IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Data sharing using a global data registry: on a place to discover global structured time series, macro and micro data[1]

Matt Nelson and Glenn Tice,
Bank for International Settlements

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Data sharing using a global data registry
On a place to discover global structured time series, macro and micro data

Matt Nelson and Glenn Tice

Abstract

The increasing availability of data and the importance of evidence-based decision making have led to an urgent need for better data sharing practices. A 'global data registry' may provide a practical solution to enhance data discoverability, accessibility, and interoperability. This paper discusses the specific needs of data scientists and presents a potential solution for a global registry of structured statistical, parametric and unit record datasets using the Statistical Data and Metadata Exchange (SDMX) international standard ISO 17369:2013 and open-source software. Ultimately, we argue that a global registry for structured data can play a key role in enabling more effective use of datasets already available for policy making, research, and innovation by easing the burden of data acquisition on data scientists.

Keywords: data discovery, structured data, data science, SDMX

JEL classification: C80

Contents

1. Navigating the labyrinth: Challenges in accessing and utilising public data for data scientists

Data science, a discipline fuelled by the analysis and interpretation of data, relies heavily on access to diverse and comprehensive datasets for modelling, analysis, and innovation. While there exists a vast repository of structured public time series, macro, and micro data, sourced from international, regional, and national institutions, data scientists encounter formidable challenges in harnessing this information for their work.

Primarily, the obstacle lies in the fragmented nature of these datasets, scattered across disparate data portals hosted by different organisations. The data ecosystem lacks a unified platform or central repository, compelling users to navigate through numerous websites, each with its interface, to seek and retrieve relevant data. This dispersion and lack of a centralised resource significantly hampers the efficiency and productivity of data scientists, consuming valuable time and effort in the search for appropriate datasets.

Compounding this issue is the inadequacy of structural metadata accompanying these datasets. Often, the metadata—information describing the dataset's structure, attributes, and relationships—remains insufficient or poorly defined. Ambiguous or missing definitions of variables within the datasets lead to challenges in interpretation and utilisation. This ambiguity not only impedes the understanding of the dataset but also complicates the integration of these datasets into analytical models, hindering the accuracy and reliability of data-driven insights.

Furthermore, the absence of a standardised format or clear taxonomy across these datasets exacerbates the problem. Datasets from different sources may adopt varying conventions, making it challenging for data scientists to harmonise disparate data for meaningful analysis. The lack of a common data language or standardisation amplifies the complexities in data integration and hinders the seamless utilisation of these resources.

Consequently, the absence of a centralised repository, coupled with scattered data across multiple portals, inadequate structural metadata, and a lack of standardised formats, poses a significant barrier for data scientists. This fragmented landscape not only impedes the discovery and accessibility of data but also hampers the efficiency and effectiveness of data analysis, limiting the potential for innovation and insights derived from these public resources.

In essence, the critical need for a unified, easily accessible repository of structured public data, equipped with comprehensive metadata and standardised formats, becomes apparent. Such a platform would streamline the process of discovering and accessing global data, empowering data scientists to focus more on analysis and innovation rather than grappling with the challenges of data collection and interpretation across multifarious sources.

2. Existing landscape of public data portals: A mosaic of regional and domain-specific resources

The data landscape teems with a multitude of data portals, serving as repositories for a wealth of information curated by various entities ranging from international organisations to national governments. These portals, designed to offer access to a diverse array of data, play a pivotal role in facilitating data-driven research, analysis, and decision-making. However, their effectiveness in meeting the structured data needs of data scientists remains somewhat fragmented.

One category of these portals comprises open data repositories, epitomised by stalwarts such as the EU Open Data Portal and the US data.gov, among numerous other national-level services worldwide. These platforms boast extensive repositories spanning a broad spectrum of data types, encompassing everything from highly unstructured data like images and document scans to geospatial information and semi-structured datasets. While these portals offer a vast array of information, they encounter a challenge when it comes to structured data. Despite hosting structured tabular datasets, the metadata outlining the structure of these datasets often remains limited. For data scientists reliant on structured data as their primary working medium, this scarcity of comprehensive metadata poses a hurdle, impeding their ability to interpret data reliably and compare it with other datasets—an essential aspect of rigorous analysis and cross-comparison.

Another category comprises statistical data portals, prominently represented by institutions like the European Central Bank (ECB), Bank for International Settlements (BIS), Organisation for Economic Co-operation and Development (OECD), along with individual National Central Banks (NCBs) and National Statistical Offices (NSOs). These portals excel in providing exclusively structured data accompanied by robust metadata. However, their drawback lies in often being regional or domain specific. While they offer high-quality structured datasets with comprehensive metadata, their focus on specific regions or domains limits the breadth of data accessible to data scientists with diverse research interests or global-scale analyses.

Amidst these established portals, emerging initiatives like Google Data Commons strive to furnish a global data catalogue. However, these initiatives, while promising, are still in their nascent stages, grappling with immaturity in terms of comprehensiveness, usability, and accessibility.
The coexistence of these diverse data portals underscores the rich tapestry of available resources. Yet, their regional, domain-specific nature or their immaturity in catering to structured data needs presents challenges for data scientists seeking comprehensive, globally relevant structured datasets.

In essence, while these data portals serve as valuable repositories of information, the lack of a singular, all-encompassing platform for structured data presents a significant gap. Data scientists, reliant on structured data for their analyses, face the arduous task of navigating this mosaic of portals, often finding themselves constrained by the regional or domain-specific limitations, and grappling with inadequate metadata crucial for robust data interpretation and cross-comparison.

3. Data Scientists' needs for structured data discoverability and access

Delving deeper into the quest for structured data discoverability on a global scale unveils a series of critical prerequisites essential for data scientists. Foremost among these requirements is the necessity for a comprehensive view of global dataset availability. In a digital landscape dominated by ubiquitous search engines like Google and Bing, the rapid discoverability of general web content transcends geographical boundaries. However, attempts to reliably discover datasets in a similar manner remains challenging. While the finer granularity of availability information at series and data point levels is invaluable, a fundamental requirement emerges—a global index of availability at the dataset level. This foundational perspective offers substantial benefits for data scientists and serves as an initial stepping stone for the development of more sophisticated and refined discovery services.

Critical to the success of any discovery service is the establishment of a robust metadata framework. Such a framework must comprehensively describe the structure of each dataset (structural metadata) while also exposing referential metadata provided by the data product provider. For data scientists

seeking datasets pertinent to their applications, this metadata serves as a compass, guiding them towards candidate datasets and aiding in evaluating their suitability. Clarity in metadata is paramount; data scientists require precise definitions of the concepts encapsulated within the dataset variables. A dataset merely labelling an independent variable as 'country' without accompanying metadata diminishes its utility compared to a dataset explicitly defining the variable's conformity to international standards like the ISO 3166 country classification or providing a domain-specific taxonomy supplemented by geographic shape information and outlining precise regional boundaries.

Once datasets are discovered, ease of retrieval becomes paramount. That involves ensuring datasets are not only available in machine-consumable formats but also conform to well-understood schemas. Classically format means CSV, XML, JSON and Excel, but as the scale of granular data increases others such as Apache Parquet become increasingly important. Compatibility with commonly available connectors or data parsers streamlines the process, negating the need for bespoke code to transform incoming data into usable structures. Additionally, the transfer mechanism plays a pivotal role; facilitating easy movement of data from source to the point of computation and analysis with minimal overhead. Data published in formats like PDF or accessible only through a complex download procedure, while not inherently unusable, necessitate unwelcome pre-processing efforts before data scientists can delve into the core aspects of data analysis, impeding efficiency and agility in the data science workflow.

In essence, the efficacy of a global data discovery service for data scientists hinges on the provision of a comprehensive dataset view, robust metadata frameworks, machine-consumable formats, and seamless data transfer mechanisms—factors crucial in facilitating efficient, precise, and unencumbered access to structured datasets across global domains.

4. Single access points

As its name suggests, the planned European Single Access Point solves the problems of data discoverability and accessibility in the European public financial and sustainability domain by consolidating the data into a centralised warehouse and providing a single search and retrieval interface.

For the data consumer, the single access point approach provides a convenient abstraction hiding the complexity of querying possibly multiple heterogeneous data sources and even allowing analytics to be taken to the data, for instance using SQL to execute joins and aggregations within the data warehouse.

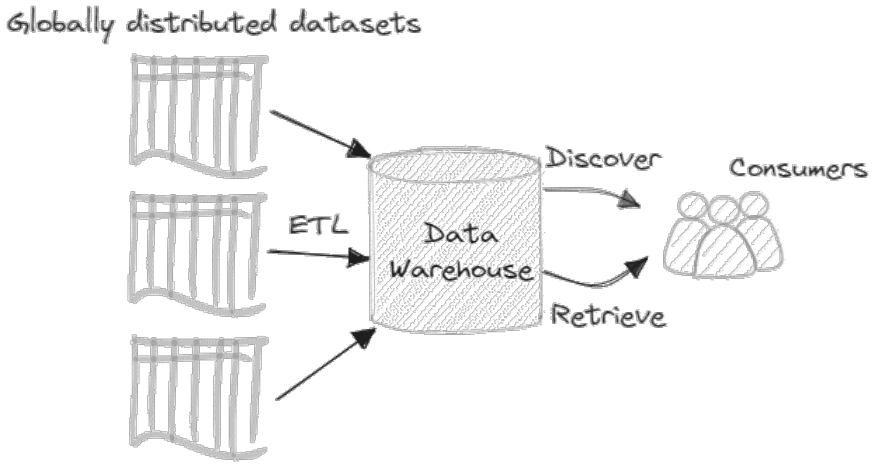


Figure 1 Single access point with data consolidated into a central warehouse.

Federation provides a variation on the approach where the data remains at source but is retrieved, integrated, transformed and transmitted to the data consumer via the single access point on demand.

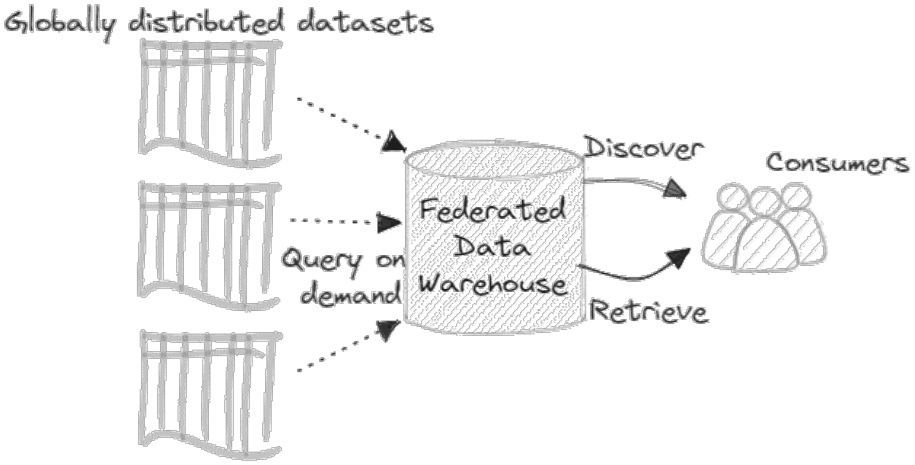


Figure 2 Single access point with a federated data warehouse.

For the single access point provider however, such services are often complex and costly to both build and operate particularly in term of storage and compute demand which invariably grows over time.

Establishing the data collection framework is also a significant challenge. The single access point approach is feasible where robust data collection pipelines are already in place, or at least data exchange protocols are in place limiting the scope of the problem to one of a technical implementation exercise.

5. Data registry: Paving the way for an evolved global data portal

As we have seen, single access points can be justified where the value of the data products and the benefits for data consumers outweigh what can often be the considerable cost of building and operating such services.

In the pursuit of a more streamlined and efficient global data ecosystem, the concept of a data registry emerges as a cost-effective alternative, reshaping the landscape of data access and utilisation. Unlike conventional data repositories that store the entirety of datasets, the data registry approach opts for a more agile and efficient methodology—centralising only the metadata detailing the structure of each dataset, along with pointers directing users to their actual locations.

This approach alleviates the burden of managing and handling the actual data, focusing instead on curating rich metadata that acts as a comprehensive catalogue of available datasets. Essentially, the registry functions as a compass, guiding data scientists to the precise locations of datasets without shouldering the weight of hosting the entire dataset repository.

Conceptually akin to existing services such as the EU Open Data Portal and similar data catalogues, this evolved data registry distinguishes itself by its emphasis on enriched metadata. It transcends mere pointers by providing sufficiently detailed and structured metadata, enabling effective discovery, evaluation, and interpretation of datasets. With comprehensive metadata elucidating the structure, attributes, and relationships within each dataset, data scientists gain the crucial insights necessary for confident and informed data utilisation.

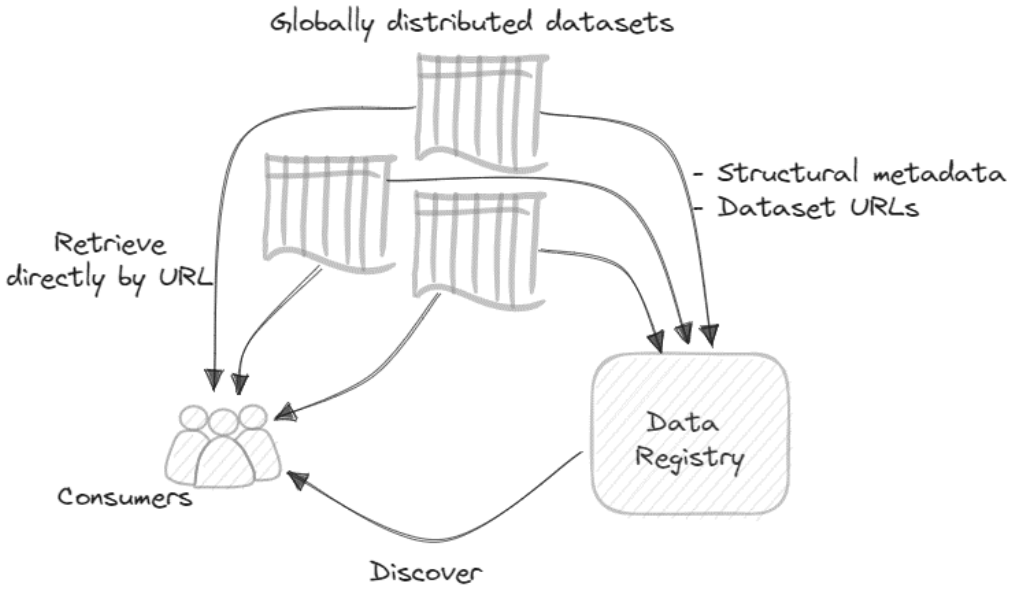


Figure 3 The data registry approach centralises only the data discovery and access metadata

Moreover, this model aligns harmoniously with the emerging paradigm of the Data Mesh—an approach geared towards organising and facilitating access to decentralised data across intricate data ecosystems. Two fundamental principles underpin the Data Mesh concept, resonating with the ethos of the data registry approach:

(a) Domain-Oriented Ownership: In this paradigm, each domain assumes responsibility for managing, curating, and granting access to its data. The data registry, by cataloguing metadata and pointers

rather than actual datasets, empowers individual domains to maintain control over their data while ensuring its discoverability on a global scale.

(b) Data Products as Standardised Interfaces: Embracing the philosophy of treating data as a product, the data registry fosters the creation of standardised interfaces—data products—by dedicated data teams within each domain. These interfaces provide uniform and standardised methods for accessing and consuming data, enhancing interoperability and simplifying the process of utilising diverse datasets across domains.

In essence, the data registry approach embodies a paradigm shift—a departure from the conventional handling of data repositories towards a leaner, metadata-focused model. By cataloguing rich metadata and employing pointers, it not only streamlines data access but also aligns seamlessly with the progressive principles of the Data Mesh, fostering domain-oriented ownership and standardised interfaces for enhanced data utilisation across complex and decentralised data ecosystems.

In summary, a data registry provides a light-weight and agile alternative where the case for a centralised or federated data warehouse and single access point cannot be argued.

6. SDMX 3.0: A standard metadata solution for a global data registry

As already noted, the utility of a global registry for structured data is dependent on a robust metadata framework. Version 3.0 of the ISO 17369 SDMX metamodel is a good candidate in that it provides specific artefacts for (a) data modelling, (b) data discovery and, (c) data registrations – pointers in the form of URL links to the actual data.

Modifications to the SDMX metamodel in version 3.0 improved the ability to model micro-datasets by adding support for multiple measures and array values for measures and attributes. These improvements mean that unlike previous versions, SDMX 3.0 is flexible enough to model most tall and wide microdata, plus macro, star schema and time series. Essentially, any dataset that can be managed as a data frame, but with the backing of the rich structural metadata essential for effective discoverability and interpretability.

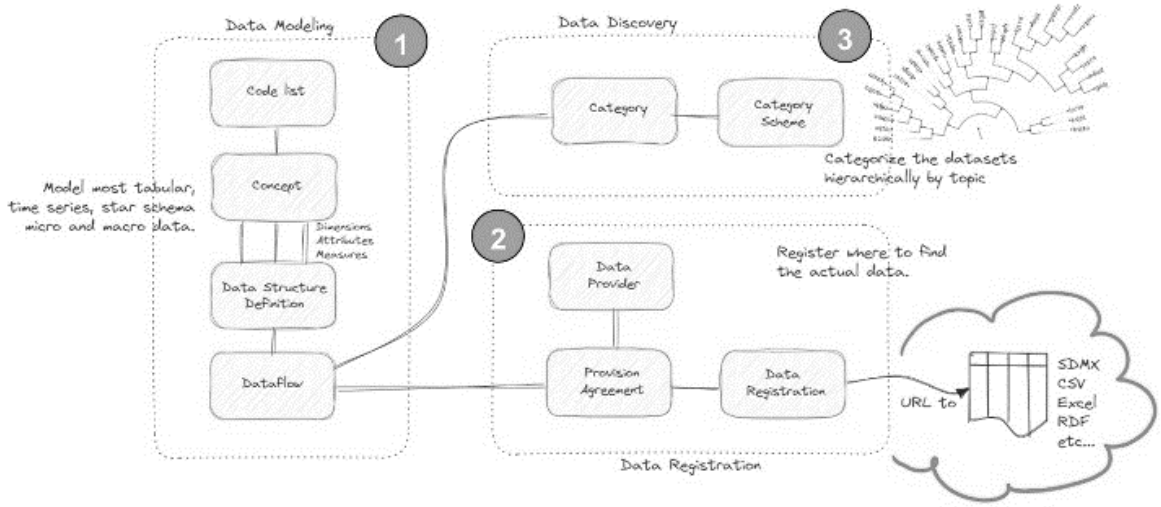


Figure 4 The SDMX 3.0 metamodel supports data modelling, discovery and registration metadata.

Examination of the SDMX 3.0 metamodel reveals several metadata artefacts specifically for data modelling. The four most important are the Code List, Concept, Data Structure Definition (DSD) and the Dataflow as illustrated in group (1) in Figure 4. An explanation of the precise purpose of these artefacts is beyond the scope of this paper, but the main principles are that the Dataflow is analogous to a fact table definition in a star schema data warehouse, and Code Lists define precise enumerations of values for discrete variables.

Data discovery is assisted by categorising datasets using hierarchical tree structures. In SDMX 3.0 the Category and Category Scheme illustrated in group (3) in Figure 4 support this purpose.

Finally, pointers to the actual data are managed as Data Registration artefacts (group (2), Figure 4) materialised as URLs. These are specifically decoupled from the Dataflow in the SDMX 3.0 metamodel to support the use case where multiple providers each contribute separate subsets towards a wider dataset. A common example being a dataset such as 'population' with global coverage where each country contributes observations for their own specific region.

For maximum accessibility, the URLs are expected to use the http:// or https:// protocols, but other more exotic protocols could also be offered, for instance s3:// referring to resources in the Amazon Web Services 'S3' object store.

7. A minimum viable implementation using open source software

While SDMX 3.0 provides a standard theoretical basis for a global data registry, software tooling is needed to make it a reality. For illustrative purposes, we present a reference solution based on the free and open-source Fusion Metadata Registry 11 (FMR) application which implements the SDMX 3.0 metamodel, critically including support for data registrations.

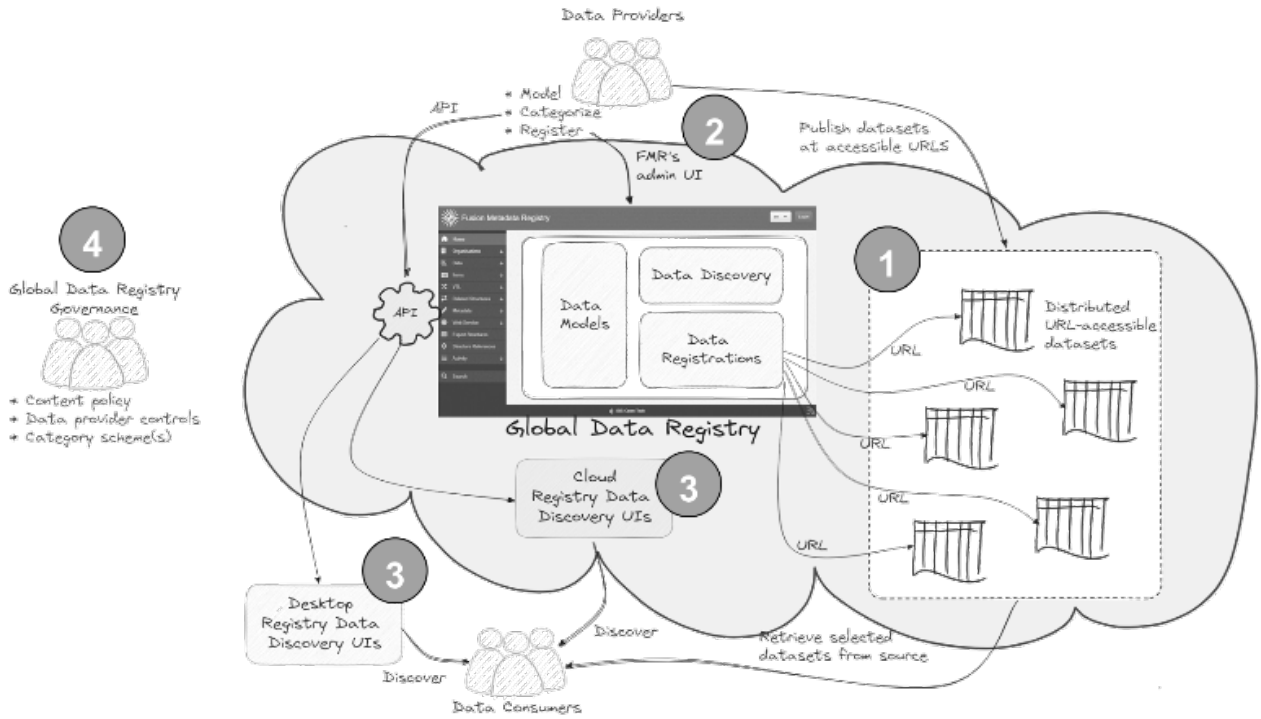


Figure 5 A minimum viable product using FMR 11

Structured datasets are first required, the only hard requirement being that they are individually resolvable by URLs. Next, the datasets need to be modelled, categorised and registered. Generally the data providers or the data product owners are best placed to undertake that task being those in possession of both the definitive structural and referential metadata. In practice the modelling task may best involve independent curators to help ensure consistency and maximise harmonisation of concepts within the global data registry, thus improving interoperability between datasets.

As previously noted, data scientists typically require both user and application programming interfaces. FMR's standardised SDMX REST API adequately supports the latter.

A cloud-hosted web user interface could be offered for simple use cases, although desktop applications perhaps implemented in Python could be conceived for more complex applications where data scientists need seamless integration of datasets into their workflow. Irrespective of the precise nature of the user interfaces, the key architectural principle is that all are loosely coupled to the central global data registry interfacing exclusively through its REST API.

Finally, a governance model is required to exercise control over the category tree(s) and other aspects of the registry including defining and enforcing content policy and administering the community of data providers.

8. Conclusion

While formal research is lacking on the subject, desktop analysis and anecdotal evidence indicates that data scientists would benefit from a place to discover global structured time series, macro and micro data. We observed that a large corpus of structured data is published by institutions globally, but the effort needed to find, retrieve and interpret them could be reduced.

The option of providing a 'single access point' provides attractive benefits for data scientists as the end users of such a discovery service, but the relatively high implementation and operating costs mean it is hard to make an investment case. The 'data registry' approach presents a more practical option in this respect as storage and compute, the main cost drivers for the single access point option, are contained and rise slowly as the service scales.

To test the feasibility, we examined a concrete reference implementation using the SDMX 3.0 international standard for the metadata framework and the Fusion Metadata Registry 11 software tool as the main data registry engine concluding that the key components for a minimum viable product exist.

Data sharing using a global data registry

On a place to discover global structured time series, macro and micro data

3rd IFC Workshop on Data Science in Central Banking

Glenn Tice & Matt Nelson (BIS MED IT)

The requirement for global structured data discoverability and access

The Problem

Public structured data is hard to find

- Data scientists need data
- Large global resource of structured data published by institutions worldwide

But

- No single place to discover what's available and how to get access to it

Existing Landscape

Many excellent data services already exist, but are variously regional, domain specific or not tuned for structured data

- EU open data portal
- US data.gov
- Institutional data portals
- European Single Access Portal (planned)
- Emerging initiatives (e.g. Google Data Commons)

The requirement for structured data discoverability and access

Global

Easily discover what datasets are available

Sufficient information to evaluate the suitability of a dataset

Confidently interpret and use datasets

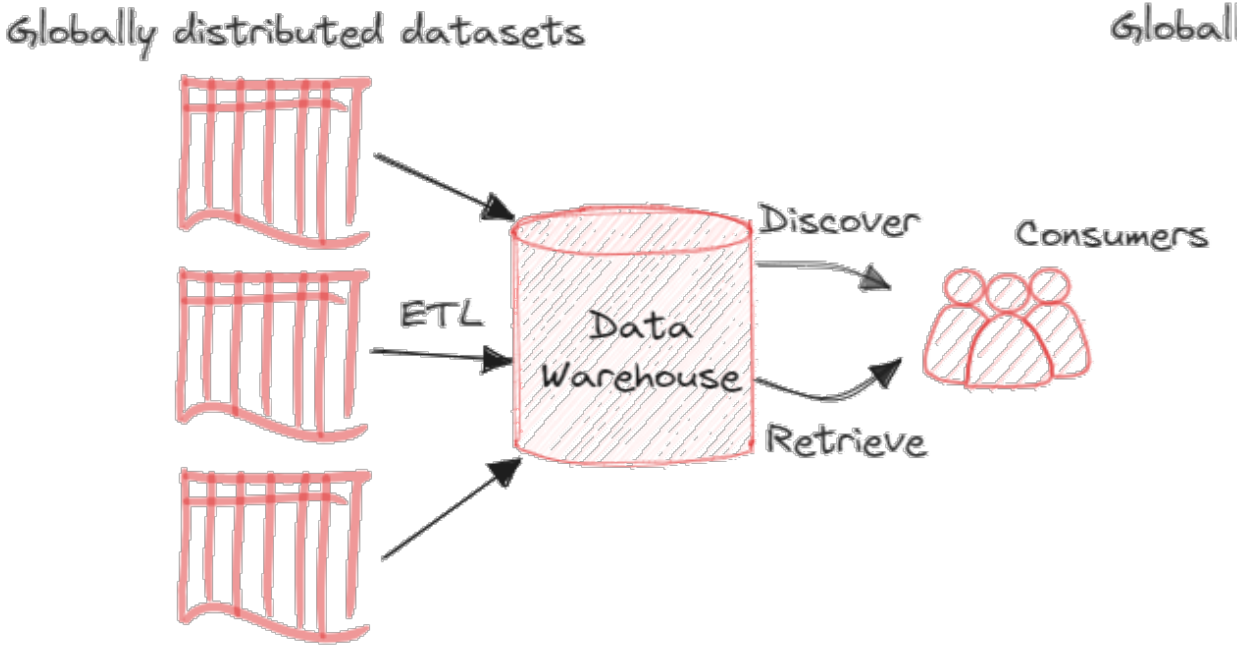
Tuned for structured datasets

Easily retrieve selected datasets

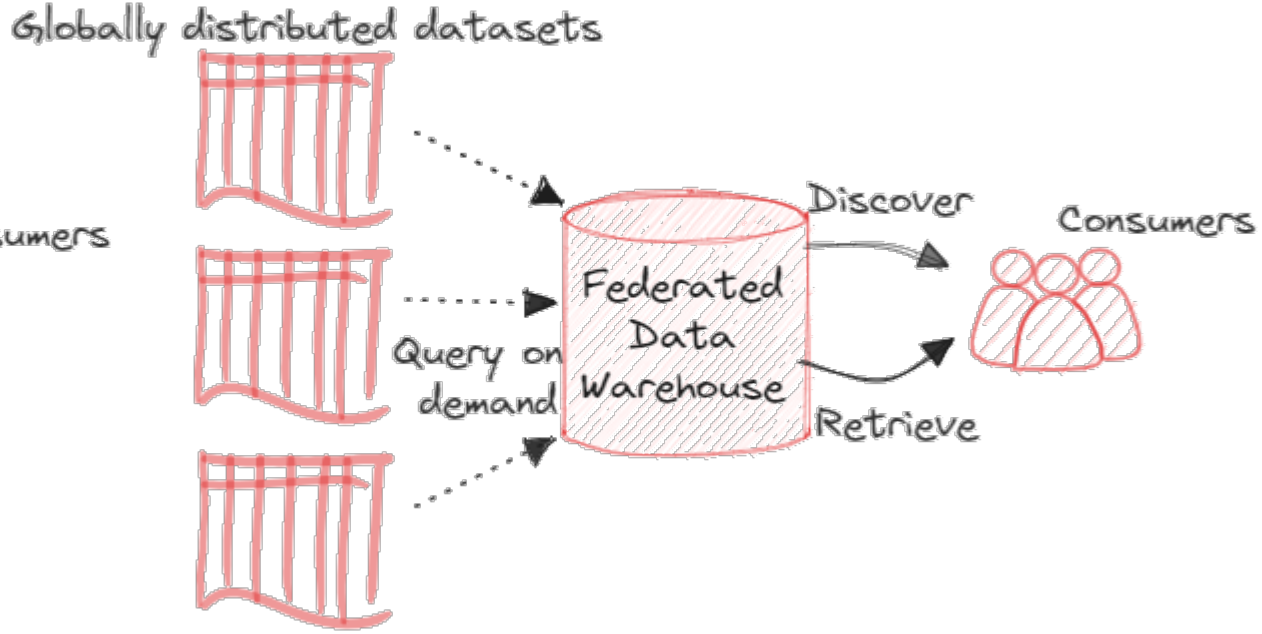
UI for interactive use

API for automation

Options – single access point



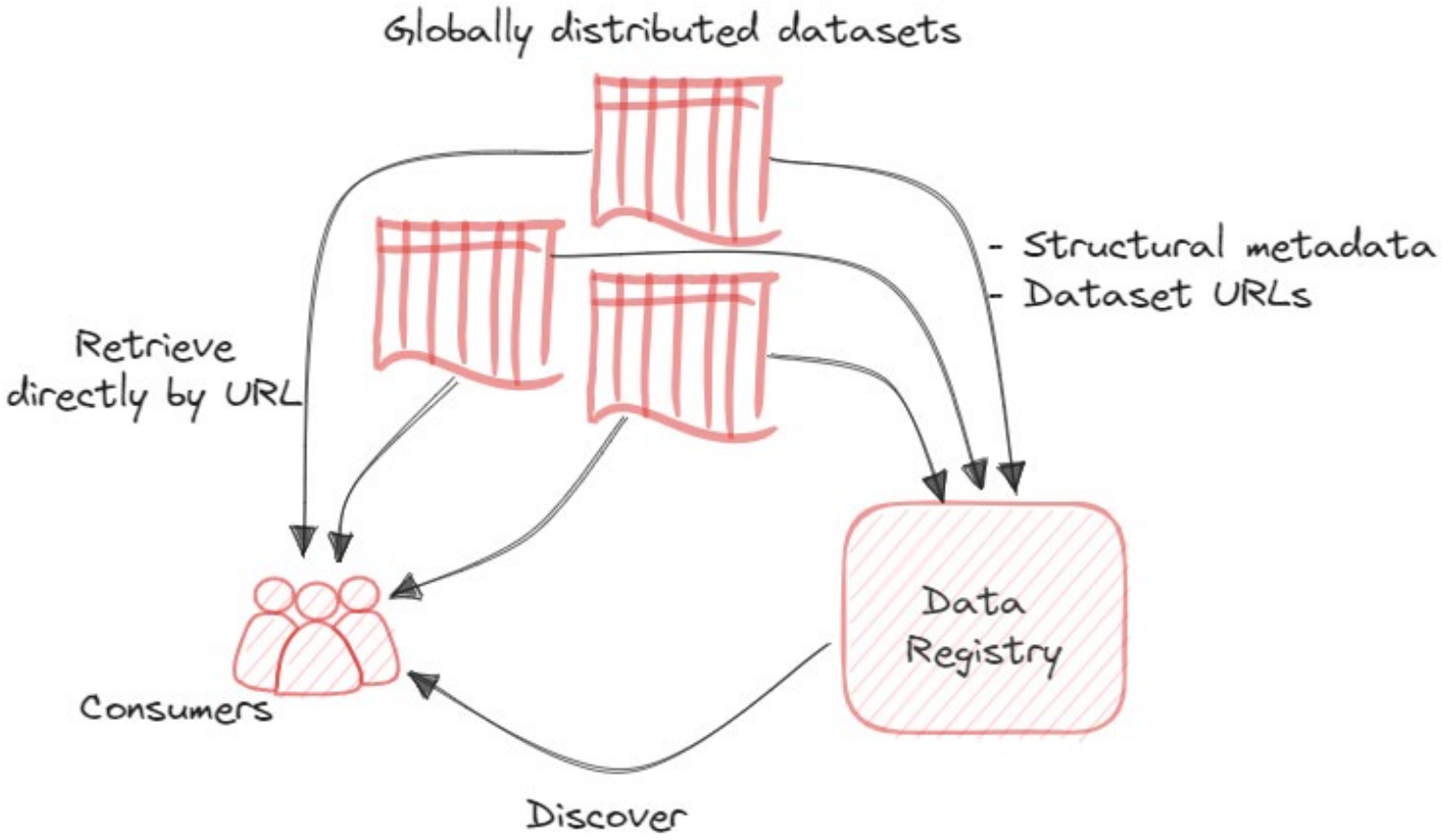
Data consolidated into a central warehouse



Federated data warehouse

A single access point is attractive but potentially complex and costly to implement

A ‘data registry’ provides a simpler alternative approach



- Catalogue of datasets and where to find them
- Each dataset has:
  - Detailed structural model
  - Data URL
- Decentralised – follows Data Mesh principles
- Technically simpler
- Legally simpler - risk of third party IP or personal data privacy claims reduced

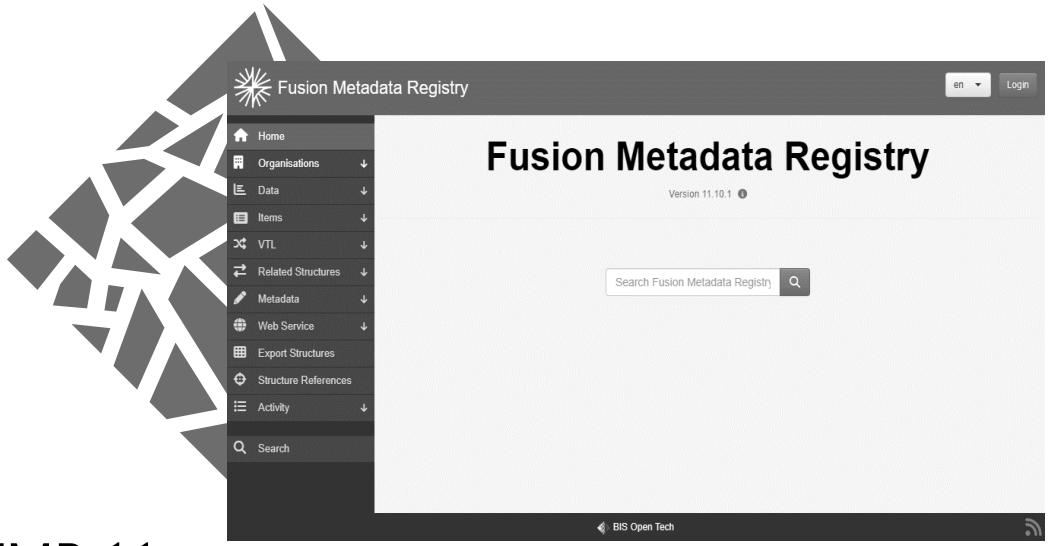
Concrete solutions

Data modelling

SDMX 3.0

- Data modelling
- Data discovery
- Data registrations – URLs link to the data

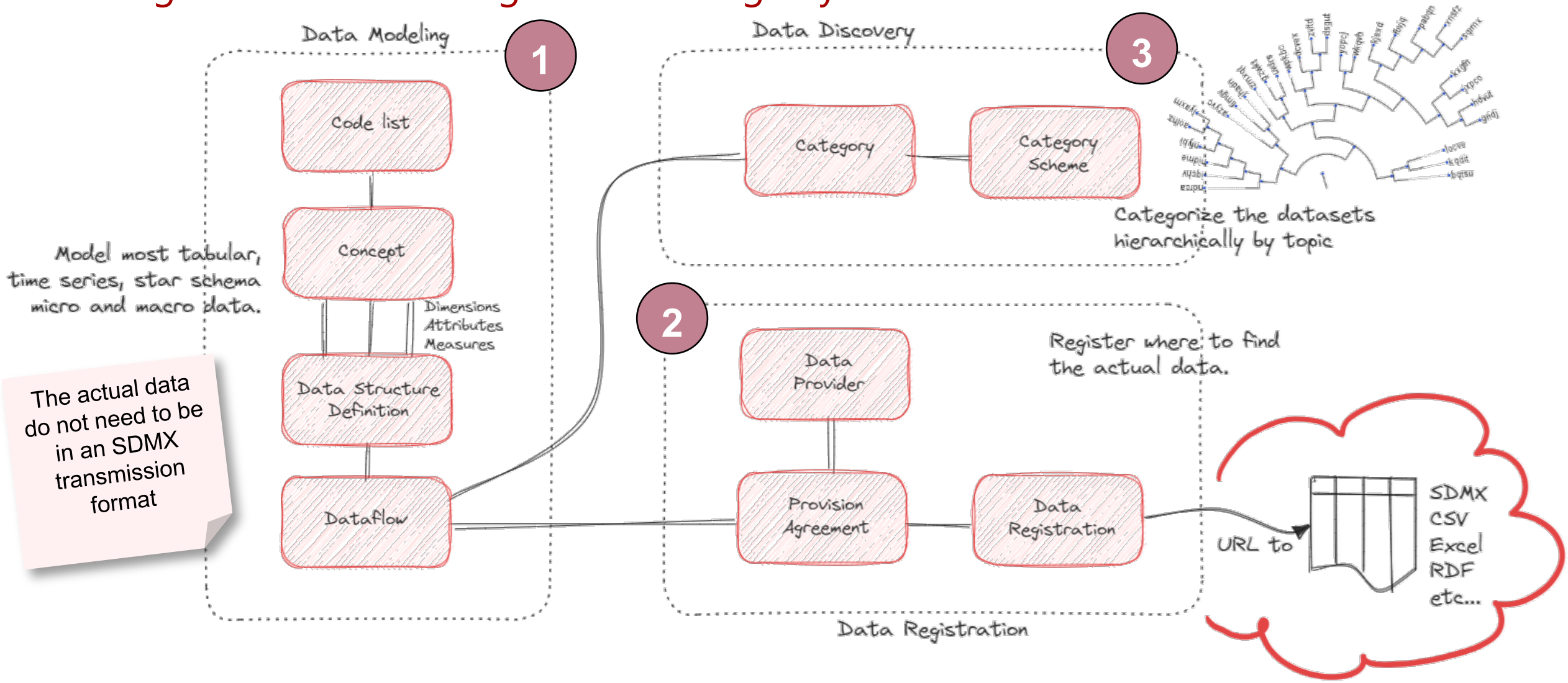
Software



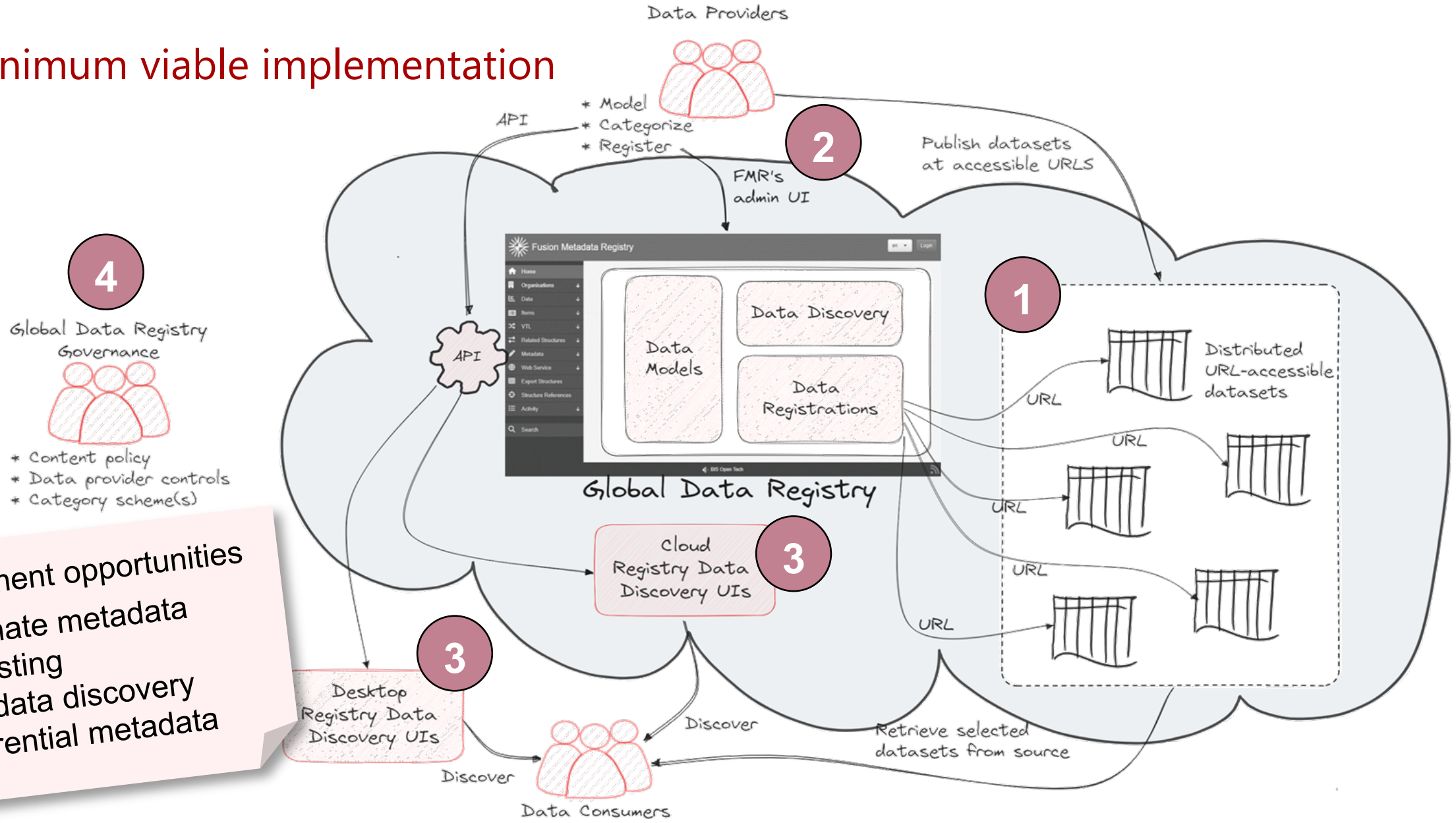
FMR 11

- SDMX 3.0 structural metadata registry
- Free and open source
- Cloud native
- BIS owned and maintained

Using SDMX 3.0 for the global data registry use case



A minimum viable implementation



Enhancement opportunities

- Automate metadata harvesting
- Metadata discovery
- Referential metadata

Conclusion

Plenty of public structured datasets published by institutions globally, but **hard to find and retrieve.**

The **global data registry** approach provides a centralised place to discover and retrieve datasets without the complexity of a single access point.

A practical implementation could be based on **SDMX 3.0** and existing software tools like **Fusion Metadata Registry.**

Is there **demand** from data consumers?
More consultation needed.

Thank you

Glenn Tice glennphilip.tice@bis.org
Matt Nelson matt.nelson@bis.org

BIS MED IT

References

About Fusion Metadata Registry (FMR)	Link
FMR quick start using Docker	Link
European open data portal data.europa.eu	Link
Banca d'Italia public Aggregated Data – Statistical Database	Link
US open data portal data.gov	Link
Google Data Commons	Link
European Single Access Point (ESAP) regulation	PDF Link
SDMX data registrations API query example (returns XML)	Link