07/12/2023

A solera for data processing

The promised land is a journey not a destination (Version 1.0)

Gordon Guthrie
RESEARCH FELLOW AT SCOTTISH GOVERNMENT

BIUS WORKING PAPER NO 6

(THIS DOCUMENT DOES NOT REFLECT THE VIEWS OF SCOTTISH GOVERNMENT)

Table of Contents

1 Introduction

1.1 What is a solera and why do we need one?

A solera is a stack of barrels that are used to mature sherry. Each year a new batch of raw wine is made. The same amount of sherry is drawn from the bottom barrel in the stack and bottled. That barrel is topped up from the barrel directly above it, and up and up until there is space at the topmost barrel – which the new wine goes in. The wine is blended and gradually matured is it is taken down the stack.

The way in which the state specifies data and the powers to hold and share data is primitive and in bad need of reform. Individual data sets are unsharable at the moment by reason of law, of data quality, of lack of metadata, of lack of technical data sharing implementations, or a mixture of these reasons.

The state has a small amount of high quality, mature data and the challenge is find a way to take our raw data and slowly step-wise mature it. This document proposes a solera to do that. The top barrel contains raw (or dirty) data which might be useful for certain purposes but brings risk and contamination. This document proposes a set of intermediate steps, barrels in the solera, in which the data can be matured. Each step requires the dataset to have been fixed in part legally, for quality, metadata and technical implementation. And as certain steps are taken the data set moves down the solera until finally it is <drinkable>.

Fixing data is at the heart of joined up government. And the fixing of it will have to be incremental because of the scale of the problem with all its entanglements. Without an incremental strategy for moving data sets up the maturity curve/down the solera, whilst still using them, joined up government will fail.

1.2 Who are you?

This is quite a technical paper, so you are a technical or data specialist with an interest in open data or a parliamentary drafter interested in data law reform[1].

1.3 Why should you read this?

The improvement of state data will require a complex implementation plan – this is not that plan, but it outlines an architecture for that plan, a path of attack on a difficult problem.

[1] see Working Paper 5 – ***Law reform for data***

2 The Blus Project

This is Working Paper No 5 of *Blus - Basic Law-Making For Legislative Computer Systems* which is a research project looking systemically at how the state creates the digital systems underpinning its services.

Working papers are being released gradually for comment:
Working Paper X – ***The heart of the beast*** (published)
Working Paper 0 – ***The locus of change*** (forthcoming)
Working Paper 1 – ***Data and the rule of law*** (published)
Working Paper 2 – ***Rules as code*** (published)
Working Paper 3 – ***The Lego state*** (published)
Working Paper 4 – ***The remixable state*** (published)
Working Paper 5 – ***Law reform for data*** (forthcoming)
Working Paper 6 – ***A solera for data cleansing*** (this document)
Working Paper 7 – ***Experimental digital legislative processes*** (forthcoming)
Working Paper 8 – ***An Enabling Act*** (published)
Working Paper 9 – ***Reading legislation with a non-functional eye*** (forthcoming)
Working Paper 10 – ***Immediate Hygienic Measures*** (published)
Working Paper 11 – ***Jeff Bezos' Memo for Government*** (published)

Blus working papers are designed to stimulate discussion about key elements of the relationship of the state to digital systems and their delivery. Your feedback, input, and particularly criticisms of this paper are most welcome. Feel free to distribute it however you wish.

Working papers are published via the *Digital Policy* SubStack.

Author/contact: gordon.guthrie@gov.scot or subscribe to Digital Policy | Gordon Guthrie | Substack[2]

The author is an independent Research Fellow at Scottish Government under the First Minister's Digital Fellowship programme. The views of this paper do not represent the views of Scottish Government.

[2] https://digitalpolicy.substack.com/

3 A target data pipeline

3.1 Overview

This paper focuses on non-person data – primarily place data. At its core is a desire to link up data based on geographical tags (Unique Property Reference Numbers – UPRNs[3] – These provide higher resolution of property data than postcode/housenumbers). Data about people – the other major category of critical data for joined up government requires different treatment and is considered more directly in Working Paper No 5 – ***Law reform for data***.

Everyone who have used Google Maps once is already familiar with joined up geographical data. The Scottish Government produces lots of geographical data – but the task of coding it and joining it up is left to the citizen, private or 3rd sector. This imposes a time tax which inhibits use of the rich data we have.

Making that data pre-joined up would lower the entry costs for many commercial sectors dramatically, make geographical statistical data available to public and private sector alike at a more fine-grained resolution. This would lead to better decision making and be a significant contribution to creating a unified market in geographical information.

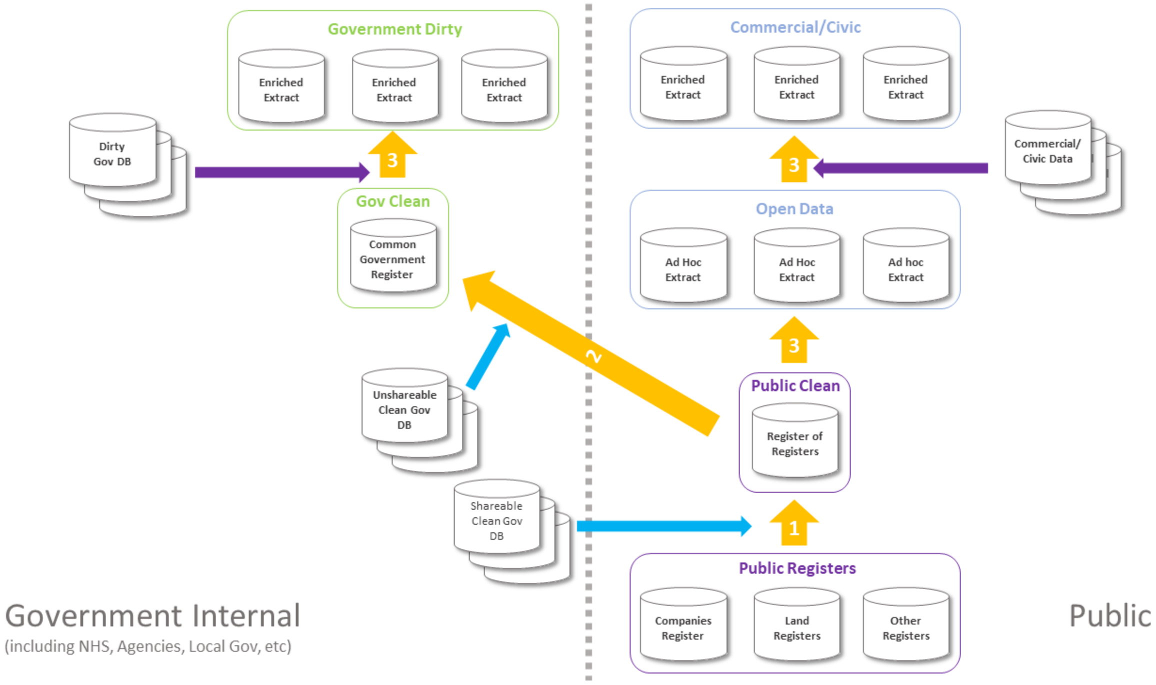
I have sketched out a data pipeline to provide context.

Sherry is aged in a system called the solera. There is a stack of barrels. Sherry is always bottled from the bottom of the stack. Each bottle extracted leaves capacity for a top-up, and each barrel is topped up with sherry from the barrel above. This capacity travels backwards up the solera and the new year's wine can be added into the top most barrel.

This pipeline is designed as a data solera – a system that keeps dirty data from clean and decouples the work of maturing datasets. Individual datasets can be matured and promoted independently. Maturation covers data cleanliness, proper time handling, technical implementation, legal permissions and so on. Different data systems will have different time schedules based on legislative slots, software upgrade cycles, clashing delivery schedules and so on.

It tries to group data logically and subsequent sections will step through it and make relevant observations.

[3] See https://www.ons.gov.uk/methodology/geography/geographicalproducts/nationalstatisticsaddressproducts for more details



The pipeline design separates geographical data into two domains – public and government internal (also shareable/unshareable). This is possibly a simplification. There are additional degrees of sensitivity (commercial sensitive, personally sensitive) that will need to be taken into account.

It splits data in to clean and dirty. Essentially all data that is not clean is dirty – so it is worth sketching out what clean means here.

Clean data MUST meet the following (not complete) criteria:

Criteria	Description
Maintained	Someone is charged with ensuring that the data is kept up to date, curated, and maintained against public data standards. In the case of Registers – this obligation is placed on the registered person or organisation to self-maintain on pain of sanction. Government systems that are maintained need to have a mandatory obligation placed on them, might be statutory, might be guidance – my recommendation would be tertiary legislation from a transformed, statutory Digital Assurance Office[4].
Timeous	The data MUST be correctly structured for handling time, statuses and attributes (created, in registration, suspended, closed down) are all time marked, historical data isn't deleted but marked, etc, etc

[4] I have a briefing paper for a consultation on an Enabling Act for digital which puts this in wider context. https://scotsconnect-my.sharepoint.com/:w:/r/personal/gordon_guthrie_gov_scot/Documents/Digital%20Fellowship/Insights/Enabling%20Act/Enabling%20Act.docx?d=w3af6067a3aee4e37ae2ad49223ea0d96&csf=1&web=1&e=xQGZDg

Criteria	Description
Immutable	Data structures exported MUST be ledger-based (CR) and not CRUD based[5] (the internal structure MAY be CRUD but the Deltas MUST be transcribed to CR/Ledgers).
Dump/Deltas	Data from a clean database MUST be available as a starting Dump and on-going Deltas – the goal is real-time update flows.
Keyed	The data MUST be supplied with appropriate keys (UPRNs, USRNs, etc). This is to ensure that import to a graph database MUST NOT require AI or fuzzy matching
Shareable (public or gov domain)	Clean data is either public or pan-state shareable – the necessity to maintain RBAC (Role Based Access Control) on a departmental or agency level make it administratively dirty.
Documented	The data has meta-data of the appropriate quality – the data must not just be findable and (legally) usable but able to be used.

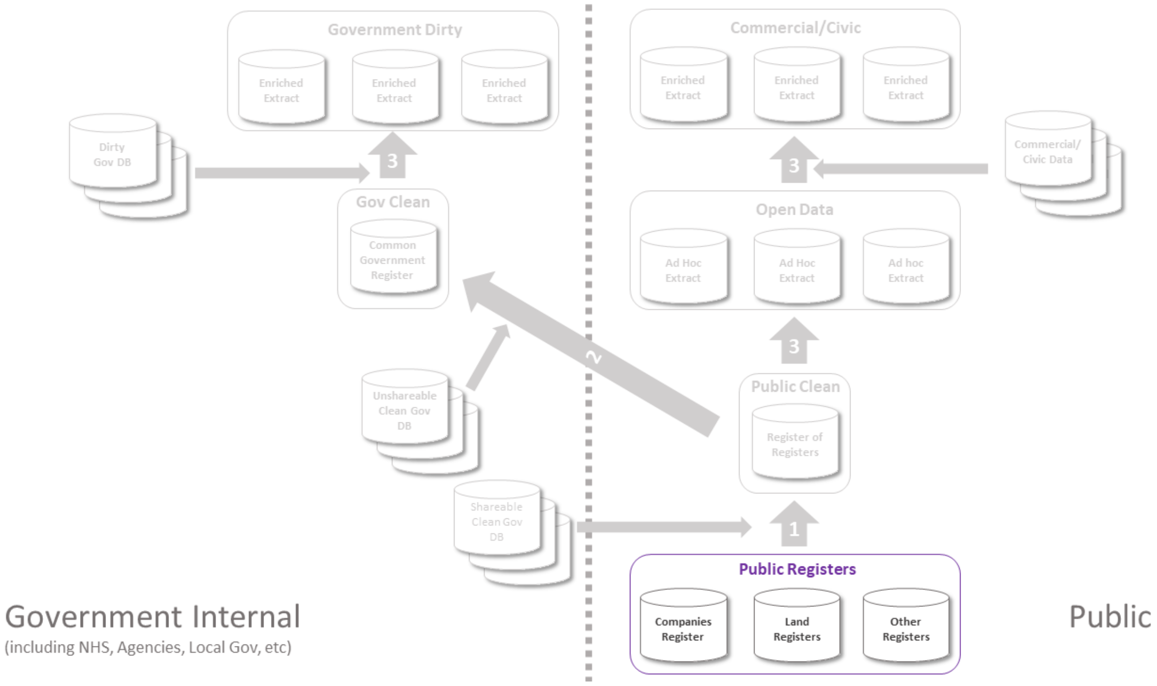
Each step of this pipeline will now be discussed:

- Public Registers
- Public Clean – the register of registers
- Government Clean
- Government Dirty
- Open Data and Commercial/Civic registers

Each stage can ingest data with different characteristics.

3.2 Public Registers

[5] See Working Paper No 1 – ***Data and the rule of law***. It draws on an extensive review of Universal Credit by the Child Poverty Action Group. It focusses on directly judiciable administrative data used to make decisions, as opposed to analytical data used to inform policy. As such it is not directly relevant, but it does explore civil servant's duty of care with respect of decision making. The forthcoming Working Paper No 5 – ***Law reform for data*** explores the issue in more detail.



There have been at least 2 attempts to put the various land registers onto a single GIS/Land Information System.

The respected former Green MSP and Land Expert Andy Wightman was commissioned by the David Hume Institute to write a report[6] this year on making the 3rd attempt happen.

The report is well worth reading as it summarises the history of ScotlandLIS dating back to the 1990s, through the 2007 creation of Unifi Scotland as the delivery vehicle and 2015 commitment of the then DFM John Swinney to deliver it. Registers of Scotland produced a proposal[7] which was never implemented.

The David Hume report lays down the following principles which is says are vital to success:

1. *There needs to be a firm agreement and commitment to deliver ScotLIS by Scottish Government and the wider public sector. Ministers have a key leadership role here.*
2. *Necessary policy and legislative changes to permit the development of ScotLIS need to be agreed in principle.*
3. *Agreed protocols on data, access, technical design and data use need to be developed.*
4. *There needs to be a suitable governance framework in order to direct and monitor development of ScotLIS with agreed timescales, milestones and final delivery.*
5. *Any necessary finance needs to be in place.*

[6] https://davidhumeinstitute.org/latest-news/2023/3/6/press-release-siloed-land-information-is-holding-back-scotland

[7] It is perhaps indicative that this proposal is not to be found (or at least not by me) on public government sites but only on Andy Wightman's personal archive http://www.andywightman.com/docs/Digital-land-and-property-information-system-report-July-2015.pdf

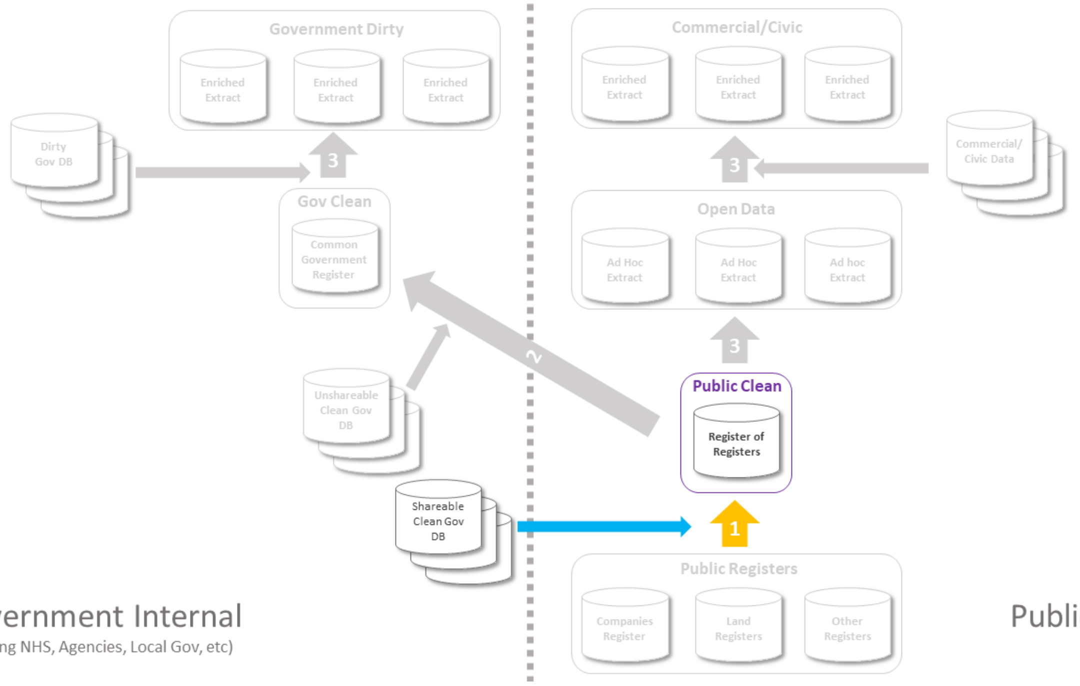
Andy Wightman was interviewed as part of my research project Blus – and most of the interview was spent on this issue. One of the core reasons for continuous failure to deliver new registers in a joined up manner was the absence of a <brain> in the Scottish Public Sector with all the requisite capabilities to hold and drive strategic work in this sector (finance/resource, technical chops, access to the legislative timetable, political/administrative authority across organisations and so on).

Currently the land registers are a mixture of:

- GIS data attached to maps
- Simple tabular data
- Ad-hoc data (like PDFs) with some searchability

Harmonisation of the land-based registers is a foundational task of this entire proposal, without there is no possibility of it working.

3.3 Public Clean



In this view the Public Clean step in the pipeline is a database accessible by GraphQL, run and maintained on a statutory basis to which the public has a defined right of access.

The choice of graph database and front-end to use it MUST be an operational decision – but the prejudice SHOULD be towards an open-source implementation.

There are some caveats – not all the existing legislation-based registers are suitable to be added to it in their current form (data might not be maintained well enough).

In an ideal world these registers would enable a point-in-time data dump followed by (batchable) deltas on change to allow the Public Clean to be updated on an appropriate cycle

– the expectation might be to start with monthly releases and move to dynamic/soft-realtime updating.

In the absence of deltas updates will need to be done on a read-and-match basis – and as many of the registers have an income stream funding model this is a potential killer.

Fixing the pricing/obligation to provide updates for existing registers won't be a quick process and might require legislative changes[8]. Some of the registers (eg Companies House) are not devolved and putting the screws on them would involve pushing legislation to Westminster via the Joint Ministerial Committee structure (good luck with that[9]!).

Appropriate oversight for the Public Clean MUST be established. It SHOULD at least rhyme with standard internet governance models like the Apache Foundation[10]. It MUST have a route to bring forward appropriate law reform to make it happen.

The organisation model for managing non-functional requirements is outlined elsewhere in Working Paper X – ***The heart of the beast*** and Working Paper 0 – ***The locus of change***.

This work is also a prime candidate for remedial work to address its funding model and so on via an Enabling Act – see Working Paper No 8 – ***An Enabling Act***.

The discussion in Working Paper No 11 – ***Jeff Bezos's API Mandate but for Government*** is also relevant.

The oversight body MUST have ownership and responsibility for all the open source artefacts to be developed to implement the pipeline (at this and higher stages in the solera)

Because all the data in the Public Clean is open data, there MUST be an ability to filter, query and extract datasets from it for use the Open Data and Commercial/Civic parts of the pipeline. That capability (expressed as software) MUST be appropriately open sourced so that it can be reused within government and by commercial suppliers to build augmented (and possibly dirty) data on top of the clean register data.

The extract process from Public Clean MUST itself conform to the clean data standards so that the pipelines built on it can benefit from continuous update.

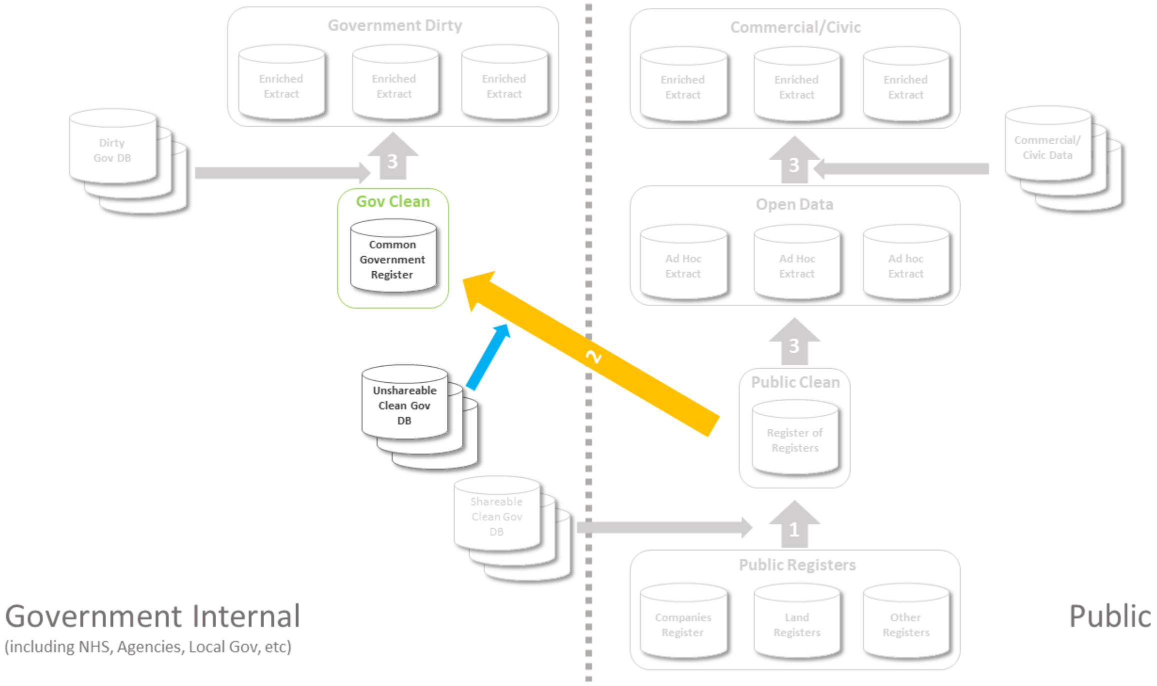
Once the Public Clean databases start being fed with soft-real time deltas instead of batched updates they in turn will be able to offer soft-real time deltas and commercial providers can build API backed services on the open data – building a commercial landscape of service providers backed by excellent and comprehensive pre-joined up and excellent government data. This will dramatically reduce the barriers to entry for data-driven businesses in the Scottish economy – all that manual cleaning and joining.

[8] The private sector must seduce but the parliament can compel….

[9] Caveat Lector: I know the square root of bugger all about the JMC procedures – but anecdata suggests its not going great, if it is at all existent, at Minister/Minister. But work at civil-servant/civil-servant might be going swimmingly [shrugs].

[10] https://apache.org/foundation/governance/

3.4 Government Clean



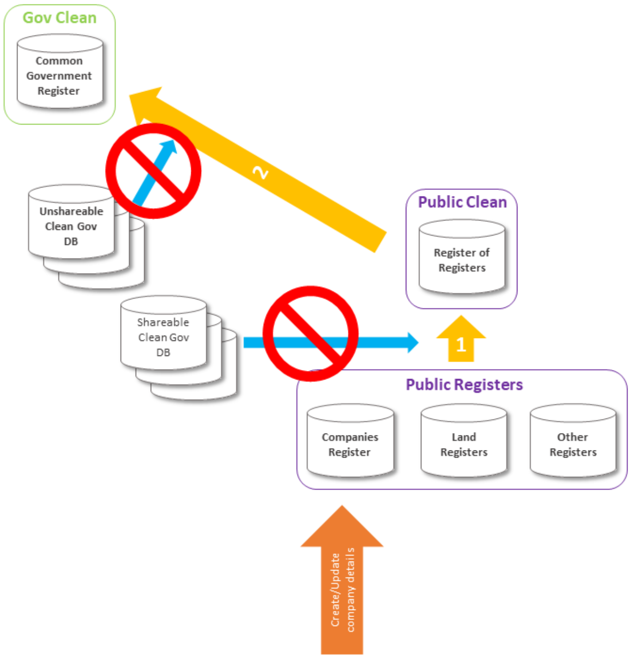
The Government Clean database takes Public Clean data and augments it with shareable, clean government data to create the pan-Government platform for internal use.

Unshareable in this instance means not to be shared with the public, but able to be shared over all state bodies (Scot Gov, NGAs, Agencies, Local Authorities, Health Boards). Data that is confidential to an agency department is defined as administratively dirty.

The platform for Government Clean and Public Clean SHOULD be the same, both at a database and query/front end level and it SHOULD run on the same logical platform and it SHOULD be managed by the same service provider. It perhaps goes without saying that the Public Clean and Government Clean systems will be running in different security zones and MUST be separate to the degree laid down by the appropriate security standards.

It MUST have the same data standards/acceptance criteria as the Public Clean.

It is critical to understand that this pipeline MUST be based on immutable enrichment and not data reconciliation. Enrichment is adding additional data which shares a common key (in property cases a UPRN) of the base date. Editing of data cannot be done in the pipeline – changes do not propagate back down. Data must be consistent. That means changes MUST flow from the data source up the pipeline:

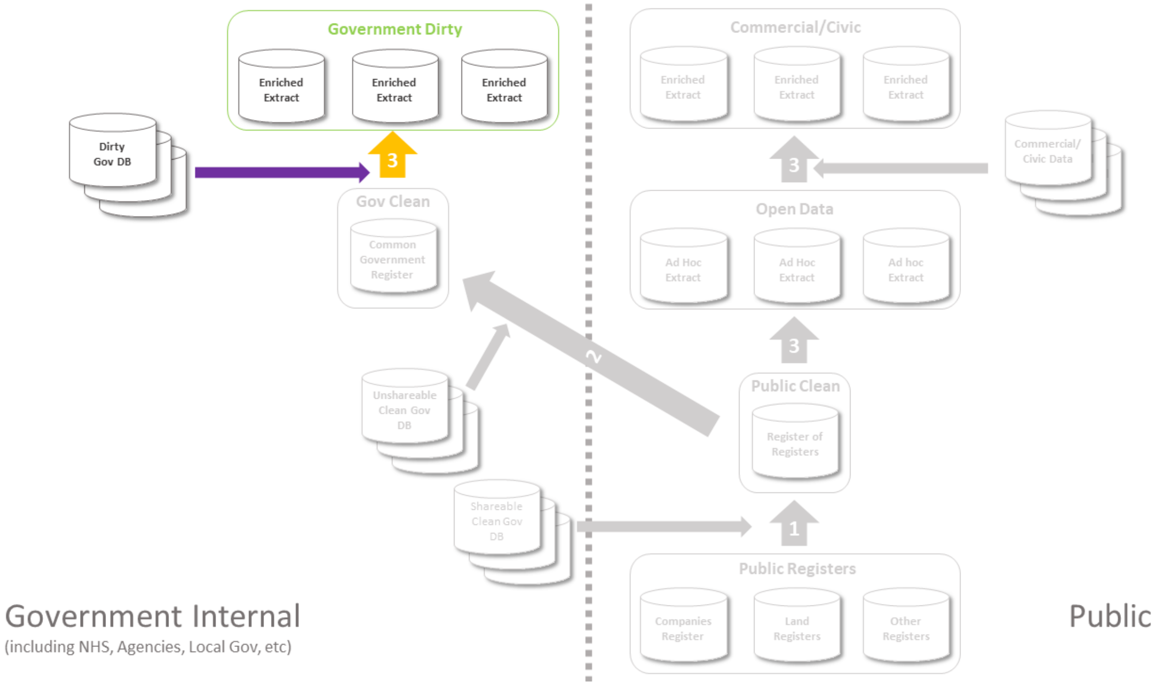


This is ***the critical point*** in the choice of a solera design for the pipeline. Allowing multiple points of update generates manual reconciliation and clean up. This can often appear reasonable, proportionate and contained in a pilot/prototype – but is a highway to hell and ends in a rotting data environment.

The data refresh rate at each stage is critical to the smooth operation of the entire pipeline – an NBR with a yearly base update from Companies House would be unacceptable. Ideally we want to get to near-realtime but the relevant extracts for augmentation would need to be checkpointed and hold-back as-yet-uncreated-upstream information. (An upstream data source might well know about new company creation before the changes flow upstream from Companies House – but the enrichment process MUST only have the ability add enrichment to existing data but NOT create new or missing instances.

Like Public Clean, the Government Clean data will enable the creation of API based data services for use across the entire public sector – with built in security for commercial-in-confidence information.

3.5 Government Dirty



It should be expected that most of the day to day work would be done at a Government Dirty level – at least initially – for analytical and not operational work.

In this world, the end user is wanting to do some analytical work and orders a snapshot of the dataset they require and it arrives locally for them to do as they see fit.

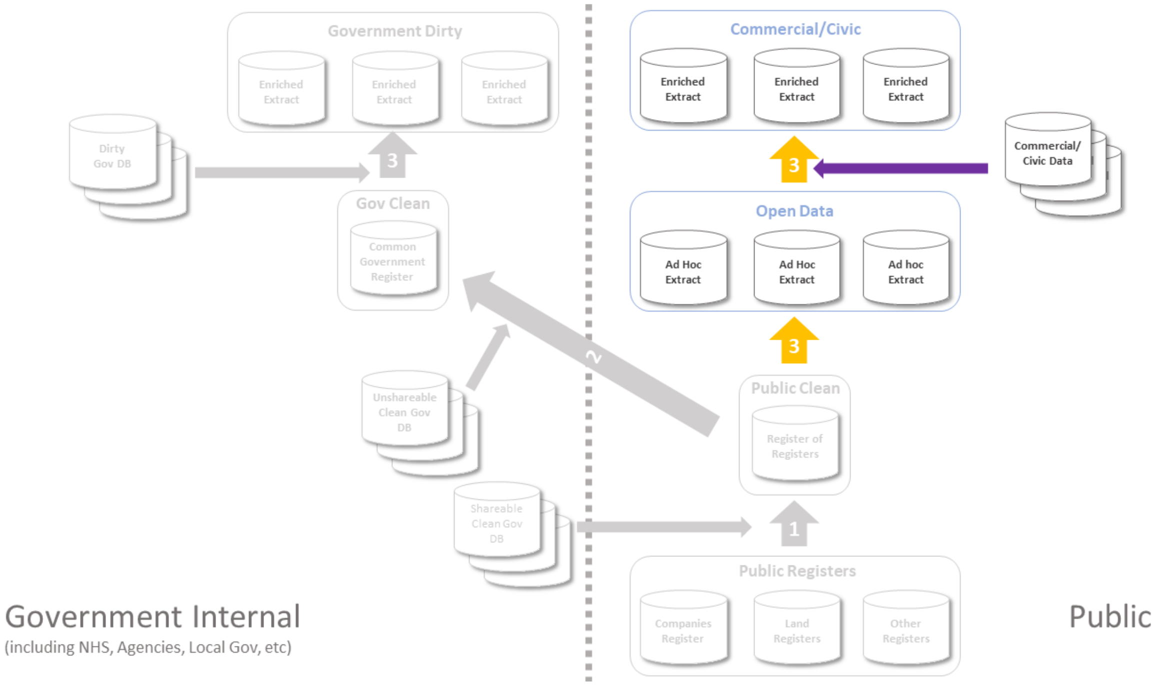
In this model:

- a user selects a filtered subset of the Gov Clean data
- starts up a containerised solution
  - pre-populated with the snapshot Government Clean subset
  - registered to receive updates from Government Clean
  - with the ability to apply those updates
- is supported in adding their own dirty data set
  - training
  - tools
  - hiring
  - support/help desk
- has power to invite/share within the government domain

The containerisation software, filter/select, transform/update software MUST be Open Source and under the ownership/control/governance of the body established for Public Clean. Commercial solutions MAY be purchased by Scottish Government to provide both graph DB and query/use tooling – but the containerisation pipeline MUST be constructed so as to allow them to be swapped out. The container solutions MUST be portable over cloud providers.

The deficiencies in this data model - the reason it is dirty – is that it can't be operationalised as a service-over-API. Getting data from Public Dirty to Public Clean and making the API environment richer is necessary for better government outputs.

3.6 Open Data and Commercial/Civic



The Open Data and Commercial Civic pipelines are not under the control of Scottish Government. They are run by civic or commercial organisations on commercial clouds or on on-prem as they see fit.

However to enable and support those communities to grow cheaply and effectively the entire software stack that is used in Public Registers -> Public Clean -> Government Clean -> Government Dirty should be managed as a single project under the oversight body (approximately rhyming with an Apache Foundation project in governance terms, with open standards, open road maps, open software, etc, etc, etc).

Needless to say, civil and commercial organisations are under no obligation to use the open source solutions – and steps should be taken to ensure that Scot Gov's choice of analytical software to run against the graph data does not constrain civic/commercial users.

It is good for the economy for public and third sector parties to build APIs and services (and possibly charge for them) by taking open data delivered in soft-realtime deltas, enriching it with clean data and exposing it. It is also good for the economy for public and third sector parties to do ad-hoc work by joining their own data, dirty or clean, privately to good clean, joined up public sector open data.

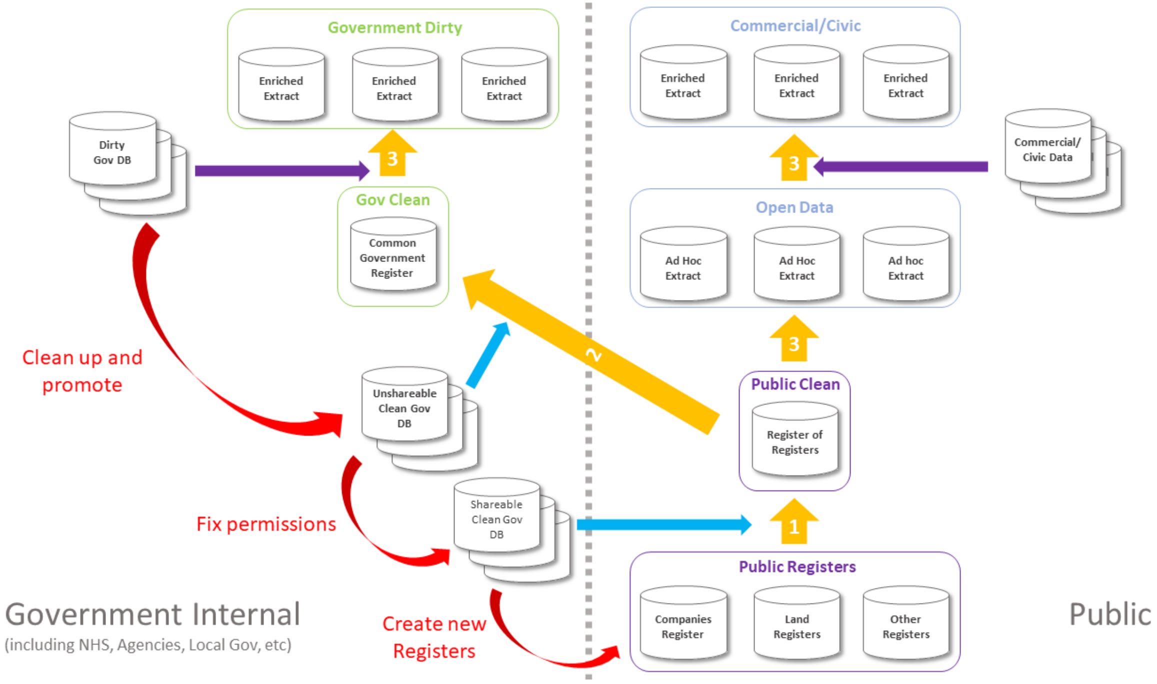
4 The Data Solera

In this section we show how the various data sets can be independently matured by moving them down the solera so they are injected into the customer pipeline at different stages.

In the early days the expectation is that most data sets are Dirty Gov. Gradually they can be fixed (legally, technically, in terms of documentation, etc, etc) and moved down a step.

In the final state the majority of government data is Shareable Clean Gov data and the number of base registers has increased substantially

By implementing a data solera we can decouple the necessary aging and improvement processes for each data supplier from each other:



Failure to decouple improvement programmes across Scottish Government will ensure that the entire project fails (like the long history of ScotlandLIS initiatives). A solera MUST be built.

In the world of Sherry the customer is interested in the lowest barrel, the finest sherry, but in the data solera the end-user cases, the insight, the buy-in, come from the top barrel – the dirty data.

Lets look at the 3 steps of cleaning in a bit more detail. The Solera breaks down the 7 criteria

	Step	Notes
1	Clean up and promote	In this stage the data set is fixed up in regard of the following criteria: • Maintained • Immutable • Dump/Deltas • Keyed • Documented These are all attributes which fall within the control of a particular department responsible for a database – they can get on with fixing it. It also requires addressing the technical basis for: • Timeousness This stage fixes the ***means***.
2	Fix permissions	This addressed the legal basis for sharing: • Shareable (public or gov domain) Fixing this might require primary or secondary legislation – or getting citizen/client permissions This stage fixes the ***will***.
3	Create new registers	In this final stage responsibility for running a data service is transferred from a government department of body to the Registers of Scotland – a dedicated governmental body that has provided data-as-a-service in the most foundational way for 4 centuries now. Responsibility for data maintenance remains with the Registree: • Timeousness This stage ***institutionalises*** the provision of this data class.

We can move fastest at the dirty level (we don't need to retrofix institutions and legislation).

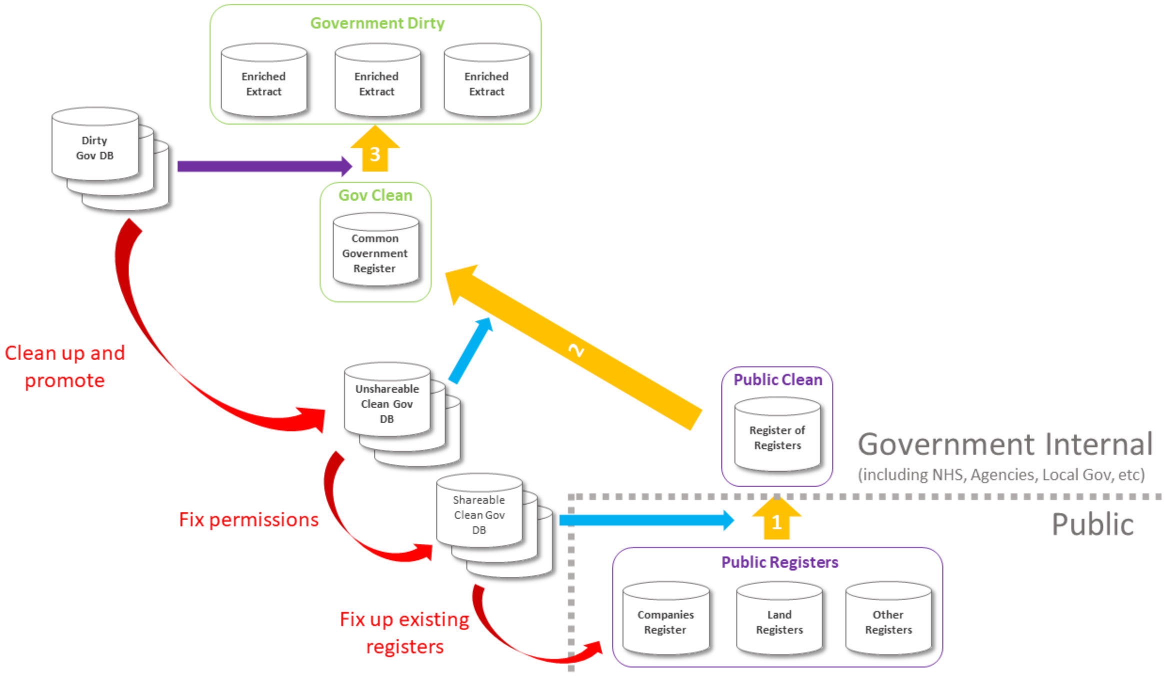
It is important to get end-users up and running in the Government Dirty as soon as possible. Delay in getting there will result in increasing pressure to relax the data standards criteria until the inevitable "*the Minister wants this to happen* (*just one time, just one time*)" happens and the pass is sold.

This is an instance of the gearbox problem with is discussed extensively in my blog series[11]:
Part 1 – we need a gearbox (blogs.gov.scot)
Part 2 – Frankenstein Bill (blogs.gov.scot)
Part 3 – technical pattern books (blogs.gov.scot)

[11] https://blogs.gov.scot/digital/2023/08/28/basic-law-making-for-legislative-computer-systems-part-1/
https://blogs.gov.scot/digital/2023/09/04/basic-law-making-for-legislative-computer-systems-part-2/
https://blogs.gov.scot/digital/2023/09/11/basic-law-making-for-legislative-computer-systems-part-3/
https://blogs.gov.scot/digital/2023/09/25/basic-law-making-for-legislative-computer-systems-part-4/
https://blogs.gov.scot/digital/2023/10/02/basic-law-making-for-legislative-computer-systems-part-5/

Part 4 – a legislative architecture (blogs.gov.scot)
Part 5 – testing the proposals (blogs.gov.scot)

In order to decouple the project from the legislative engine the solera MUST be built within Scottish Government and then transitioned when the appropriate quality mechanisms/legislative changes are in place:



The implication is that the Public Clean environment will first be deployed inside Scottish Government before being made public.

This enables us to continue learning from end-users throughout the process.

5 Conclusion

Working Paper No 5 – ***Law reform for data*** focusses extensively on the ***will*** to do joined up government.

This paper is a companion piece which focusses on the ***means*** to do it – and it builds on Working Paper No 1 – ***Data and the rule of law***.

The institutional basis for the creation of a solera and the proposing, designing and scheduling of the work (which will take not a couple of months or a couple of years, but rather be an continuous on-going project) is dealt with in a variety of working papers:
Working Paper X – ***The heart of the beast***
Working Paper 0 – ***The locus of change***

Some of the technical elements of it are described in Working Paper No 8 – ***An Enabling Act***.