Flexible dissemination software for the 2021 England & Wales Census

15th April 2022

IAOS 2022

The Sensible Code Company

Statistics Poland

Contents

- Introduction
- Background
- Technical challenges & solutions
- Benefits of flexible dissemination
- What’s next?
- Questions

Introduction

Summary

The development, over the last five years, of a **flexible dissemination service** to support the publication of results from the 2021 England and Wales census, in partnership with the Office for National Statistics.

The Sensible Code Company

We make software products that modernise the processing and publication of data.

ScraperWiki

Code | Business | About | Sign Up | Log In

New Scraper | New View | Browse Scrapers | Premium Accounts | Help & Docs | Search code...

Write code that gets data

or

Ask us to get it for you

Write code | Request data

ScraperWiki. Trusted by some of the biggest names in media and government.

Dispatches | theguardian | informa Agra | CabinetOffice

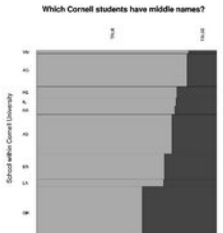
POPULAR TAGS

usa | opencorporates | government | uk | companies list | companies | ireland | finance | parliament | australia | planning | locations | alphagov | germany | twitter | people | weather | scraperwiki

FROM THE BLOG

Blogged: Middle Names in the United States over Time

I was wondering what proportion of people have middle names, so I asked the Census. Recently you requested personal assistance from our on-line support center. Below is a



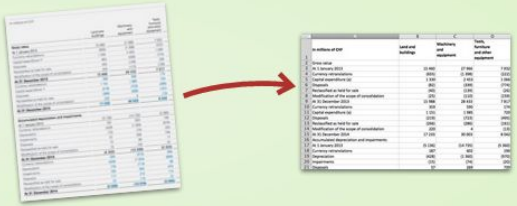
PDFTables Quick Trial Pricing Enterprise API Join CONVERT A PDF LOGIN PDF association MEMBER

Accurately convert PDF to Excel

Try our PDF to Excel converter for free!

No more time consuming and error prone copying and pasting. Convert PDF to Excel, CSV, XML or HTML.

CONVERT A PDF



Read reviews on TrustPilot

How to use — For Business — Blog — Questions?

Fast and efficient

Copypasting or transcribing large datasets by hand is very time-consuming. Free up hundreds of hours of work with PDFTables.

Excel and API

Effortlessly convert PDF to XLSX online. Or CSV, XML or HTML. If you're a coder, automate it using the PDFTables web API.

Cloud and on-premises

Use our website, powered by Amazon Web Services, or install our standalone Linux binary on your own infrastructure.

Secure and private

Support

Scalable

DataBaker

DataBaker is a Python library that helps you wrangle complex spreadsheets into **clean, normalised data tables**

Get the code | Read the tutorials

How it works

DataBaker is built to integrate with Jupyter and Pandas. It allows users to iteratively build up intuitive recipes that describe the structure of a spreadsheet. These recipes translate the spreadsheet into flat tables of data that can be used by Pandas and other data analysis libraries or saved as CSVs.

Examples

Cantabular

Blog Jobs

New! Public demo • We've republished the 1911 Irish census using Cantabular →

Real-time data publication with built-in privacy protection

Automate privacy protection and production of tabular data to ensure repeatability of outputs and enable flexible dissemination with our powerful API, Python tools and user interfaces.

Request a demo → Get in touch →

Why Cantabular?

Statistical disclosure control at speed and scale

Unlock the value of your data by publishing it faster

Significantly reduce the delay between data collection and publication with fast, automated privacy protection and tabulation.

Increase productivity through automation and repeatability

Free up your statisticians' time by automating tasks with our API and Python tools and speeding up testing of outputs and privacy methods.

Keep control of your own data and privacy approaches

Use our flexible configuration options and Disclosure Rules Language to fully control privacy techniques.

Background

Drivers for change

Growing user demand for data

Increase in volume of data

Relentless advances in technical capabilities

Challenges for delivery

Security and privacy requirements

Legacy systems and software

Governance and existing processes

ONS vision for 2021 Census

- **Flexible:** users will be able to define their own, more detailed outputs
- **Timely:** automation of cross-tabulation and statistical disclosure controls will allow results to be published sooner
- **Accessible:** census data will all be accessible from one location

2011

- 1000s of static tables
- Limited customisation of tables
- Manual review of every table released
- 4-5 years to release all outputs

2021

- Hundreds of millions of possible tables
- Build your own table from scratch
- Automated table checks
- 18-24 months to release all outputs

Technical challenges & solutions

Challenges

- Automating perturbation algorithms in real-time
- Giving ONS control of automated disclosure checks
- Helping users build their own tables
- Flexible metadata to complement data
- Bringing together everything into a single API

Challenge #1

Automating perturbation algorithms in real-time

Need

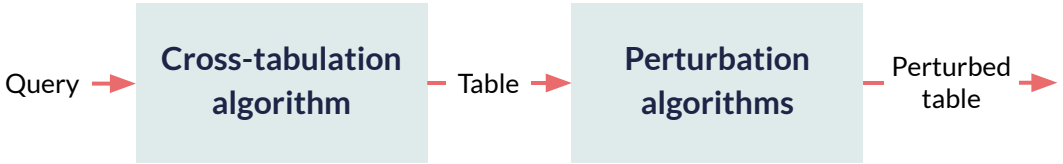
Build cross-tabulations from confidential microdata and apply perturbation algorithms in real-time, in response to a user's query.

Perturbation approach

- **Cell-key perturbation** of frequency counts
- Independent **perturbation of zeros** in frequency counts
- Preservation of **structural zeros**

Note: Source data will already have been aggregated and row-swapped.

How it works


Query
Cross-tabulation algorithm
Table
Perturbation algorithms
Perturbed table

Technical approach

- **Data changes infrequently:** forgo complicated database software and implement our own algorithms and data structures, keeping things simpler and more easily scalable
- **Data is small:** 10GB of CSV can be stored in 1GB of RAM and scanned in place, eliminating slow operations like disk or network access

Results

- Query for Age by Sex by Output Area (low level geography)
  - 60 million rows of input data
  - 3 million cells of output data
  - **Takes ~0.5 seconds**
- Outputs validated independently for correctness

Demo

Challenge #2

Giving ONS control of automated disclosure checks

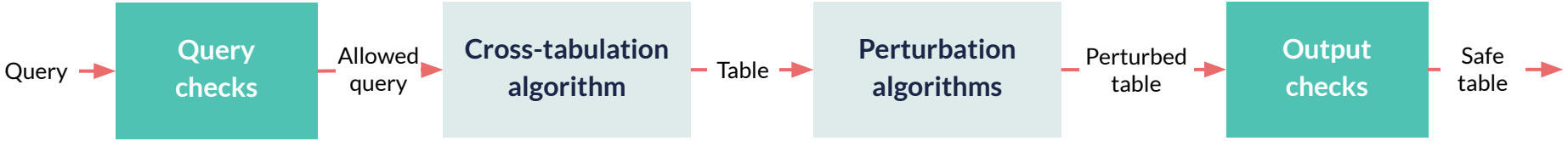
Need

Create the capability to allow disclosure checks to be specified and automated, and for new checks to be created without requiring software changes.

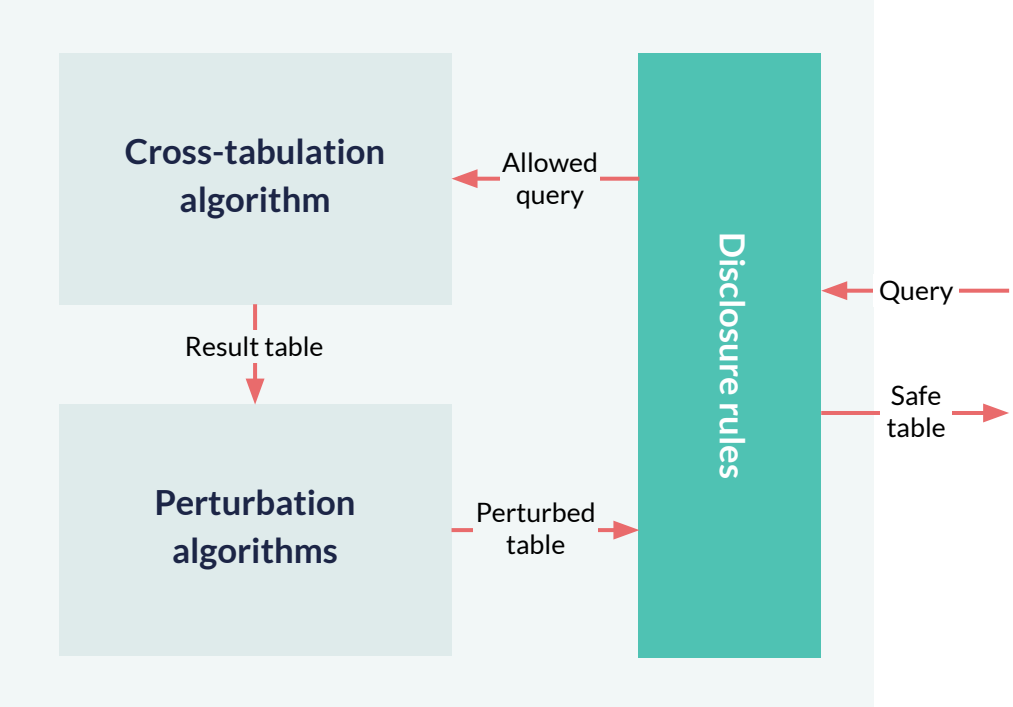
Illustrative disclosure checks

- **Set maximum variables:** block queries that will lead to overly sparse outputs before they are run
- **Attribute disclosure:** individual or group attribute disclosure in a table can be detected suppressed
- **Identity disclosure:** tables containing too many values of one can also be blocked

How it works


Query
Query checks
Allowed query
Cross-tabulation algorithm
Table
Perturbation algorithms
Perturbed table
Output checks
Safe table

How it works: single software component



A domain specific language

- Simple imperative language for specifying disclosure rules as basic algorithms
- Limited syntax makes it easier to learn and read
- Rules can be created and modified independently
- Rules can be kept secret from software engineers

Example rule

```
querytest withinMaxVars(max)
  if "OA" sourceof query.vars[0].name
    // any geographic query
    fail if (len query.vars) > (max + 1)
  else
    // non geographic query
    fail if (len query.vars) > max
  end
end
```

Results

- **Successfully implemented language:** ONS tested and confirmed results and ability to write their own rules
- **Mitigated any impact on performance:** system performance following implementation actually improved
- **Facilitated use cases beyond initial design:** ONS using rules to gradually open access to more queries

Demo

Challenge #3

Helping users build their own tables

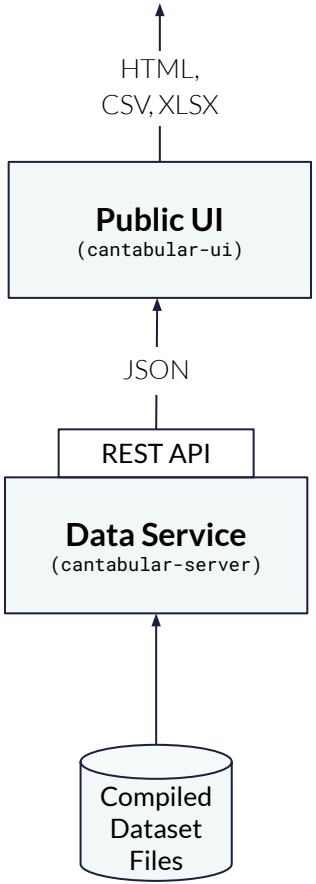
Need

Explore how to design a user interface that helps users build their own table from a microdata-based dataset.

Design constraints

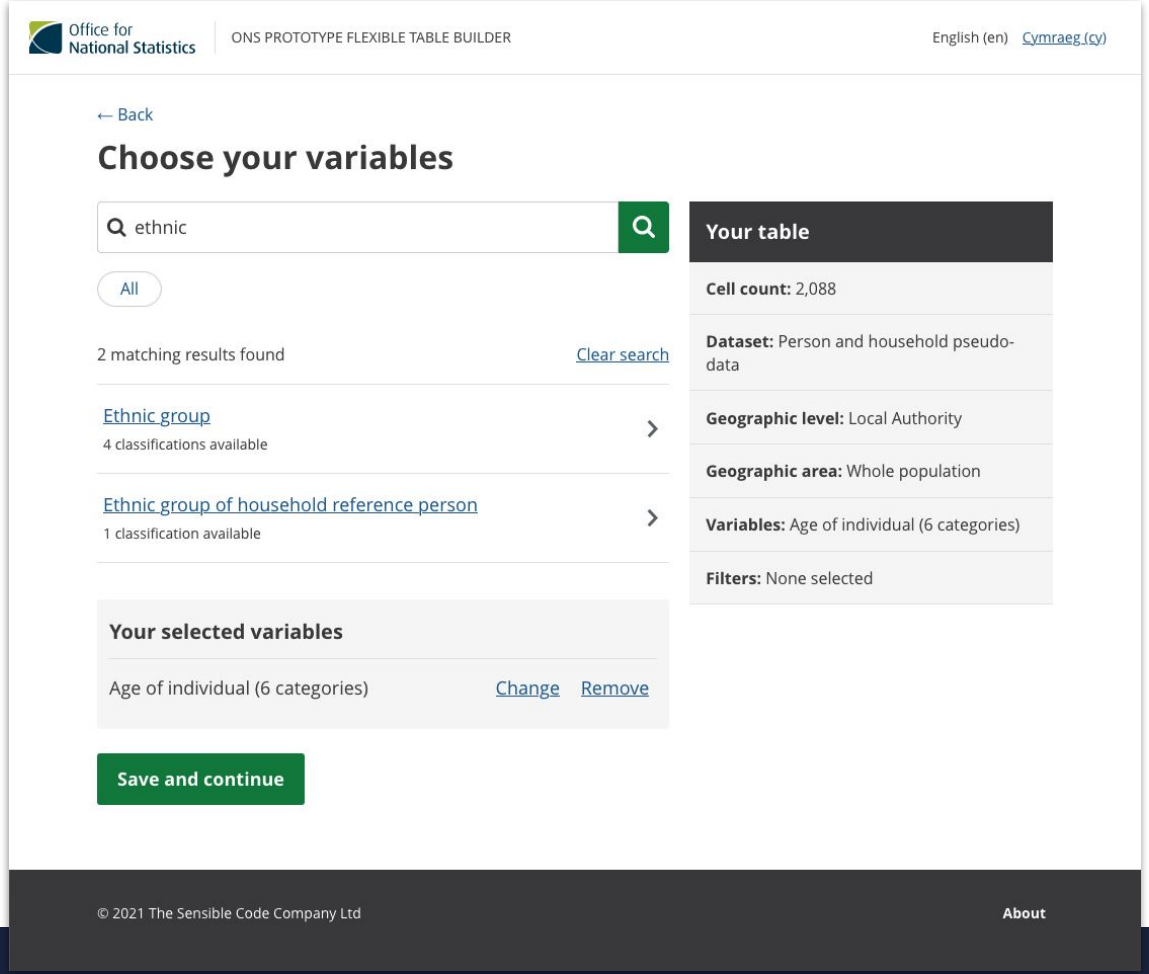
- **Lots of dimensions:** hundreds of possible variables, some variables may have tens of classifications.
- **Disclosure control conditions:** the number of dimensions you can choose is limited, and choosing lower level geographies limits it still further.

How it works


HTML,
CSV, XLSX
Public UI
(cantabular-ui)
JSON
REST API
Data Service
(cantabular-server)
Compiled
Dataset
Files

Approach

- **Developed alternative prototypes** which ONS tested with their users and used in consultations.
- **Implemented separate user interface service** and continued to develop it as a product independent of ONS
- **Now supporting ONS** to develop their own user interface, using similar design patterns, built on top of our software


Office for National Statistics
ONS PROTOTYPE FLEXIBLE TABLE BUILDER
English (en) Cymraeg (cy)
← Back
Choose your variables
ethnic
All
2 matching results found
Clear search
Ethnic group
4 classifications available
Ethnic group of household reference person
1 classification available
Your selected variables
Age of individual (6 categories)
Change
Remove
Save and continue
Your table
Cell count: 2,088
Dataset: Person and household pseudo-data
Geographic level: Local Authority
Geographic area: Whole population
Variables: Age of individual (6 categories)
Filters: None selected
© 2021 The Sensible Code Company Ltd
About

Demo

Challenge #4

Complementing data with flexible metadata

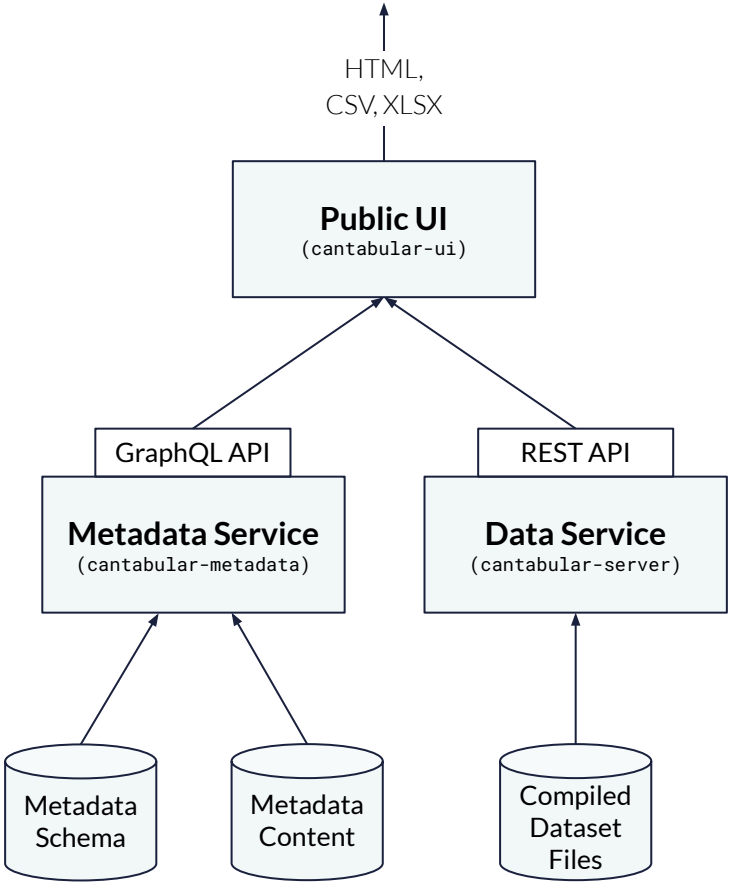
Need

Develop a capability to allow multi-lingual reference metadata to be associated with flexibly created outputs.

Constraints

- **Unknown metadata schema:** at the time of its creation, the ONS metadata model had not been completed
- **Schema/vocabulary agnostic:** different organisations adopt different approaches so we needed a flexible solution

How it works


HTML,
CSV, XLSX
Public UI
(cantabular-ui)
GraphQL API
REST API
Metadata Service
(cantabular-metadata)
Data Service
(cantabular-server)
Metadata Schema
Metadata Content
Compiled Dataset Files

Technical approach

- **Minimal built-in schema:** simple hierarchical schema of Service > Dataset > Variable
- **User-defined schema:** allow specification of arbitrary fields to be associated with different built-in concepts, using a user-defined schema parsed at runtime
- **Simple data loading and storage:** all metadata is specified as JSON files and stored by the service in-memory

Demo

Challenge #5

Bringing everything together in a single API (application programming interface)

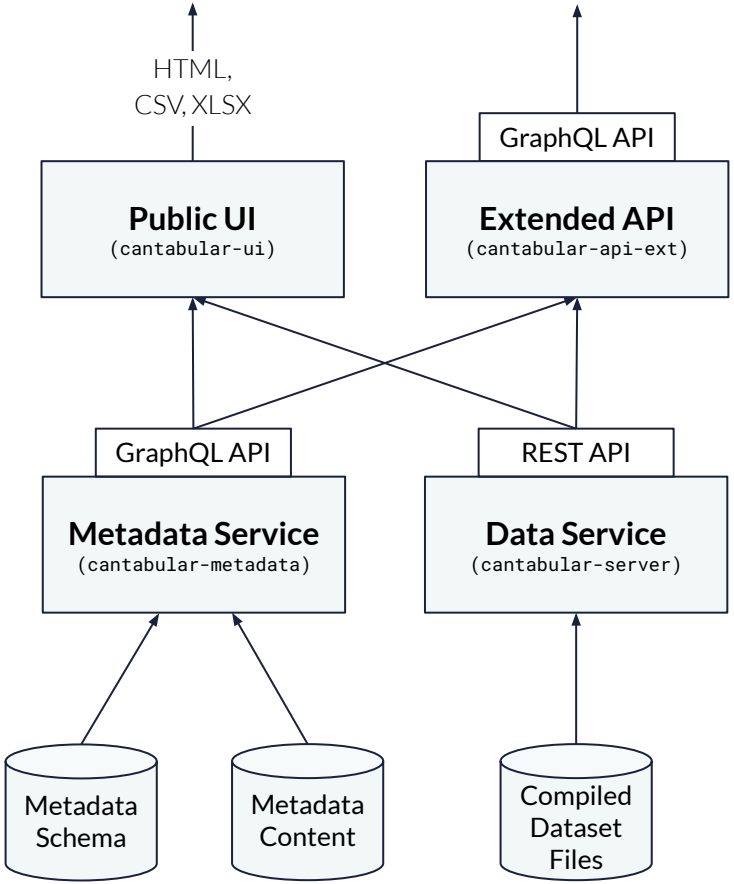
Need

Provide an easy-to-use API that integrates data and metadata into a single combined interface for use in ONS digital products.

Technical approach

- **Single source of data & metadata:** combine data, structural metadata and reference metadata into one integrated API
- **Support multiple languages:** Census outputs need to be in English and Welsh; use metadata service to translate all metadata
- **GraphQL API for flexibility:** use a GraphQL rather than REST API to allow complete flexibility in what can be queried

How it works


HTML,
CSV, XLSX
GraphQL API
Public UI
(cantabular-ui)
Extended API
(cantabular-api-ext)
GraphQL API
REST API
Metadata Service
(cantabular-metadata)
Data Service
(cantabular-server)
Metadata
Schema
Metadata
Content
Compiled
Dataset
Files

Results

- Being used by ONS to power a range of different census products:
  - Dataset search and discovery
  - Custom table user interface
  - Geographic area profiles
  - Data visualisations
  - Data dictionary

Demo

Benefits of flexible dissemination

Benefits

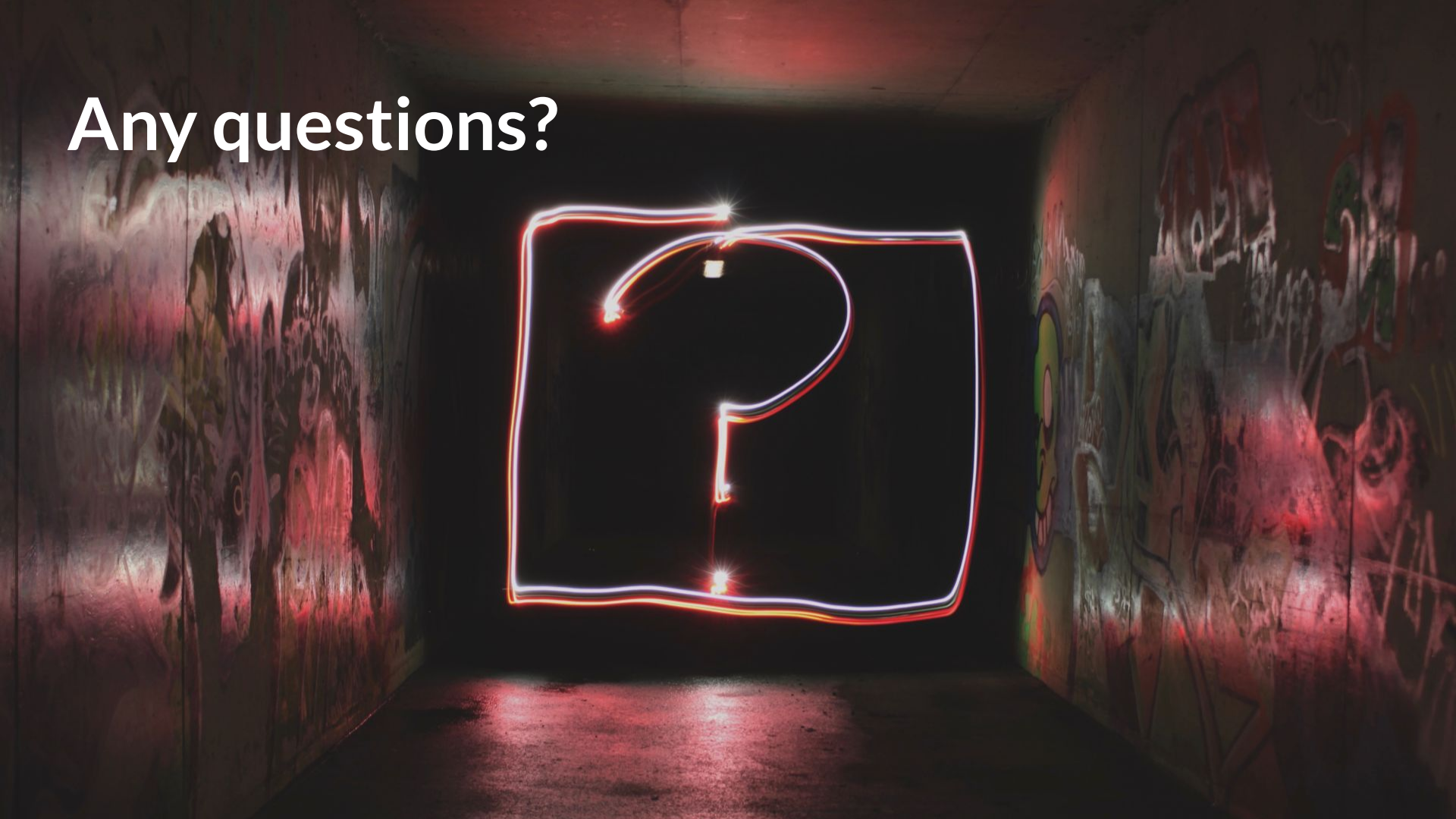
- **Publish more data:** flexible dissemination means the range of possible outputs is huge and users can self-serve
- **Publish more quickly:** automation of SDC checks means time taken to release everything will be compressed
- **Improve reliability and reproducibility:** More automation reduces opportunities for human error to creep in
- **Multiple language support:** provide structural and reference metadata in multiple languages

What’s next?

What's next?

- **Supporting flow data:** allowing cross-tabulation of migration and commuting patterns data, which are often very large tables
- **Supporting magnitude data:** extending disclosure control approaches to magnitude data (with NSI support on methodology)
- **Adding visualisation tools:** allowing some exploratory visualisation and mapping in the user interface

Any questions?



Thanks!

mike@sensiblecode.io