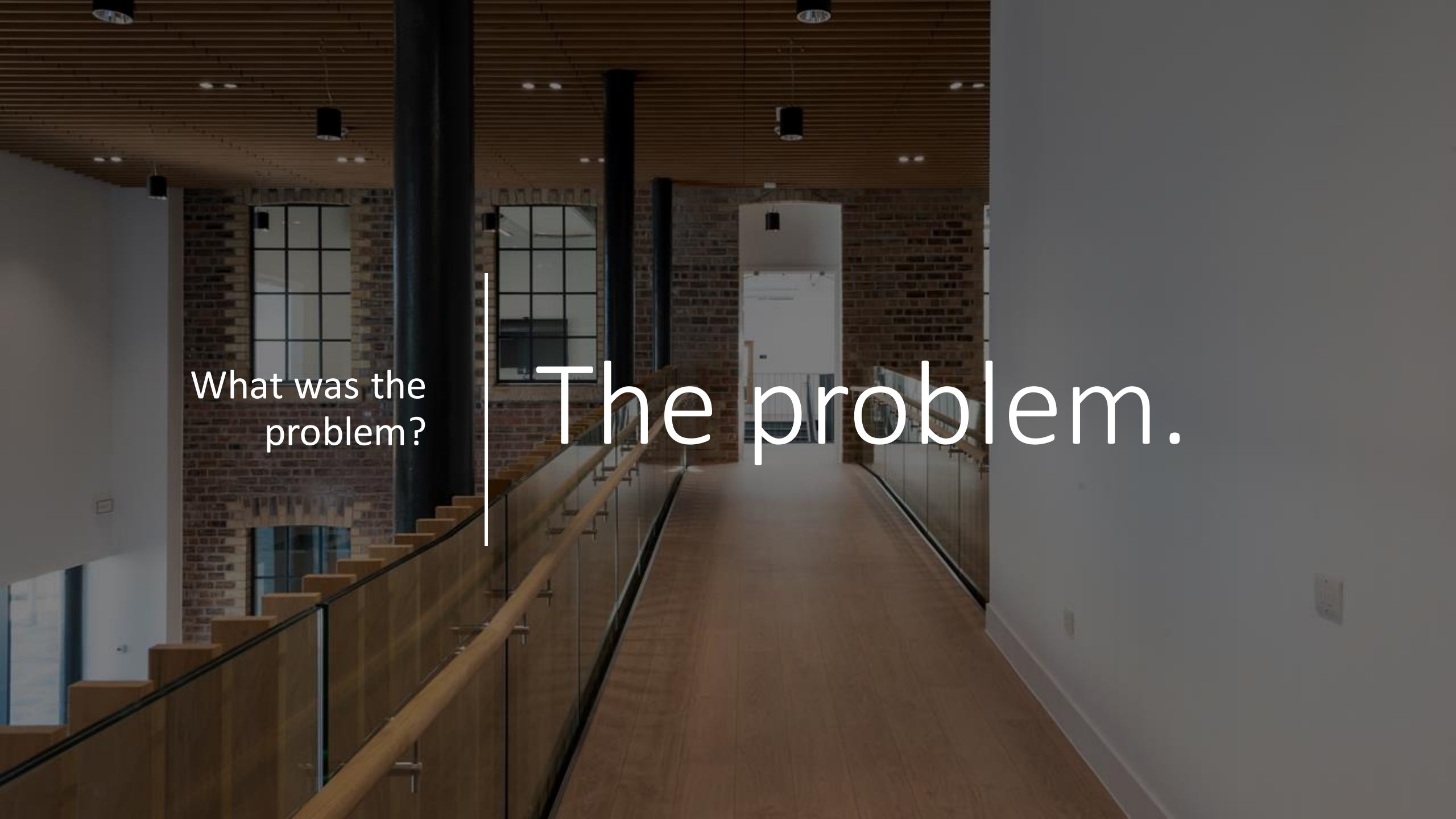


Registry Internship

Gregor Soutar



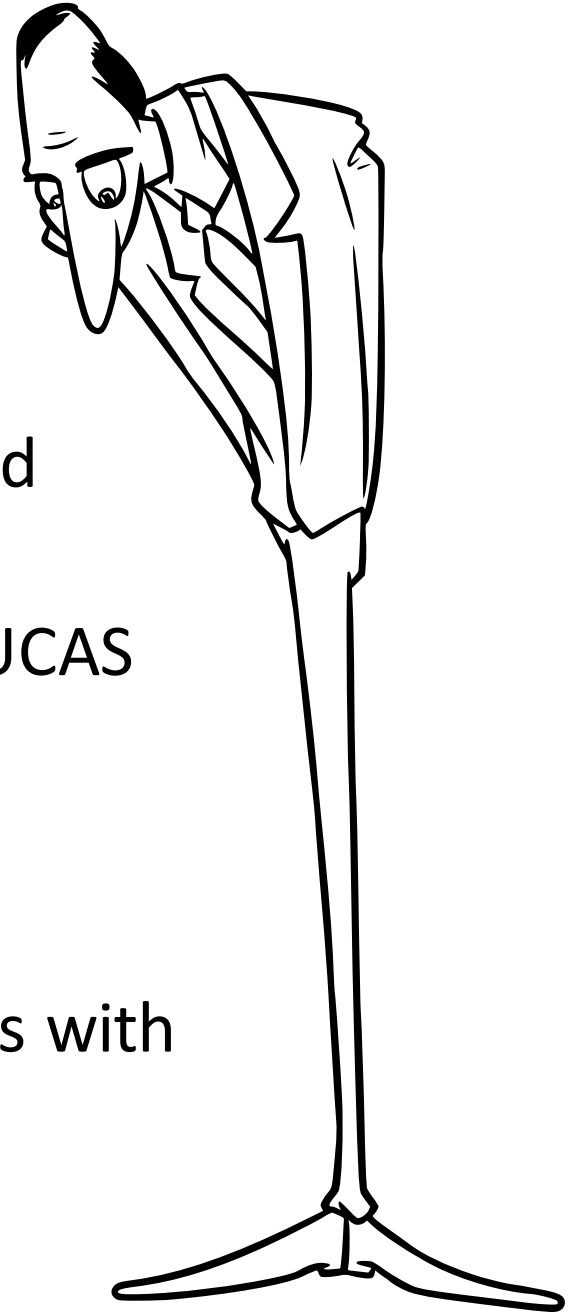
A photograph of a modern interior space, likely a hallway or staircase. The walls are made of exposed brick, and the floor is polished wood. Large windows with black frames are visible on the left. A wooden handrail with glass panels runs along the left side. The ceiling has exposed wooden beams and modern lighting fixtures. The text "What was the problem?" is overlaid on the left side of the image.

What was the
problem?

The problem.

An Overview

- Two datasets: UCAS and SCL.
- If a school is not in the UCAS data, a school is created and given an internal ID.
- Over time, more and more international schools adopt UCAS and with that the UCAS data is updated.
- To keep the SCL data up to date, the UCAS data must be regularly imported (in theory).
- But with this comes the potential for duplication, schools with an internal ID may now have a corresponding UCAS ID.

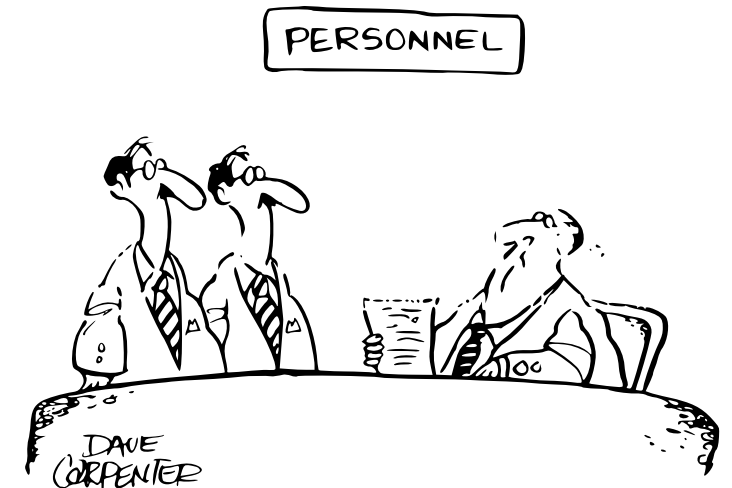


An Example

A school that now has a corresponding UCAS code:


Data A: SCL	
School code	ES0341
Full name	Escola IPSI De Barcelona
Address Line 1	Carrer Del Comte Borrell 243
Address Line 2	
Address Line 3	08029 Barcelona
Address Line 4	
Postcode	8029

Data B: UCAS (HCV)	
School	42231
School Name	IPSI
Address line 1	Comte Borrell, 243-249
Address line 2	L'equerra De L'exemple
Address line 3	Barcelona 08029
Address line 4	
Postcode	



"WE FOUND BOTH OF YOU EQUALLY QUALIFIED FOR THE POSITION..."

Importing the new ucas record for this school would result in the school having two records associated with it, a duplicate. But how can duplicates, including obscure examples such as this, be found?

A photograph of a modern interior space, likely a hallway or staircase. The walls are made of exposed brick, and the floor is made of light-colored wood. There are large windows with black frames on the left wall. A wooden handrail with glass panels runs along the left side of the hallway. The ceiling is made of dark wood slats. The overall atmosphere is clean and industrial.

How was the
problem solved?

The solution.

Registry Data Tool

- A terminal-based python  program that can detect matching schools across UCAS, SCL data and more.

```
----- Registry Data Tool -----  
  
1. Match records between two files.  
2. Detect duplicates within a single file.  
3. Verify links in file.  
4. Check for differences between records with same ID accross two files.  
5. Exit  
  
Please enter selection [1-5]: █
```

Match Records Across Files

- Can take any two delimited files and search for records that are present in both datasets.
- In the case of UCAS and SCL data, this option can be used to find schools in both datasets with different school codes.
- This option produces three files containing:
 - The potential matches
 - Schools in the first dataset only
 - Schools in the second dataset only
- There will be false positives/things missed, largely based on training.

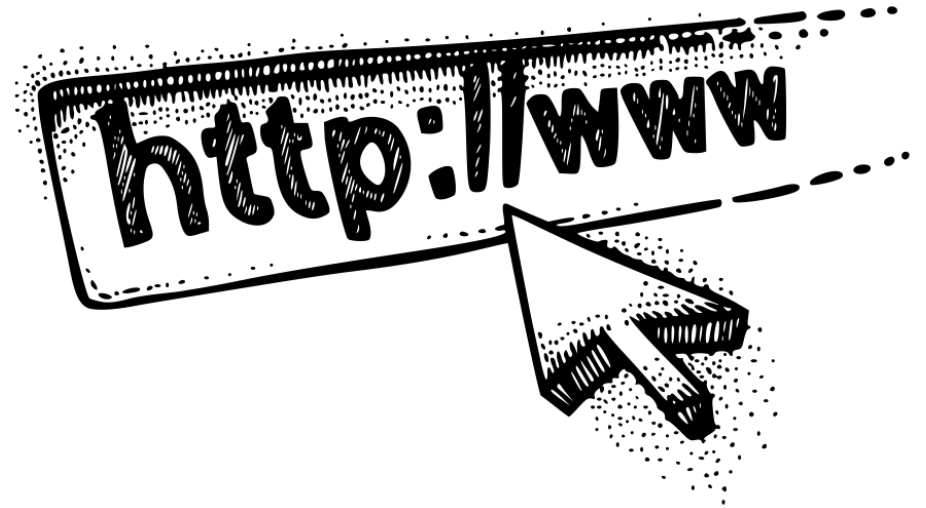
Find Duplicates Within a File

- Can take any delimited file and search for duplicates within itself.
- This option could be used to discover any duplicates currently in the SCL data.
- This option produces a single file containing the original data with an additional column, cluster-id.

Cluster-ID	Forename	Surname	Address Line 1
1	Gregor	Soutar	1 Planet Earth
1	Greg ^a r	Sout ^e r	Planet Earth, ¹
2	Another	Person	2 Planet Mars

Tests Links in a File

- Can take any delimited file and automatically test each link.
- Sends a request to a website. If the site responds with 200 this means 'OK', showing the link to work.
- If the site doesn't respond with 200 (or at all) then the link is likely invalid.
- There will be many sites that work that are thought to be invalid by the program.




Check for differences Across Files

- Can take any two files that contain the same unique identifier for records and discover where differences lie.
- For example, you could compare the school status column between ucas and scl data and get a list of schools where that column has changed.
- A single file is produced by this option containing a list of the schools where differences were found.

Documentation

- Instructional documentation was produced to try and ensure the maintainability and usability of the program.
- It outlines:
 - Installation of the program
 - Usage of each of the four options
 - Configuration files
 - Training



A photograph of a modern, multi-story brick building with large, dark-framed windows. The building is constructed of red brick with some white mortar. The windows are arranged in a grid pattern. A paved sidewalk runs along the front of the building. The sky is overcast and grey. The text "Do you have any questions?" is overlaid on the left side of the image in white, sans-serif font. A thin white vertical line is positioned to the right of the text.

Do you have any
questions?

The End.