

Behind the Scenes: Crafting Hungary's New Census Database

Melinda Oparin-Salamon Ph.D¹

¹*Hungarian Central Statistical Office, Hungary*

Abstract

The Hungarian Central Statistical Office launched a new census database in 2023. The paper delves into the technical aspects of this in-house development, accentuating the use of open-source technologies and the benefits of adopting SDMX 3.0 for data description. A departure from the old methods of disseminating Hungarian census data through Excel files and a constrained array of multidimensional datasets within an outdated database sets the stage for the shift. The inherent limitations of past practices hindered the ability to provide comprehensive insights to users. In pilot phase previous census data were seamlessly integrated into the new system, undergoing user testing. In response to user feedback and the identification of other issues, the system underwent refinement. As a result, the census database now including the latest census data provides a user-friendly interface that significantly elevates overall usability. Noteworthy among its features is the absence of limitations on dimensions in multidimensional datasets (although there is a non-technical limit of dimensions which can be comprehensible for users). The pre-execution of all the calculations (including statistical disclosure control) ensures rapid query execution. Users benefit from instantaneous presentation of chosen dataset data, allowing subsequent application of filters—a functionality that further enhance the system's advantages. The usability improvements and a notable increase in user downloads underscore the success of the new census database. It presents now an inspiring model for renewing the Hungarian dissemination database as well. This paper narrates the story of a transformative endeavour, highlighting the practical implementation of a user-focused and data-driven approach to statistical dissemination.

Keywords: census database, SDMX, dissemination

1. Introduction

The Hungarian Central Statistical Office (HCSO) embarked on a transformative journey to modernize its data dissemination methods, culminating in the launch of a new census database in 2023. This paper details the technical processes involved in this endeavour, highlighting the adoption of open-source technologies and the significant role of SDMX 3.0.

2. Background

After the previous census in 2021, Hungarian census data were disseminated through Excel files and a few multidimensional datasets were created for the dissemination database, limiting the accessibility and comprehensiveness of insights provided to users. The need for the development of the census database arose from several aspects. Firstly, the Population

Census and Demographic Statistics Department did not wish to repeat the dissemination of 2011 census data across hundreds of Excel sheets for the 2022 census, neither from the user's nor from the production perspective. Secondly, a load test on the dissemination database conducted in 2021 revealed that the twenty-year-old database, which was already cumbersome for modern users, would eventually need significant improvement or even replacement in the future. To address these two issues simultaneously, our organizational unit, the Data Visualization and Web Editing Section, proposed creating a possible new dissemination database pilot where census data would be integrated into a new database as multidimensional datasets.

HCSO voted for the in-house development as it plays a pivotal role in ensuring the success and sustainability of projects, especially in complex ones such as the creation of a census database. There are several key reasons why in-house development is invaluable:

- **Tailored Solutions:** In-house development allows creating solutions specifically tailored to the organization's unique needs and requirements. This level of customization ensures that the final product aligns perfectly with the objectives and workflows of the organization, maximizing efficiency and effectiveness.
- **Control and Ownership:** Developing solutions in-house grants full control and ownership over the entire development process, from conception to implementation.
- **Flexibility and Adaptability:** In-house development provides the flexibility to adapt and evolve the solution over time as needs change or new requirements emerge.
- **Cost-Effectiveness:** Compared to off-the-shelf solutions, in-house development often proves to be more cost-effective in the long run.
- **Expertise Utilization:** Leveraging in-house expertise enables organizations to tap into the knowledge and skills of their own team members. This not only fosters professional growth and development but also ensures that the solution is developed by individuals intimately familiar with the organization's operations and requirements.
- **Data Security and Confidentiality:** Building solutions in-house provides greater control over data security and confidentiality. Organizations can implement robust security measures tailored to their specific needs, reducing the risk of data breaches or unauthorized access.

In conclusion, the in-house development of the census database not only offers numerous benefits but also represents a strategic investment in the organization's long-term success. By leveraging internal expertise, customizing solutions to fit specific needs, and maintaining control over the entire development process, organizations can create robust, efficient, and scalable solutions that meet the evolving demands of data management and dissemination.

3. Adoption of SDMX

SDMX 3.0 served as a cornerstone in the development of the census database, providing standardized tools, definitions, and recommendations for data management and exchange.

SDMX (Statistical Data and Metadata eXchange) serves as an important framework for standardizing statistical data management and exchange. It offers a comprehensive suite of standard tools and definitions tailored to the needs of statistical offices and data users. These tools encompass a range of functionalities, including guidelines for URL patterns, APIs, and strategies to overcome inherent technological limitations in data handling and distribution.

Moreover, SDMX offers a robust set of definitions, encompassing concepts such as data structure definitions, code lists, and metadata. These standardized definitions not only enhance the clarity and consistency of data representation but also promote interoperability and data comparability across different domains and jurisdictions.

SDMX is designed to streamline the automated exchange and processing of data and metadata among organizations, focusing on a typical data structure that can be mapped by various counterparties involved in the exchange. Data structure definitions (DSDs) and metadata structure definitions (MSDs) specify the concepts used and their representation, with XML schemas generated from structural metadata. SDMX provides equivalent formats for different technical use cases, ensuring versatility in data exchange. Salon & Sosnovsky (2010)

SDMX offers a diverse range of technical tools that, when utilized effectively, yield positive outcomes. Beyond being a standard for data and metadata exchange, SDMX serves as a framework for data discovery through web services, facilitating query, visualization, database creation, and metadata retrieval. It also provides guidelines for applying harmonized statistical concepts, codes, and classifications, fostering consistency and interoperability. SDMX's model enables the organization and classification of statistical metadata and data, supporting search, linking related material, and providing context. By structuring statistical material in standard formats, SDMX lays a strong foundation for automated tools to operate on data and metadata, enhancing data dissemination and accessibility. In essence, SDMX plays a key role in creating a powerful and flexible environment for statistical data dissemination and utilization. Eurostat (2015)

The SDMX information model, based on a star schema, represents datasets as multidimensional cubes, describing data structures through structural metadata, including dimensions, attributes, statistical concepts, and data structure definitions (DSDs). Structural metadata are administered and maintained by institutions managing the data. As Salon and

Sosnovsky (2010) mentioned in their paper the example of the ECB's Statistical Information Services Division for ESCB data. SDMX facilitates standardization and harmonization efforts in statistical data management, contributing to greater efficiency and interoperability within the statistical community.

Given the advantages highlighted earlier regarding SDMX, it was determined at the project's outset that SDMX 3.0 would furnish standard tools, definitions, and recommendations for technical implementation. Consequently, the development process encompassed not only the utilization of standard tools, such as SDMX guidelines, but also involved optimizing URL patterns, leveraging APIs, and implementing recommendations to overcome technological limitations.

4. Development of the database

Following the inception of the development concept in the latter half of 2022, strategic priorities were established. Foremost among these priorities were ensuring rapid query execution and crafting a user-friendly interface to enhance user satisfaction. Adopting open-source technologies was considered essential for enhancing economic efficiency, aligning with the utilization of standard tools, as previously discussed. Addressing limitations inherent in the Hungarian Central Statistical Office's existing dissemination database was also a focal point, particularly concerning constraints on the number of dimensions.

Achieving swift query executions emerged as a paramount objective, guiding the development process. Consequently, all calculations and aggregations were pre-executed prior to data loading into the census database. Efforts were made to maximize the number of dimensions, a crucial step toward enhancing user satisfaction. Technical constraints dictated a ceiling of approximately 14 dimensions, as the storage requirements exponentially escalate with additional dimensions. However, considering users' interpretative capacity, datasets were structured with a maximum of 10 dimensions, although certain datasets included up to 14 dimensions.

By early 2023, the initial version of the census database had been released, incorporating 2011 census data, prompting users to test the system and provide feedback. Based on feedback from both colleagues and external users, we iteratively refined the system to ensure optimal presentation of the 2022 census data in a user-friendly manner. Insights garnered from this feedback informed subsequent developments, shaping the evolution of the database.

5. Data pipeline

In our data pipeline, raw data is initially stored and processed using SAS rather than our production database. While the Hungarian Central Statistical Office maintains a metadata database as well containing census metadata, data and metadata for the 2022 census are transmitted in Excel files from the Population Census and Demographic Statistics Department. Upon receipt, metadata, already formatted in JSON, is leveraged to generate codelists and dataset definitions for backend processing. Notably, Node.js is employed for metadata conversions, ensuring efficient and seamless transformation processes.

This phase also encompasses defining attributes, executing various data conversions, and conducting data transformations using Python. Subsequently, data is aggregated using predefined scripts, with templates for operations, dimensions, and dataflows established in advance for streamlined processing.

In addition, our backend processing is powered by Java Springboot, providing robust functionality and scalability. Post-aggregation, indexing is applied to optimize database search times, further enhancing efficiency. Deployment of the processed data is then facilitated through scripted processes, with Docker utilized for efficient execution and deployment management.


6. Launch and Usability Enhancements

Through intensive collaboration with the Population Census and Demographic Statistics Department, the official Census Database was ready for the first data release in June 2023. The Population Census Themes Section compiled 17 datasets for settlements data publication, from which users downloaded over 17,000 tables in the first month alone. The second data release in September, which included personal data (including ethnicity and religion), further enhanced the database. Users could now choose from 45 datasets, resulting in over 47,000 downloads in September 2023 and over 50,000 in October. Since the initial release in June, there have been over 200,000 table downloads at the beginning of 2024.

The new census database boasts several usability enhancements, including quasi unlimited dimensions in multidimensional datasets and pre-execution of calculations for rapid query execution. Users benefit from instantaneous presentation of chosen dataset data, further facilitated by intuitive filtering functionalities (Figure 1).

The success of the new census database is evidenced by a notable increase in user downloads, surpassing 200,000 tables from 45 datasets by January 2024. This underscores the efficacy of the user-focused approach adopted in its development.

Figure 1: Census database

Census database [Home](#) 



Population by sex, age and district

Sex ▼

Age group, itemized ▼

District, county, region ▼³

Census year ▼

Transpose ▼ Clear filters ↺ Reset ⌵ Download ⌵   [Feedback](#)

Age group, itemized	Total	Under 15 years	Under 5 years	Under 1 year	1 year	2 years	3 years	4 years	5 to 9 years
District, county, region									
Census year: 2022									
- Budapest	1 685 342	209 733	67 166	13 504	13 630	13 724	13 158	13 150	69 64
- Pest county	1 333 533	225 399	73 911	13 842	14 824	15 383	14 817	15 045	75 51
- Tolna county	207 931	29 172	9 593	1 861	1 945	1 955	1 947	1 885	9 90
Census year: 2011									
- Budapest	1 729 040	210 640	80 067	16 078	16 154	16 485	16 006	15 344	68 65
- Pest county	1 217 476	204 269	68 754	12 294	13 324	14 187	14 745	14 204	69 50
- Tolna county	230 361	32 394	10 162	1 787	1 908	2 172	2 134	2 161	10 80
Census year: 2001									
- Budapest	1 777 921	227 622	67 894	14 364	13 506	13 196	13 103	13 725	76 58
- Pest county	1 083 877	193 067	57 731	11 439	11 035	11 344	11 558	12 355	67 60
- Tolna county	249 683	42 537	11 723	2 378	2 230	2 250	2 352	2 513	14 60

Initial feedback has been overwhelmingly positive, but we still plan to make minor to major improvements to the current system. Additionally, we plan to create predefined tables to alleviate the burden on users during table compilation. This can be envisioned as a menu where users can select table names. Clicking on the selected table name will display the requested table, which can then be used or modified by the user as if they had compiled the table themselves. One of the current shortcomings of the database is the display of metadata, which is currently only available in PDF format. In a future development, navigating to dimension names will display concise, understandable descriptions of the concepts.

7. Future Directions

Future plans include further developments on the existing census database and the creation of a new dissemination database based on its framework. This poses a significant challenge due to the vast number of datasets involved, estimated at around 1,000. Currently, it contains over 500 regularly updated datasets, but users can query about 1,000 datasets. Further development of the Census Database could also provide an opportunity to review existing datasets. For example, incorporating more dimensions is feasible, and there may not necessarily be a need to create new datasets in the event of methodological breaks as it is the case in our old dissemination database. The idea is to gradually include newer datasets through close collaboration with the statistical departments, progressing by statistical themes.

8. Conclusion

The launch of Hungary's new census database represents a transformative endeavour, showcasing the practical implementation of a user-focused and data-driven approach to statistical dissemination. By embracing open-source technologies, adhering to SDMX standards, and opting for in-house development, the HCSO has established a benchmark for modernizing data dissemination practices. The decision to pursue in-house development played an important role in ensuring the success and sustainability of the project. By leveraging internal expertise and resources, the HCSO was able to tailor the development process to meet specific needs and challenges.

Furthermore, the experience gained from the development of the census database provides valuable insights for future projects, particularly the renewal of the Dissemination Database. Building upon the foundation laid by the census database project, the HCSO plans to apply similar principles to the renewal process, all within the framework of SDMX standards.

In conclusion, the success of the census database project highlights the importance of in-house development, coupled with SDMX adoption, for advancing statistical dissemination practices. This experience provides valuable insights and opportunities for further innovation in this field.

References

- Eurostat (2015) Understanding SDMX. <https://sdmx.org/wp-content/uploads/SDMX-intro-MP-2015.pdf> , 2015.
- Gregory, A., & Heus, P. (2007). DDI and SDMX: Complementary, Not Competing, Standards. https://www.odaf.org/papers/DDI_and_SDMX.pdf , 2007.
- Salon, G., & Sosnovsky, X (2010). SDMX as the logical foundation of the data and metadata model at the ECB. IFC Bulletin No. 33, Irving Fisher Committee on Central Bank Statistics, July 2010.