Strengthening data governance with SDMX

OCTOBER 30, 2023

Allen Boddie

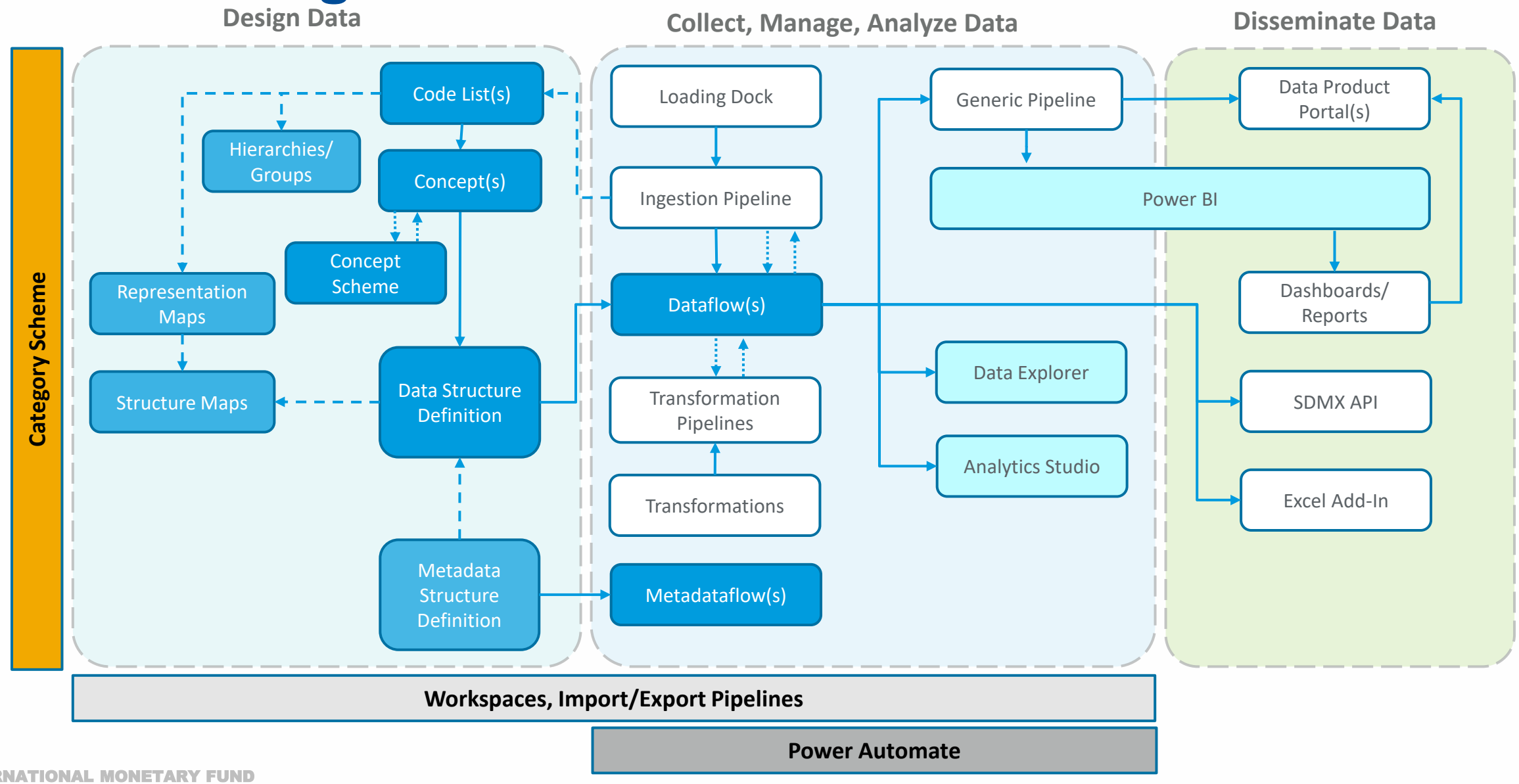
The Vision, Benefits, and Opportunities of iData

The iData project was established to mitigate the risk that IMF's legacy system (EcOS) fails. It is also an opportunity to modernize data management and dissemination at the Fund.

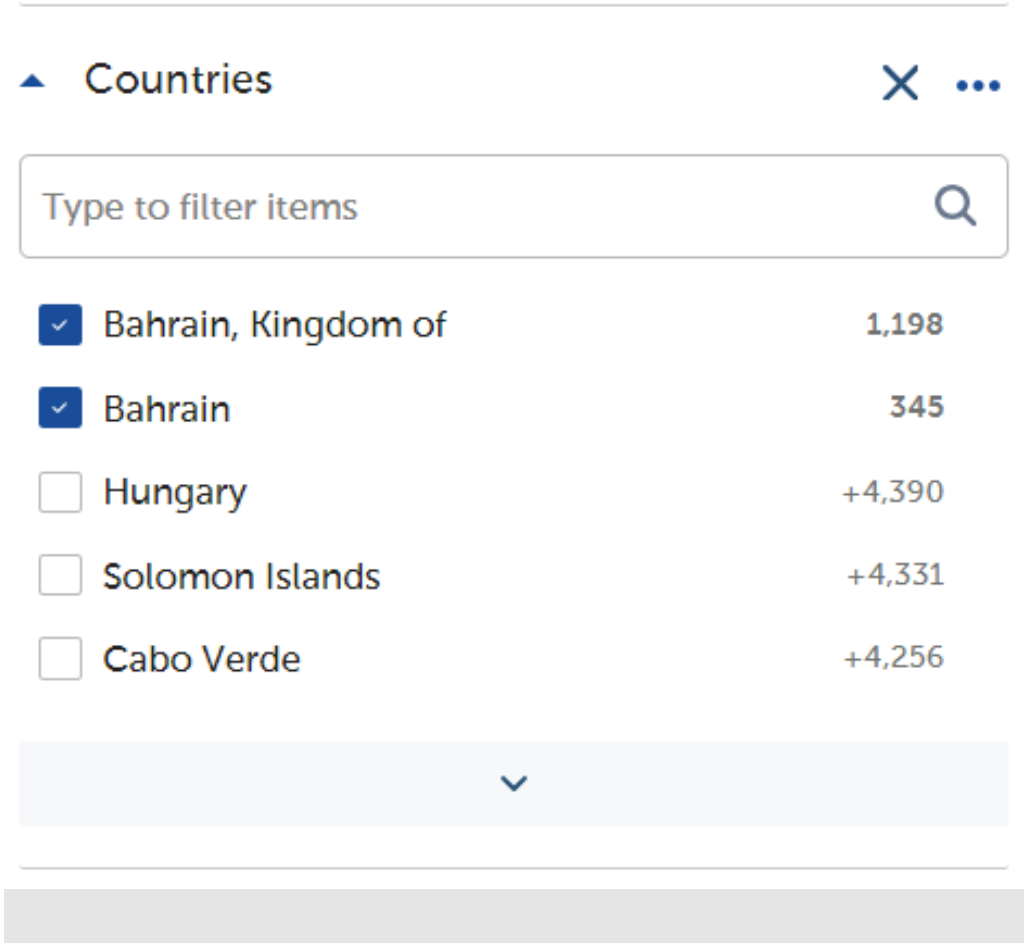
- Key Benefits to **Data Producers** include:
  - Tools supporting enhanced collaboration
  - Integration with analytic tools
  - Off-the-shelf solutions providing access to a large ecosystem of resources

- Key Benefits to **Users** include:
  - Enhanced data access and discovery
  - Industry-leading visualization
  - Multilingual access

iData Big Picture



Problem: Inconstant Names Across Datasets



Dataset owners maintain their own list of countries relevant to their dataset. This includes country groupings specific to that dataset.

This results in:

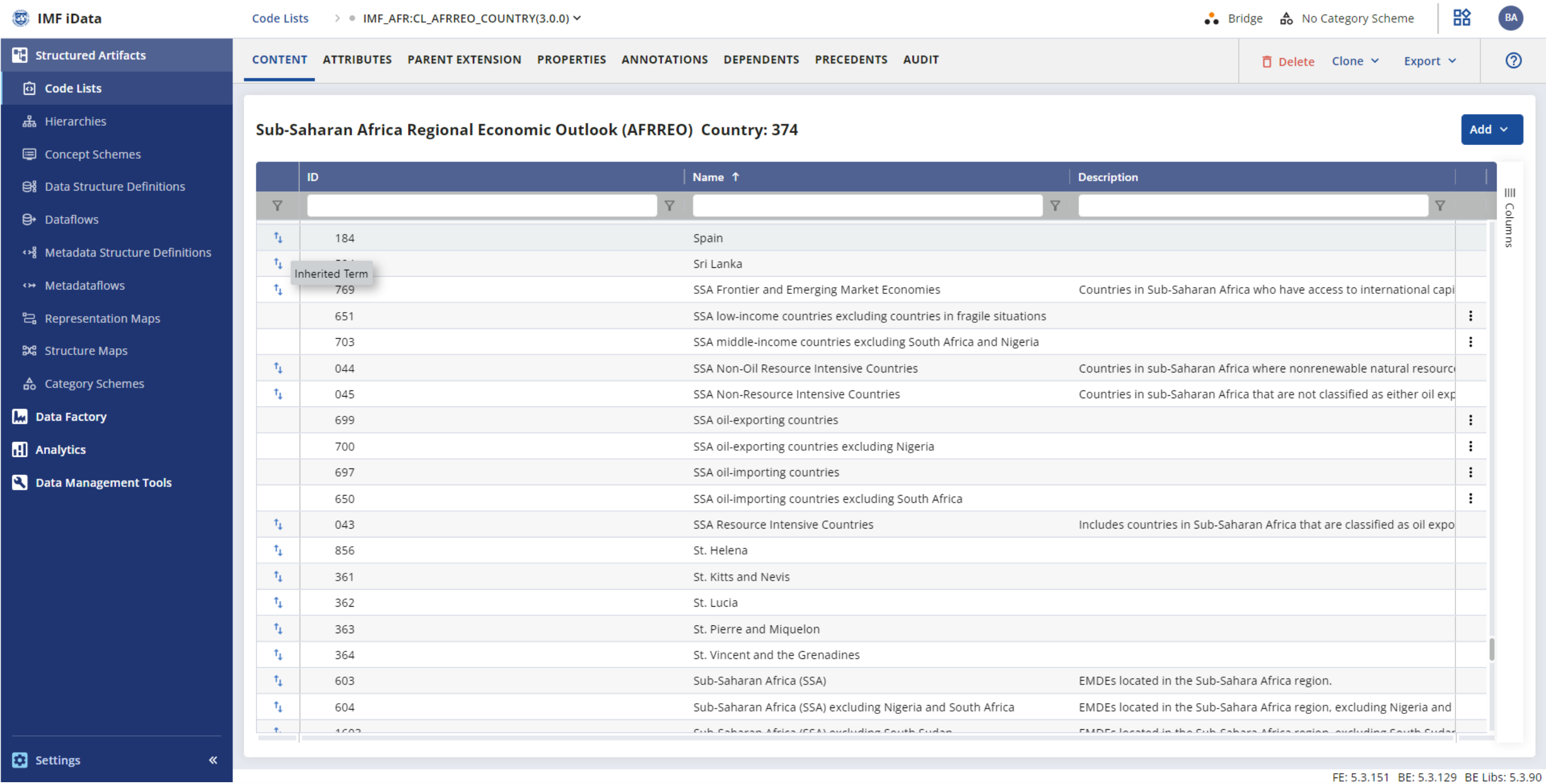
- Deviation from IMF's official country names
- A poor experience for data users
- High level of effort for:
  - Name changes
  - Country groupings change

Solution: Inherited Code Lists

Official country code list with official names:

- Maintained by a central team
- Used as basis for cross-dataset portal facets
- Datasets which need additional terms can inherited from the official code list
- Datasets directly using official code list can reuse hierarchies built by central team

- Remaining Issue: Can’t reuse hierarchies for inherited code lists



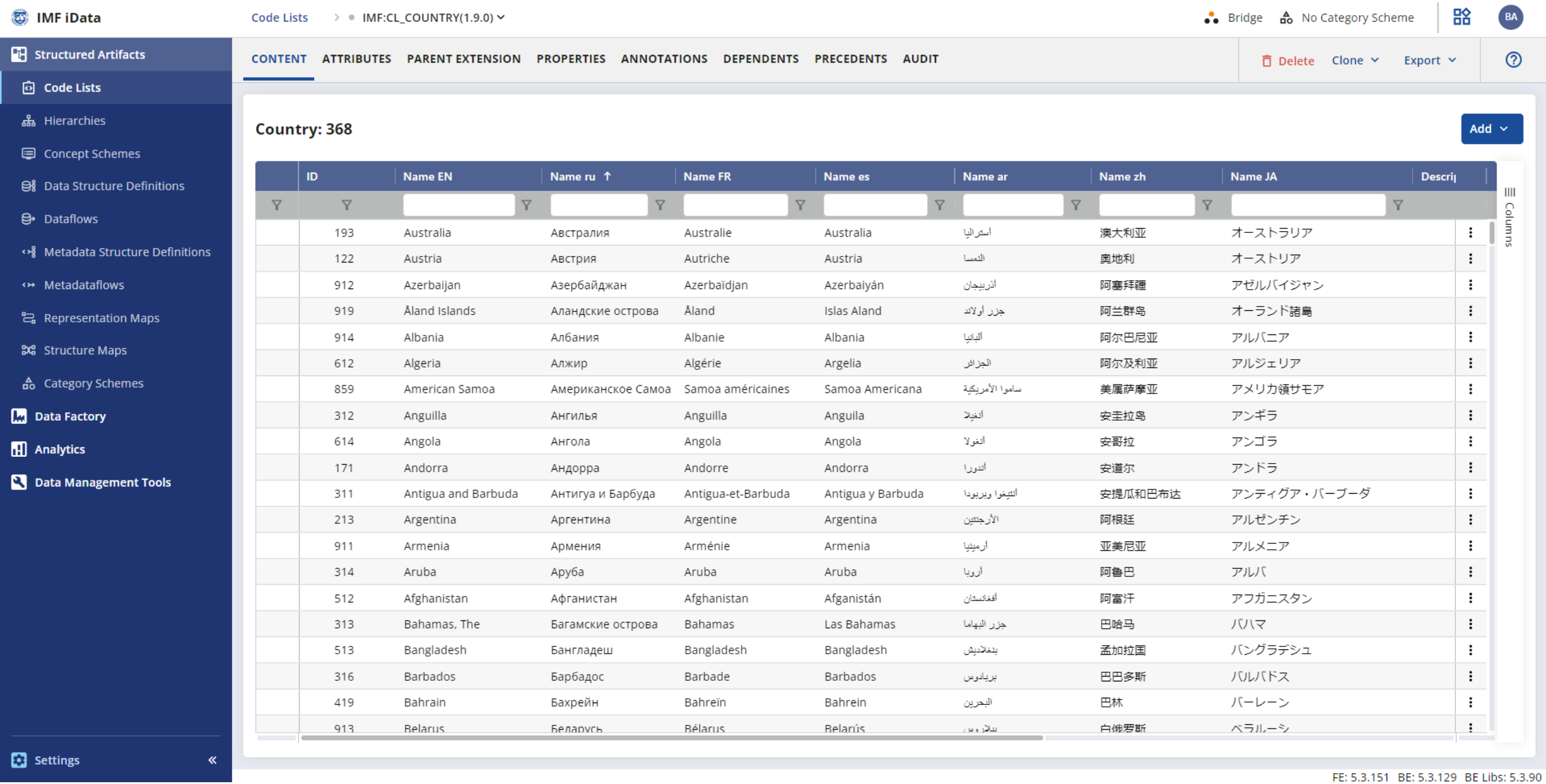
Inherited Code List for AFR Regional Economic Outlook

Problem: Limited Global Usage

Country	Users ↓	New Users	Sessions
	6,898,004 % of Total: 100.00% (6,898,004)	6,922,374 % of Total: 100.00% (6,922,374)	10,612,588 % of Total: 100.00% (10,612,588)
1. United States	**3,411,242** (49.29%)	3,402,701 (49.16%)	3,938,798 (37.11%)
2. India	**278,729** (4.03%)	280,132 (4.05%)	458,542 (4.32%)
3. United Kingdom	**258,097** (3.73%)	257,927 (3.73%)	451,113 (4.25%)
4. China	**252,831** (3.65%)	253,575 (3.66%)	530,834 (5.00%)
5. Japan	**191,840** (2.77%)	193,596 (2.80%)	344,525 (3.25%)
6. Germany	**121,990** (1.76%)	122,044 (1.76%)	224,239 (2.11%)
7. South Korea	**110,216** (1.59%)	111,190 (1.61%)	199,534 (1.88%)
8. Russia	**95,072** (1.37%)	95,743 (1.38%)	193,310 (1.82%)
9. Spain	**90,696** (1.31%)	91,791 (1.33%)	161,653 (1.52%)

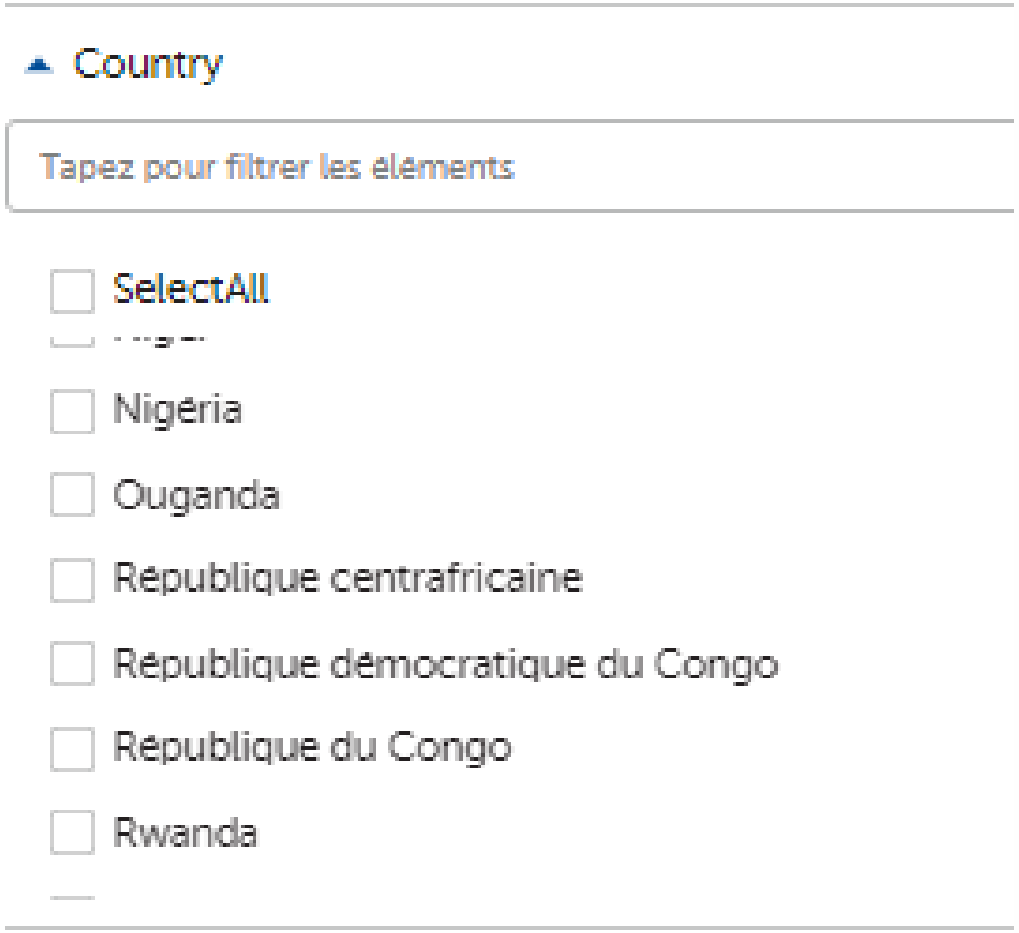
Despite the IMF severing a global membership, the usage of our data portal is limited:

- Around 50% of our users are US based
- Current portal content only available in English
- Current portal very slow away from Eastern United States



Solution: Multilingual Artifacts

Solution: Multilingual Artifacts



Remaining Issue: The SDMX standard seems to only allow for downloading all localizations of an artifacts.

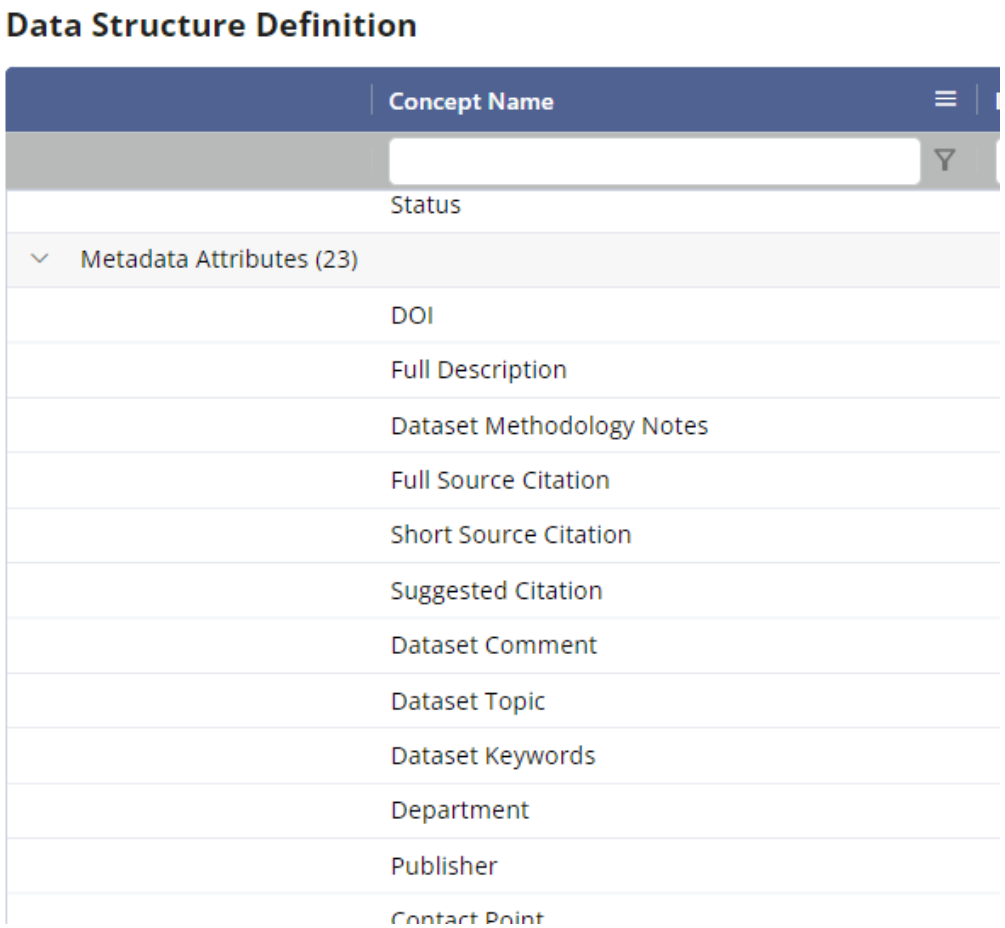
- This means the API will take longer to return information needed on the portal, where a user is looking at one language, than needed
- We implemented the ability to return a single localization to improve performance, but this is not in the standard

Using Akamai CDN to improve performance for non-US based users

Problem: No Consistent Metadata for Datasets

To enable portal users to search through datasets and improve data governance around datasets we want to enforce a common set of required metadata for all published datasets.

Solution: MSD Attached to DSD



Centrally managed MSD with all required metadata fields.

- Require population of these fields for published datasets
- Information can be displayed on portal in a uniform way for all datasets

Remaining issues: Harder to maintain for datasets which need their own metadata attributes

Enforcing Use of Central Artifacts

The solutions given require different teams managing artifacts uses to support the same dataset.

To get this all to work together we use semantic versioning and wild carding:

- All artifacts in iData, even non-SDMX artifacts used in data production, follow semantic versioning.
- A non-backward compatible change is a major version change
- For most artifacts, we do not allow wildcarding of major versions

Multi-dataset Queries

We are exploring using multi-dataset queries to eliminate repetitive datasets while still allowing that selection of data to be accessible.

Multi-dataset queries are also used in the watch list on our portals

Remaining issues: The format for multidata quires does not allow for keys to be passed for individual datasets resulting in extra data being displayed.

- We use a POST request where an array of dataset objects with their own filters are passed
- Fussion Registry modifies the GET request to prefix filters with reference to dataflow

Questions?