

OPEN DATABASES FOR THE RESEARCH OF EDUCATION STATISTICS USING AI TECHNIQUES

There is a large number of existing databases, containing data on similar subjects. These datasets partially intersect, but also often have some data, that others lack. It is thus often necessary to search, collect and compile data from different datasets. However, different datasets, as well as access to them can be organized in significantly different ways. Let's look at the structure of some databases, relevant to the problems of decision-making in education:

EuroStat is a collection of datasets of statistics for EU countries, that has a lot of data about society, economy, environment, industry and other fields, and also some data on EU policy. The datasets relevant to education can be found in the "population and social conditions" category, and the datasets mostly show the number of people in a certain group, like students of a certain education level, academic personnel, people with a certain degree, or of some age group, etc. The data within these datasets is separated by a number of criteria and marked accordingly, to allow for grouping by a number of parameters, like sex, age, field of education and some others depending on a dataset. There are also datasets on financing and expenditures, that are organized in a similar way, and allow for filtering by education level, program orientation, sourcing, and other dataset-specific options. All of these datasets are accessed through a tree of categories leading to specific topics.

Organization for Economic Cooperation and Development or OECD Has a database that includes many of the same data categories as Eurostat's, but is structured in a different way, with more field being a separate top-level category. Relevantly for us, education is a top-level category, and not part of population and societal conditions, like in Eurostat's database. Also, OECD's data is not limited to EU countries. There is a lot of the similar data about students and staff and their distribution by various metrics, and data on the financial side: expenditure for education, both for the sources of funding and for the distribution of funds between educational institutions, about use of said funding, distributions and trends for salaries, etc. However, the data is organized in a different way. Instead of a tree of categories, leading to specific data tables that can be downloaded entirely or viewed on the website, specific datasets can be found on the website using tags and filters, and downloaded either entirely, or in a subset filtered by certain metrics.

European tertiary education register, or ETER is focused, as the name suggests, specifically on tertiary education. For that it contains data on the same parameters of institutions mentioned above, plus more data on scientific research

and specific programs like Erasmus, and generally more data on the tertiary education specifically. At the same time, the narrow focus means, that this register doesn't include the data about other levels of education, or anything outside of it. And once again, the data is organized in a different way. Unlike separate data tables in previous examples, ETER's data is in one big dataset, and their website offers a variety of categorized variables, which can be selected to pull specific data to show in the website's data viewer. User can find specific data through these variables and download a resulting table, or download the entire dataset.

As evident from these examples, similar data and user access to it can be organized in a variety of different ways. That can make simply searching for specific data overwhelming and time consuming. And even more so searching a dataset to find what data is not covered by others, and compiling data from different datasets into one. This is where artificial intelligence tools can be used. AI can be trained to comb through datasets, find and cross-reference the intersecting data, and use that to connect the unique parts of the datasets, compiling the required parts into one set.

REFERENCES

1. Eurostat's database: <https://ec.europa.eu/eurostat/web/main/data/database>
2. OECD Data explorer: <https://data-explorer.oecd.org/>
3. ETER Database: <https://eter-project.com/data/data-for-download-and-visualisations/database/>