Assessing Fitness for Integration – a Metadata-driven Approach

Thomas Gottron, Andrea Novello, Ilias Aarab, Bernadette Lauro[1]

European Central Bank, Sonnemannstrasse 20, 60314 Frankfurt am Main, Germany[1]

Abstract

Modern data landscapes are composed of a large number of diverse but complementing datasets. For insightful analytics these datasets need to be semantically and technically integrated (i.e. actual data records need to be combined). Integration, however, poses several challenges, including determining which datasets are compatible for integration, understanding the technical methods for achieving integration, and assessing the extent of integrability and linkage rate among various datasets.

To support users in using and combing a large variety of datasets, we maintain comprehensive metadata repositories at the European Central Bank (ECB). Metadata describing data products, storage and access roles are used to support data discoverability and accessibility. Metadata describing concepts, data models, transformation rules and mappings are used to support users in dataset integration and analysis. By leveraging extensively such metadata, we designed a *fitness for integration dashboard*. The dashboard aims to inform users about available data, illustrating how it can be integrated and the degree to which data aligns across common dimensions. However, the dashboard represents just the visual component of a broader solution. The centrepiece of the solution is a metadata-driven and fully automated four-step process for populating the dashboard with relevant information. In a first step, the process utilises metadata to identify semantic dimensions and suitable identifiers for data integration and aggregation. Subsequently, by leveraging logical inferencing, we ascertain which datasets can be integrated, determine their technical storage locations, and identify attributes for slicing data into semantically valuable aggregates. The third step is to query the underlying actual data and to compile various integration metrics to assess integrability, linkage rate and linkage weighted by relevant business indicators. These metrics are calculated at different levels of detail and aggregation, such as per country, over time, or based on other relevant breakdowns sourced from taxonomy-driven reference information. Finally, an interactive dashboard retrieves and visualizes these pre-computed metrics, along with additional business metadata related to the datasets. This dashboard efficiently serves ECB analysts and researchers, enabling them to make more informed decisions on utilizing the ECB's data for their analytical and research purposes.

To showcase the feasibility of our approach, we implemented a prototype leveraging the ECB's data dictionary, an in-house Hadoop based data and analytics platform, and RShiny to construct the dashboard. Furthermore, the prototype demonstrates the benefit of high-quality and semantically well modelled metadata for supporting users in exploring and understanding the data landscape.

Keywords: data integration, quality indicators, metadata

1. Introduction

More and more often, data analytics is based on combining granular datasets. The flexibility to integrate diverse datasets, focus on relevant aspects and aggregating data as needed

[1] This paper should not be reported as representing the views of the European Central Bank (ECB). The views expressed are those of the authors and do not necessarily reflect those of the ECB.

provides the basis for new insights. However, before diving into the analytics, users of modern, distributed data landscapes face the challenge to identify which datasets are compatible for integration. From our experience in a large ECB-internal project on data integration and data analytics, we found that most investigations started by looking into three questions: which datasets can be integrated, how to perform this integration in a semantically meaningful way and how well and complete the actual observations in the datasets can be linked with each other to obtain meaningful insights, i.e. how good is the linkage rate.

In practice, we found that there is a prevalent approach to answering these questions: users leverage information on structural metadata (i.e. metadata on data models, transformation rules and mappings etc.) to identify common identifiers within datasets' objects or dimensions. Common identifiers and time dimensions (reference or validity period) are used to link and join data. The resulting joint dataset is subsequently probed and explored, often looking at different breakdowns and relevant measures to check the linkage rate. However, the tools employed in this approach are used in isolation, investigations are done manually, and the same task has to be reiterated for every combination of datasets.

This motivated us to investigate the potential of automating this approach, leveraging sound metadata descriptions for the datasets, scalable infrastructure for querying and computing data aggregates as well as for visualising pre-computed integration quality metrics. The vision was to build a *fitness for integration dashboard*. This dashboard aims to inform users about available data, illustrating how data can be integrated and the degree to which data aligns across common dimensions. We implemented a prototype using available information systems and infrastructure. The prototype demonstrated the feasibility of the approach, provided a basis for discussion with users and indicated clear benefits from such a solution.

2. Background and Related Work

Quality assurance for statistical data has a long-standing tradition. Accordingly, there is a wide range of indicators used to measure and quantify data quality under various aspects (Sidi, et al., 2012). Metrics to indicate specifically how well datasets can be integrated, however, are sparse. The overview of quality indicators for the *GSBPM* (United Nations Economic Commission for Europe, 2017) refers to the linkage rate as one "very important measure of the quality" for linked or integrated datasets. However, the report also explicitly mentions that further work needs to be done to develop indicators for describing the accuracy and reliability of linkage variables. Other sources develop concepts similar to the linkage rate but refer to it with different terminology like "coverage" or "support" (Explorium, 2023). We will specify and formalise key metrics for our approach in Section 3.

The basis for our metadata-driven approach is a rich semantic ontology, which is specified using the SMCube information model (European Central Bank, 2024), also underlying the *Single Data Dictionary* (SDD) and the *Banks Integrated Reporting Dictionary* BIRD (European Central Bank, 2024). This model allows for obtaining detailed structural and semantic information about several kinds of datasets (aggregated, granular, template-based). Using structural information for guiding users on diverse, potentially distributed, and semi-structured data has been analysed in various domains. *Data guides* have been found to provide accurate schema descriptions that can support users in browsing and exploring datasets (Goldman & Widom, 1997). The method of capturing links between clusters of the same semantic nature and to indicate possible linkage and queries is inspired by work on *Linked Open Data* (Konrath, Gottron, Staab, & Scherp, 2012). Measuring or estimating distributions of how well subsets of data can be interlinked can provide valuable insights for data exploration and can be achieved efficiently and with relatively high precision (Gottron & Gottron, 2014).

3. Metrics to Assess the Fitness for Integration

To provide our users with the relevant information, we look primarily at three different metrics: *integrability*, *linkage rate* and *weighted linkage rate*. As existing literature does not agree on terminology and does not provide precise and harmonised definitions, we formalise the metrics described in this paper.

From a notation point of view, we use A and B to denote *datasets* of tabular form with different attributes. We use $A.z$ to refer to *attribute* z in dataset A. With $|A.z|$ we denote the cardinality of z, i.e. the number of distinct values that are observed for z. An *aggregate function* f applied to the values of an attribute (e.g. computing the sum or mean value) is described by $f(A.z)$. With $A \bowtie_z B$ we denote the *inner join* between A and B on a common attribute z. Whenever the join attribute is obvious from the context we will simply write $A \bowtie B$. Furthermore, we allow for formulating *constraints* $A|_{z=\alpha}$ to indicate all elements in A where attribute z has a value equal to α. All notations can be combined in an intuitive way, e.g. $f\big(((A \bowtie_z B)|_{x=\alpha}).y\big)$ describing an aggregation function f applied to attribute y on the inner join of datasets A and B which has been constrained to elements, where attribute x has a value equal to α.

Furthermore, we use some specific letters to mark attributes of particular types. We use i to denote *identifiers*, which can be used to (uniquely) identify elements in a dataset. Identifiers typically serve as primary or foreign keys in data and are natural candidates for computing joins. Examples for identifiers are legal entity identifiers, transaction identifiers or technical identifiers. An attribute t indicates a *time dimension*, e.g. reference periods or validity ranges for observations. Time needs to be considered for a semantically correct integration of

observations, potentially requiring a conversion of time values to facilitate matching. A *stratification attribute* s provides a natural source for constraining a dataset to subgroups of observations or elements and can serve as basis for comparing how well subsets of the data can be integrated. Examples are geographic regions, economic sectors, or types of financial instruments. Finally, u indicates a *unit of measurement* which provides the basis for aggregating information. Units of measurements can be identifiers used to count unique elements, but also numeric attributes can indicate a measure, e.g. outstanding nominal amount for loans or the number of employees for a legal entity. Please note, that a structured dataset may be composed of more than one table, containing more than one type of identifier, time dimensions, stratification attributes or units of measurement. Based on these formal notations and definitions, we define the following metrics we use for assessing the fitness for integration:

Integrability $I(A,B)$: Indicates if two datasets can be integrated at all. Integrability is a binary metric with a value of 1 if A and B share a common identifier suitable for linking and integrating the data. Vice versa, $I(A,B) = 0$ if no such identifier is available. Integrability is obviously symmetric, i.e. $I(A,B) = I(B,A)$. Note, that time dimensions are not considered for integrability. While time might play an important role in a semantically correct integration, it is not a precondition for the technical ability to integrate data.

Linkage rate $L_i(A,B)$: Let A and B be integrable on identifier i, then the linkage rate is defined as the ratio of identifier values present in A that can also be found in B. Formally we define $L_i(A,B) = \frac{|(A \bowtie B).i|}{|A.i|}$ with values in the interval $[0,1]$. Note, that the linkage rate can easily be computed on a constrained subset, e.g. using a stratification attribute s to select a subgroup of elements. Due to the deliberate choice of the normalisation factor being based on only one of the two dataset sizes, this definition of linkage rate is asymmetric and in general we may expect to observe $L_i(A,B) \neq L_i(B,A)$.

Weighted linkage rate $L_{u,f}(A,B)$: For an attribute u indicating a measurable unit and an aggregation function f, we define the weighted linkage rate as $L_{u,f}(A,B) = \frac{f((A \bowtie B).u)}{f(A.u)}$. This provides a weighted view of the linkage rate, e.g. reflecting linkage according to market values, outstanding nominal amount etc.

4. Approach

Our approach is based on four steps: (1) retrieving relevant metadata information, (2) determining integrability by inferring suitable dimensions for integrating, (3) retrieving actual data to compute the linkage rate quality metrics and (4) visualising the precomputed metrics. In the following we will provide further details for each of these steps.

4.1 Retrieving Metadata

The entire process is designed to be metadata-driven. We start with retrieving metadata from a common repository covering a wide range of datasets available in the institution: the ECBs SDD, based on the SMCube metadata (information) model. The SDD covers different aspects of structural metadata, common code lists, technical metadata, and business metadata.

Structural metadata is of particular relevance for us, as it describes tables and their attributes. Moreover, the SDD captures the semantics (meanings) of the attributes. This is a key information to identify common identifiers and time dimensions for integrating data, for detecting stratification attributes and units of measurements. Stratification attributes with a hierarchically organised code-list offer themselves additionally for aggregation at different levels and depths of the taxonomy. Furthermore, the structural metadata also gives insights into definitions of time and validity ranges, which are crucial for a harmonised and correct integration of data with a historicity component. Technical metadata provides insights into where data is physically stored and how it can be accessed. Business metadata provides additional context and descriptions of the datasets.

To ensure scalability and to avoid an excessive combination of all possible attributes, we used a central configuration setup to denote relevant concepts which represent identifiers, time dimensions, stratification attributes and measurements units. In theory, using self-descriptive metadata, e.g. including meta-metadata, would allow for embedding such information in the metadata system and facilitate further automation. For the sake of our prototype, we did not investigate establishing such a bootstrapping setup.

4.2 Inferencing Common Identifiers and Determining Integrability

The structure of datasets and their attributes serves as basis for determining integrability. Given the semantic nature of the metadata, we can be sure to identify attributes with the same type of content and that equivalent values have the same meaning. By computing an inverted index (Baeza-Yates & Ribeiro-Neto, 1999) to map semantically equivalent identifiers to dataset attributes and corresponding datasets, we can easily infer which datasets can be integrated based on common identifiers. This essentially provides the integrability indictors $I(A, B)$ for all pairs of datasets A and B.

4.3 Retrieval of Data to Compute Quality Metrics

Integrability provides the basis for retrieving and joining data from various datasets. At this point we leverage information on time attributes to ensure correct integration along the time dimensions. Bringing together the data allows for further determining cardinality and other

types of weighted aggregates which provide the basis for computing our linkage rate metrics (i.e. expressing how linkable are two datasets). The stratification attributes provide a vector for semantic grouping and selection of subgroups of elements and observations. The attributes identified as units of measurement provide the values for weighted aggregation (e.g. by counting, summing or averaging). This directly gives all relevant information to compute linkage rates and weighted linkage rates.

4.4 Deliver Metrics via an Interactive Dashboard

Once the metrics have been computed on all subgroups and on a global level, the results are stored with clear information on what datasets, tables, attributes, aggregates and reference periods are considered. Operating on pre-computed metrics has two distinct advantages. First of all time efficiency: pre-computed values do not need to rely on complex live calculations involving joins of potentially large and distributed tables. Instead, the values can be retrieved and displayed with hardly any computational overhead. Secondly, access rights: there is no need for the visualisation of the output component to have access to the underlying (and potentially confidential) granular information. In this way we can simplify the access rights requirements for users to a permission to see aggregated and high-level integration metrics.

5. Implementation of a Prototype

To demonstrate the feasibility of our approach and test the metrics with users we implemented a prototype. The implementation leveraged available platforms to facilitate a scalable solution.

5.1 Processing

For the processing component of the solution, we used Python which retrieves metadata information on datasets via the SDD API interface. This provided us with the semantic representation of attributes, allowing us to identify common dimensions. For the purpose of the prototype we defined a range of relevant attributes such as identifiers, measures, stratification or time attributes which are suitable for the link, in a configuration file. The inverted index to compute intersections of identifier attributes and datasets was implemented in-memory. Any index entry referring to more than one dataset indicated integrability for those datasets.

Alignments along time dimensions required to harmonise formats and suitable points for connecting datasets. Based on the availability of data we chose monthly or quarterly breakdowns. Master data on legal entities is defined over validity periods (e.g. the address of an entity being valid from 7. July 2008 until 23. September 2012). Such time periods needed to be checked for intersection with the reference periods to ensure correct integration.

Subsequently we constructed SQL queries to join tables based on common identifiers and time dimensions. The count and aggregation information needed for linkage rate and weighted linkage rate as well as any constraints formulations to select subgroups of elements based on stratification attribute values are directly translated to SQL elements in the queries.

5.2 Storage and Data Handling

Eventually, all pre-computed metrics were stored in a dedicated, Hadoop based database to serve as background for generating the visualisation dashboard. Each measurement for a metric is clearly and uniquely identified by: (a) type of the computed metric, i.e. integrability, linkage rate or weighted linkage rate, (b) integrated datasets or tables, including the identifiers used for integration, (c) time dimension used for integration and relevant reference periods for which the metrics have been computed, e.g. December 2022, (d) stratification value used to constrain the data to a subset, and the observed value used for the constraint, e.g. Portugal as a specific country, (e) measurable unit and aggregation function used for the weighted linkage rates and (f) actual value of the metric, e.g. a weighted linkage rate of 0.826.

5.3 Visualisation

The visualisation component was implemented as an RShiny dashboard. The dashboard connects with the backend database to retrieve and display the pre-computed metrics. The only business logic implemented in the dashboard is related to browsing, selecting and visualising information as requested by the user.

The default view provides an entry point to the investigation of all datasets. To this end, it presents the binary integrability metric using a graph visualisation. Each dataset is represented as a node in the graph, with edges connecting those datasets that can be integrated. Context information for the edges denotes which identifiers have been used to perform the integration. A second high-level visualisation is based on a heat-map to indicate the linkage rate on a global dataset level (i.e. without any breakdowns using the stratification information).

Once the user selects an individual dataset, the dashboard offers general information on the specific data set. This includes business metadata and descriptions as well as indicators on volume, composition, and structure of the dataset. Furthermore, based on the integrability, the user can select a secondary dataset that can be used for integration. This selection helps in retrieving all pre-computed linkage and weighted linkage rates involving both datasets. In general, indicators are presented in a tabular format and in aggregated form as bar-plots. Additionally, geographical information (e.g. based on a stratification by geographic regions) is

visualised as map. If a time dimension is used, the evolution of linkage rate over the last 12 month is visualised as time series.

An additional feature is the provision of the SQL queries used for integration and that are underlying the provided metrics. In this way users can start off further analysis by building on and extending the pre-defined integration (assuming they have access to the original data).

6. Findings

The prototype setup demonstrated that it is feasible to implement an automated and metadata-driven dashboard that provides insights on the fitness for integration of various granular datasets. The metadata assessment, the computation and storage of metrics and the visualisation component demonstrate furthermore, that such a loosely coupled approach is easier to maintain, more performant and more stable for end users.

Interviews with test users provided several insights. Most important was the perceived benefit of the overview generated by the dashboard. Users confirmed the value of the insights based on integrability, linkage rate and weighted linkage rate metrics. This was valid for high-level overviews as well as for more detailed breakdowns on subsets of stratified data. An additional benefit was seen in the different access rights modality for the dashboard. As no detailed data is presented, but only high-level quality metrics, users were not required to have access to the original data. Overall, the dashboard provides a fast alternative and explorative approach to assess if a deeper analysis based on integrated data from different datasets looks promising and if it is worth to invest more time and resources in a fully-fledged analysis.

7. Summary and Conclusions

In this paper we presented an automated and metadata-driven dashboard to assess the fitness for integration of granular datasets at the ECB. The approach is leveraging automatic inferencing on rich metadata of the ECB's Single Data Dictionary to detect which datasets can be integrated, infer a suitable semantic and technical integration, translate this into SQL queries to feed metrics of relevance and interest for users. A prototype dashboard has been implemented and tested with users to confirm the feasibility and added value of the approach.

In future work, we plan to assess integration of additional datasets, investigate options for leveraging further metadata sources and how to establish links with data sources using other, standardised metadata models. A further extension would be to include experimental approaches for interlinking datasets, e.g. based on fuzzy matching.

References

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval.* Addison-Wesley Longman.

European Central Bank. (2024, March 26). *BIRD project*. Retrieved from The Banks' Integrated Reporting Dictionary: https://bird.ecb.europa.eu/projectDefinition

European Central Bank. (2024, March 26). *SMCube Information Model*. Retrieved from European Central Bank: https://www.ecb.europa.eu/stats/ecb_statistics/co-operation_and_standards/html/smcube_model.en.html

Explorium. (2023, Aug 6). *Support and Coverage – Data Integration Metrics You Should Know*. Retrieved from Explorium AI: https://www.explorium.ai/blog/external-data/support-and-coverage-data-integration-metrics-you-should-know/

Goldman, R., & Widom, J. (1997). Dataguides: Enabling query formulation and optimization in semistructured databases. *VLDB*(97), 436-445.

Gottron, T., & Gottron, C. (2014). Perplexity of index models over evolving linked data. *ESWC 2014: The Semantic Web: Trends and Challenges: 11th International Conference* (pp. 25-29). Springer International Publishing.

Konrath, M., Gottron, T., Staab, S., & Scherp, A. (2012). Schemex—efficient construction of a data catalogue by stream-based indexing of linked data. *Journal of Web Semantics*(16), 52-58.

Sidi, F., Panahy, P., Affendey, L., Jabar, M., Ibrahim, H., & Mustapha, A. (2012). Data quality: A survey of data quality dimensions. *International Conference on Information Retrieval & Knowledge Management* (pp. 300-304). IEEE.

United Nations Economic Commission for Europe. (2017). *Quality Indicators for the Generic Statistical Business Process Model (GSBPM) - For Statistics derived from Surveys and Administrative Data Sources.* United Nations.