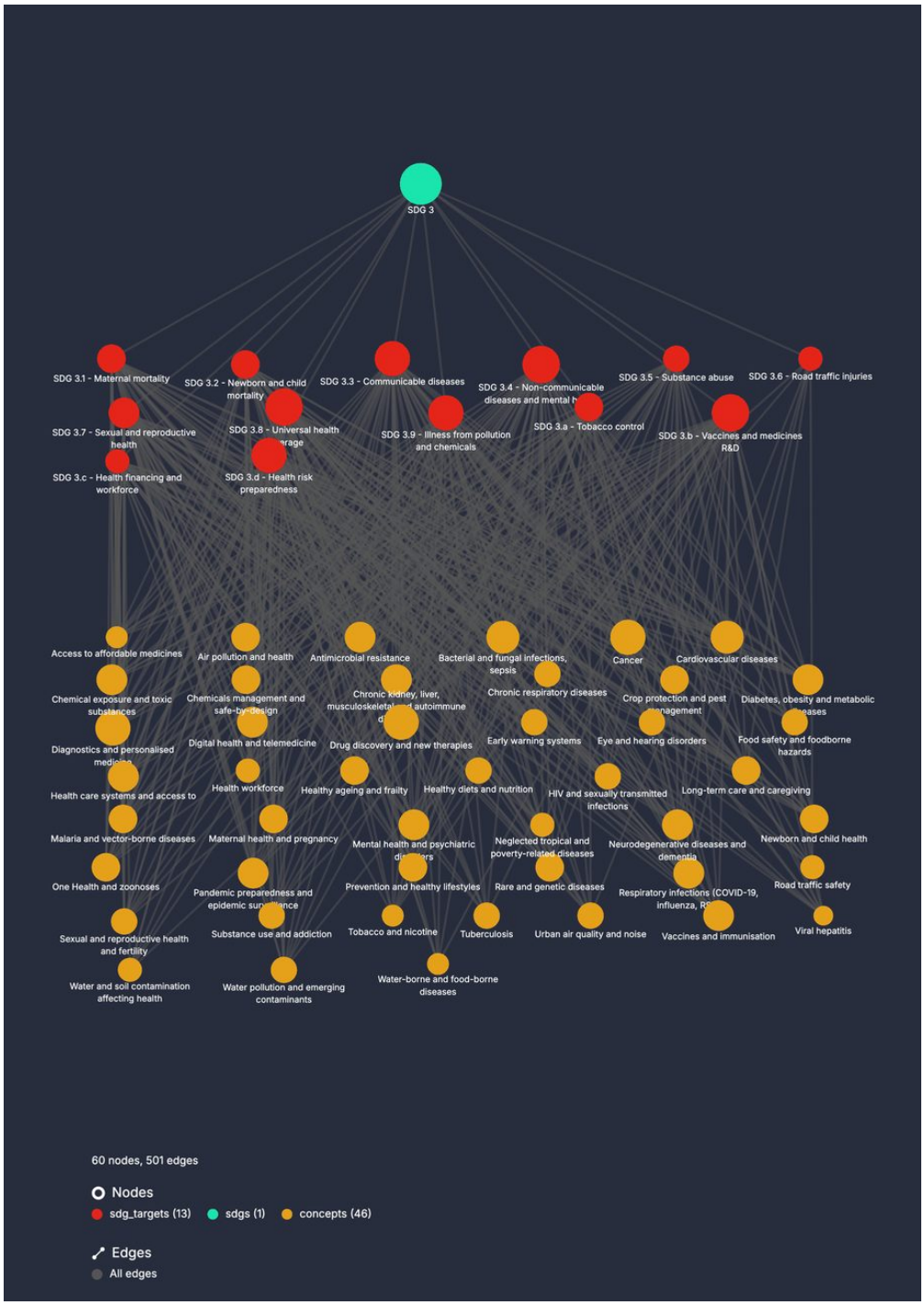


Figure 6. *SDG 3, its 13 targets and their 46 concepts in the Dtangle platform: the three layers of the reading grid on one screen (60 nodes, 501 edges).*

Targets and their concepts

Three ways to weigh an edge. **All projects**: every project where the concept is found counts 1. **Very pertinent projects**: only projects where the concept scores 2.00 or more count, i.e. the project is about

it. **Sum of scores**: each project counts for its concept score, so INNOAQUA's *aquaculture* (2.00) weighs more than a mention in the keywords alone (0.80). Some concepts carry many projects that only mention them; the last 2 views show where the substance is.

Projects, concepts and funding

Each project's EU contribution is shared across its concepts in proportion to their scores (section 4, Funding), so the amounts add up to the €49.9 bn granted to the projects with a concept link. The 20 best-funded concepts carry 42 % of it. The number of projects and the money do not rank the same way: *Aviation and maritime decarbonisation* has 91 projects but €740 m, about €8 m per project, the mark of large industrial partnerships.

The other end of the scale, concepts with few, shallow, concentrated or ending projects, is the **white space**: see section 4, White space analysis.

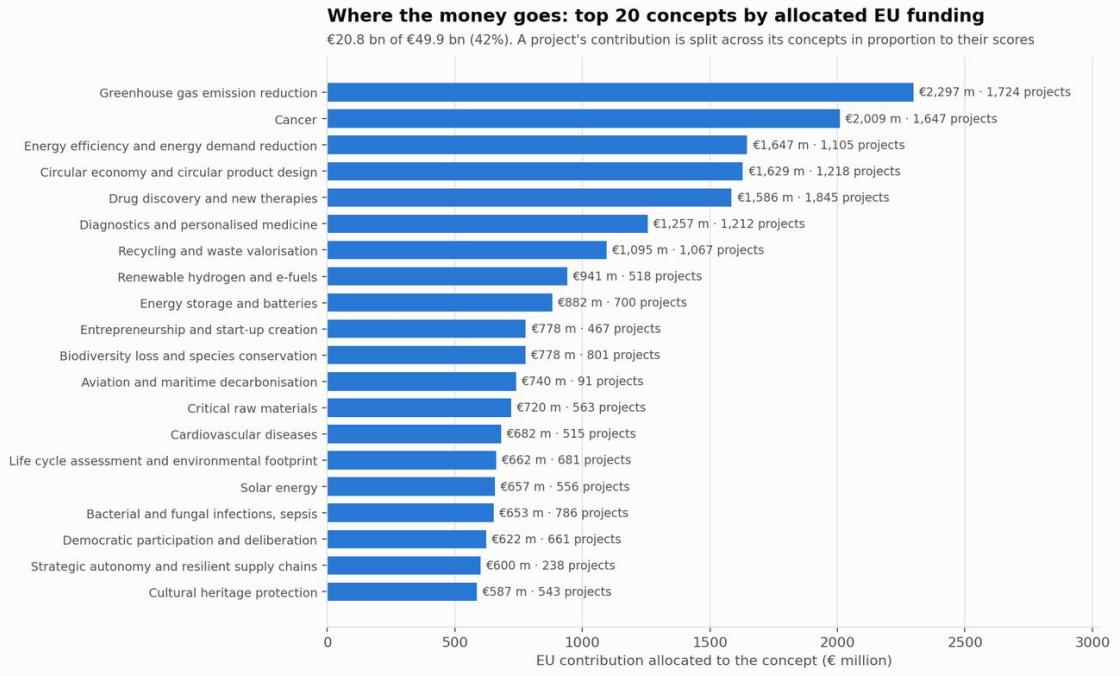


Figure 7. *Top 20 concepts by allocated EU funding.*

In the Dtangle platform

Select a handful of concepts and ask which goals they reach. Greenhouse gas emission reduction alone connects SDG 1, 7, 8, 9, 10, 11, 12, 13 and 15; circular economy reaches SDG 8, 9, 11 and 12. This is the graph behind the pairs of goals in section 1: many of them are made by one concept serving two goals at once.

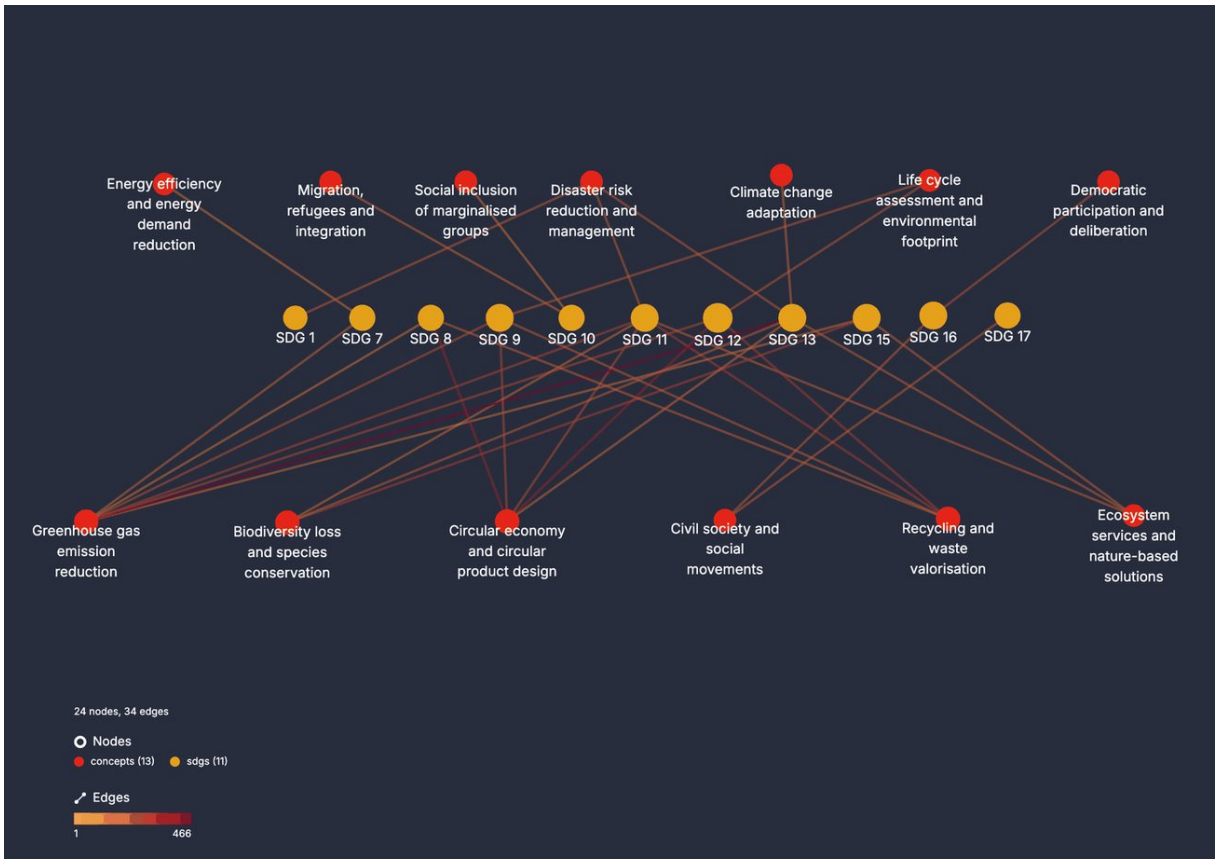


Figure 8. *Thirteen cross-cutting concepts and the 11 goals they reach, edges coloured by number of projects (1 to 466), hierarchical layout.*

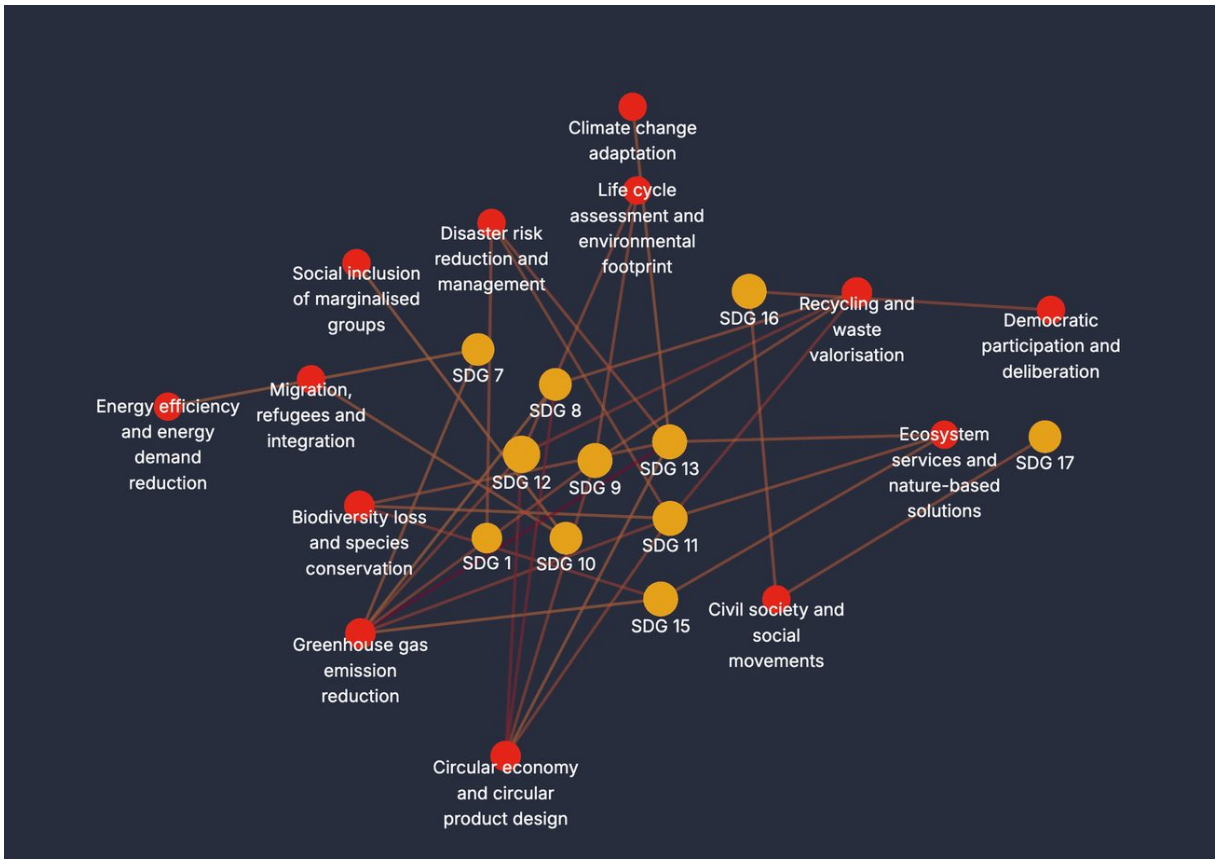


Figure 9. *The same selection in a force-directed layout: the most connected goals (SDG 12, 13, 11 and 9) are pulled to the centre.*

4 · Evidence and coverage

Where the evidence comes from

Most textual evidence comes from where a project states what it will do.	**Best location of the evidence**	**Links**
Share	Aim sentence	16,184
34.9 %	Method sentence	10,716
23.1 %	Title	9,989
21.6 %	Keywords	5,610
12.1 %	Impact sentence	3,655
7.9 %	Context sentence only	93
0.2 %	Call topic title only	68
0.1 %		

Projects without an SDG link

6,014 projects (26 %) are linked to no goal at all: no concept is found in their text, and no structural signal is strong enough to place them on its own. The 17,104 projects with a concept link and 333 placed by a structural signal alone make 17,437 linked projects.	**Funding scheme**	**Not linked**	**Share of the scheme**
Share of the not linked	MSCA fellowships and doctoral networks	2,495	29 %
41 %	ERC frontier research	2,492	39 %
41 %	Coordination and support actions (CSA)	313	19 %
5 %	European Innovation Council (EIC)	278	19 %
5 %	Research and innovation actions (RIA)	245	9 %
4 %	Joint undertakings	121	14 %
2 %	Innovation actions (IA)	45	4 %
1 %	Euratom	19	31 %
0.3 %	Other schemes	6	8 %
0.1 %	All projects	6,014	26 %

100 %

Mostly fundamental research. 82 % of them are ERC or MSCA projects, such as “Flexoelectric control of magnetic textures in 2D oxide membranes” or “Mechanisms of co-translational assembly of multi-protein complexes”. They advance knowledge without addressing a goal, and giving them one would force the link. This is also why target 9.5 (enhance scientific research) is excluded: it would link every project and mean nothing. Applied projects are almost all linked: only 4 % of innovation actions and 9 % of research and innovation actions are not.

Some are genuine misses. “Perfect Vision”, adaptive eyewear for myopia, could belong to health (SDG 3), and “SCORPION”, on the safety of nuclear fuel cladding, arguably to energy (SDG 7). Their summaries use words that no alias covers, the limit described in section 5.

In funding, the projects without a link account for €11.7 bn of the €62.4 bn of EU contribution. The other €50.7 bn is what the funding allocation of section 4 shares out.

5 · Go further

Studies in Dtangle and CSV

Everything above can be explored live. The studies behind the graph views are ready in the Dtangle visual analytics tool: open trials.dtangle.ch and sign in with the login `cordis@dtangle.ch` and the password `sdgmap`.	**Study**	**One row =**
Shows	1 · Goals, 3 series	a project with its goals; a goal with the goals it shares projects with
projects per goal, correlated goals (shared projects and lift, independent and raw; pairs also at broad level)	2 · Targets, 3 series	a project with its targets; a target with the targets it shares projects with
correlated targets (independent and raw)	3 · Targets and concepts	a target with its concepts, weighted by all projects, very pertinent projects and sum of scores (fine and broad)

the vocabulary of each goal

To build a custom file, the Explorer tab of the prototype lets anyone choose goals, filters and an analysis angle, and download a CSV ready for Dtangle. How every figure is computed and checked is in section 4.

4. Methodology

Every Horizon Europe project in CORDIS is read for evidence of 251 SDG concepts. Each concept belongs to one or more SDG targets, so each piece of evidence connects a project to a target and to its goal. The method is deterministic and every link can be traced back to the words in the project text that produced it.

At a glance

The run in 8 figures.	**Figure**
What it counts	**23,451**
Horizon Europe projects read	**166 / 169**
SDG targets in scope	**251 · 128**
fine · broad concepts	**3,437**
searchable formulations (aliases)	**46,315**
project × concept links	**17,104**
projects with ≥ 1 concept link (73 %)	**199**
structural rules (trackers, programme, EuroSciVoc)	**132,369**

occurrences, each kept in a lossless evidence index

Data sources

The CORDIS Horizon Europe dataset of the European Data Portal (extract of 2026-08-06) and the EU SDG taxonomy. No personal data.

- **CORDIS Horizon Europe projects**, from the European Data Portal dataset "CORDIS – EU research projects under Horizon Europe (2021-2027)", https://data.europa.eu/data/datasets/cordis-eu-research-projects-under-horizon-europe-2021-2027 (bulk extract of 2026-08-06: project, organisation, topic, legal basis and euroSciVoc files). The text read for SDG evidence is the project **title**, its **keywords** and its **objective** (the public summary).
- **SDG taxonomy** from EU Vocabularies (SKOS, version 1.0, 2020): 17 goals, 169 targets and 232 indicators. Target wording, indicator wording and indicator series names were all used to build the concepts.
- **Organisations** appear only as legal entities (name, country, type, role). No personal data is processed.

Pipeline

The grid is built from the taxonomy once; each run matches, adds structural signals and scores.

The reading grid comes from the taxonomy alone: nothing in the corpus can create a concept. The 23,451 projects enter twice only: once to supply the wording under which a concept is recognised, once as the text to be matched.

Steps 1 and 2 below describe the grid: the concepts and the word lists that recognise them. Steps 3 to 6 follow the 4 stages of each run:

1. **Matching:** title, keywords, call topic, acronym, objective sentences, weighted by field
2. **Structural signals:** EU trackers, programme parts, missions, EuroSciVoc
3. **Scores:** weighted sum per SDG and per target, then rankings
4. **Funding:** each grant, gross and shared across its goals by score

Step 7 then reads the result the other way round: for each concept, how little the projects cover it.

Step 1 · The SDG → Target → Concept layer

166 targets broken down into 251 concepts that a research project would actually write about.

A target such as 7.2 ("increase substantially the share of renewable energy") is too abstract to search for directly. It is broken down into **concepts**: things a research project would actually write about, such as *Solar energy*, *Wind energy* or *Bioenergy and biofuels*.

- Concepts were derived from everything the taxonomy says about each target (target, indicator and indicator-series wording), enriched with knowledge of how research and innovation projects address that target.
- There are 2 nested levels: **251 fine concepts**, each rolled up into one of **128 broad concepts**. The prototype counts at either level.
- Each concept records the targets it serves (a concept can serve several targets, in several goals), a written **rationale** for those links, and a **role**: an *outcome* the target wants (123 concepts), a *means* such as a technology or intervention (91), or a *domain* of study (37).

- Three targets are **excluded** because every Horizon Europe project would qualify: 9.5 (enhance scientific research), 17.6 (science and technology cooperation) and 17.16 (global partnership). The remaining 166 targets are all covered by at least one concept.
- The concept list and first drafts of the aliases were written with the help of a large language model (Claude, by Anthropic), then reviewed and corrected against samples of CORDIS projects. No model is used when projects are matched: the matching below is a fixed, repeatable program.

Step 2 · Aliases and guards

3,437 wordings recognise the concepts; exclusions stop words that mean something else.

Each concept has a list of **aliases**, the plain-language formulations under which it is recognised (3,437 in total, median 12 per concept). For *Poverty and multidimensional deprivation*, for example: “poverty”, “deprivation”, “low-income households”, “poor communities”, “at risk of poverty”.

Exclusions and required context

- **Exclusions** stop a term from counting where it means something else. “Energy poverty” and “sleep deprivation” do not count as poverty, “solar wind” does not count as solar energy, and “cell migration” does not count as human migration.
- **Required context** binds some concepts to an application: “broadband”, for instance, only counts when the project also speaks about connectivity.

Where the aliases come from

Part of each list is written from the concept itself; the rest is learnt from how CORDIS projects actually write. This was done once, in 2 passes. Cancer shows how.

1. **Pass 1: the first search patterns.** Each concept started as a few search patterns written from the target wording. Some are open-ended: `metasta\w*` means “any word beginning with *metasta*”. Such a pattern finds things, but it is not a list of words, and it catches more than intended.
2. **Count what the patterns caught.** `scan_surface_forms.py` ran them over the titles, keywords and objectives and counted every word they hit, nothing more (no weights, no sentence roles, no scores). For `metasta\w*`: *metastasis* 282 times, *metastatic* 266, *metastases* 91, but also *metastable* 49 times, a term of physics.
3. **Turn the catch into a clean list.** `derive_aliases.py` kept the words seen at least twice (a typo such as *metastasic*, or a glued *metastaticdissemination*, seen once, dropped out) and removed a short hand-written list of false friends (*metastable*, *thrombospondin*…). Cancer ends up with 27 plain wordings, among them *metastasis*, *metastatic* and *metastasize*.
4. **Pass 2: every run since.** The matching of step 3 uses those lists only. Pass 1 is not repeated: the lists are now the reference (`sdg_concept_aliases.py`), can be corrected by hand, and anyone re-running the pipeline gets the published result from them alone.

The projects therefore teach the grid *how* a concept is written; they never add a concept, so the grid stays independent of the corpus it reads.

Step 3 · Matching and weights

Where a concept appears decides what it is worth: the title and aim sentences weigh most.

Reading a project, step by step

For each project, the program:

1. splits the title, the keywords and the objective into words, and the objective into sentences;
2. normalises every word (case, hyphen or space, British or American spelling, plural of the last word) and looks up every run of 1 to *n* words in the alias index, which is an instant dictionary look-up;
3. gives each objective sentence a role from its wording: *aim* ("this project aims to", "we will"), *impact* ("will enable", "contribute to"), *speculative* ("could", "in the long term"), *context* (the opening sentence, "currently", "worldwide", "however"), otherwise *method*;
4. ignores a mention that is negated ("not", "without", "rather than" just before it) or that falls inside an exclusion;
5. gives each occurrence a weight that depends on **which field** it sits in and, inside the objective, on **what the sentence does**:

Field weights	**Field**	**Who wrote it**
Weight	Project title	the applicant, deliberately
1.00	Objective · aim sentence	the applicant: what the project will do
1.00	Objective · expected-impact sentence	the applicant: what it should change
0.90	Objective · method sentence	the applicant: how
0.80	Author keywords	the applicant, without context
0.80	Objective · context sentence	the applicant: background, not commitment
0.70	Call topic title	the Commission: describes the call, not the project
0.70	Objective · speculative sentence	"could", "in the long term"
0.60	Acronym	only when the acronym *is* the concept

0.60

Repetition, saturated

Saying the same thing in several sentences is evidence, up to a point. On top of the value of its best sentence, a concept earns +0.25 when it appears in 2 or 3 distinct objective sentences and +0.50 from 4 on; speculative sentences do not count, and a word repeated inside one sentence counts once. Thirty mentions can never outweigh a title, and the role of each sentence is kept. 11,249 concept links receive the first step, 4,487 the second.

When a concept counts, and its score

A concept counts as **found** when its best occurrence is worth 0.80 or more, or when it appears in at least 2 distinct places worth 0.70 each. Its **score** is then the sum of its best occurrence *in each field*: a concept stated in the title, promised in an aim sentence and listed in the keywords scores 1.00 + 1.00 + 0.80 = 2.80, while the same concept mentioned once in a method sentence scores 0.80. Two thirds of the links rest on a single field; 1,172 rest on 4 or 5.

The score as an encoding

The score is also an **encoding**: `score_encoded` = 3.30284 means 3.30 of weight and component code 284, and 284 read in mixed radix (title 2 · objective 6 · repetition 3 · keywords 2 · call topic 2 · acronym 2) gives back title yes, objective = aim, 4 or more sentences, keywords yes, topic no, acronym no. The weight itself is recomputable from the code, so the two halves of the number check each other. Nothing is lost by publishing a single figure.

Beside the number, the same fact is written in plain words: `evidence_fields` is a Python list of the dimensions the concept was actually found in, such as `['title','aim_sentence','sentences_4plus','keywords']`, so anyone can filter the tables by hand, in a spreadsheet or a one-line script, without decoding anything.

Every link keeps its evidence: the matched terms, the number of places, the best location and the sentence that supports it (`evidence_quote`).

How it runs in about a minute

The matching runs over the 23,451 projects in about a minute. The obvious implementation (251 concepts, each a regular expression, run over every abstract) costs one scan of the corpus per concept. This one never scans for a concept at all.		**One regular expression per concept**
One dictionary of keys	Passes over the corpus	one per concept (251)
one	Work per word of text	grows with the number of concepts and aliases
at most 7 hash look-ups	Spelling and plural variants	one more pattern each

folded into the key, free

One dictionary of normalised keys

Every alias is reduced once to a **normalised key**: lower case, hyphen or space treated alike, British and American spelling folded together, the last word singularised. The 3,437 formulations collapse to 3,425 keys in a single dictionary, the longest being 7 words.

At most 7 look-ups per word

A project's text is cut into words once, each word normalised once, and then, at each position, the program looks up the key made of the next 1 to 7 words: at most 7 hash look-ups per word, whatever the number of concepts.

The same lookup runs at every word; here it starts at "solar" and stops after 4 windows because the sentence ends. The window "solar cells" is looked up as "solar cell": the plural of the last word is part of the key, so it meets the alias *solar cells* of Solar energy without any extra pattern. (One word earlier, at "perovskite", the one-word window is itself an alias of the same concept.)

Cost is linear in the text and flat in the vocabulary: adding a thousand aliases tomorrow changes nothing at scan time. Tolerance is free for the same reason: “photovoltaic cells”, “Photovoltaic-cell” and “photovoltaic cell” are the same key, so no variant needs its own pass.

Regular expressions only around a hit

Regular expressions are still used, but only around a hit: to check an exclusion, to look for a negation in the 40 characters before it, or to type the sentence. It always works on one sentence, never on the corpus. Sentence roles are computed once per project rather than once per concept.

8 shards in parallel

The corpus is finally split into 8 shards run in parallel (`--shard i/8`), which is what brings the whole pass to about a minute on an ordinary laptop.

Step 4 · Structural signals

What the Commission records about a project (trackers, programme, EuroSciVoc) adds at most 1.20.

A project is more than its prose. CORDIS also records what the Commission itself asserts about it: that its budget counts towards the EU **clean-air** or **climate** spending target, that it sits in a given part of the work programme or in an EU **Mission**, and which **EuroSciVoc** science categories it was classified under. Those claims are evidence of a different nature from a sentence in an abstract, so they are scored separately and can be switched off in the prototype.

The 3 families	Family	CORDIS field	Rules
Weight	EU spending tracker (climate, biodiversity, clean air)	`policy_*_pct`, coefficients 0 / 40 / 100 % from SWD(2022)225	3
0.50 at 100 %, 0.20 at 40 %	Programme part or EU Mission	`legal_basis`, `topics`	36
0.40 direct, 0.20 partial	EuroSciVoc science category	`euroscivoc_paths`	160

0.30 direct, 0.15 partial

One rule per family

Only the **strongest rule of each family** counts for a given SDG, so the structural part never exceeds 1.20, about half a well-attested concept. The rules come from the earlier structured-signal mapping, each with its rationale, and travel with the results (`SDG_Structural_Rules`). A programme code inherits an SDG from the destinations below it only when *every* documented destination agrees, which is computed rather than decided by hand; programme coverage is therefore deliberately partial.

Text earns the entry

A structural signal alone puts a project in a ranking only at full strength (a tracker at 100 % or a “direct” programme rule), and such a row is marked `Structural signals only`. A tracker at 40 % or a science category merely corroborates a project the text already placed. Across the 17 goals this adds 3,880 entries, almost all to SDG 13 (+1,686), SDG 3 (+924) and SDG 11 (+740).

What is never evidence

Organisations, countries, dates and budget are never evidence: a partner's nationality says nothing about what a project works on. They stay filters and aggregation axes.

Step 5 · Scores and rankings

Scores add up per SDG and per target; projects are ranked on their pertinence key. Funding is kept out of both.

Score of a project for an SDG

The score of a project for an SDG is the sum of the scores of the distinct concepts linked to a target of that SDG (`score_concepts`) + the structural part (`score_structural`). The same logic applies within each target.

Ranking by pertinence

Projects are ranked on their **pertinence**, a fixed-width key that carries the whole calculation in one number: the total in hundredths (7 digits), the structural part (4), then in 3-digit slices the number of concepts and how many of them are attested by the title, an aim / impact / method / context / speculative sentence, two-to-three or four-plus distinct sentences, the keywords, the topic and the acronym, and finally the points from each signal family. Sorting that number reproduces the ranking, and total − structural gives the textual part back. Equal keys share a rank (1, 2, 2, 4…). The interface never shows it raw: it is the quantity the *pertinence* filter cuts on, and `evidence_dimensions` lists the same content in plain words.

One rank per SDG

A project touching several SDGs has one rank per SDG. In the prototype, ranks are recomputed within the current selection, and a cap (by default the top 50 per SDG) keeps graphs readable.

Coverage

Coverage (`coverage`) sits next to the score, not inside it: the share of the concepts defined for the SDG (or the target) that the project touches. It separates a project that says one thing very strongly from one that spans the goal. At target level it is coarse: 66 of the 166 targets have a single concept.

The previous scoring, kept for comparison

The previous scoring is kept alongside: `n_concepts` (the number of distinct concepts, which was the score) and `strength_sum` (the former 1–5 strength). Moving to the weighted score changes the order substantially. SDG 7 and SDG 13 keep 5 of their previous top 10, SDG 3 keeps 3, because a concept written into the title and the aim now outweighs one merely repeated.

What the score measures

The score measures how much of an SDG a project *talks about* and how the Commission classified it, not the impact it will have.

Step 6 · Funding: gross and allocated

A grant never raises pertinence. It is reported twice: the funding that touches a goal, and the share that can be attributed to it.

Why funding stays out of the score

A €20 m project is not more pertinent than a €2 m one: the size of a grant says how much work is funded, not how closely that work serves a goal. The score and the pertinence key are therefore computed from the text and the structural signals only, and the EU contribution is reported beside them, in 2 forms that answer 2 different questions.

Gross funding

Write G_P for the maximum EU contribution of project *P* (`ec_max_contribution_eur`). The gross funding of project *P* for goal *g* is simply its whole grant, for every goal it is linked to:

$$G^{\text{gross}}_{P,g} = G_P$$

Gross funding answers "how much funded activity touches this goal?". It is **not additive** across goals: a project linked to 3 goals is counted 3 times. Summed over the 17 goals, the €50.7 bn granted to the linked projects would appear as €164.9 bn.

Allocated funding

Write $R_{P,g}$ for the score of project *P* for goal *g* (text plus structural part, step 5). The grant is shared across the project's goals in proportion to that score:

$$G^{\text{alloc}}_{P,g} = G_P \cdot R_{P,g} \,/\, \Sigma_h R_{P,h}$$

where *h* runs over the goals linked to *P*. The shares of one project add up to its grant,

$$\Sigma_g G^{\text{alloc}}_{P,g} = G_P$$

so allocated amounts can be summed across projects and goals without double counting: over the 17 goals they add up to the €50.7 bn actually granted. Allocated funding answers "how much funding can be attributed to this goal?". The same rule is applied at target level with the target scores $R_{P,t}$. The two levels are separate allocations: each adds up to G_P on its own, and they are never mixed.

Worked example: ENDURE

ENDURE receives €2,496,306 and is linked to 3 goals. Its scores add up to 7.60, and each goal receives the same share of the grant as of that total.	**Goal**	**Score *R***	**Share**	**Gross**
Allocated	SDG 11 · Sustainable cities and communities	4.35	57.2 %	€2,496,306
€1,428,806.72	SDG 14 · Life below water	1.60	21.1 %	€2,496,306

ENDURE receives €2,496,306 and is linked to 3 goals. Its scores add up to 7.60, and each goal receives the same share of the grant as of that total.	Goal	Score *R*	Share	Gross
€525,538.11	SDG 15 · Life on land (1.50 text + 0.15 EuroSciVoc)	1.65	21.7 %	€2,496,306
€541,961.17	Total	7.60	100 %	€7,488,918

€2,496,306.00

The gross column triples the grant; the allocated column returns it exactly.

Two questions, two measures	Question	Measure	SDG 13
Target 2.4	How much funded activity touches it?	gross	€22.70 bn
€4.86 bn	How much funding can be attributed to it?	allocated	€5.90 bn
€0.91 bn	Projects linked		5,827

1,386

The gap between the 2 lines measures how shared a goal is. Target 2.4 (sustainable food production) is touched by €4.86 bn of projects, but those projects also serve other targets, and only a fifth of their funding can be attributed to it.

Gross and allocated funding by goal

All 17 goals, published run, € bn; the percentage is allocated over gross.	Goal	Gross (€ bn)	Allocated (€ bn)
Allocated / gross	SDG 1 · No poverty	2.23	0.31
14 %	SDG 2 · Zero hunger	6.54	1.79
27 %	SDG 3 · Good health and well-being	20.04	12.25
61 %	SDG 4 · Quality education	3.17	0.87
27 %	SDG 5 · Gender equality	2.31	0.53
23 %	SDG 6 · Clean water and sanitation	4.17	0.74
18 %	SDG 7 · Affordable and clean energy	12.15	5.12
42 %	SDG 8 · Decent work and economic growth	10.33	2.10
20 %	SDG 9 · Industry, innovation and infrastructure	17.14	4.57

All 17 goals, published run, € bn; the percentage is allocated over gross.	Goal	Gross (€ bn)	Allocated (€ bn)
27 %	SDG 10 · Reduced inequalities	3.94	1.07
27 %	SDG 11 · Sustainable cities and communities	16.85	4.19
25 %	SDG 12 · Responsible consumption and production	16.47	4.38
27 %	SDG 13 · Climate action	22.70	5.90
26 %	SDG 14 · Life below water	8.05	1.66
21 %	SDG 15 · Life on land	8.55	2.23
26 %	SDG 16 · Peace, justice and strong institutions	7.29	2.47
34 %	SDG 17 · Partnerships for the goals	3.01	0.50
17 %	All goals	164.9	50.7

31 %

Health projects rarely serve another goal, so SDG 3 keeps 61 % of its gross amount. At the other end, SDG 1 keeps 14 % and SDG 17 keeps 17 %: poverty and partnerships are mostly touched by projects whose main subject lies elsewhere.

Rounding and checks

Shares are computed in integer cents and rounded half up, $(2 \cdot G \cdot R + \Sigma R) \operatorname{div} (2 \cdot \Sigma R)$, identically in the Python pipeline and in the prototype, so both give the same amount to the cent. Because each share is rounded, the shares of a project can differ from its grant by a few cents; `verify_scores.py` checks every project at both levels. In the prototype the denominator follows the current filters (minimum concept score, structural signals), so a filter that removes a weak link hands its share to the project's other goals.

What allocation is not

Allocation is a reading convention. It does not say how a consortium spends its budget: a project's share for a goal follows how strongly the project writes about that goal, not an audit of its work packages. Gross funding, for its part, overstates any single goal by construction. Read together, gross funding gives the upper bound of what touches a goal, and allocated funding a fair share of it.

Step 7 · White space analysis

The same graph read the other way round: which concepts the projects leave uncovered, measured 6 ways.

The question

The rankings answer "which projects serve this goal?". White space answers the reverse: which concepts of a goal are left with few projects, with projects that only mention them, with funding carried by one or

two grants, or with projects about to end. Those are the places where a new call or a new project would add the most.

The project–concept graph

The links of step 3 form a bipartite graph: projects on one side, concepts on the other, and an edge wherever a concept is found in a project, weighted by its score $s_{P,c}$. The simplest measure of how much a concept is covered is its degree, the number of projects it is linked to:

$k_c = \#\{ P : s_{P,c} > 0 \}$

Weighting the same edges differently (by score, by funding, by date) gives the other flavours of white space below. Each is computed on the current selection of projects, so the same concept can be well covered across Europe and white space in one country's portfolio.

6 measures	Flavour	Definition, for concept *c*
Column	Presence	number of projects linked to *c*
`n_projects`	Intensity	sum of the scores $s_{P,c}$ over those projects
`strength`	Depth	projects where *c* scores 2.00 or more, i.e. the project is about it (for example title + aim sentence)
`n_projects_deep`	Funding	EU contribution attributed to *c*: each project shares its grant across all its concepts in proportion to their scores, the rule of step 6 applied at concept level
`ec_allocated_concept_eur`	Robustness	effective number of projects carrying that funding, $(\Sigma f)^2 / \Sigma f^2$: 1 when a single grant carries it all
`n_effective_projects`	Future	projects still running on 1 January 2028
`n_projects_active_2028`		

The concept-level funding shares of a project add up to its grant, like the SDG shares: over all concepts they total €49.9 bn, the contribution of the projects with a concept link.

From 6 measures to one index

Each measure is turned into a percentile among the concepts of the selection, $p_m(c)$ between 0 (lowest) and 1 (highest), ties sharing their middle rank. The white-space index is the complement of their mean:

$WS_c = 100 \cdot (1 - (1/6) \Sigma_m p_m(c))$

100 means the least covered concept of the selection on all 6 measures, 0 the best covered. The 6 percentiles are published beside the index (`pct_presence` … `pct_future`), so every index can be taken apart, and concepts are ranked by it (`white_space_rank`).

Gap types

The index says how much white space there is; the gap type says what kind. The first rule that applies gives the type:	Gap type	Rule
Concepts	Empty	2 projects or fewer

The index says how much white space there is; the gap type says what kind. The first rule that applies gives the type:	Gap type	Rule
6	Shallow	10 or more projects, fewer than 5 % of them about the concept
16	Concentrated	5 or more projects, funding carried by fewer than 3 effective projects
4	Fading	5 or more projects, fewer than 30 % still running in 2028
5	Thin	white-space index of 75 or more
38	Covered	none of the above

182

Counts for the 251 fine concepts and all projects of the published run.

Most white space in the published run	Rank	Concept	Index	Gap type
Projects	1	Aid for trade and export capacity of developing countries	100	Empty
0	2	Harmful practices (FGM, child marriage)	99	Empty
1	3	Country ownership and aid effectiveness	99	Empty
1	4	Child poverty	99	Empty
2	5	Unemployment, disability and family benefits	98	Empty
2	6	Sustainable infrastructure in developing countries	97	Thin
3	7	Access to basic services	97	Thin
3	8	Fossil-fuel subsidies and phase-out	96	Empty
2	9	Water and sanitation cooperation with developing countries	96	Thin
3	10	Remittances	96	Thin
3	11	Pro-poor and gender-sensitive development strategies	96	Thin
5	12	Debt sustainability and investment in developing countries	96	Thin
6	13	Gender and technology (digital gender gap)	95	Concentrated
9	14	Women’s economic empowerment	94	Thin
6	15	Ocean acidification	94	Shallow

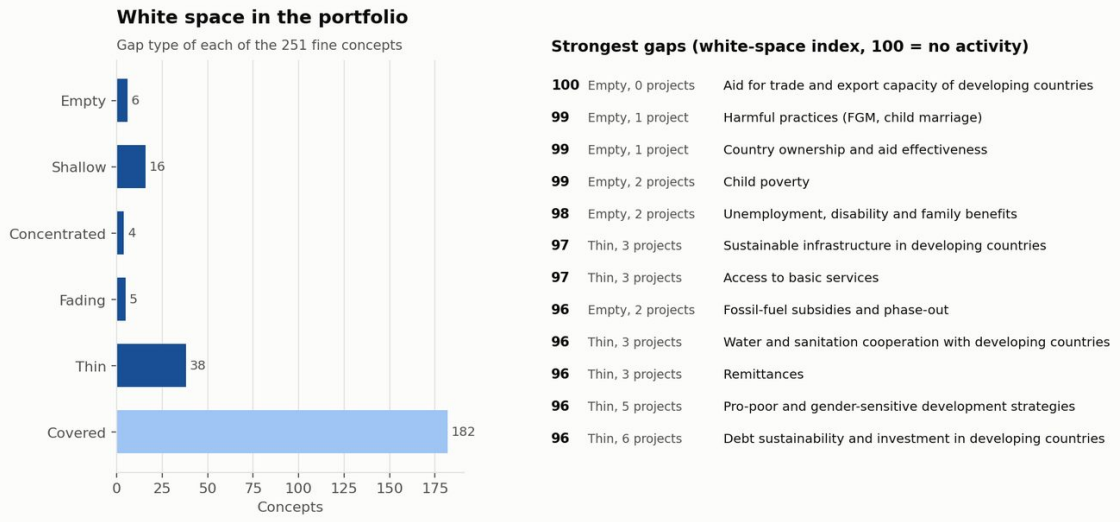


Figure 10. *White space: gap type of each concept and the strongest gaps.*

Beyond the empty concepts, each type points at a different need:

- **Shallow.** *Ocean acidification* appears in 12 projects and is the subject of none; so do *Food prices and agricultural markets* (15 projects), *Accessible and inclusive transport* (17) and *Access to affordable medicines* (18). The topic is present as context, not as research.
- **Concentrated.** *Deep sea and seabed ecosystems* has 94 projects, but its funding rests on the equivalent of 2.7 of them; *Small-scale and artisanal fisheries* has 9 projects and 2.4 effective ones.
- **Fading.** Of the 50 projects on *Road traffic safety*, 11 will still run in 2028; *Viral hepatitis* keeps 4 of 16 and *Underwater noise, shipping and oil pollution* 13 of 45.

Reading with care

Little coverage does not always mean a need for research. Three causes are mixed in a high index:

1. **Not a research topic.** The top of the ranking is dominated by development finance (aid for trade, remittances, debt, aid effectiveness). These targets are served by other EU instruments than Horizon Europe, and their white space here is expected.
2. **A vocabulary miss.** A concept can be studied in words its aliases do not cover (step 2). A high index is a reason to read the projects, not a verdict.
3. **A true gap.** What remains once the first two are ruled out is the white space worth supporting.

The index is relative: it compares the concepts of the current selection with each other. Selecting one goal, one country or one programme gives the white space of that portfolio.

Combinations, exploratory

A second kind of white space lies between concepts: 2 well-covered concepts of the same goal that rarely meet in one project. With N linked projects, independence would predict $k_a \cdot k_b / N$ projects combining a and b. In SDG 3, *Drug discovery and new therapies* × *Air pollution and health* meet once where 25 were expected, *Cancer* × *One Health and zoonoses* once for 16, *Cancer* × *Chemicals*

management and safe-by-design twice for 15. Many other low pairs are simply unrelated topics, so combinations need a human reading and are not part of the index.

In the prototype

The **White space** analysis angle gives one row per concept (or broad concept), most white space first, with the index, its rank, the gap type, the 6 measures and their percentiles. Everything is recomputed on the current selection: goals, targets, countries, programmes, organisation types and the minimum concept score. The published run is written by `white_space.py` as `CORDIS_SDG_WhiteSpace_250` and `_128`; the prototype, with every goal selected and the default filters, produces the same 2 files to the byte, and `verify_scores.py` rebuilds every figure in them independently.

Two BigNum codes

One code ranks the projects, the other stores every occurrence without loss.

The method uses the **BigNum** approach of the SGDL group (several values packed into one arbitrary-precision integer that decodes back exactly) in 2 places, for 2 different jobs. They are kept apart on purpose: a code built to *rank* projects and a code built to *identify* evidence obey different mathematics.

Pertinence ranks and decodes

It is a mixed-radix positional code with the score in its most significant digits: comparing 2 keys compares the scores first, then the counts that explain them. Sorting the column is the ranking, and every component is read back by position. The interface never shows it raw: it is the quantity the *pertinence* filter cuts on, and `evidence_dimensions` gives the same content as a plain Python list.

The evidence index identifies, without loss

Every occurrence of a concept in a project is written as 7 integers `[c, f, s, p, l, a, q]`: concept, field, objective sentence, character offset, length, alias and sentence role. The occurrences behind an assignment are put in a canonical order, laid end to end and coded as a single list with the hyper-pyramid indexing of the SGDL group, described in patent **US 9,900,024 B2** (Parallel Geometry Inc.; inventor Jean-François Rotgé). A list $[x_0 \dots x_n]$ is a point of the lattice T(m, n), whose degree m is the sum of the integers and whose dimension n is the length minus one; the lattice holds C(m + n, n) points, the list receives its rank X inside it, and the triplet $\langle X, m, n \rangle$ is folded into one integer by 2 Cantor pairings, with the patent's formula $(a + b)(a + b + 1)/2 + a$. Decoding reverses each step and returns every occurrence: the `evidence_index` column holds the complete textual evidence of a row in one cell.

Compact triplet form, not the dense global index

The patent's global index counts every point of every earlier lattice (about 2^{m+n}), so one occurrence at character 2,000 would already need 2,000 bits, and a list of them could not be computed at all; the triplet grows with the logarithm of C(m + n, n) instead. The price is that the code is one-to-one but not onto: every evidence list has its own integer, but not every integer is a valid code, and the decoder says so.

Never used to rank

Its magnitude grows with the sum of the coordinates (offsets, concept numbers), which says nothing about pertinence. Ranking is the pertinence key's job.

Size in practice

In practice the numbers stay modest: 57 digits at the median for a project × SDG row, 1,368 at most, about 5 million characters for the whole SDG ranking. The codec (`scripts/evidence_index.py`) was written by Dtangle from the public text of the patent, with a lexicographic order inside each lattice (the patent offers several). The prototype carries the same codec in JavaScript and reproduces every index of the published tables.

Worked example

ProACCT, from its summary to a score of 5.30 for SDG 3.

Project **ProACCT**, "Proteomic Analysis of Cell communication in Tumors":

The links found	Concept	Target	Found in	Score
Evidence	Cancer	SDG 3.4	title 1.00 + aim sentence 1.00 + 9 distinct sentences 0.50 + keywords 0.80	3.30
"…Cell communication in Tumors"	Diagnostics and personalised medicine	SDG 3.8	impact sentence	0.90
"…will provide the basis for advanced personalized treatment."	Drug discovery and new therapies	SDG 3.b	method sentence	0.80
"…critical for the development of novel therapeutic approaches."	EuroSciVoc *clinical medicine / oncology / skin cancer / melanoma*	SDG 3.4	structural signal, direct	0.30

category assigned to the project in CORDIS

Reading the result

For SDG 3 the project scores **5.30** = 5.00 of text + 0.30 of structural signal, 1,101st of the 7,918 projects linked to SDG 3. Its pertinence reads `0000530 0030 003 001 001 001 001 000 000 000 001 001 000 000 000 000 030`: 5.30 in total, 0.30 structural, 3 concepts (1 in the title, 1 in an aim sentence, 1 in an impact sentence, 1 in a method sentence, 1 repeated in 4 or more sentences, 1 in the keywords) and 0.30 from the EuroSciVoc family. Its coverage is 0.065 (3 of the 46 concepts of SDG 3), and its €2.0 m EU contribution is allocated entirely to SDG 3, the only goal it touches. Its evidence index, 666 digits long, decodes into the 27 occurrences behind the score; the decoder of the prototype reads it back. The example also shows a limit: the project is mostly fundamental proteomics, and a human reader would call its link to health real but indirect.

Traceability

Every figure can be rebuilt from the published evidence tables, and decoded in the prototype.

Every figure published here can be taken apart, at the granularity where the evidence is produced:

Evidence tables	Table	One row =
Carries	CORDIS_Concept_Occurrences	one occurrence (132,369)
field, sentence, offset, length, alias and sentence role, the matched text itself, and the 7-integer tuple the evidence index is built from	CORDIS_Concept_Evidence	project × concept × field (86,592)
the best occurrence per field with its points and its sentence, plus one row for the repetition bonus; the rows of a link add up to its score	CORDIS_Structural_Signals	project × rule × SDG (82,658)
the raw CORDIS value (e.g. policy_clean_air_pct = 100 %), the rule and its rationale, whether it was counted, whether it can create an entry	SDG_Structural_Rules	one of the 199 rules

family, code, SDGs, strength, weight, rationale

Independent verification

A verification script, verify_scores.py, rebuilds **every** published figure from those tables alone, never from the pipeline's internal state, the way a reviewer would. On the current run it checks 46,315 concept scores, 48,290 SDG rows and 73,893 target rows with zero mismatch; it confirms that each of the 132,369 recorded (sentence, offset, length) positions really points at the recorded alias in the project text; it re-encodes all 168,498 evidence indexes from the occurrence table and decodes 4,500 of them in full, without a single difference; it checks that every project's allocated funding adds up to its EU contribution; and it rebuilds the 2 white-space tables of step 7 with exact fractions, from the project × concept links and the project data, without a single difference in counts, funding, index or gap type. It earns its keep: on its first run it caught 2 rules of the same family with equal weight both being counted, which inflated 9,038 rows.

Prototype parity

The prototype is held to the same standard. It recomputes scores, ranks, coverage, allocated funding and evidence indexes in the browser rather than reading the published tables, and a parity test runs its own code over the full corpus: all 48,290 SDG rows and 73,893 target rows match the published CSVs exactly, down to the cent and to the last digit of each evidence index.

Reproducibility

Plain Python scripts, run in a fixed order, produce every published file.

Scripts, in run order

All steps are plain Python scripts, run in this order:	**Script**
Role	sdg_concepts.py
Concept layer: concepts, broad concepts, roles, targets, rationales, exclusions, required context, excluded targets	sdg_concept_aliases.py

All steps are plain Python scripts, run in this order:	**Script**
Editable alias lists (built by `derive_aliases.py` with `scan_surface_forms.py`)	`alias_regex.py`
Normalisation and the alias index	`sdg_score.py`
The weights, the repetition bands, the encoded score and the pertinence key: encode, decode and self-test	`evidence_index.py`
The evidence index after US 9,900,024 B2 (compact triplet form): encode, decode and self-test	`sdg_rules.py` · `sdg_structural.py`
The 199 structural rules (trackers, programme parts, missions, EuroSciVoc) and how they apply to a project	`match_concepts.py --shard i/8`
Matching, sentence roles, per-field weights and evidence (8 shards, run in parallel)	`build_concept_layer.py`
Assembles the mapping tables, the rankings, coverage, allocated funding, the evidence tables and the evidence indexes	`white_space.py`
The white-space tables, one row per fine and per broad concept (step 7)	`results_report.py`
The figures of the Results tab and the Dtangle studies behind its graph views (goals, targets, concepts; 3 series)	`verify_scores.py`
Rebuilds every score from the evidence tables, re-encodes and decodes the evidence indexes, checks positions, ranking order and funding, and rebuilds the white-space tables	`mkxlsx_concept_layer.py`
The 3 reading workbooks	`sdg_short_names.py`
Short names of the 169 targets (editorial)	`explorer/pack.py`

Packs the tables into the data files used by the prototype

Outputs

Outputs: the concept tables (fine and broad), the target table, the structural rule table, the project × concept mapping, the occurrence and evidence tables, the structural signals, the SDG and target rankings, the white-space tables (fine and broad), and a project data table with funding, participants and countries. All are UTF-8 CSV with list cells written as `['a','b']`.

5. Validation, limitations and uncertain cases

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.

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.

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.

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.

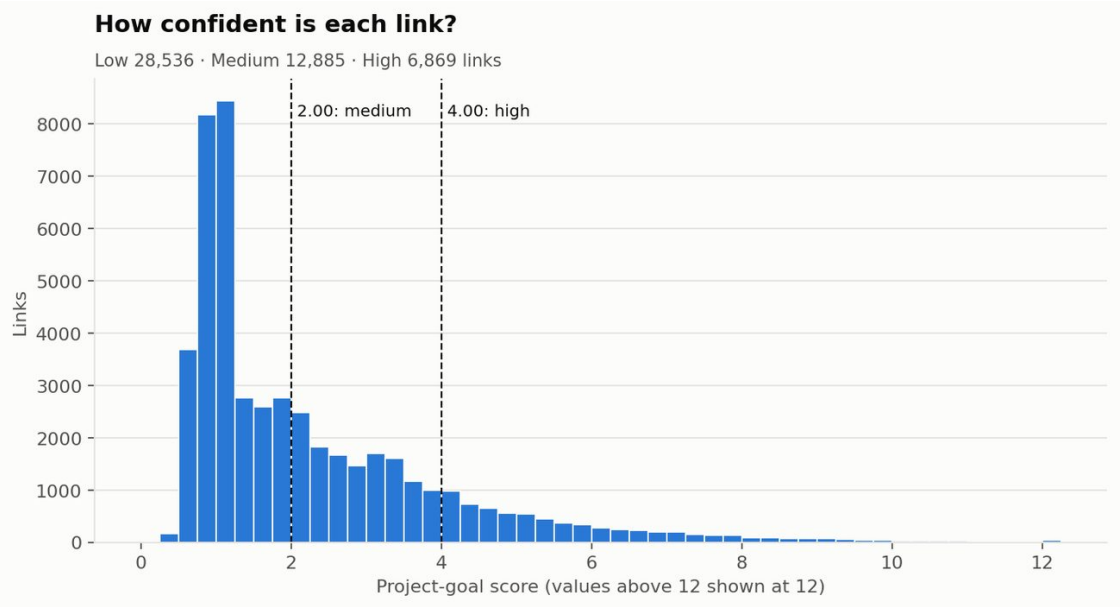


Figure 11. *Distribution of the project-goal scores and the 2 confidence thresholds.*

Earlier iterations and integrity checks

Iterations

A first version used structured CORDIS signals only (topics, programme parts): it linked 9,781 projects and found nothing for SDG 1. A second, target-level version was tuned on a pilot of 150 projects and tested on a blind hold-out of 60: agreement with a close reading of those projects rose from F1 0.55 to 0.72 at goal level. The current concept layer replaces both.

Integrity checks

Integrity checks on every output: parallel lists (targets ↔ goals, concept ids ↔ labels) are aligned on every row, and all 13 files pass the Dtangle CSV validator. Every published score is rebuilt from the evidence tables by `verify_scores.py`, and the prototype reproduces the published rankings exactly (section 4, Traceability).

Evaluated and set aside

Three refinements were tested and not adopted. A specificity weight per concept (fewer targets, more weight) barely varies here: the concepts were built target-specific, so 246 of the 251 would fall in the 2 top bands. Letting structured CORDIS signals *nominate* SDGs and text only confirm them would lose most of the corpus: ERC and MSCA projects (15,060, 64 %) carry no thematic programme, and an earlier version built that way reached 9,781 projects and none for SDG 1, against 17,104 now. Ranking on the size of a hyper-pyramid code was rejected for the reason given in section 4 (Two BigNum codes): its magnitude follows the character positions, not the pertinence.

Partial structural coverage

Programme rules only cover the destinations that were documented, and a cluster inherits an SDG only on unanimity, so about 3,100 projects carry a programme signal. The digital-agenda and AI trackers are left unmapped: their only plausible targets (9.5, 9.c) are excluded or too broad to be informative.

Editorial choices

Target short names and concept boundaries are the project's own and can be discussed. They are kept in plain files so they can be revised.

6. The project-to-SDG mapping table

One row per project-goal assignment: 48,290 rows, 17,437 projects. The complete table is published as one CSV file (UTF-8, 38 MB) at https://cordis.dtangle.ch/downloads/CORDIS_SDG_Mapping_Table_260924.csv, with a compressed copy and a 1,710-row sample (the 100 top-ranked projects of each goal), all listed in the Data tab of the prototype.	**Column**
Content	project_id, acronym, project_title, cordis_url
CORDIS identifier, acronym, title, link	sdg, sdg_label, sdg_uri
assigned SDG concept and its EU Vocabularies URI	sdg_targets, sdg_target_uris
narrower SDG concepts (targets) and their URIs	thematic_concepts
the concepts found, which connect the project to those targets	confidence_score, confidence_level
score, and High (4.00 or more) / Medium (2.00 or more) / Low	rank_in_sdg
rank of the project among those linked to the goal	explanation
how the score is made: each concept with its matched terms and fields, each structural signal	evidence_excerpt
the best sentence (or the title) supporting the assignment	source_fields
CORDIS fields used: title, acronym, objective, keywords, topics, legalBasis, euroSciVoc, policy trackers	evidence_basis

concepts only / concepts and structural signals / structural signals only

Example: SMOOTHER, SDG 7, High, score 11.70 = concepts 11.40 + structural 0.30. Explanation: *Electric mobility and charging infrastructure [3.05]: 'charging stations', 'electric vehicle' in title, objective (aim), objective (repeated), keywords; Energy storage and batteries [2.80]: …; EuroSciVoc renewable energy (+0.30)*. Targets 7.1, 7.2, 7.3.

A separate target-level ranking gives 73,893 project-target assignments with the same columns.

The table is written by `build_submission_table.py` from the same published tables as the prototype, so it carries exactly the same scores. It opens directly in a spreadsheet.

7. The prototype: CORDIS SDG Explorer

The prototype is a single web page with no server code. It loads the published tables and recomputes scores, ranks, coverage and allocated funding in the browser; a parity test confirms that it reproduces all 48,290 SDG rows and 73,893 target rows of the published tables exactly. It is prepared by Dtangle and carries the Dtangle logo.	**Tab**
What it offers	About us
Dtangle, its platform and its team	Results
the findings of section 3, interactive, with 3 series (all links, score ≥ 2.00, score ≥ 4.00), a fine / broad concept switch, network views of goals and targets, and an In Dtangle section with eight graph views from the Dtangle platform, each opening the study in Dtangle	Explorer
pick goals and targets, filter (countries, programmes, organisation types, minimum score, structural signals on or off), choose an analysis angle (projects, project × concept, SDG or target ranking, organisations, white space, concept reference), pick columns, preview and export a CSV ready for the Dtangle platform; every ranking can be traced back to its evidence	Data
the complete mapping table (CSV, zip, sample), the column reference, the source datasets, the JSON tables the page loads, and the licence	Methodology
the method of section 4, with a decoder for the evidence codes	Future work

the next code, section 8

The exported studies open in the Dtangle visual analytics platform (https://trials.dtangle.ch/, login `cordis@dtangle.ch`, password `sdgmap`), where projects, goals, targets and concepts become live graphs.

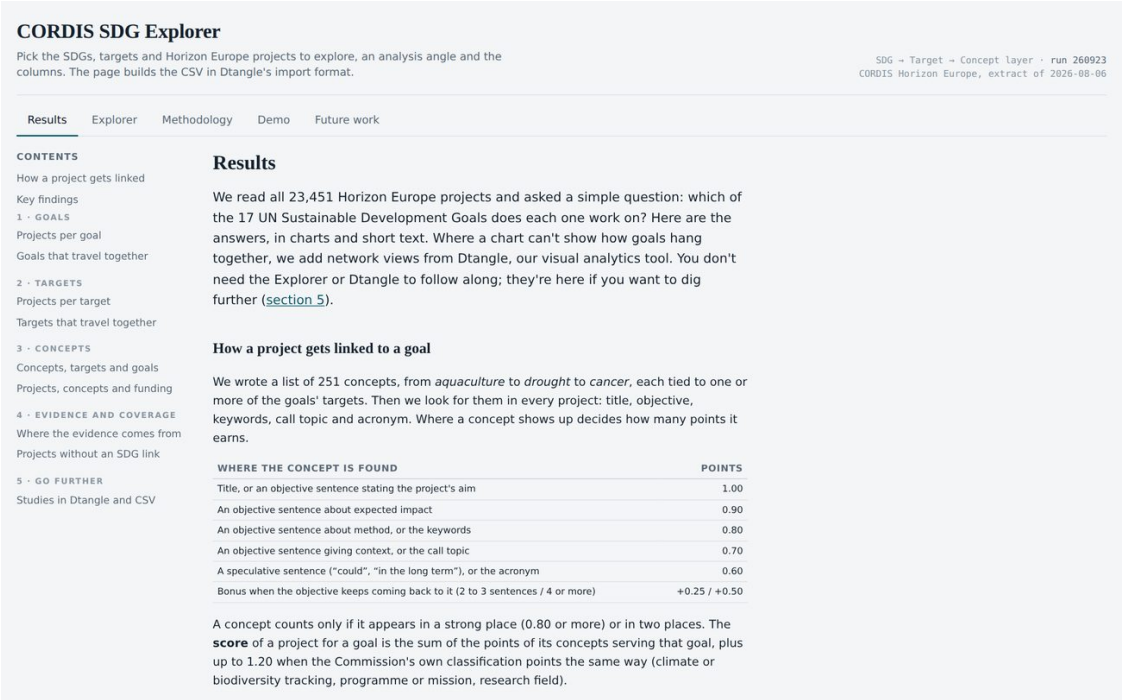


Figure 12. *Results tab: how a project gets linked to a goal.*

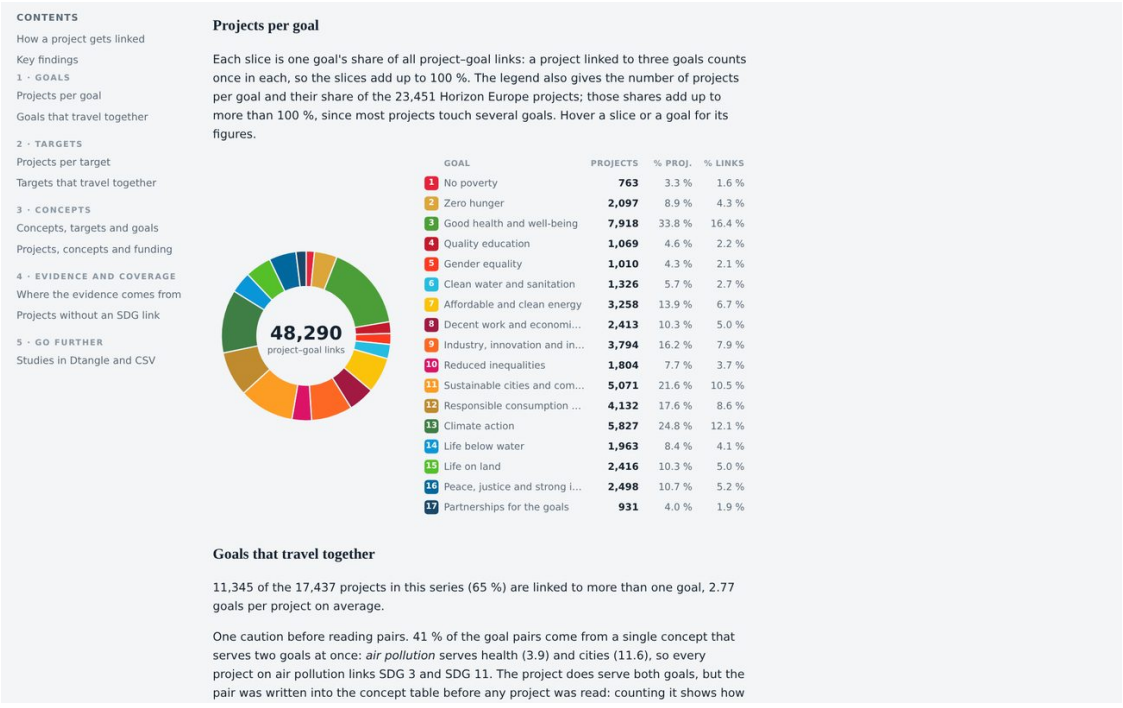


Figure 13. *Results tab: projects per goal.*

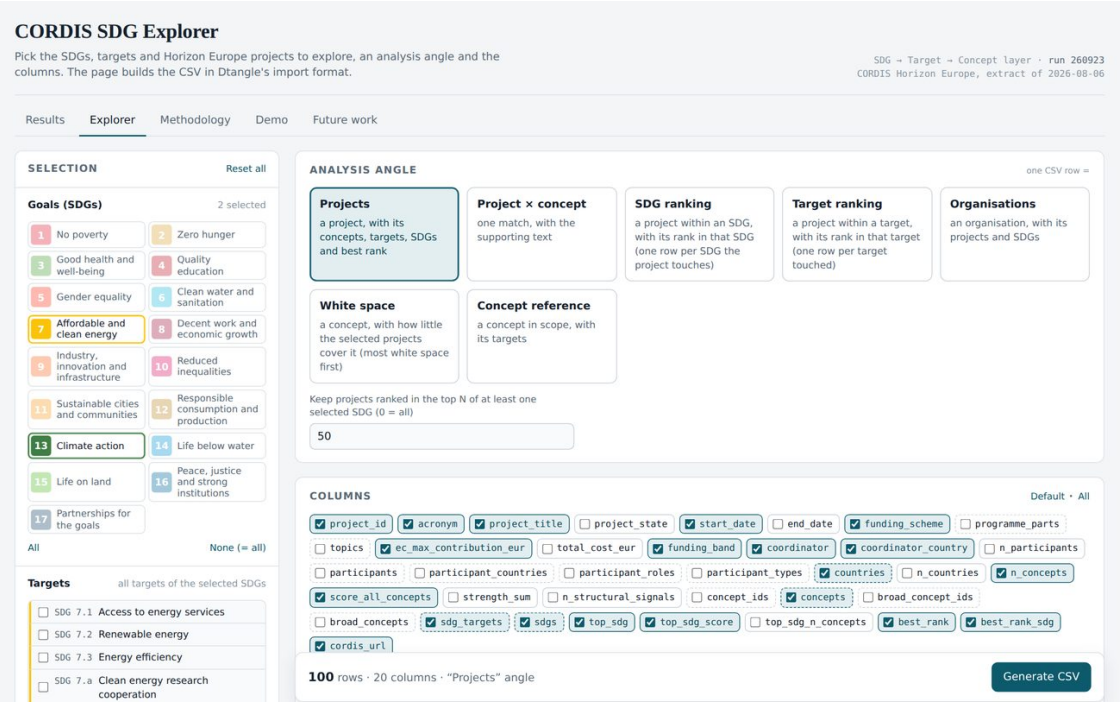


Figure 14. *Explorer tab: goals selected, analysis angle and columns.*

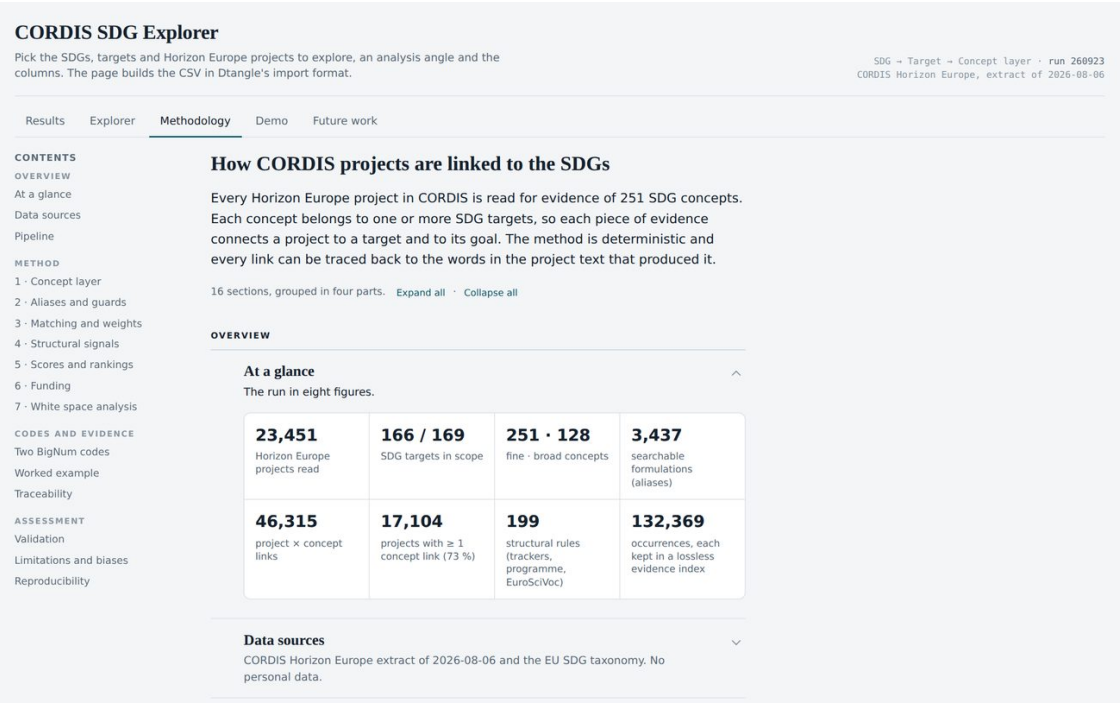


Figure 15. *Methodology tab.*

8. Future work

1 · The question the 2 codes leave open

The method already writes 2 codes for every project and goal (section 4, Two BigNum codes). Each answers one question:	**Code**	**Question it answers**
What it guarantees	**Pertinence key**	Which project is most pertinent for this goal?
Sorting the keys is the ranking; every component decodes back by position.	**Evidence index**	Exactly why was the project assigned?
One-to-one: decoding returns every occurrence, down to the character offset.	**Neighbourhood index** proposed	Which projects have a similar evidence profile?

Projects with close indexes have close profiles (the converse does not hold, see section 3).

Neither existing code answers the third question. Two projects can share the same pertinence score for very different reasons, and 2 evidence indexes that differ by one occurrence can be far apart as integers.

2 · A third code: the neighbourhood index

Dtangle plans to use n-dimensional indexing: an n-dimensional space is meshed into cells, the cells are chained by a space-filling metacurve, and every point receives the position of its cell along that curve. We would apply it to a short profile of each project and goal:

x_P = (title evidence, concept coverage, objective evidence, structural evidence, target specificity, funding band, …)

Each axis is normalised to the unit interval, the mesh is fixed, and the position of the cell holding x_P along the metacurve is the neighbourhood index L_P. It is one integer, like the other two codes, and it can be written as one more column in the CSV.

3 · What it cannot promise

- **Locality holds one way.** Close indexes mean close profiles, but 2 close profiles can land far apart on the curve when they sit on either side of a cell boundary. The effect grows with the number of axes and becomes marked beyond 4 or 5.
- **It is reversible only to the cell.** Decoding L_P returns the cell, not the exact profile, and the profile is itself a summary. The complete evidence stays the job of the evidence index.
- **The neighbours depend on choices.** The axes mix continuous scores and ordered bands. How each axis is scaled and weighted, and how fine the mesh is, decides which projects count as neighbours. These choices would be published with the index.
- **It does not beat a direct comparison.** With 23,451 projects, comparing one profile directly with all the others (exact nearest neighbours) takes a few milliseconds. The index does not find better neighbours; it finds them from a single sorted column.

4 · What it would be for

The value is practical rather than analytical: a compact, stable key that puts similar projects next to each other.

- Sorting a CSV by L_P in the Dtangle platform or a spreadsheet places projects with similar profiles in adjacent rows, with no extra tool.
- A range of indexes selects a region of the profile space, which a database can serve from an ordinary index.
- Blocks of consecutive indexes give ready-made groups to browse or compare across runs, as long as the mesh and the scaling stay fixed.

5 · How we would test it

Before publishing the column we would measure how well it keeps neighbours together. For each project, take its k nearest neighbours by direct comparison of profiles, then its k nearest by index, and count the overlap (recall at k). Repeat for 2 to 8 axes and several mesh sizes. The index goes into the export only for the settings where recall stays high, and the page would report the figure next to the column.

Licence. The mapping tables, the validation and uncertain-cases files, the figures and this document are released under Creative Commons Attribution 4.0 (CC BY 4.0): CORDIS and the European Data Portal may reuse them freely, with attribution to Dtangle SA. The prototype and the Python scripts remain © Dtangle SA 2026; CORDIS and the European Data Portal may reuse them for their own purposes, and the source is available to the evaluators on request. Third-party data (CORDIS, EU Vocabularies) keep their own licences.

Contact

Dtangle SA

La Voie-du-Coin 28

1218 Le Grand-Saconnex, Geneva, Switzerland

contact@dtangle.ch