The case study is the thing you are reading

Twenty datasets from seventeen seeded scripts that regenerate byte-identically, twenty test files and 4,212 assertions, seven checks that fail the build, and a schema that refuses non-synthetic data. Every claim in this course is checkable in this repository.

Pierre Robentz Cassion

Lesson 8 of 8

What this lesson covers

- Why the platform is the example
- Raw data that is generated, not collected
- Committed output, for a reason that names its condition
- Checks in three places, and the constraint that decided where
- The check that a reference graph cannot express
- What the repository refuses to hold
- Reproducibility at the level of the language
- What this repository does not do
- Report it whole
- What comes next

Why the platform is the example

- Every other course here uses a dataset.

Raw data that is generated, not collected — Shell

```
pnpm datasets:generate
git status --short apps/data-analysis/public/datasets/files
# (nothing)
```

Raw data that is generated, not collected

- Twenty CSV files, seventeen generator scripts, and regenerating all of them changes nothing — Each generator is seeded...
- That is a stronger property than "the pipeline reruns" — It means rerunning is a *verification* step: if `git status`...
- And the defects are deliberate — Every entry in a dataset's `knownIssues` list corresponds to a block in its generator...

Committed output, for a reason that names its condition

- Because the deploy runs on Cloudflare Pages with no TeX, no Python and no pandoc — The build copies static files, so...
- That is the general rule from lesson 1, applied — Commit a generated artefact when rebuilding it is not available to...
- And it works because the output is deterministic — `SOURCE_DATE_EPOCH` pins pdfTeX's timestamp and pandoc's, so a...

Checks in three places, and the constraint that decided where

Where	What	Why there
Zod schema	Field types, enums, required fields	Runs on one entry, no filesystem
`astro build`	Seven graph-level checks	Needs the whole collection
`pnpm test`	Filesystem checks	The build has no `node:fs`

Checks in three places, and the constraint that decided where

- The Astro build prerenders inside a Cloudflare worker with no `node:fs` — so a check that has to look at a file on disk...
- Put the check where it can run — That is a real engineering constraint, and it produced a better arrangement than...

The check that a reference graph cannot express — Example

```
REFERENCE_RULES verifies that a declared slug resolves.
It cannot verify that an entry is pointed at.
```

The check that a reference graph cannot express

- So a course could ship with no lab and no exercise, and every gate would stay green — That is what...
- Any "every X has at least one Y" rule needs a test of that shape — and recognising which of your constraints are of...

What the repository refuses to hold — Example

```
.refine((d) => d.dataQuality.synthetic === true, {
  message: "Only synthetic or fully de-identified datasets may be published.",
})
```

What the repository refuses to hold

- A dataset that is not declared synthetic fails the build — This audience models beneficiary, protection, GBV and...
- It pairs with generating the data rather than collecting it — The guarantee is structural at both ends: the generator...

Reproducibility at the level of the language — Example

```
{ "packageManager": "pnpm@11.9.0", "engines": { "node": ">=22" } }
```

Reproducibility at the level of the language

- The package manager is pinned, not just the packages — A different pnpm resolves the lockfile differently, and that is...
- TypeScript is held at 6.x with the reason written down — TypeScript 7 does not yet expose the programmatic API ‘astro...

What this repository does not do

- No container — A lock file plus a pinned package manager reproduces the JavaScript side; the Python and TeX sides…
- No end-to-end determinism test in CI — The dataset regeneration is deterministic and verified by hand; there is no CI…
- Handover documentation is thin — `CLAUDE.md` and `docs/DEFINITION_OF_DONE.md` carry the decisions, which is more than…
- Naming the gaps is the point — A case study that claims everything is a case study nobody can learn from, and the three…

Report it whole — Example (cont.)

Reproducibility of this platform

Data	20 CSV files generated by 17 seeded scripts. Regenerating all of them produces byte-identical output; `pnpm datasets:generate` followed by `git status` is the verification.
Checks	7 fail the build (schema, references, locale, translation parity, topic-sector, synthetic-only, programme spine). 20 test files and 4,212 assertions run first in CI.
Output	Course PDFs, slide decks, figures and notebooks are committed and deterministic. SOURCE_DATE_EPOCH pins pdfTeX and pandoc.
Environment	pnpm 11.9.0 and Node 22 pinned. TypeScript held at 6.x with the reason and the removal condition documented.

Report it whole — Example (cont.)

Refused	A dataset not declared synthetic fails the build.
Not done	No container; no CI job that regenerates everything and fails on a diff; handover documentation thinner than this course asks for.

Report it whole

- The last row is the one to copy into your own version — A reproducibility statement without a "not done" section is a...

What comes next

- The thread through all twenty is the same — A number in a programme report is a claim, and everything this platform...
- The last of those is the one that makes the others checkable — An analysis nobody can rerun is an analysis whose other...

Read the full lesson, with runnable code

https://data-analysis.cassion.dev/courses/reproducible-workflows/repro-08-this-repository