Enabling data linkage on pseudonymised address data

(GeoDirectory to Geography)

Justin McGurk - November 7th 2023

UN-GGIM Webinar: Linked Data in the European spatial data infrastructure – an operational view

Statistical Geospatial Framework (GSGF)
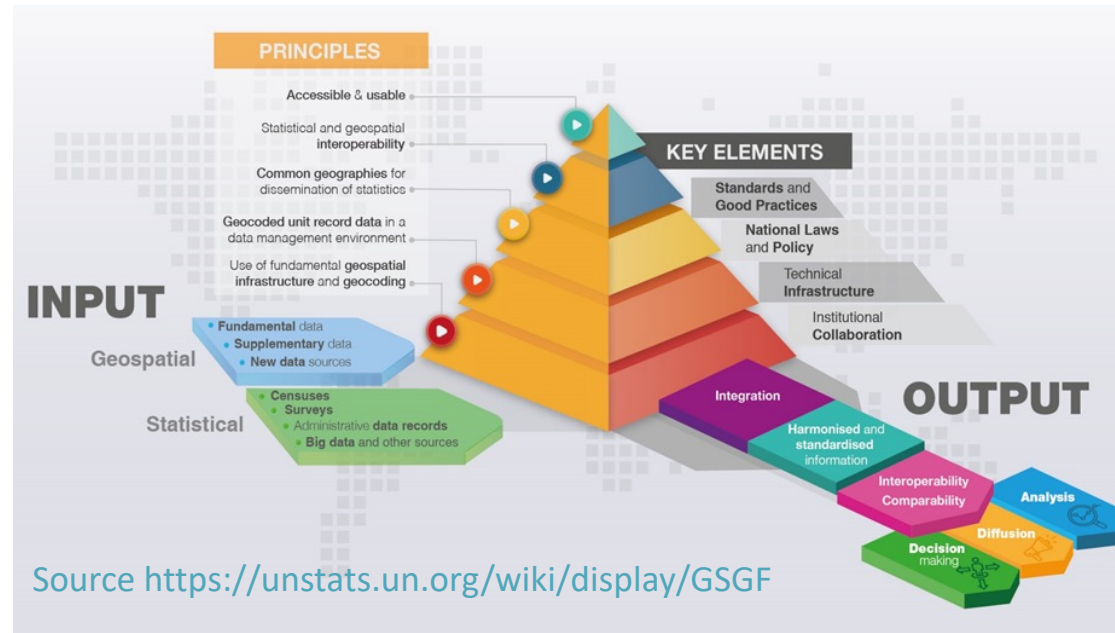
Aim

- Facilitate the integration of statistical and geospatial information

Eircode

- Seems like a candidate for this use

- Administrative data that contain Eircodes

- In use since 2015

INSPIRE Spatial Data Themes

- Addresses, (ad)

- Statistical units, (su)

- Administrative units, (au)

Source https://unstats.un.org/wiki/display/GSGF

Some Definitions and Terminology (CSO)

ADC – Administrative Data Centre

- The ADC is responsible for ingesting, processing, transforming, pseudonymizing, and storing records for statistical analysis within CSO.

Source Tier Data

- Data in ADC as supplied. Access strictly controlled and not to be used for statistical analysis.

Analysis Tier Data

- Data in ADC that has been pseudonymised and made available for statistical analysis within the CSO.

Some Definitions and Terminology (Data)

LEA – Local Electoral Area

- Implementation of Local Administrative Units (LAUs) in Republic of Ireland, currently 166 in total.

ED – Electoral Division

- Smallest legally defined administrative area in Republic of Ireland. Traditionally 3440 or 3441, now 3455 due to Local Authority boundary changes in 2019

GeoDirectory

- Address database in Ireland that is provided by a company with ownership split between An Post (51% - Postal service) and Tailte Éireann (49% - National mapping agency)

Eircode

- Postcodes in Ireland: Unique to postal address – seven-character alpha-numeric. Example for CSO office in Cork. T12X00E

User Story

What User Wanted to Do

Produce statistics by LEA from Analysis Tier Data using admin data:

- Vacant Dwelling Indicators based on Metered Electricity Consumption

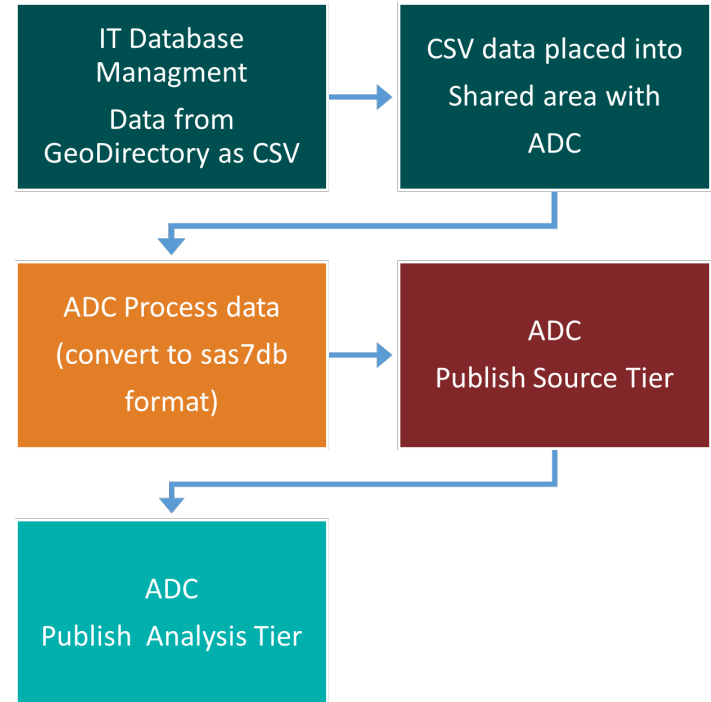However, this admin data does not have LEA information

What is the issue?

What were we doing – Address data

- Regular quarterly update

- Standard process in place for number of years

```
IT Database Managment
Data from GeoDirectory as CSV
        →
CSV data placed into Shared area with ADC
        ↓
ADC Process data (convert to sas7db format)
        →
ADC Publish Source Tier
        ↓
ADC Publish  Analysis Tier
```

What is the issue – Part 1

Process in place for number of years

- Not ingesting EIRCODE_LOOKUP table consistently

- Does not as a matter of course add value or post process data

10.5	October 2015	Moving to 139 Principal Post-towns, adding Eircodes to addresses, new table called Eircode_Lookup and removing rules for creating the Official Postal Address.

EIRCODE_LOOKUP	
	BUILDING_ID
	ADDRESS_POINT_ID
	ADDRESS_REFERENCE
	ECAF_ID
	EIRCODE

Source: GeoDirectory Technical Guide V10.doc

GeoDirectory – Structure

Building Table

- Only place with x-y or lat-long data.

Eircode_Lookup

- Buildings, Addresses, and Eircode's are not synonyms.

BUILDINGS
BUILDING_ID
GROUP_ID
THORFARE_ID
PRINCIPAL_POST_TOWN_ID
DATA_SRC_ID
CHANGED_DATE
POSTAIM_PRESORT_61
POSTAIM_PRESORT_152
ED_ID
NAME
NO
BUILDING_USE
DERELICT
VACANT
INVALID
UNDER_CONSTRUCTION
RESIDENTIAL_DELIVERY_POINTS
COMMERCIAL_DELIVERY_POINTS
COUNTY_ID
TLAND_ID
EAST
NORTH
LOCALITY_ID
SECONDARY_LOCALITY_ID
VERIFIED
ITM_EAST
ITM_NORTH
QUALITY_CONTROL
PPZ_ID
HOLIDAY_HOME
DEMOLISHED
CREATE_DATE
LATITUDE
LONGITUDE
BUILDING_NUMBER_ALPHA
LA_ID
VO_ID
LOCALITY_POSTALLY_REQ
SMALL_AREA_ID

EIRCODE_LOOKUP
BUILDING_ID
ADDRESS_POINT_ID
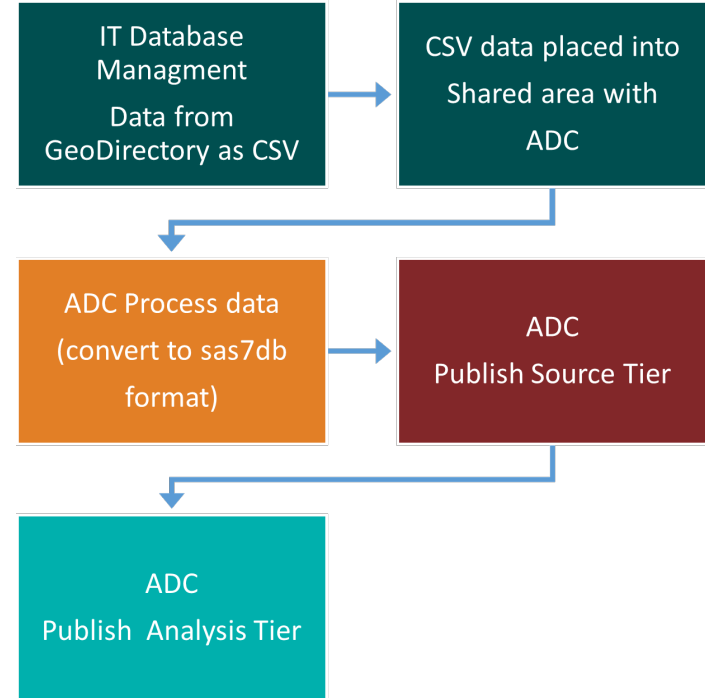ADDRESS_REFERENCE
ECAF_ID
EIRCODE

What is the issue – Part 2

Source Tier

- Trivial to create point for unique Eircode.

Analysis Tier

- Used to link administrative data.

- Pseudonymisation process carried out to create it.

IT Database Managment

Data from GeoDirectory as CSV

→

CSV data placed into Shared area with ADC

ADC Process data (convert to sas7db format)

→

ADC Publish Source Tier

ADC Publish Analysis Tier

What is the issue – Part 2

Source Tier view of CSO Cork office

BUILDING_ID	ADDRESS_POINT_ID	ECAD_ID	ADDRESS_REFERENCE	EIRCODE
54094840	33251725	1900192711	5409484033251720	T12X00E
54094840	36645882	1900192712	5409484036645880	T12X00E

BUILDING_ID	EAST	NORTH	ITM_EAST	ITM_NORTH	LATITUDE	LONGITUDE	LA_ID	VO_ID	LOCALITY_POSTAL	SMALL_AREA_ID
54094840	171869.9	70880.13	571826.326	570942.539	51.8899345	-8.409300049	33	Y		7802

GeoDirectory: 2023 Q02

The Root Problem

Pseudonymisation destroys the possibility of spatial analysis

Co-ordinates are encrypted and data matching/linking is not possible

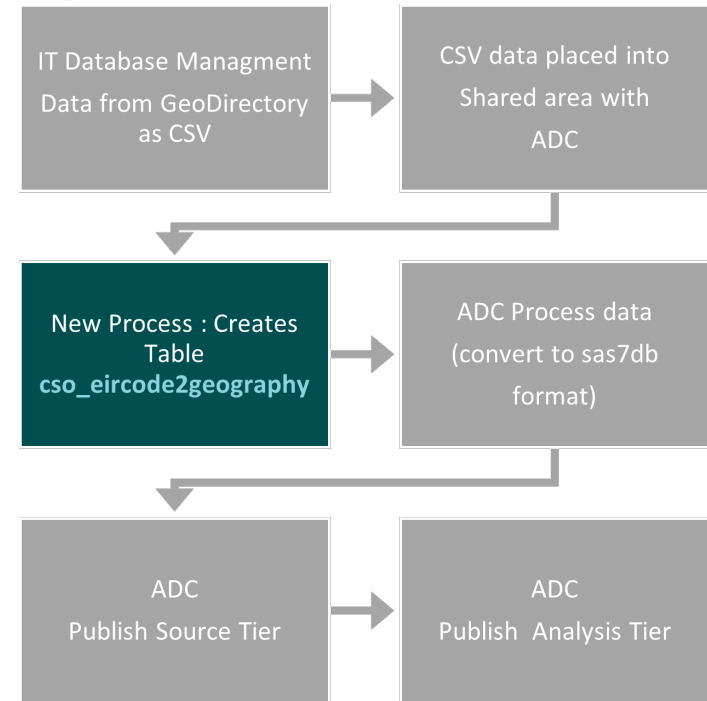| Building_IDPIK	LA_ID	SMALL_AREA_ID	EastPIK	NorthPIK	LatitudePIK	LongitudePIK
000010491ACF5CADF5FA4243052ED63E5BF	28	4350	A973043CE	EDB79B7982	0F377265CFAC	EF2F84C3E5F518FC0
000011C385DC0EB0359135481393E0B1EC43	12	11792	2086B2905	30A74E56AA	BAC548BE8AFE	1D11DAC0464B8DFE
000018173EDD940E05EAFEEF89468EB44E31	18	13878	54F0C2FEC	AE7DA0CB58	335816847E802	96E2DFD381FED105E
00002767C988190CD830A01601B54141D0E0	13	12027	ACCC7528:	E17019EF520	7BCF06817BF0(	A9DF71FF1306A569(
00003826C625392E8912E11F5DE16DA974C8	1	NA	09E04CE87	F46AE7C35F	93E9969FEC8C/	C7AEB69C4F6B18321

Revised Process

Proposed Process – Adding a step

Post process data and add value

- Create Eircode point
- Link point to an area
  - Census 2022 boundaries
  - European Grid
  - etc

IT Database Managment Data from GeoDirectory as CSV → CSV data placed into Shared area with ADC

New Process : Creates Table **cso_eircode2geography** → ADC Process data (convert to sas7db format)

ADC Publish Source Tier → ADC Publish Analysis Tier

Data Hub - 2023

Key year for onboarding ADC data

- Opportunity to improve data & metadata quality

- ADC consumers will be re-directed to Data Hub

ADC SAS files → Data & metadata enhancements → Data Hub Hive Db tables

Data Hub Portal – Now live

- Data owners

  - Setup key metadata and approval processes before release of data

  - Manage access requests

- Consumers of data in Data Hub

  - Search metadata

  - Request access to data – more granular – by variable & reference period

- Data Office

  - Improve integration with Data Office systems/processes
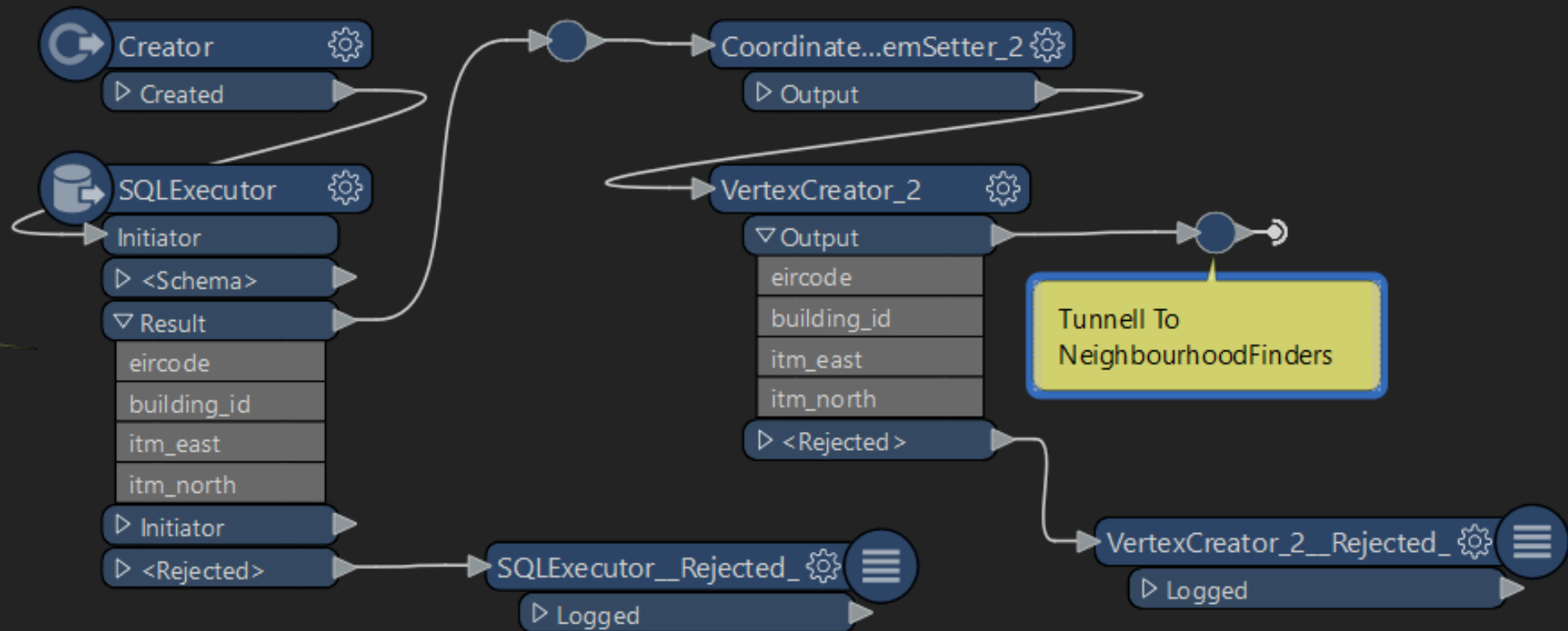
CSO Data Hub – SQL like environment

Use Data Hub to do the work

- Temporary table – non persistent data

- Join BUILDINGS to EIRCODE_LOOKUP

- Get only distinct values of Eircode, building, co-ordinates

Fetch Parameterisation

- Table names are based on version of GeoDirectory

  - YYYY

  - q0X {q01,q02, q03, q04}
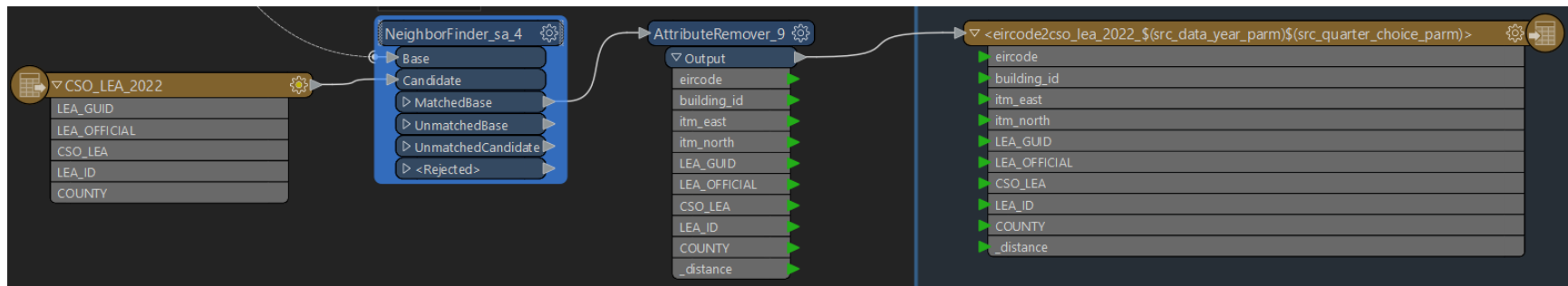
- Process is the same only the input differs

FME – SQLExecutor statement

```
FME_SQL_DELIMITER ;
-- parameterised from database :$(src_data_base)
-- parameterised temptable  from database : $(src_data_base).temp_fme_x
-- parameterised fetch Eircodes:
$(src_data_base).eircodes$(src_data_year_parm)$(src_quarter_choice_parm)_src
-- parameterised fetch Buildings:
$(src_data_base).buildings$(src_data_year_parm)$(src_quarter_choice_parm)_src
CREATE TEMPORARY TABLE $(src_data_base).temp_fme_x AS
SELECT DISTINCT g.*
FROM (
    SELECT e.eircode, e.building_id, b.itm_east, b.itm_north
    FROM adc_geodirqrtly_source.eircodes$(src_data_year_parm)$(src_quarter_choice_parm)_src as e
    INNER JOIN adc_geodirqrtly_source.buildings$(src_data_year_parm)$(src_quarter_choice_parm)_src  as b
    on e.building_id = b.building_id
    ) as  g;
SELECT t.eircode, t.building_id, t.itm_east, t.itm_north
FROM $(src_data_base).temp_fme_x as t
```

Using points to find nearest neighbor

For each point find closest neighbouring area

- Using Open data for boundary (Tailte Éireann Open data portal for source for 2022 census boundaries)

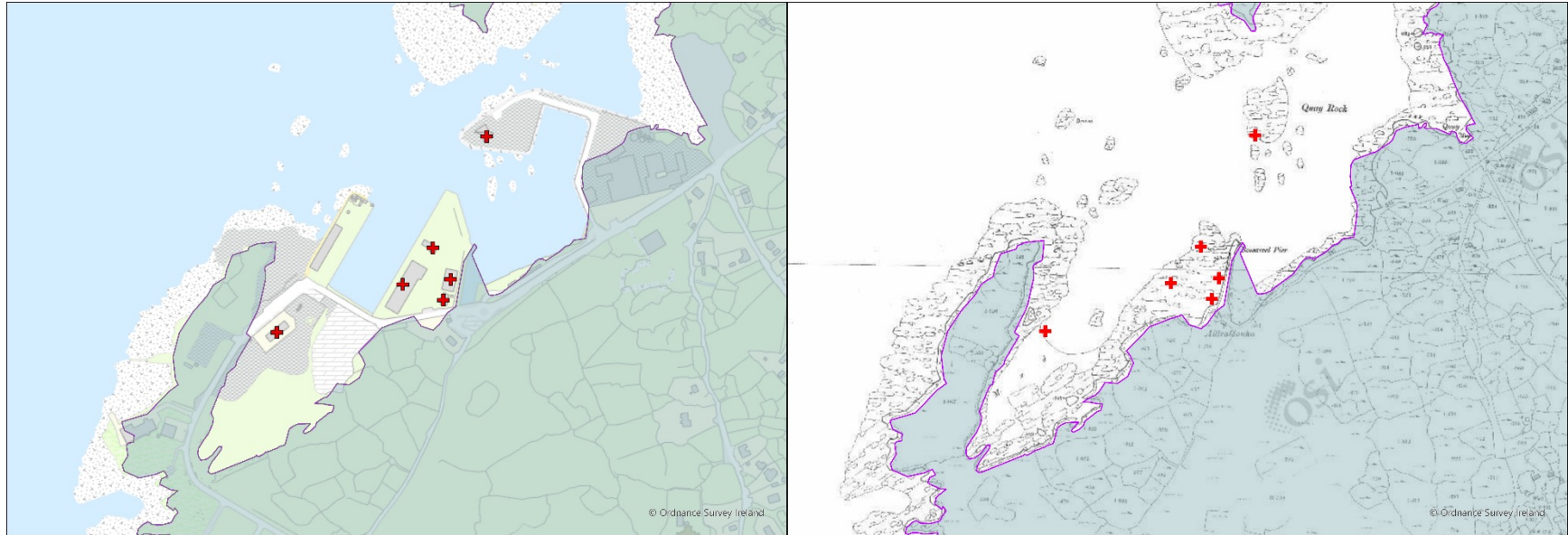- Writes results into CSV – ADC ingest and Pseudonymisation

Outcome: CSV file with linked data and...

Distance

- Value of zero means Eircode is within geography

- Nonzero value is distance to closest geography in metres

- Can be done for _**any**_ arbitrary geography

Near enough is good enough: example

© Ordnance Survey Ireland

© Ordnance Survey Ireland

Outcome – CSV file for each level of Geography used

Data Hub
ADC

ADC's role

Data Custodian

- Grant access and ensure governance procedures

Pseudonymise data: Replace unique identifiers with Protected Identifier Keys (PIKs).

- Eircode, Building ID, latitude and longitude.

- Geography data and identifiers left 'as is'.

Data Hub

- Metadata specifying the file contents, and Data Hub ingest the data

- Users can apply for access to the (Analysis Tier) data via the Data Hub Portal

- Now have data that allows linking administrative data to statistical geographies via PIK Eircode
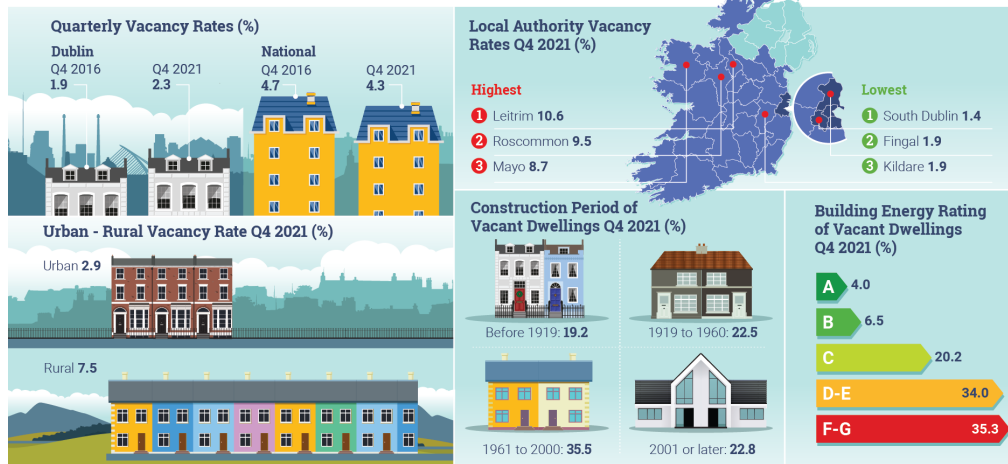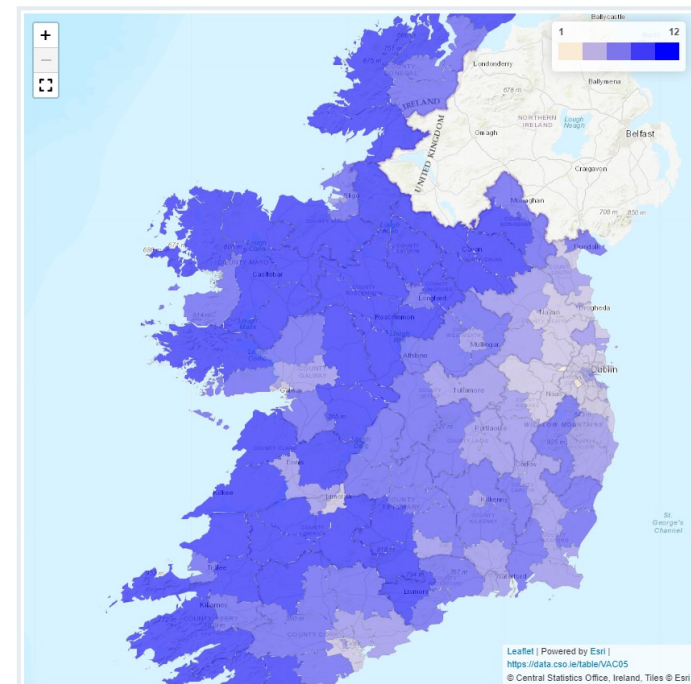
Outcome

Vacant Dwelling Indicators based on Metered Electricity Consumption 2021

Vacant Dwelling Indicators based on Metered Electricity Consumption 2021

Map 1.2 Local Electoral Area Vacancy Rates - Q4 2021

Quarterly Vacancy Rates (%)

Dublin
Q4 2016 **1.9**
Q4 2021 **2.3**

National
Q4 2016 **4.7**
Q4 2021 **4.3**

Urban - Rural Vacancy Rate Q4 2021 (%)

Urban **2.9**

Rural **7.5**

Local Authority Vacancy Rates Q4 2021 (%)

Highest
1. Leitrim **10.6**
2. Roscommon **9.5**
3. Mayo **8.7**

Lowest
1. South Dublin **1.4**
2. Fingal **1.9**
3. Kildare **1.9**

Construction Period of Vacant Dwellings Q4 2021 (%)

Before 1919: **19.2**
1919 to 1960: **22.5**
1961 to 2000: **35.5**
2001 or later: **22.8**

Building Energy Rating of Vacant Dwellings Q4 2021 (%)

A **4.0**
B **6.5**
C **20.2**
D-E **34.0**
F-G **35.3**

Source:
https://www.cso.ie/en/releasesandpublications/fp/fp-viec/vacantdwellingindicatorsbasedonmeteredelectricityconsumption2021/

www.cso.ie

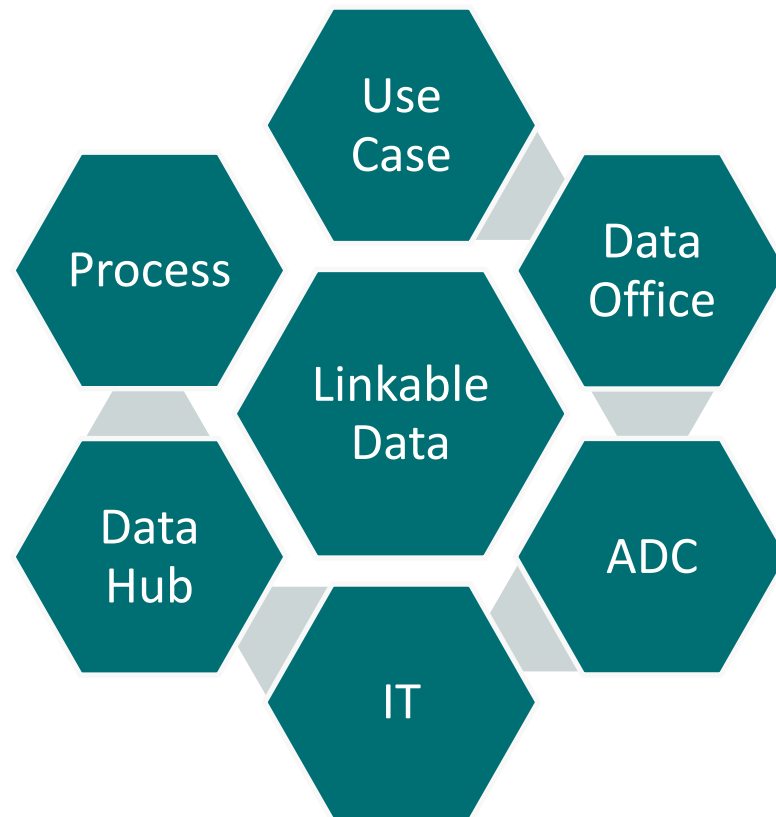# Summary

There are many actors involved.

Developing a process is itself a process.

It will take longer than you think.

Use Case

Data Office

Process

Linkable Data

Data Hub

ADC

IT

Any Questions?

We will take them at the end of the webinar