Solving the right problems:

Requirements capture for large-scale, evolving research projects

Gabrielle M. Schroeder

Newcastle University RSE Team

RSECon24

Embarking on a new RSE project

Goal: Map the course of the project

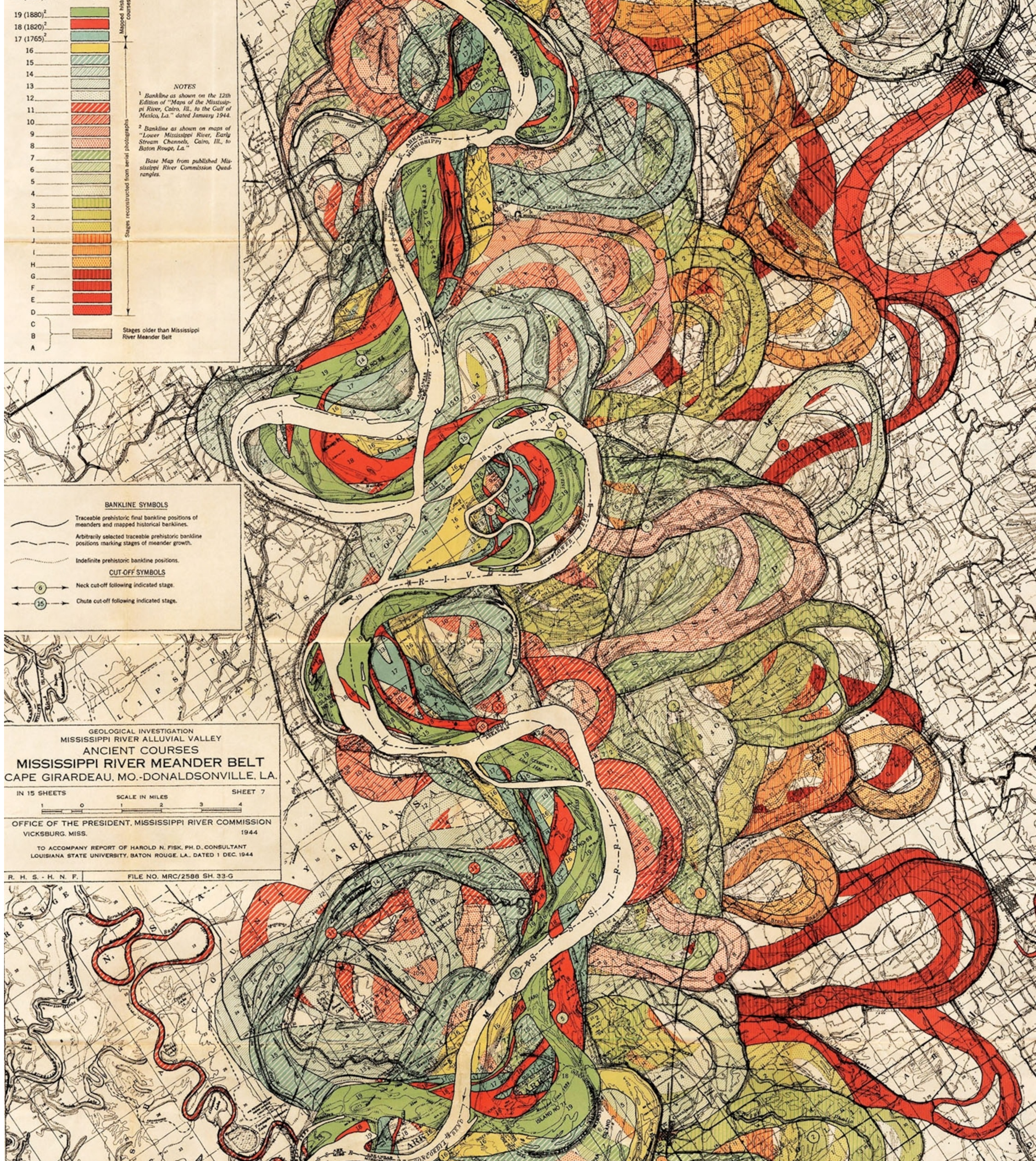
Image: Tabea Schimpf (Unsplash)

Embarking on a new RSE project

Goal: Map the course of the project

Challenge: Large, evolving research projects

“The nature and origin of the Alluvial Valley of the Lower Mississippi River” by Harold Fisk (1944)



19 (1880)¹
18 (1820)²
17 (1765)²
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
J
I
H
G
F
E
D
C
B
A
Stages reconstructed from aerial photographs
Stages older than Mississippi River Meander Belt
NOTES
¹ Bankline as shown on the 12th Edition of "Maps of the Mississippi River, Cairo, Ill., to the Gulf of Mexico, La." dated January 1944.
² Bankline as shown on maps of "Lower Mississippi River, Early Stream Channels, Cairo, Ill., to Baton Rouge, La."
Base Map from published Mississippi River Commission Quadrangles.
BANKLINE SYMBOLS
Traceable prehistoric final bankline positions of meanders and mapped historical banklines.
Arbitrarily selected traceable prehistoric bankline positions marking stages of meander growth.
Indefinite prehistoric bankline positions.
CUT-OFF SYMBOLS
6
Neck cut-off following indicated stage.
15
Chute cut-off following indicated stage.
GEOLOGICAL INVESTIGATION
MISSISSIPPI RIVER ALLUVIAL VALLEY
ANCIENT COURSES
MISSISSIPPI RIVER MEANDER BELT
CAPE GIRARDEAU, MO.-DONALDSONVILLE, LA.
IN 15 SHEETS
SCALE IN MILES
1 0 1 2 3 4
SHEET 7
OFFICE OF THE PRESIDENT, MISSISSIPPI RIVER COMMISSION
VICKSBURG, MISS.
1944
TO ACCOMPANY REPORT OF HAROLD N. FISK, PH. D., CONSULTANT
LOUISIANA STATE UNIVERSITY, BATON ROUGE, LA., DATED 1 DEC. 1944
R. H. S. - H. N. F.
FILE NO. MRC/2588 SH. 33-G
M-I-S-S-I-S-S-I-P-P-I R-I-V-E-R

OpenScan

Driving **horizon scanning** research at the National Institute for Health and Care Research **Innovation Observatory** (NIHRIO)

Image: Clemens van Lay (Unsplash)

Capturing OpenScan's requirements

Assessment stage

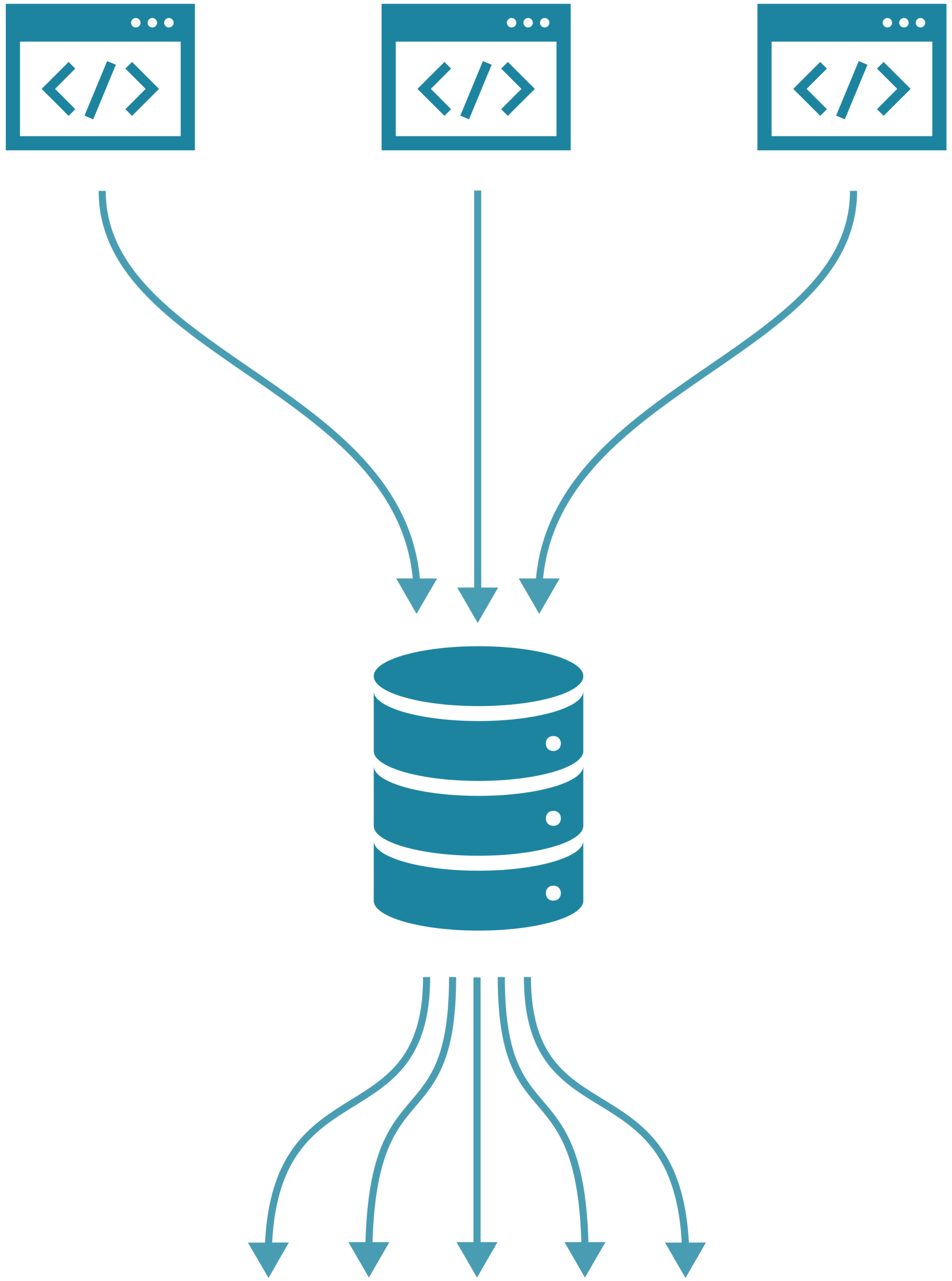
Regular meetings

Document current state

Document planned work

Example work package

Re-designing the web scraper system



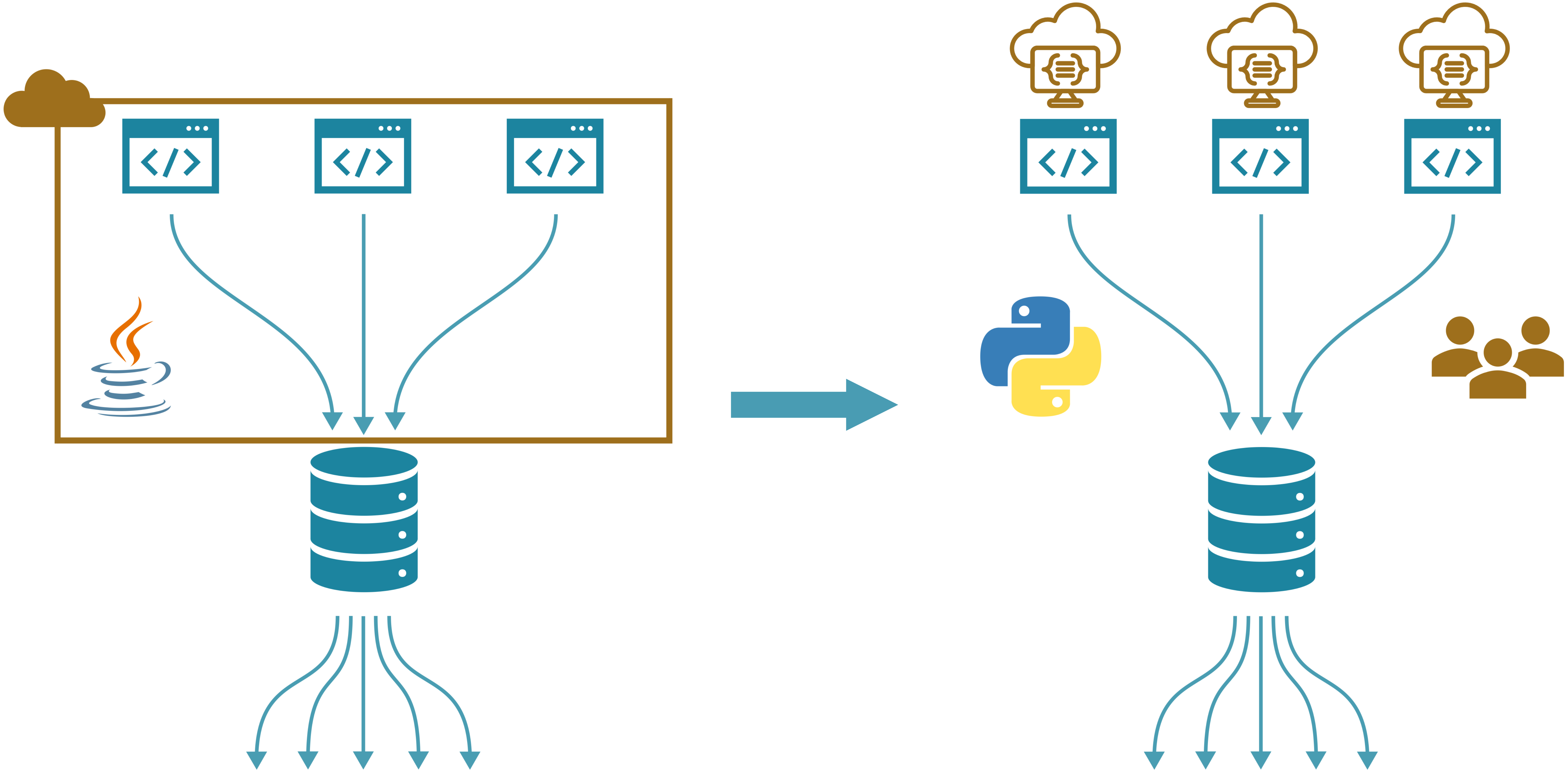
Scrape clinical trial records

Process records

Store in database

Analyse and disseminate

Requested changes



Serverless icon: Vecteezy.com

Strategy: User stories

As a **<user>** I want to **<goal>** so that **<benefit>**.

Strategy: User stories

As a **<user>** I want to **<goal>** so that **<benefit>**.

Goals should not include the technical solution:

I want to easily train researchers to write web scrapers

rather than

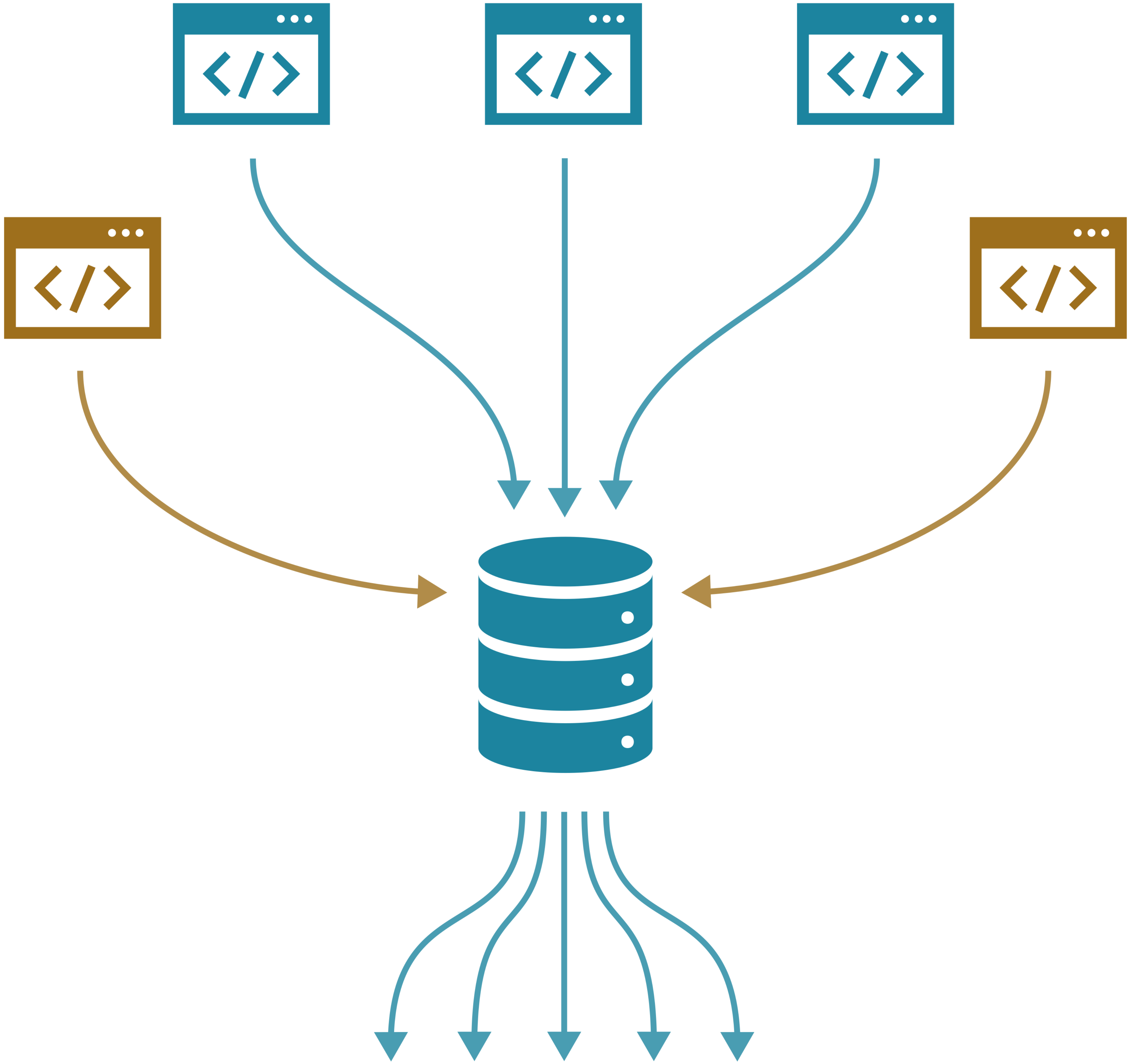
I want to write web scrapers using Python

User stories

As an **NIHRIO developer**,

I want to **easily add new web scrapers**

so that **we can expand OpenScan and develop similar projects**.

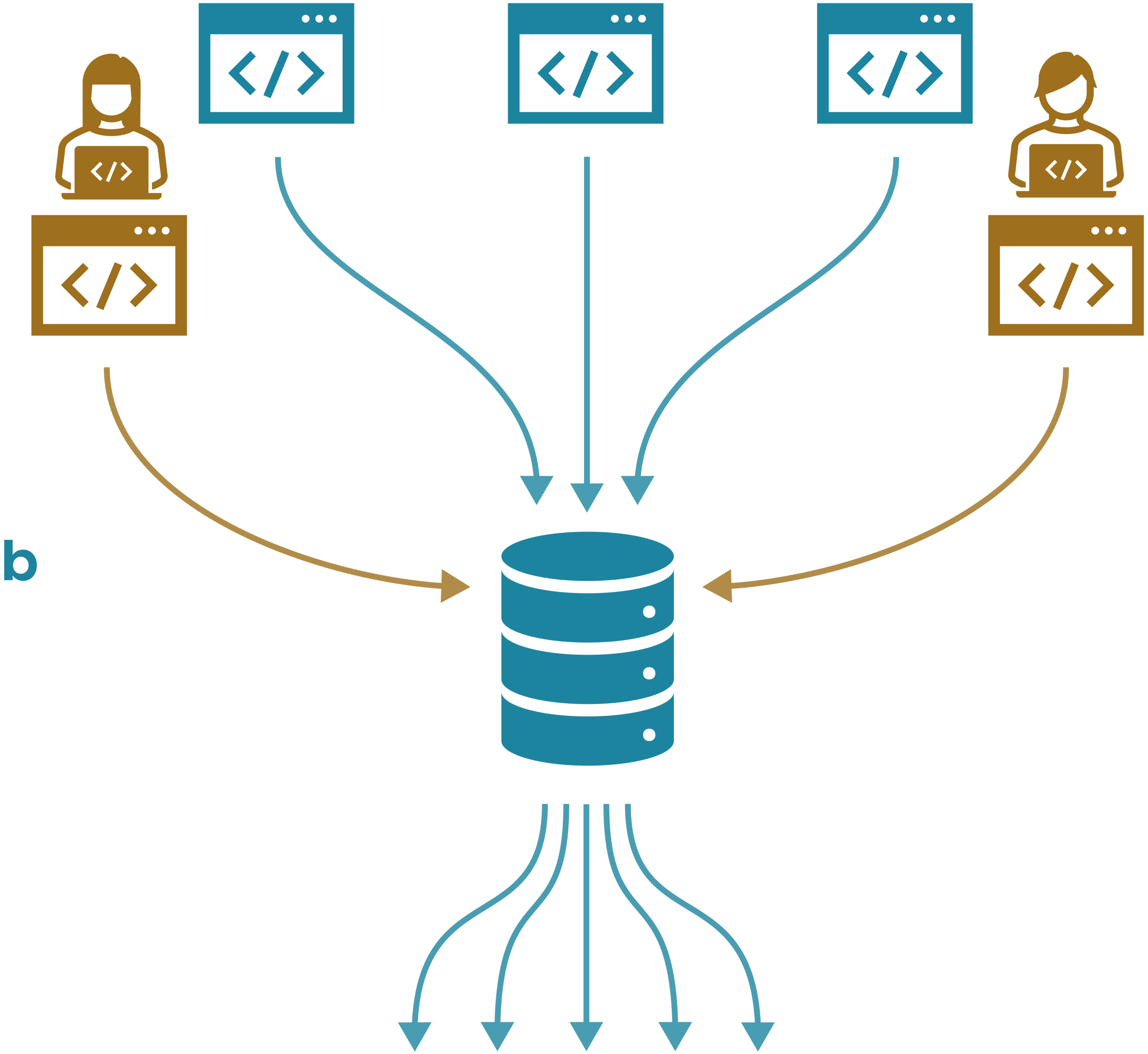


User stories

As an **NIHRIO developer**,

I want to **collaborate on web scrapers**

so that **multiple people can maintain and develop OpenScan**.

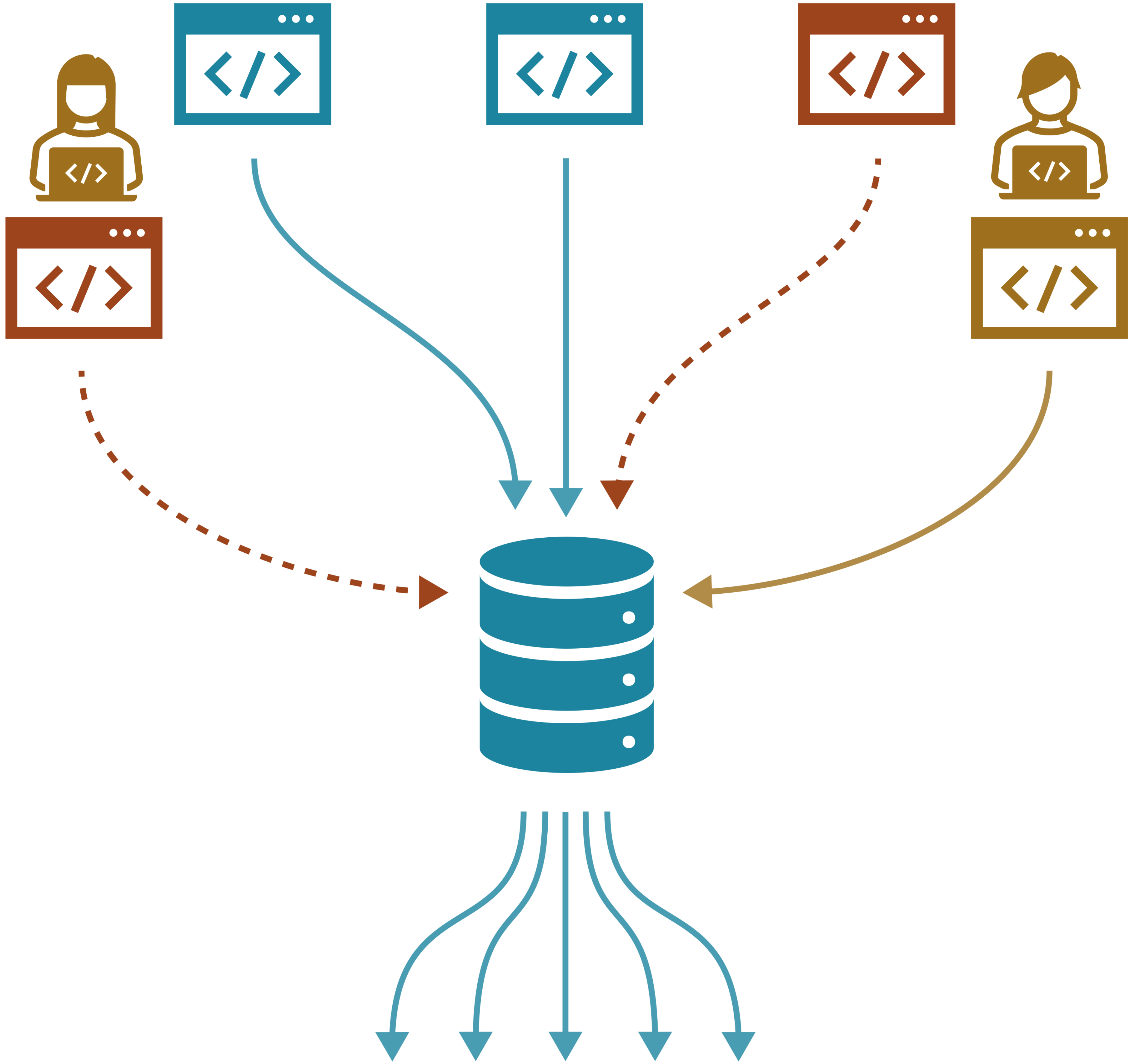


User stories

As an **NIHRIO developer**,

I want to **ensure that different developers do not break other's web scrapers**

so that **our pipelines are reliable and robust**.



Strategy: Investigate current workflows

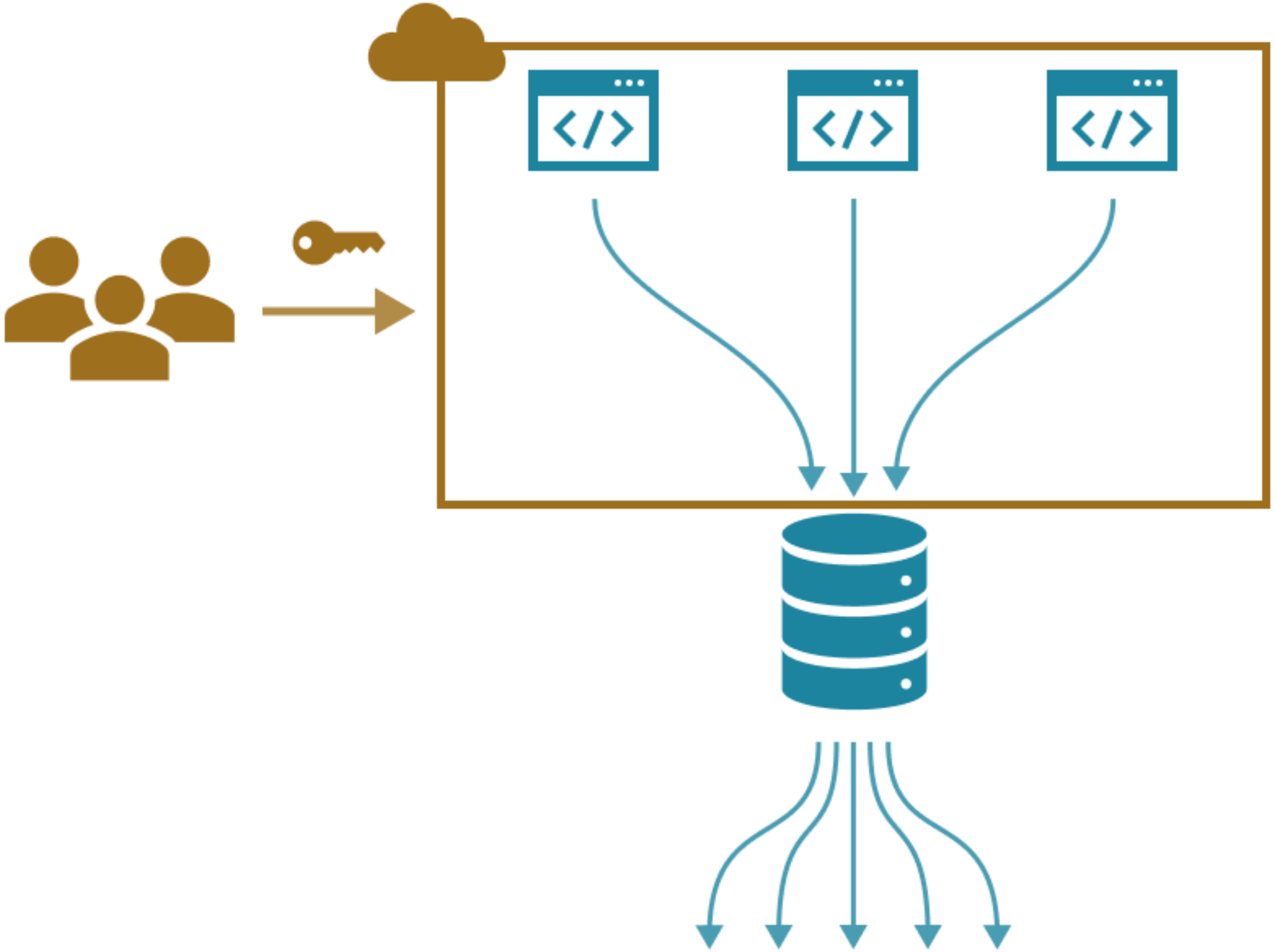
Strategy: Investigate current workflows

One login for server!

→ Difficult to deploy

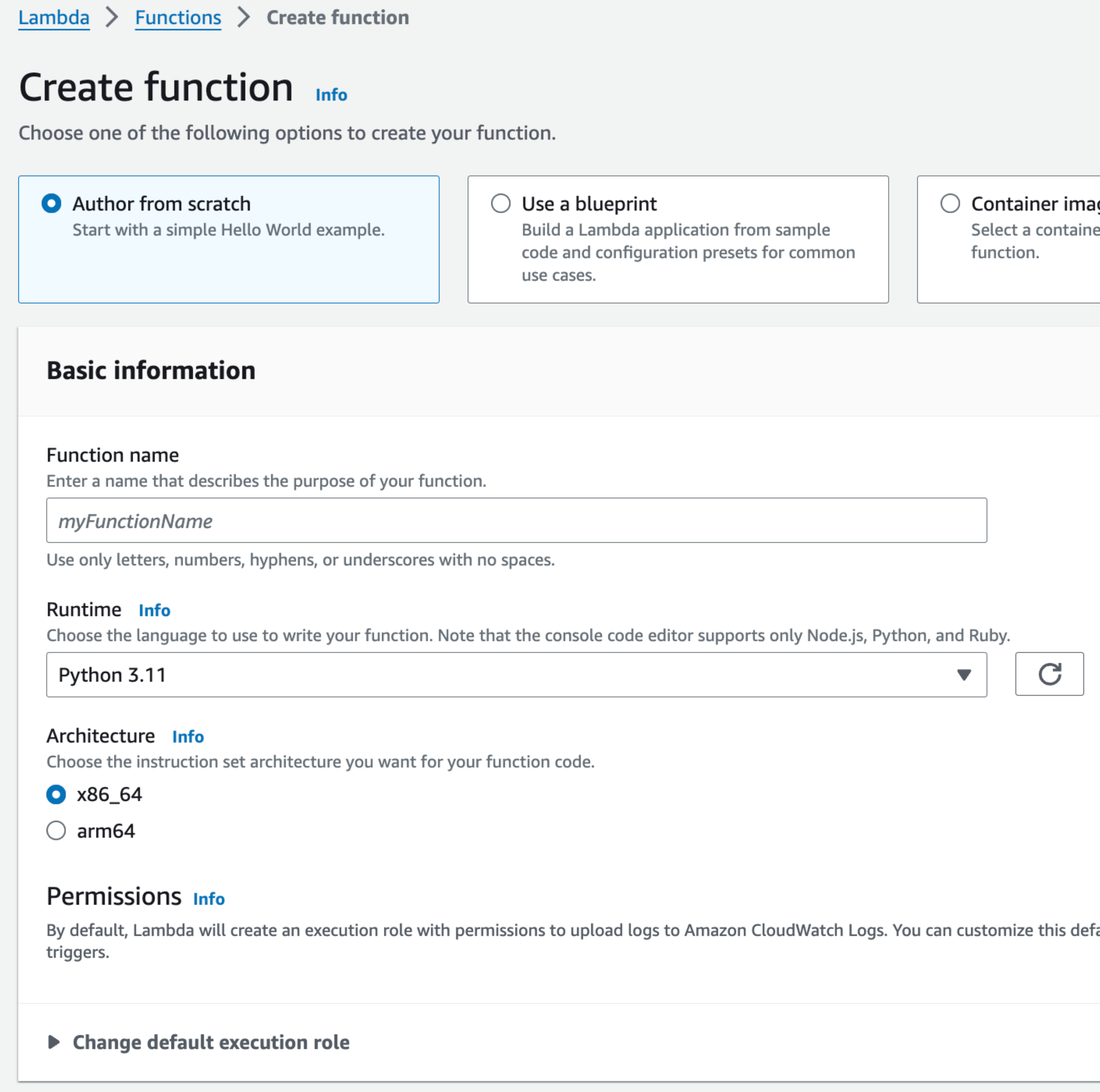
→ Poor version control

→ Fragile



New requirement

Need **easy-to-use, reproducible, and traceable workflows** for developing web scrapers.



Lambda > Functions > Create function

Create function Info

Choose one of the following options to create your function.

● Author from scratch
Start with a simple Hello World example.

○ Use a blueprint
Build a Lambda application from sample code and configuration presets for common use cases.

○ Container imag
Select a containe
function.

Basic information

Function name
Enter a name that describes the purpose of your function.
myFunctionName
Use only letters, numbers, hyphens, or underscores with no spaces.

Runtime Info
Choose the language to use to write your function. Note that the console code editor supports only Node.js, Python, and Ruby.
Python 3.11

Architecture Info
Choose the instruction set architecture you want for your function code.
● x86_64
○ arm64

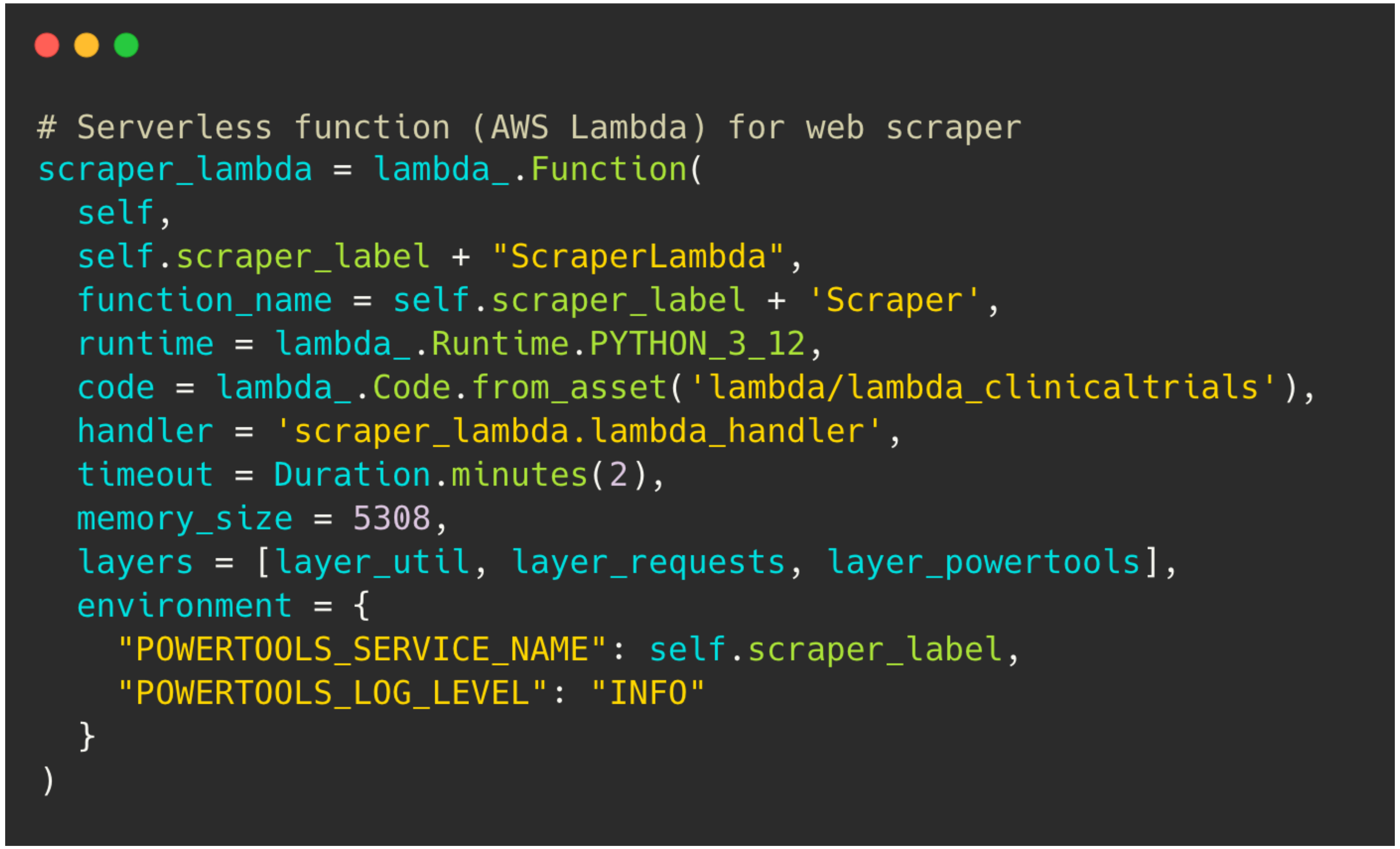
Permissions Info
By default, Lambda will create an execution role with permissions to upload logs to Amazon CloudWatch Logs. You can customize this defa
triggers.

▶ Change default execution role

New requirement

Need **easy-to-use, reproducible, and traceable workflows** for adding web scrapers.

Solution:
Infrastructure-as-Code
(in Python)



```
# Serverless function (AWS Lambda) for web scraper
scraper_lambda = lambda_.Function(
  self,
  self.scraper_label + "ScraperLambda",
  function_name = self.scraper_label + 'Scraper',
  runtime = lambda_.Runtime.PYTHON_3_12,
  code = lambda_.Code.from_asset('lambda/lambda_clinicaltrials'),
  handler = 'scraper_lambda.lambda_handler',
  timeout = Duration.minutes(2),
  memory_size = 5308,
  layers = [layer_util, layer_requests, layer_powertools],
  environment = {
    "POWERTOOLS_SERVICE_NAME": self.scraper_label,
    "POWERTOOLS_LOG_LEVEL": "INFO"
  }
)
```

Summary

We identified additional requirements for OpenScan's redesign.

Careful requirements capture will help make OpenScan maintainable and reproducible.

Maintainable

Easy to adapt and to correct faults

Reproducible

Enable trust in research

RSECon2024 Conference Themes

Lessons for RSE projects

- Dedicate time for requirements capture.
- Investigate the current state and pain points.
- Document requirements to formalise your thoughts.
- Plan for long-term needs.

Thank you

Contact

gabrielle.schroeder@ncl.ac.uk

www.linkedin.com/in/
gabrielleschroeder

Robin
Wardle

Gabrielle
Schroeder

Kate
Court

Mark
Turner

Chris
Marshall

Saleh
Mohamed

Amey
Vedpathak

Hongbo
Bo

Kieran
McDonough

Karim
Elkobrossy

Teresa
Fortune

Dawn
Craig