What CoARA Action Plans Commit To

Ricardo Hartley Belmar[1], Veronika Schöpf[2]

1. Independent researcher, Linz, Austria, ORCID 0000-0001-5058-9309, ricdho@duck.com (corresponding author)
2. IT:U Interdisciplinary Transformation University Austria, Linz, Austria, ORCID 0000-0003-1937-4216, veronika.schoepf@it-u.at

Keywords: research assessment reform; CoARA; action plans; policy document analysis.

Submitted version (preprint) – not yet peer reviewed

This is a preprint published in fteval Preprints, the open repository of the Austrian Platform for Research and Technology Policy Evaluation (fteval). It has been screened for formal quality and thematic fit, but has not been peer reviewed. The content is the sole responsibility of the author(s).

Open review: This preprint has been submitted to the fteval Journal for Research and Technology Policy Evaluation and is undergoing open peer review. Review reports are published with their own DOIs and linked to this preprint. If the article is accepted, the final version will appear in the journal under a separate DOI and will be linked here. This preprint remains available in any case, together with its review reports.

How to cite this version:

Hartley Belmar, Ricardo and Schöpf, Veronika (2026) *What CoARA Action Plans Commit To*. Working Paper. fteval Preprints. (Submitted)

Please check the DOI for the most recent version before citing: 10.22163/fteval.2026.770

License: CC BY 4.0

Abstract

The Agreement on Reforming Research Assessment asks each signatory to share an action plan. We analysed the 335 CoARA action plans that a search of Zenodo returned up to 6 August 2026. The reading is descriptive: it records which actions each plan commits to and whether each is in place or planned. Language-model agents followed a documented procedure: one extracted the actions a plan commits to, each with a verbatim quote; a second matched them to the 62-action catalogue of an open planning tool; a third tried to refute every record; a fourth settled the disputed matches; a script checked every quote against the plan text. The coding rules were set after 50 records of an earlier reading of the same plans were coded by hand. In all, 8,870 records passed the checks of quote, commitment and wording, and are counted. The most common action is to redesign evaluation rubrics (64.8% of the plans).

Within the commitment to abandon inappropriate uses of journal- and publication-based metrics, the median action is taken by 9.0% of the plans, the lowest of the ten commitments. Of these records, 67% are planned rather than in place, most of all under the commitment to communicate progress. Three items of DORA, the Leiden Manifesto and SCOPE are linked only to actions that fewer than one plan in ten takes, and where keyword counts and reading can be compared, they disagree in both directions: the keywords find more plans for seven of 22 actions and fewer for 15. No sample coded by people measures this reading's accuracy yet; we report its internal checks and a second run on 30 plans. The catalogue and the counts are built into the Research Assessment Reform Planner, an open tool in which an institution assesses itself against the ten commitments and drafts an editable plan from the actions it proposes, each shown with the share of plans that take it.

1. Introduction

The Agreement on Reforming Research Assessment, dated 20 July 2022, asks the organisations that sign it to change how they assess research, researchers and research units (CoARA, 2022). It sets out ten commitments, four core and six supporting. Each signatory is to share an action plan with milestones by the end of 2023 or within one year of signing, and to have worked through at least one cycle of reviewing its assessment criteria by a touch point at the end of 2027 or five years after signing. The Agreement builds on earlier texts: it names DORA (2013), the Leiden Manifesto (Hicks et al., 2015) and the Hong Kong Principles (Moher et al., 2020) as the texts that inspired the reforms already under way.

The Agreement does not say that the plans must be public; CoARA's guidelines invite signatories to deposit them in Zenodo's coara_action_plans community (CoARA, 2023, p. 2). Many plans are on Zenodo, so anyone can read what those institutions have promised. The plans follow no template: the guidelines offer guiding questions "as a resource and suggestion" (CoARA, 2023, p. 2). They differ in length, language and structure, and a reader who wants to know what the signatories, taken together, commit to has to read hundreds of documents written in different ways.

Earlier studies have looked at the plans, and at who takes part in CoARA, from several angles. Rushforth et al. (2026) counted open science terms in 248 plans and found most terms in fewer than one plan in five, with value terms outnumbering operational ones among the most mentioned. Pölönen and Rushforth (2025) measured how many European higher education institutions take part in CoARA. An analysis commissioned by the European Commission counted 263 action plans submitted by 8 January 2026 (265 less a duplicate and a corrected version), extracted information on each commitment from 235 plans with AI and on implementation methods and tools from 165 by hand, and interviewed fifteen signatories (European Commission, 2026); its first results were presented to CoARA's General Assembly in June 2026 (Stroobants, 2026), and a research-in-progress conference paper of September 2026 draws on them (Gadd et al., 2026). In the literature we found, no study reads every plan at the level of the actions it commits to, records whether each action is in place or only planned, and makes the records available so that others can check them.

We analyse the plans in that way, and ask four questions of them:

1. What do the plans commit to: which reform actions, how often, and under which commitments?
2. How firmly: is each action in place, planned, or unclear?
3. What is missing: which actions recommended by DORA, the Leiden Manifesto and SCOPE do almost no plans take, and how far does counting by keyword differ from reading?
4. How far can a reading done by language models be trusted, when no person has yet coded a sample of its records?

Section 3 describes the reading. Its unit is the catalogue of reform actions of the Research Assessment Reform Planner, an open self-assessment tool for CoARA signatories that the first author built and that an earlier reading of the plans revised (Hartley Belmar, 2026). The article contributes three things: the counts and what they show (Section 4), a dataset of every record with its match, its verdict and, where the plan's licence allows, its quote, together with the instructions each agent received (see Data and code availability), and the planner, which turns the counts into guidance for an institution writing or revising its plan (Section 5).

2. Background

2.1 Reform frameworks and their critics

Research assessment reform has a decade of reference texts. DORA (2013) asks institutions and funders not to use journal-based metrics as a surrogate for the quality of individual work. The Leiden Manifesto (Hicks et al., 2015) sets out ten principles for the use of research metrics, and The Metric Tide (Wilsdon, 2015) reviewed their role in research assessment in the United Kingdom. The Hong Kong Principles (Moher et al., 2020) tie assessment to research integrity. SCOPE (Himanen et al., 2024) turns these ideals into a five-stage process for designing an evaluation, and the Barcelona Declaration on Open Research Information (2024) asks organisations to make openness the default for the research information they use. CoARA gathers signatories around ten commitments that build on DORA, the Leiden Manifesto and the Hong Kong Principles (CoARA, 2022).

The reform also has critics. Abramo (2024) questions whether qualitative judgement should prevail over quantitative methods in every context. Bornmann (2026) weighs what signing gives a research organisation, legitimacy and a push to modernise, against what it costs. Morgan-Thomas et al. (2024) ask whether declarations change outcomes: in the business and management sub-panel of the United Kingdom’s 2021 assessment exercise, the association between journal rankings and expert evaluations did not change despite institutional endorsements of DORA. We treat a plan as a statement of intent, not as evidence of practice.

2.2 Language models as readers of documents

Reading hundreds of plans by hand takes time, and keyword counts miss context. Language models can read at scale, and a growing literature measures how well they do it. Gilardi et al. (2023) found that a language model outperformed crowd workers on several annotation tasks for tweets and news articles. Ziems et al. (2024) tested many models on social science benchmarks: on classification they fell short of fine-tuned models but reached fair agreement with human coders. Chae and Davidson (2026) compared models and training regimes for stance detection, and Gunes and Florczak (2025) applied them to the classification of policy documents. Two concerns run through this work. Models produce fluent text that is not grounded in the source (Huang et al., 2025), and their results are usually checked against labels that people assigned (Pangakis & Wolken, 2025; Ziems et al., 2024).

The method answers the first concern with a quote that a script looks up in the plan for every record, so a record the source does not contain cannot count. It answers the second only in part: people's coding of an earlier reading set its rules, but no sample coded by people measures the reading reported here. Section 3.4 sets out what the internal checks show instead, and Section 7 what they do not.

3. Data and methods

3.1 Corpus

The corpus is the set of CoARA action plans that searches of Zenodo returned up to the last harvest, on 6 August 2026: the records that match the text query "coara action plan" and the records in Zenodo's coara_action_plans community. The planner's extraction script took the text from each deposited PDF or DOCX file, using pdftotext for PDF files. Two harvested records yielded no text and are not analysed. Two of the plans are versions of one Zenodo record, Technological University Dublin's of September 2024 and July 2026; both are kept, so that institution counts twice. The corpus therefore holds 335 plans, published on Zenodo between 4 October 2023 and 4 August 2026. Plans deposited after the harvest are not part of it.

The extraction agent (Section 3.3) recorded each plan's institution, type of organisation and language. Table 1 gives the types. The plans are written in nine languages, and 301 of them in English only. Section 3.5 sets out the plans' licences and what the dataset may quote from them.

Table 1: Plans in the corpus by type of organisation, as recorded by the extraction agent.

Type of organisation	**Plans**	**Share**
University	220	65.7%

Type of organisation	Plans	Share
Other research-performing organisation	35	10.4%
Network or association	34	10.1%
Funder	34	10.1%
Agency	9	2.7%
Other	3	0.9%

3.2 The catalogue of reform actions

The reading used the planner's catalogue of reform actions (Hartley Belmar, 2026). Its first questions and actions came from reading 15 published plans by hand. The catalogue was then revised during an earlier reading of all the plans by agents of the same model, on 28 and 29 September 2026. That reading started from 47 actions; agents grouped the records that matched no action, a sceptical agent judged each group that four or more of the first 40 plans took, and the groups that entered the catalogue were decided by hand. After that reading, actions that almost no plan takes were folded into their neighbours and frequent practices the catalogue did not cover were added, which gave 61 actions. Coding records from that reading by hand (Section 3.3) added one more: to tell researchers what will count, and encourage them to show it. The reading reported here used these 62 actions.

Each action belongs to one of the ten commitments of the Agreement. Of the 62 actions, 42 are linked to the items of DORA, the Leiden Manifesto or SCOPE that they put into practice. The planner publishes this catalogue, with the counts of this reading (Section 5). Supplementary file S1 lists the 62 actions with their commitment, the framework items they put into practice and their uptake.

3.3 Reading by language-model agents

Language-model agents read the plans, following a written procedure designed for this study. Each agent was an instance of Claude, model claude-opus-5-5, started through the Workflow tool of Claude Code (version 2.1.309, on a laptop with an Apple M1 Pro processor and macOS 26.6) with a fixed instruction and, where the step required one, a fixed output schema. The model ran on Anthropic's servers; the Workflow tool offers no setting for temperature or seed, so a run cannot be repeated exactly, and Section 3.4 reports a second run. Nothing else reached an agent: it read the plan text and the catalogue from the files its instruction named, and some steps also received earlier records written into the instruction itself. The instructions, schemas and scripts are part of the dataset (see Data and code availability), in the order they ran, on 27 and 28 October 2026. The procedure was revised before this reading. The earlier reading of the same plans (Section 3.2) had told its agents to give no action when in doubt.

From that reading, 50 records, drawn at random (seed 20261027) with at most one per plan, were coded by hand, and 45 were completed. Of the 15 records it had left without an action, a catalogue action was judged to fit 14; its status was confirmed for 32 of 41 records (Cohen’s κ 0.43). These records, and four pilots on ten plans whose output was compared with them, gave eight coding rules: when an activity is implemented, planned or unclear; what counts as a commitment; how close a match must be; where steps about the plan itself go; when training is an assessment action; and that a status the checker rejects is replaced rather than the record dropped. Because these records were used to write the rules, they cannot measure the accuracy of the reading that followed. The reading had six steps, numbered in Figure 1

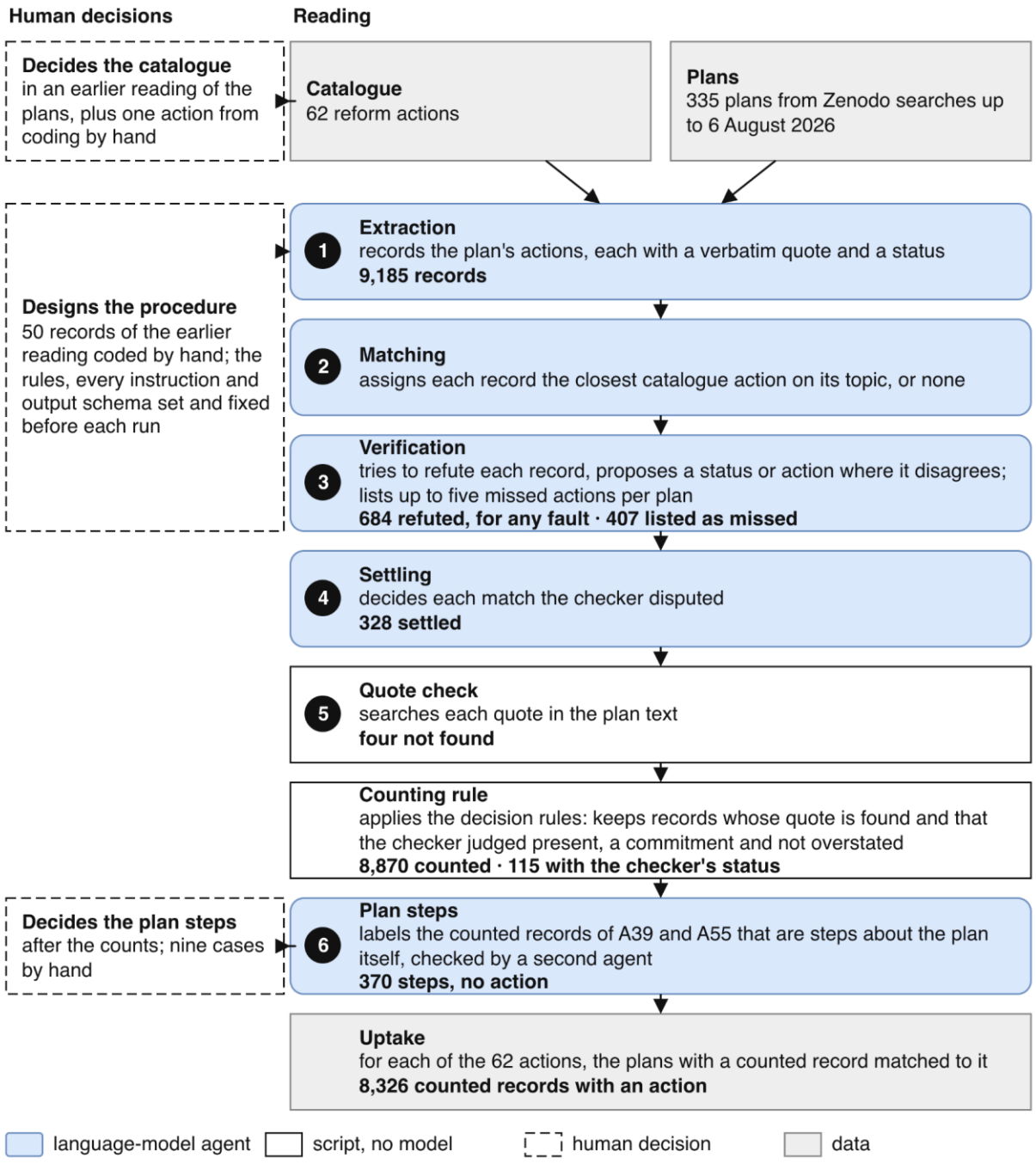


Figure 1: The reading procedure, with the human decisions on the left. The numbers are the six steps described below; the counting rule applies the decision rules. Every agent is the same model, claude-opus-5-5. Counts from the dataset.

1. Extraction. For each plan, one agent read the full text, with the instruction to record every action the plan commits to, and recorded each with a verbatim quote, the commitment it serves and the mechanism, and, where the plan states them, the

responsible unit, the timeframe, an indicator and a budget. It gave each record a status: implemented, planned or unclear.
2. Matching. A second agent matched each record to the closest catalogue action on the same topic, or to none when no action is on that topic.
3. Verification. A third agent, instructed to refute a record unless it was clearly supported, read the plan again and judged each record: whether the quote is in the text, whether it is a commitment, whether the record says more than the quote, whether the status is defensible, and whether the match is right, proposing a status or an action where it disagreed. It also listed, with quotes, up to five actions per plan that no record captured.
4. Settling. For each match the checker disputed, another agent chose the matcher's action, the checker's, another, or none.
5. Quote check. A script searched every quote in the plan text, exactly or after collapsing whitespace. No model took part in this step.
6. Steps about the plan itself. When the counts were first computed, most of the counted records matched to two actions, to publish and deposit the plan and to write the reform into HRS4R and existing strategy cycles, only described drafting the plan, called it a living document or scheduled its review. After these counts were seen, such records were set to count for no action. An agent labelled each of the 638 counted records matched to the two actions and a second agent checked every label, after the first had labelled 16 planted cases correctly; nine cases, where the agents disagreed or the record named an HRS4R process and nothing more, were decided by hand.

The full reading ran as one workflow. Usage limits, and once the workflow's cap on the number of agents in one run, stopped it five times; each time it was resumed without running finished agents again. The full reading, the pilots, the second run on 30 plans (Section 3.4) and the labelling of step 6 used 121,811,757 tokens in all.

The procedure, its decision rules and the catalogue were set by hand; the instruction texts and scripts were drafted by Claude Code under human direction and fixed before each run. The agents judged each record within those instructions; they did not set the catalogue, the rules or the count. The use of generative AI in writing the article is described in the Appendix.

Decision rules. A record counts when its quote is found in the plan text and the checker found the quote present, the record a commitment and nothing overstated. Where the checker found the status indefensible, the status it proposed replaces the extracted one. A record's action is the matcher's when the checker accepted the match and the settling agent's when it did not; a step about the plan itself has none. The uptake of an action is the number of distinct plans with at least one counted record matched to it. These are the figures the planner shows as the share of plans that take each action.

3.4 Internal reliability checks

Studies of language-model coding usually evaluate it against labels that people assigned (Pangakis & Wolken, 2025; Ziems et al., 2024). This reading has no such sample yet. Its reliability is reported from its own checks, each computed from the dataset (Table 8): the

mechanical quote check, which shows whether every counted record quotes text that is in its plan; the checker's refutations; the agreement between matcher and checker; the outcomes of settling; the checker's list of actions the extraction missed; and a second run of the whole procedure on 30 plans drawn at random (seed 20261028) from plans whose records were not coded by hand. The quote check rules out invented quotes among the counted records; the checker's verdicts and settling are meant to catch overstated records and unstable matches, and can miss some; the second run shows how much the counts move when the reading is repeated. None of the checks can show agreement with a human reader, and because every agent is the same model under different instructions, they share any blind spot of that model. An action that neither the extractor nor the checker saw is absent from every figure. The numbers in this article are computed from the dataset when the text is produced. Those that come from the reading were also recounted from the reading database, with queries written separately from the analysis code, and the recount agrees with the text.

3.5 Licences and quotations

Each plan keeps the licence its depositor chose. The licence of every plan was retrieved from the Zenodo REST API (metadata.license.id) on 29 September 2026 (Table 2); all 335 plans are open access. The reading used the full text of every plan, but the dataset does not republish the plans: every record points to its plan by DOI.

A record does reproduce part of a plan, its verbatim quote, so the dataset keeps quotes only where the plan's licence allows reuse on terms a CC BY 4.0 dataset can meet. CC BY 4.0 asks for attribution, which the dataset gives through each record's DOI and the plan's title, institution and licence in its list of plans; CC0 asks for nothing. We treated a share-alike licence as requiring the dataset to carry the same licence, a non-commercial licence as restricting how the dataset can be reused, and a plan without a licence as granting no reuse beyond what copyright law allows. For the 20 plans under those terms, the dataset leaves empty the quote fields of their 588 records, and of the missed actions the checker listed for them. For every one of these records it keeps the character position at which the script found the quote, so a reader can find the passage through the DOI. The remaining 8,597 records keep their quotes. The dataset is licensed CC BY 4.0, and its scripts MIT.

Table 2: Licences of the plans, as their Zenodo records state them, and whether the dataset keeps the verbatim quotes of their records.

Licence	**Plans**	**Quotes in the dataset**
CC BY 4.0	314	kept
CC BY-SA 4.0	13	removed
No licence	3	removed
CC BY-NC-SA 4.0	3	removed

Licence	Plans	Quotes in the dataset
CC0 1.0	1	kept
CC BY-NC 4.0	1	removed

4. Results

4.1 What the plans commit to

Figure 2 shows, for each commitment, how many plans take each of its actions. The most common action is to redesign evaluation rubrics, taken by 64.8% of the plans; the second is to join a regional forum or national chapter (64.5%). Three of the five most common actions belong to C8, exchanging practices: join a regional forum or national chapter, co-organise peer learning activities, and join thematic CoARA Working Groups and Action Clusters. The remaining one belongs to C7, raising awareness: to run workshops and town halls (Table 3).

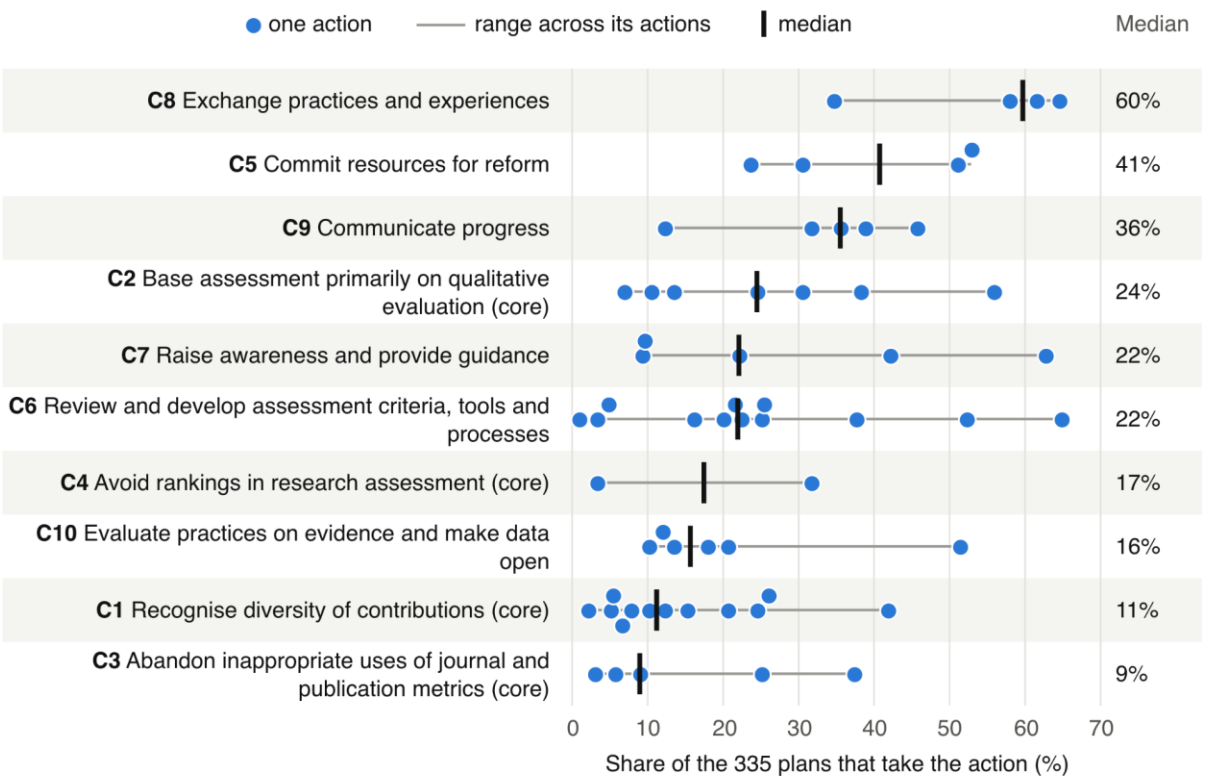


Figure 2: Share of the plans that take each of the catalogue's actions, by CoARA commitment, ordered by median uptake.

The median action under C3, abandoning inappropriate uses of journal- and publication-based metrics, is taken by 9.0% of the plans; under C8 the median is 59.7%. Three of the four core

commitments have a lower median than every supporting commitment except C10 (15.7%); the fourth, C2, basing assessment primarily on qualitative evaluation (24.5%), is above C6 and C7. Of the 62 actions, four are taken by 60% or more of the plans, 19 by 30 to 59%, 24 by 10 to 29% and 15 by fewer than 10%.

Table 3: The five actions most plans take and the five fewest take. Uptake: distinct plans with at least one counted record matched to the action.

Group	Action	Title	Plans	Share
Most taken	A13	Redesign evaluation rubrics	217	64.8%
Most taken	A35	Join a regional forum or national chapter	216	64.5%
Most taken	A16	Run workshops and town halls	210	62.7%
Most taken	A18	Co-organise peer learning activities	206	61.5%
Most taken	A17	Join thematic CoARA Working Groups and Action Clusters	194	57.9%
Fewest	A09	Identify where rankings are used	11	3.3%
Fewest	A53	Evaluate only where necessary	11	3.3%
Fewest	A08	Monitor for metric creep	10	3.0%
Fewest	A34	Audit for language bias (Helsinki Initiative on Multilingualism)	7	2.1%

Group	Action	Title	Plans	Share
Fewest	A52	Let researchers verify the data used to assess them	3	0.9%

4.2 In place or planned

Of the 8,870 counted commitments, 66.5% are planned and 29.2% in place (Table 4). A plan holds between 2 and 101 counted commitments, with a median of 24. Among the 8,326 counted commitments with a catalogue action, the share that is only planned is highest for C9, communicating progress (81%), C6 (75%) and C10 (73%), and lowest for C8 (50%) and C5 (52%). Funders report 48% of these commitments in place, universities 25% (Table 5).

Table 4: Status of the counted commitments, as the extraction agent recorded it or, where the checker found it indefensible, as the checker proposed.

Status	Records	Share
Planned	5,900	66.5%
Implemented	2,591	29.2%
Unclear	379	4.3%

Table 5: Status of the counted commitments with a catalogue action, by type of organisation, for the types with at least twenty plans.

Type	Plans	In place (%)	Planned (%)
Funder	34	48.0	45.0
Network or association	34	41.1	55.5
University	220	24.9	71.3
Other research-performing organisation	35	23.5	71.9

4.3 What is missing

Of the 25 items of DORA, the Leiden Manifesto and SCOPE that the catalogue links to actions, three are linked only to actions that fewer than 10% of the plans take (Table 6): Leiden 3 (protect excellence in locally relevant research); SCOPE P1 (principle: evaluate only where necessary); DORA 8, a recommendation to publishers (encourage responsible authorship and information about each author's contribution).

Table 6: Framework items whose linked actions are all taken by fewer than one plan in ten. Best share: the highest uptake among the actions linked to the item.

Item	What it asks	Linked actions	Best share (%)
Leiden 3	Protect excellence in locally relevant research	1	2.1
SCOPE P1	Principle: evaluate only where necessary	1	3.3
DORA 8	Encourage responsible authorship and information about each author's contribution	1	4.8

For 22 actions the planner's corpus analysis has a keyword signature of their own, so the plans it matches can be set against the plans the reading counts (Table 7). The keywords match more plans than the reading counts for seven of them, and fewer for 15. The largest gap is for the action to differentiate criteria by career stage, process and field: its keywords appear in 216 plans, and the reading counts 72.

Table 7: Plans counted by keyword against plans counted by reading, for every action with a keyword signature of its own, ordered by keyword count minus read count, from highest to lowest. Keyword: plans whose full text matches the signature in the planner's corpus analysis; read: plans with a counted record matched to the action.

Action	Title	Keyword	Read
A29	Differentiate criteria by career stage, process and field	216	72

Action	Title	Keyword	Read
A31	Link research integrity to assessment reform	115	32
A44	Create distinct career paths for all research roles	142	82
A30	Establish mentoring programmes for early-career researchers	82	31
A45	Keep a standing channel for researcher participation and feedback	92	60
A34	Audit for language bias (Helsinki Initiative on Multilingualism)	25	7
A51	Bring the reform into staff development conversations	58	54
A53	Evaluate only where necessary	10	11
A52	Let researchers verify the data used to assess them	0	3
A36	Adopt a contributor taxonomy (CRediT / CASRAI / TaDiRAH)	11	16
A47	Co-design new criteria with the people they will assess	59	75

Action	Title	Keyword	Read
A38	Get the plan formally approved by your governing bodies	84	102
A28	Upgrade infrastructure to support reformed assessment	58	85
A42	Publish your assessment criteria and give feedback to those assessed	45	84
A32	Conduct internal survey on assessment perceptions	4	45
A11	Secure dedicated funding and personnel	125	171
A40	Keep the reform on the agenda of your governing bodies	60	106
A41	Issue written guidance to evaluators and referees	74	128
A39	Publish and deposit the action plan	57	119
A49	Adopt your own responsible-assessment policy, including the use of metrics	39	125
A33	Extend reform to unit and institutional evaluation	8	126

Action	Title	Keyword	Read
A46	Evaluate the effects of reformed procedures	30	172

4.4 How far the reading can be trusted

The quote check found four of the 9,185 quotes absent from their plan; those records do not count (Table 8). The checker judged 240 records not to be commitments and 127 overstated; these do not count either. It refuted 684 records for some fault, and 373 of them still count: 261 because the fault was their match, which was then settled, 111 because it was their status, which was replaced by the checker's, and one for a fault the counting rule does not read. It accepted 96.4% of the matches. Of the 328 disputed matches, the settling agent kept the matcher's action for 148, took the checker's for 151, chose another for nine and none for 20. Of the 638 counted records matched to publishing the plan or to writing the reform into HRS4R and strategy cycles, 370, in 210 plans, are steps about the plan itself. Counted with them, publishing and depositing the plan would be taken by 205 plans instead of 119, and writing the reform into HRS4R and strategy cycles by 195 instead of 79. The checker also listed 407 actions that no record captured, and 18 plans reached the cap of five: extraction missed actions, but because the list is capped it does not measure how many.

A second run of the reading on 30 plans, with the same model, Claude Code version and instructions, and before the steps about the plan were separated, gave the same answer on whether a plan takes an action for 94.5% of the 1,860 plan-action pairs (κ 0.86). Of the 729 records that the two runs share, paired by their quotes, the same action was given to 86.8% of those with one, and the same status to 97.9%. On these plans the number taking an action differed between the runs by at most two (median one): disagreements on single plans largely offset each other in the counts.

Table 8: The reading's internal checks.

Check	Count
Records extracted	9,185
Quote found exactly	8,869
Quote found after collapsing whitespace	312
Quote not found (record not counted)	4

Check	Count
Not a commitment, for the checker (not counted)	240
Overstated, for the checker (not counted)	127
Records counted	8,870
Counted with the checker’s status	115
Refuted by the checker, for any reason	684
Counted although refuted, for the match	261
Counted although refuted, for the status	111
Matches the checker accepted	8,857 of 9,185
Settled matches	328
Settled: the matcher’s action	148
Settled: the checker’s action	151
Settled: another action	9
Settled: no action	20
Steps about the plan itself (no action)	370
Counted records with an action	8,326
Actions the checker found missing	407
Plans at the cap of 5 missing actions	18
Second run, 30 plans: plan-action pairs agreeing	94.5% (κ 0.86)

A second reader tests whether the plan-level counts can be reproduced without running the language model again. The plans' text was extracted again with Docling (Auer et al., 2024), and 8,833 of the 8,870 counted records were placed, by their quotes, on its passages. A multilingual encoder, mmBERT-small (Marone et al., 2025), was fine-tuned on those labels for 248 plans, 27 more set its decision thresholds, and it then labelled the 60 plans it had not seen. Its threshold for actions was set where the validation plans came closest to the number of plan-action pairs the reading gives them. The encoder runs on a CPU; each of the three trained models gave byte-identical output on a second machine, and one of them also when run twice on the first.

Trained three times with different seeds, on the held-out plans it counted 824 to 858 plan-action pairs against the reading's 923, ranked the 62 actions by the number of plans that take them with a Spearman correlation of 0.94 with the reading's ranking, and differed from the reading by a median of 3 plans per action. Passage by passage, Cohen's κ on whether a passage holds a commitment was 0.68 to 0.71, and the micro-averaged F1 on its actions 0.41. The encoder learnt from the reading, so this shows that a deterministic model comes close to the reading's counts, not that they are correct.

5. The planner

Institutions use the catalogue and the counts through the Research Assessment Reform Planner (Hartley Belmar, 2026). An institution answers 25 diagnostic questions, and the planner places each of the ten commitments on a maturity scale of six levels, from unaware to embedded. It then proposes the actions that fit each commitment's current level and ranks them by how many levels they would add, multiplied by their expected impact, less a penalty for effort, and adjusted by the priorities the institution sets (focus commitments, target levels, time horizon). For each action it shows the share of the 335 plans that take it, and drafts an action plan, one section per commitment, that the institution edits and downloads.

The planner runs in the browser: nothing the institution enters is uploaded. Its questions, actions and ranking rules are published as readable data, so an institution can see why an action was proposed and adapt the tool to its own context. It is available in English, German and Spanish, its code under MIT and its data under CC BY 4.0, and each release is archived on Zenodo. The planner is a guide: its levels and rankings are a prompt for reflection, not a certification, and it does not speak for CoARA, which has not endorsed it.

The share of plans next to each action is the figure this article reports. It tells an institution what its peers have committed to, not what works: that few plans take an action is no reason to skip it, as the planner's documentation says.

6. Discussion

Changing criteria is on most agendas, but as a general intention more than as specific steps. The most common action is to redesign evaluation rubrics, and three of the other four connect an institution with others. The core commitments, on recognising diverse contributions and moving away from inappropriate uses of metrics, have lower uptake than most supporting ones, and the specific steps that would take a metric out of a procedure are rare: removing metric proxies from criteria (25.1%), auditing criteria for metric references (5.7%).

Most commitments are still plans. This may partly be a matter of timing: the plans were due by the end of 2023 or within a year of signing. But the planned share is highest for the commitments that ask for new procedures or reporting (C9, C6, and C10). The finding is consistent with Rushforth et al. (2026), who found value terms outnumbering operational ones among the most mentioned, and with Morgan-Thomas et al. (2024), who found in one discipline

in the United Kingdom that the association between journal rankings and expert evaluations did not change despite institutional endorsements of DORA. Funders report more of their commitments in place than universities. The plans do not say why; one possibility is that a funder’s assessment runs through fewer, more central procedures than a university’s.

Some recommendations of the frameworks are almost absent from the plans. The actions that put into practice protecting excellence in locally relevant research, evaluating only where necessary, and recording each author’s contribution are each taken by fewer than one plan in ten. DORA addresses the last of these to publishers, so its absence from institutional plans is less surprising. For CoARA and its national chapters, the first two are places where guidance or shared examples could help. For an institution writing or revising its plan, these actions are options that its peers have mostly not taken, which is not a reason to leave them out.

Counting by keyword and reading do not agree, and the difference runs in both directions. Where the keywords find more plans, a plan can mention a theme without committing to the action. Where they find fewer, plans may commit to the action in words the signature does not catch, or the reading may match more broadly than the signature; the data do not separate the two. Monitoring the reform by counting words in plans would misstate what has been promised, and not in one predictable direction. Reading every plan at the level of its actions is slower, but with a documented procedure, quotes that a script can check, and a dataset that keeps every record, it is feasible for the full set of plans and can be repeated as new plans appear.

The two readings of the same plans also show how much such counts depend on the instructions. Under the earlier ones, which gave no action when in doubt, 6,886 counted records had an action; under the revised ones, 8,326. The rules are therefore part of the dataset, so that a reader can see what each count rests on.

The data could also support comparisons between peers. The extraction recorded each plan's type of organisation, so a new signatory could see which actions plans from organisations of its type commit to; comparing by size or region would need data the corpus does not hold. Such a comparison would show what peers have promised, not what they do.

7. Limitations

The corpus is what searches of Zenodo returned up to 6 August 2026. Plans deposited later, or elsewhere, are not in it, so a signatory with no plan in the corpus may have published one elsewhere or none at all; the data cannot tell which. One institution, whose plan was revised, counts twice. A plan’s text is not the institution’s practice: the article counts promises. The catalogue sets what can be counted; an action outside it appears only if the clustering of unmatched records in the earlier reading brought it in.

No sample coded by people measures this reading’s accuracy. The records coded by hand were used to write its rules, and every agent was the same model with different instructions, so the checks share any blind spot of that model; an action that neither the extractor nor the

checker saw is absent from every figure. The matcher gives the closest action on a record's topic, so a count may include records that fit their action only in part, and nothing corrects a record counted wrongly; no count is a lower or an upper bound. A repeated run gives slightly different counts (Section 4.4). The rule for steps about the plan was decided after the counts were seen; Section 4.4 gives the counts with and without it. The checker's list of missed actions was capped at five per plan, and so shows that the extraction is incomplete without measuring how far. The type of each organisation was recorded by the extraction agent. The comparison with keyword counts is limited to the 22 actions whose keyword signature is their own.

8. Conclusion

Read action by action, the 335 plans commit to redesigning their criteria more often than to any other action, but rarely to the specific changes that would take a metric out of a procedure, and most of what they commit to is still planned. Some recommendations of DORA, the Leiden Manifesto and SCOPE are almost absent, and counting keywords instead of reading would misstate uptake in both directions. The reading's records, rules and checks are kept as a dataset, and the planner shows the counts to institutions writing their next plans.

Data and code availability

The dataset (records, matches, verdicts, settling decisions, the catalogue and the method scripts, with a SHA-256 manifest) is available from the corresponding author, under CC BY 4.0 and its code under MIT. It does not republish the plans, and it keeps verbatim quotes only from plans licensed CC BY 4.0 or CC0 (Section 3.5). The scripts and results of the second reader (Section 4.4) and the recount script are also available from the corresponding author. The planner is open source and archived on Zenodo (Hartley Belmar, 2026).

Appendix: Use of generative AI

Language-model agents read the plans (Section 3.3); their instructions, output schemas and scripts are part of the dataset. Claude Code (Anthropic) drafted those instructions and scripts, and the scripts of the second reader (Section 4.4), under the authors' direction. At the authors' request Claude Code was used to improve the grammar, wording and citations, to improve the code for the tables and figures, and to format the references. The authors revised and edited the text over several rounds and are responsible for its ideas, findings, interpretation and conclusions.

References

Abramo, G. (2024). The forced battle between peer-review and scientometric research assessment: Why the CoARA initiative is unsound. *Research Evaluation*. https://doi.org/10.1093/reseval/rvae021

Auer, C., Lysak, M., Nassar, A., Dolfi, M., Livathinos, N., Vagenas, P., Ramis, C. B., Omenetti, M., Lindlbauer, F., Dinkla, K., Mishra, L., Kim, Y., Gupta, S., de Lima, R. T., Weber, V., Morin, L., Meijer, I., Kuropiatnyk, V., and Staar, P. W. J. (2024). Docling technical report. arXiv. https://doi.org/10.48550/ARXIV.2408.09869

Barcelona Declaration on Open Research Information. (2024). Zenodo. https://doi.org/10.5281/zenodo.10958522

Bornmann, L. (2026). To sign or not to sign initiatives to reform research evaluation: A discussion of the pros and cons. *Journal of Informetrics*, *20*(4), 101872. https://doi.org/10.1016/j.joi.2026.101872

Chae, Y., and Davidson, T. (2026). Large language models for text classification: From zero-shot learning to instruction-tuning. *Sociological Methods & Research*, *55*(2), 501–567. https://doi.org/10.1177/00491241251325243

CoARA. (2022). *Agreement on Reforming Research Assessment*. https://www.coara.org/wp-content/uploads/2025/11/2022_07_19_rra_agreement_final.pdf-3.pdf

CoARA. (2023). *Action plan guidelines: Support for CoARA signatories in the preparation of action plans*. https://www.coara.org/wp-content/uploads/2023/10/Action-plan-guidelines-October-2023_Incl.-Table.pdf

DORA. (2013). *San Francisco Declaration on Research Assessment*. https://sfdora.org/read/
European Commission: Directorate-General for Research and Innovation. (2026). Implementing research assessment reform: Insights from ARRA action plans and interviews with signatories. Publications Office of the European Union. https://doi.org/10.2777/5429449

Gadd, E., Pölönen, J., Stroobants, K., and Tóth-Czifra, E. (2026). *Implementation routes of research assessment reform: Evidence from CoARA [Conference paper]. Zenodo. https://doi.org/10.5281/zenodo.22656352*

Gilardi, F., Alizadeh, M., and Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30). https://doi.org/10.1073/pnas.2305016120

Gunes, E., and Florczak, C. K. (2025). Replacing or enhancing the human coder? Multiclass classification of policy documents with large language models. *Journal of Computational Social Science*, *8*(2). https://doi.org/10.1007/s42001-025-00362-2

Hartley Belmar, R. (2026). *Research Assessment Reform Planner: Self-assessment against the ten CoARA commitments*. Zenodo. https://doi.org/10.5281/zenodo.21492548

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., and Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, *520*(7548), 429–431. https://doi.org/10.1038/520429a

Himanen, L., Conte, E., Gauffriau, M., Strøm, T., Wolf, B., and Gadd, E. (2024). The SCOPE framework – implementing ideals of responsible research assessment. *F1000Research*, *12*, 1241. https://doi.org/10.12688/f1000research.140810.2

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), 1–55. https://doi.org/10.1145/3703155

Marone, M., Weller, O., Fleshman, W., Yang, E., Lawrie, D., and Van Durme, B. (2025). mmBERT: A modern multilingual encoder with annealed language learning. arXiv. https://doi.org/10.48550/ARXIV.2509.06888

Moher, D., Bouter, L., Kleinert, S., Glasziou, P., Sham, M. H., Barbour, V., Coriat, A.-M., Foeger, N., and Dirnagl, U. (2020). The Hong Kong Principles for assessing researchers: Fostering research integrity. *PLOS Biology*, *18*(7), e3000737. https://doi.org/10.1371/journal.pbio.3000737

Morgan-Thomas, A., Tsoukas, S., Dudau, A., and Gąska, P. (2024). Beyond declarations: Metrics, rankings and responsible assessment. *Research Policy*, *53*(10), 105093. https://doi.org/10.1016/j.respol.2024.105093

Pangakis, N., and Wolken, S. (2025). Keeping humans in the loop: Human-centered automated annotation with generative AI. *Proceedings of the International AAAI Conference on Web and Social Media*, *19*, 1471–1492. https://doi.org/10.1609/icwsm.v19i1.35883

Pölönen, J., and Rushforth, A. (2025). *CoARA by numbers: Research assessment reform uptake in European higher education*. https://doi.org/10.31235/osf.io/69vjr_v1

Rushforth, A., Gogadze, N., Skhirtladze, T., and Pölönen, J. (2026). Advancing open science through research assessment reform? Content analysis of CoARA action plans. *Scientometrics*, *131*(6), 3897–3918. https://doi.org/10.1007/s11192-026-05650-w

Stroobants, K. (2026). *General Assembly 25 June 2026: Results from first analysis of CoARA action plans [Presentation]. Zenodo. https://doi.org/10.5281/zenodo.21028639*

Wilsdon, J. (2015). *The Metric Tide: Independent review of the role of metrics in research assessment and management*. SAGE Publications Ltd. https://doi.org/10.4135/9781473978782

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., and Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, *50*(1), 237–291. https://doi.org/10.1162/coli_a_00502