Flexible dissemination software for the 2021 England & Wales Census

20th September 2022

The Sensible Code Company

The Sensible Code Company

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

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.

Contents

- Background
- Technical challenges & solutions
- Opportunities for innovation
- What’s next?
- Questions

Background

What is flexible dissemination?

Giving users the capability to create their own custom outputs directly from microdata.

2011

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

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

Benefits

- **Publish more data:** flexible dissemination means the range of possible outputs is huge and users can self-serve
- **Publish more quickly:** automation of statistical disclosure control checks means time taken to release everything will be compressed
- **Improve reliability and reproducibility:** More automation reduces opportunities for human error to creep in

Technical challenges & solutions

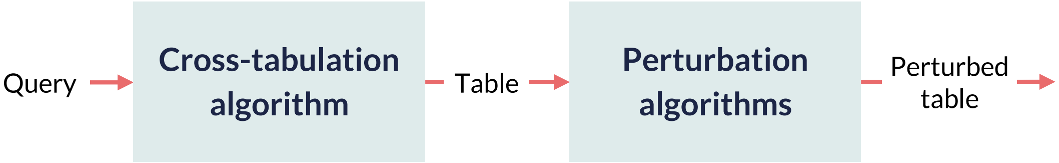
Challenge #1

Automating tabulation and perturbation in real-time

Need

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

Concept


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

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 row-swapped.

For more, see “The methodological challenges of protecting outputs from a Flexible Dissemination System” by Stephanie Blanchard.

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

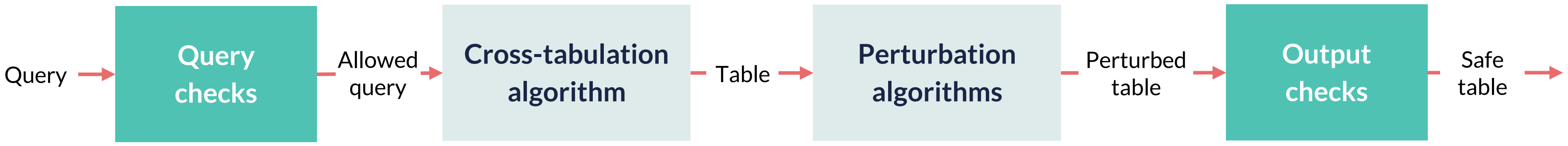
Challenge #2

Giving ONS control of automated disclosure checks

Need

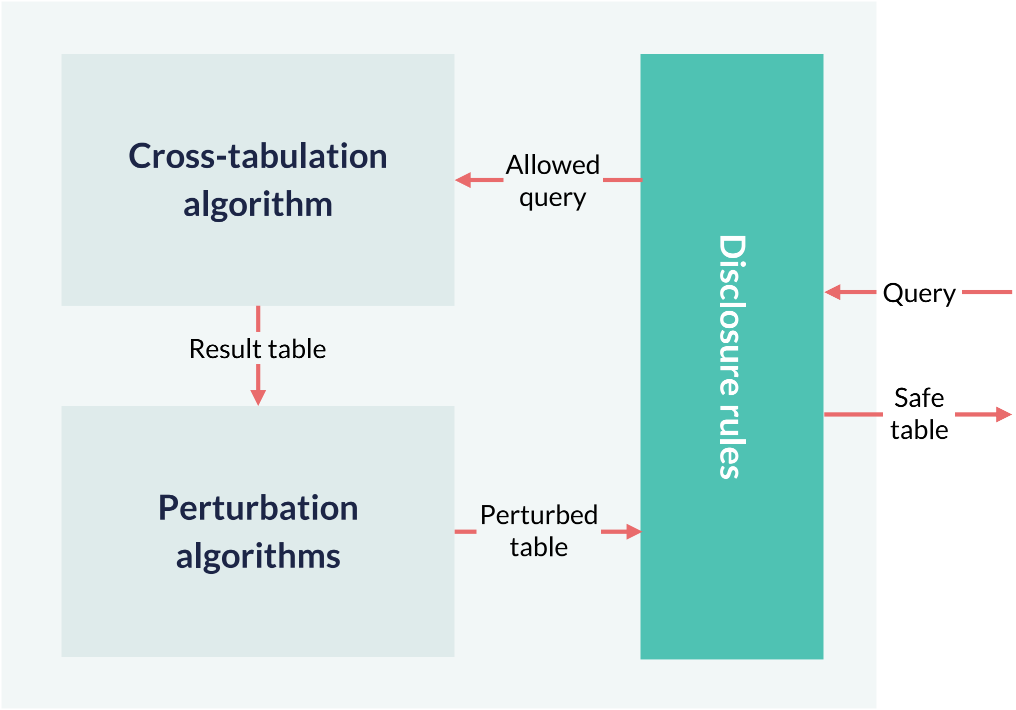
Create the capability to allow disclosure checks to be specified and automated, and for new checks to be created without requiring software changes.

How it works

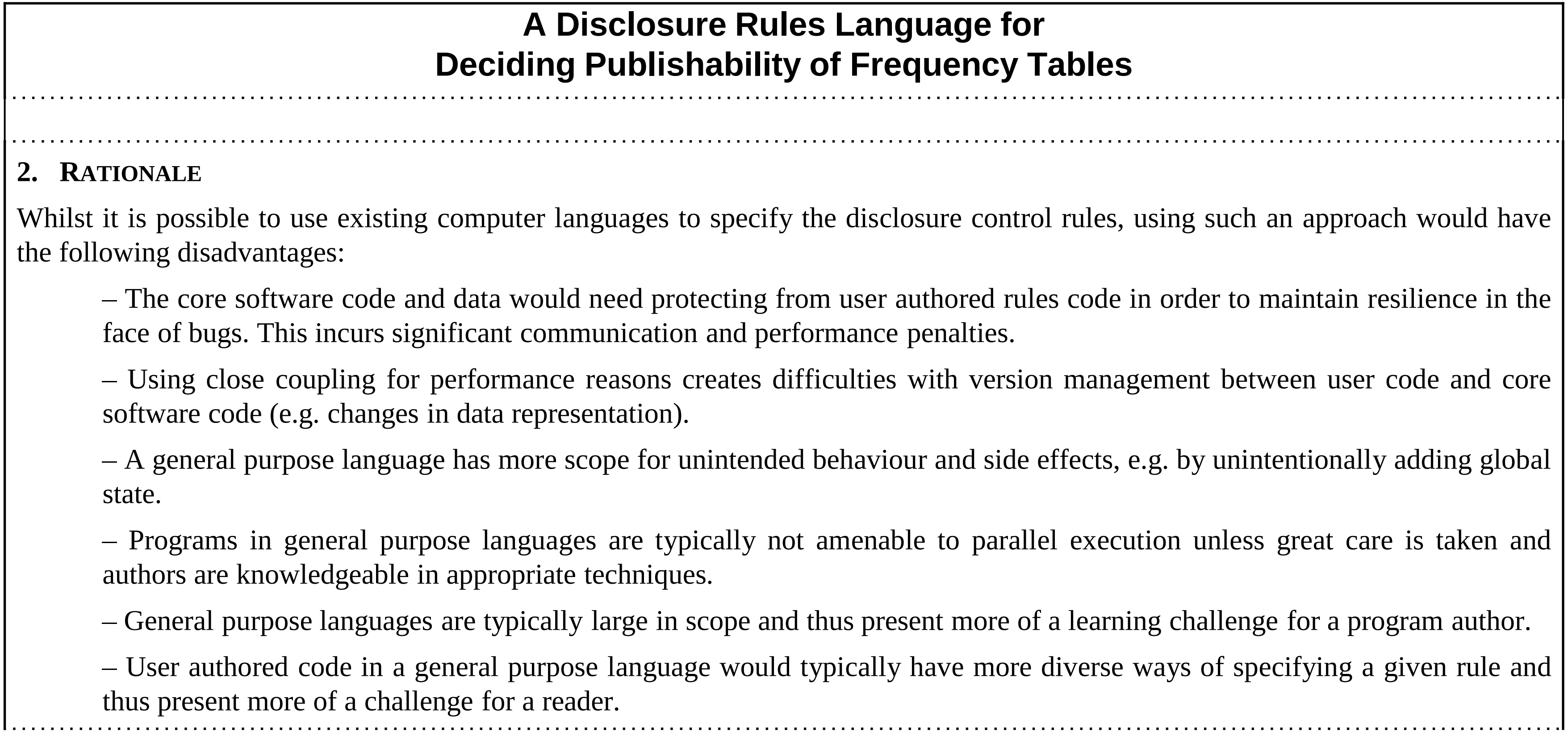


Query → Query checks → Allowed query → Cross-tabulation algorithm → Table → Perturbation algorithms → Perturbed table → Output checks → Safe table →

How it works: single software component



NTTS 2021 conference paper



A Disclosure Rules Language for Deciding Publishability of Frequency Tables

2. Rationale

Whilst it is possible to use existing computer languages to specify the disclosure control rules, using such an approach would have the following disadvantages:

– The core software code and data would need protecting from user authored rules code in order to maintain resilience in the face of bugs. This incurs significant communication and performance penalties.

– Using close coupling for performance reasons creates difficulties with version management between user code and core software code (e.g. changes in data representation).

– A general purpose language has more scope for unintended behaviour and side effects, e.g. by unintentionally adding global state.

– Programs in general purpose languages are typically not amenable to parallel execution unless great care is taken and authors are knowledgeable in appropriate techniques.

– General purpose languages are typically large in scope and thus present more of a learning challenge for a program author.

– User authored code in a general purpose language would typically have more diverse ways of specifying a given rule and thus present more of a challenge for a reader.

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 and suppressed
- **Identity disclosure:** tables containing too many values of one can also be blocked

Results

- **It worked!** ONS tested and confirmed results and ability to write their own rules
- **Facilitated use cases beyond initial design:** ONS using rules to gradually open access to more data
- **Rules can be kept secret** from software developers!

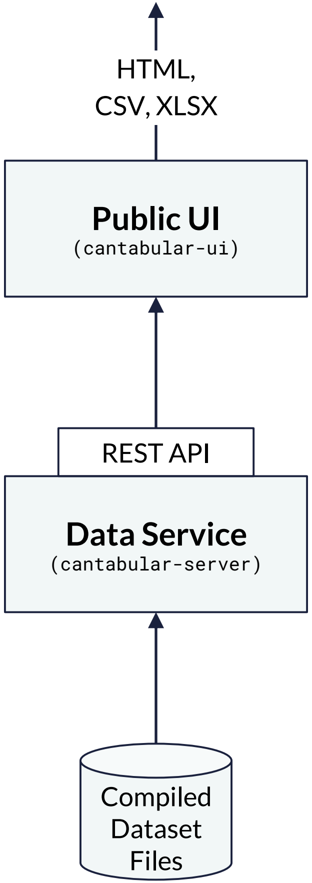
Challenge #3

Helping users build their own tables

Need

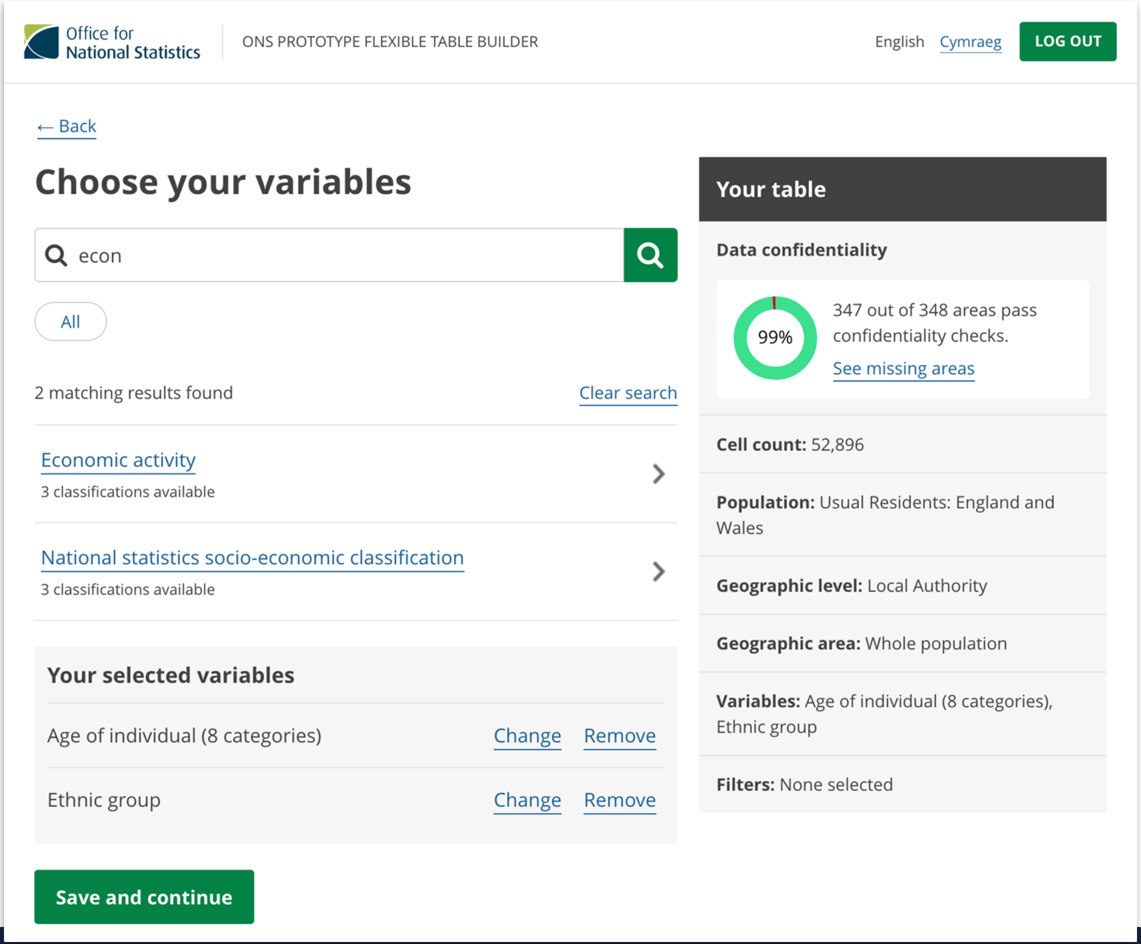
Explore how to design a user interface that helps users build their own table from a microdata-based dataset.

How it works


HTML,
CSV, XLSX
Public UI
(cantabular-ui)
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 Cymraeg **LOG OUT**

← Back

Choose your variables

econ

All

2 matching results found Clear search

Economic activity
3 classifications available

National statistics socio-economic classification
3 classifications available

Your selected variables

Age of individual (8 categories) Change Remove

Ethnic group Change Remove

Save and continue

Your table

Data confidentiality

99%

347 out of 348 areas pass confidentiality checks.

See missing areas

Cell count: 52,896

Population: Usual Residents: England and Wales

Geographic level: Local Authority

Geographic area: Whole population

Variables: Age of individual (8 categories), Ethnic group

Filters: None selected

Demo

Challenge #4

Flexible data needs flexible metadata

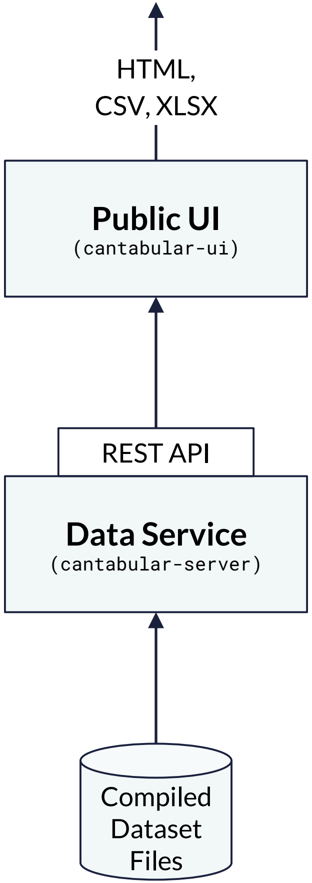
Need

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

Constraints

- **Metadata schema still in development:** at the time of its creation, the ONS metadata model was still being designed
- **Vocabulary/standard agnostic:** different organisations adopt different approaches so we needed a flexible solution

How it works



HTML,
CSV, XLSX

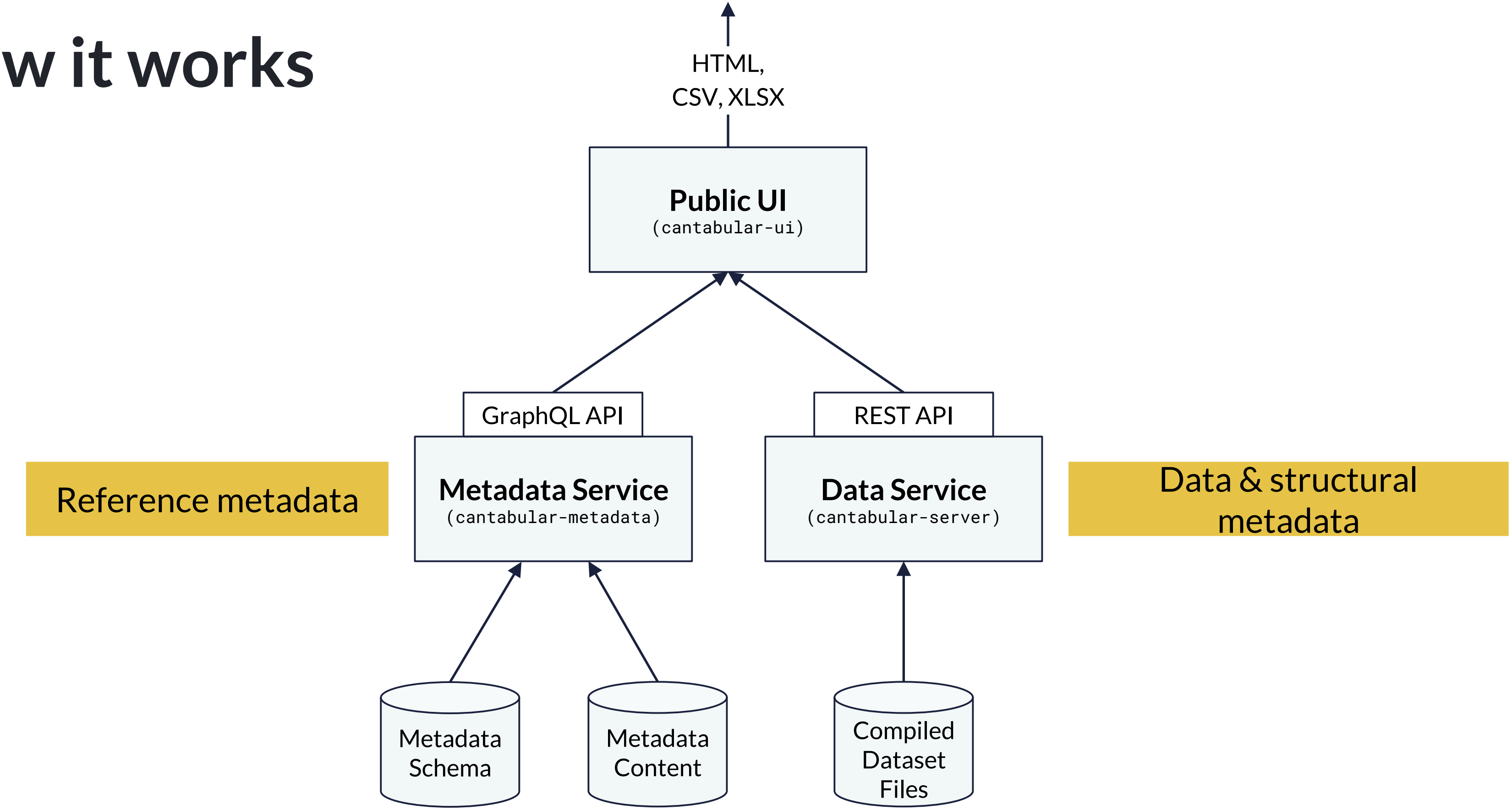
Public UI
(cantabular-ui)

REST API

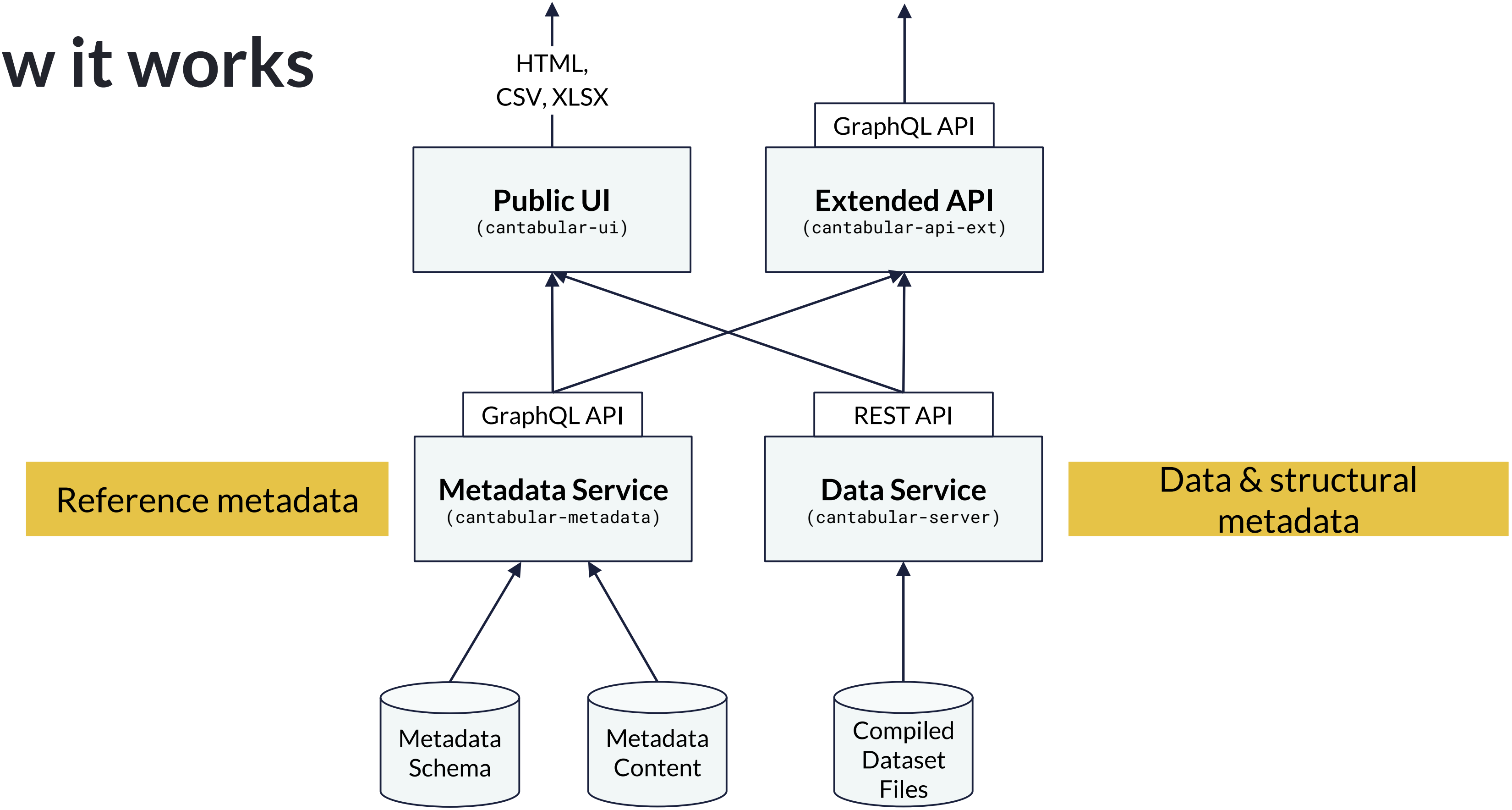
Data Service
(cantabular-server)

Compiled
Dataset
Files

How it works


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

How it works



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

Technical approach

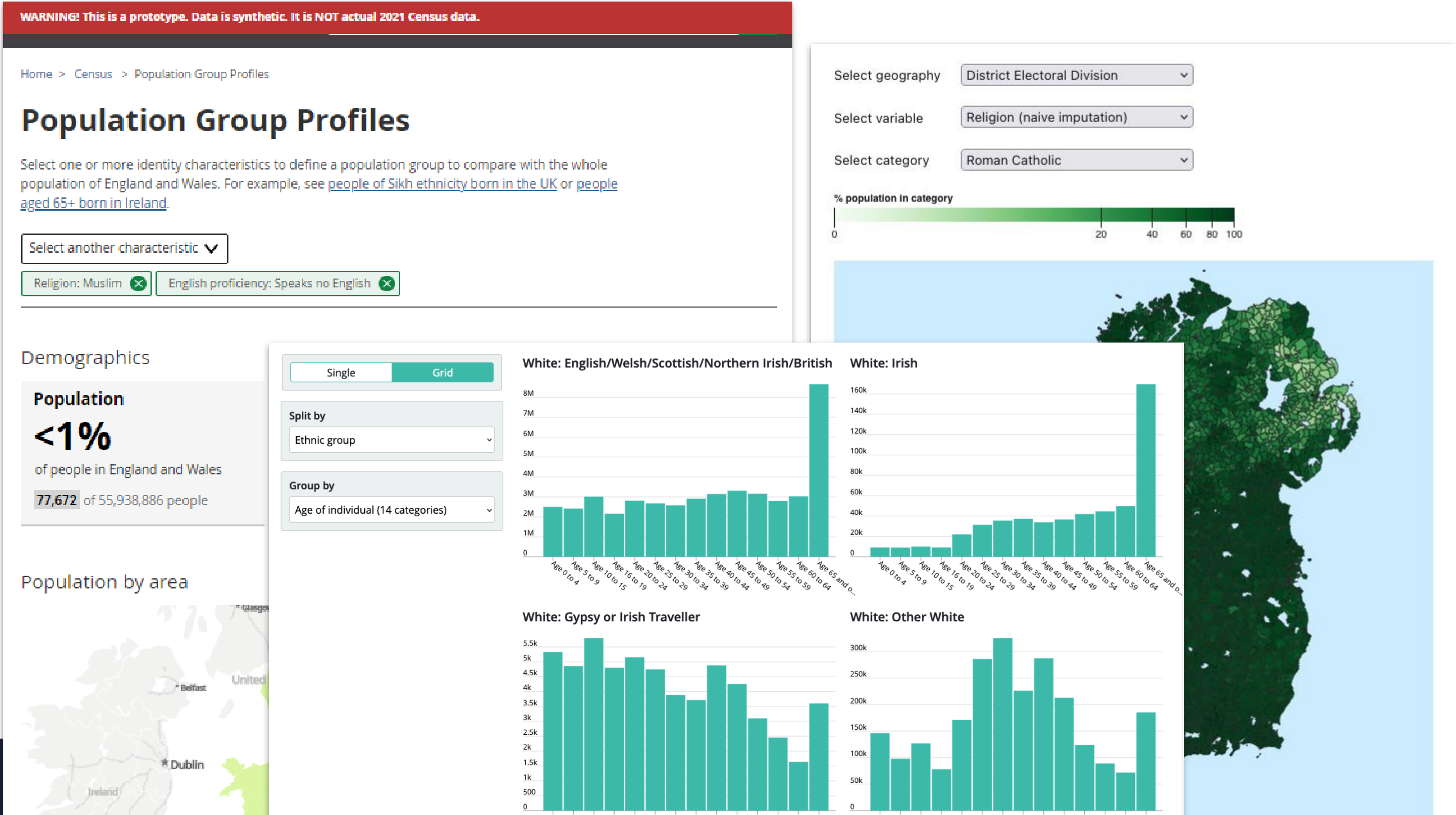
- **User-defined schema:** allow specification of arbitrary custom metadata using a user-defined schema parsed when the software is run
- **Single source of data & metadata:** combine data, structural metadata and reference metadata into one integrated API
- **Support multiple languages:** Census outputs need to be in English and Welsh; use metadata service to translate all metadata

Demo

Opportunities for innovative products

ONS census dissemination

- 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



WARNING! This is a prototype. Data is synthetic. It is NOT actual 2021 Census data.

Home > Census > Population Group Profiles

Population Group Profiles

Select one or more identity characteristics to define a population group to compare with the whole population of England and Wales. For example, see people of Sikh ethnicity born in the UK or people aged 65+ born in Ireland.

Select another characteristic

Religion: Muslim ☒ English proficiency: Speaks no English ☒

Demographics

Population

<1%

of people in England and Wales

77,672 of 55,938,886 people

Population by area

Single | Grid

Split by: Ethnic group

Group by: Age of individual (14 categories)

White: English/Welsh/Scottish/Northern Irish/British

White: Irish

White: Gypsy or Irish Traveller

White: Other White

Select geography: District Electoral Division

Select variable: Religion (naive imputation)

Select category: Roman Catholic

% population in category

Demo

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

Thanks!

mike@sensiblecode.io

https://cantabular.com

https://ireland-census-preview.cantabular.com

Cantabular™