



EUROPEAN CENTRAL BANK

EUROSYSTEM

**Searching for a needle in a haystack
- how SDMX IM can improve
searching through economic data**



29/10/2023

**Zlatina Hofmeister, Gabor Horvath, Federica Marcon
European Central Bank**

New search engine initiative

Exponential grow of the ECB disseminated data since the launch of SDW in 2005 while the core structure and appearance remained static



User experience is undermined: accessing and searching data in SDW portal feels like search for a needle in a haystack



Need for new tools aimed to improve the usability of the search engine for users

How to go from this...



... to this



The new ECB Data Portal



- EDP → **key dissemination channel for ECB statistics** Primary source of **data & metadata** for end users
- Fully designed based on the **SDMX information model (SDMX IM)**
- Access the public SDW via: <https://data.ecb.europa.eu/>

EDP and SDMX IM

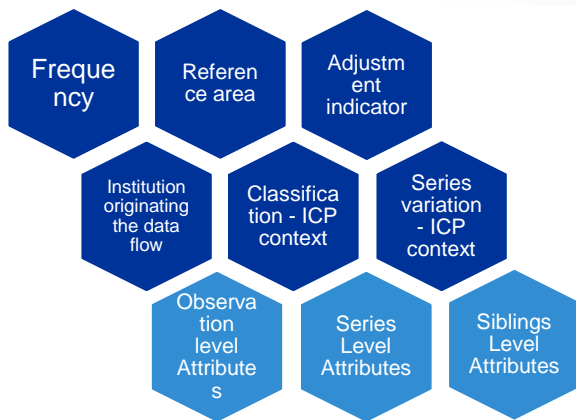
| Artefacts of the SDMX information model | Corresponding implementation in EDP |
|---|---|
| Concept scheme | Navigation hierarchy |
| Data Flow | Data Structure Identifiers |
| Data Structure Definition | SDW presentation of the series with their dimensions (series key) |
| Code list | Filters |
| Cross sectional code list | Filters –dependence of common dimensions across datasets |
| Metadata | Metadata–reference/structural |
| Provision Agreement | Provision Agreement – part of the metadata |
| Data provider | Data provider – part of metadata |

Time series disseminated in EDP



- Defined based on existing data structure definition (that includes dimensions and attributes)
- Uniquely represented by its dimensions that are composing the unique series keys
- Carries additional metadata information through observations, series and sibling level attributes
- Each dimension or attribute have conceptual meaning related to specific economics/statistics phenomenon
- Each coded dimension or attributes are linked to corresponding code lists

Opportunities of a metadata rich environment



ICP data structure definition - ECB_ICP1

The high integration of the data disseminated in the EDP and SDW with the SDMX Information Model can be leveraged in order to make the search engine more effective and user friendly



new SDW and EDP Intelligent Search

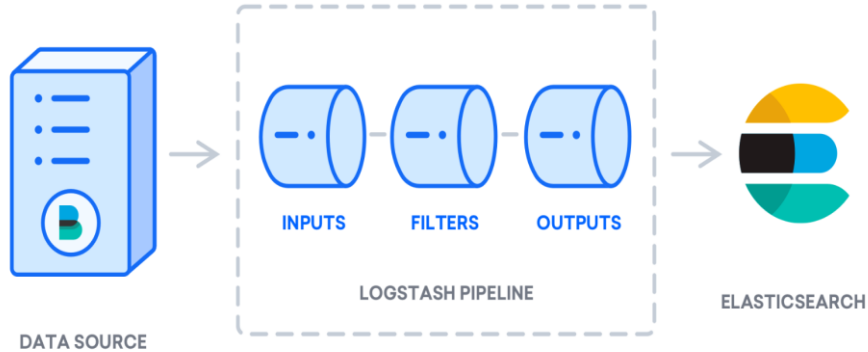


Elasticsearch: storing documents in an index

Logstash: used for processing documents e.g. from a database

Kibana: visualization tool for Elasticsearch index

SDW and EDP Intelligent Search



INPUTS: internal SDW or EDP Oracle database

FILTERS: both built-in both custom analyzers are used by IS, implementing different filters such as English possessive stemmer, English_keywords, Lowercase, Synonyms, English_stop, English_stemmer

OUTPUTS: Elasticsearch index

BOOSTING: reference series and series included in ECB publications are currently boosted

Searching for an economic concept

The screenshot shows the ECB Data Portal interface. At the top, there is a navigation bar with the ECB logo and menu items: Main figures, Browse data, Publications, Blog, Methodology, and Currency converter. A search bar contains the term 'inflation'. Below the search bar, there are two tabs: 'Data (219946)' and 'Data information'. A 'Filters' sidebar is visible on the left, with sections for Concepts, Geographical areas, Datasets, and Frequency. The main content area displays search results for 'HICP - Overall index, Euro area (changing composition), Monthly'. The first result shows a value of 4.3 (Sep 2023) with a 'Last updated' date of 29 Sep 2023 11:00 CEST. The second result shows a value of 124.44 (Sep 2023) with a 'Last updated' date of 29 Sep 2023 11:00 CEST. A third result is partially visible: 'HICP - All-items excluding energy and food, Euro area (changing composition), Monthly'.

Currently the new search engine uses:

- Dimensions meaning
- Attributes meaning
- Metadata on dataset level
- Appearance in publications
- Tagging of the series in our custom build “SDW taxonomy” based on SDW economic concept navigation (SDMX concept scheme)

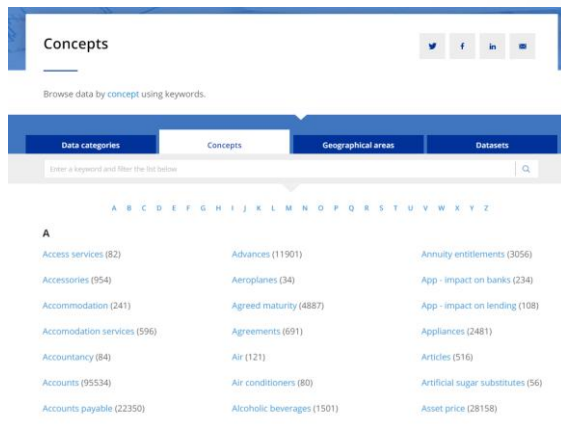
Features to be included In future releases:

- learning to rank, usage statistics, improvement of tags

EDP tagging process

Assigning meaningful key word to each disseminated series:

- Helpful for non expert users with no prior knowledge of DSD concepts/codification
- Used to improve the filtering of the search results
- More intuitive use of the search engine



NAVIGATION TAGS:

Assigned to time series based on the navigation hierarchy node (concept scheme SDMX IM) they are attached to

CODELIST TAGS:

Assigned to time series based on a meaningful codelist of the SDMX data structure definition

BUSINESS AREA DEFINED TAGS:

Defined and customized by data producers

