How Many

connecting users to data

Yves Jaques

Frontier Data and Tech Unit

Section of the Chief Data Officer

Division of Data, Analytics, Planning, and Monitoring

9th SDMX Global Conference 2023, Bahrain

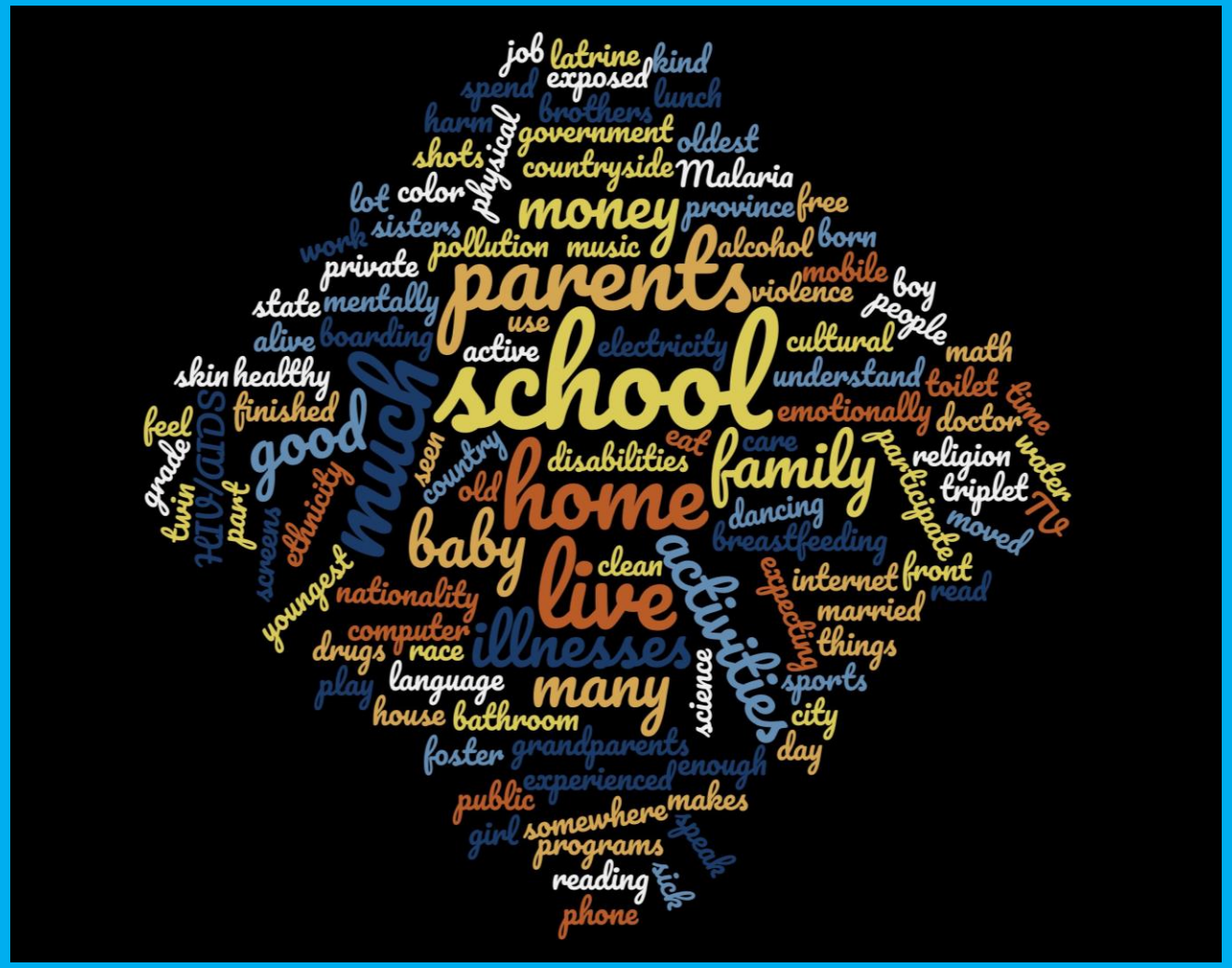
unicef | for every child

A language problem

- The statistical world uses precise but often complex terms for indicators and other statistical entities.
- Users looking for answers use common queries in plain language.
- Even if the user is used to statistical terms it may be hard for them to “guess” the exact definition.

Starting from the user side…

- We decided to analyze the keywords related to children with which users were querying search engines.
- It turned out we were missing out on a LOT of traffic.
- Users were searching for answers that we had, but we ranked low or not at all for many of their keywords.



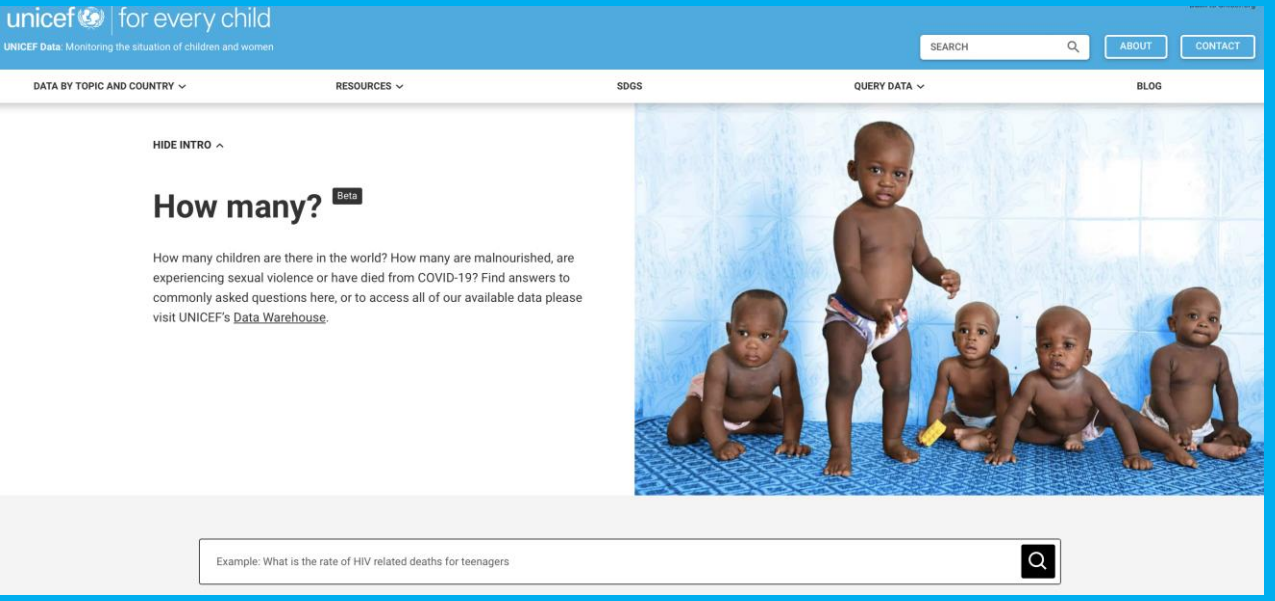
And users were missing out on our gold standard data…

- They were also often getting bad results from other sources!
- We simply couldn't answer the questions expressed by users in their own language.
- So:
  - How can we understand the user's questions and give the correct answer?
  - We needed a way to find a match between common questions and formal definitions.

Looking for solutions

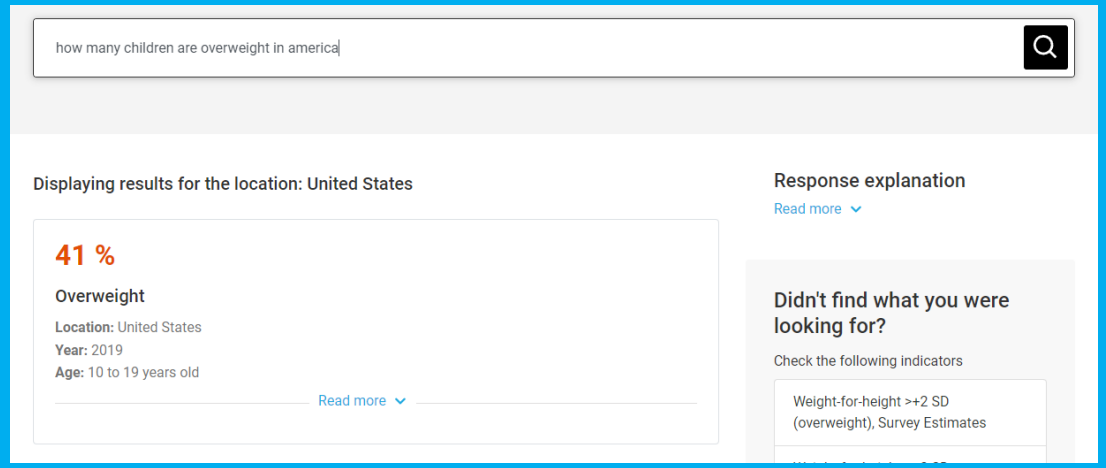
- How to get a “good enough” match between user questions and our statistical elements such as indicators, disaggregations, geographical areas…
- How to present precise and closely related answers when users “landed” on our site through a search engine?
- In summary:
  - How to capture all that traffic!
  - How to retrieve precise responses to common queries?

https://data.unicef.org/how-many/



- In short, we created the **How Many** tool
- It uses a Solr search engine with synonyms and OpenNLP (natural language processing).
- The interface is a Google-like simple text input where users can type their questions.

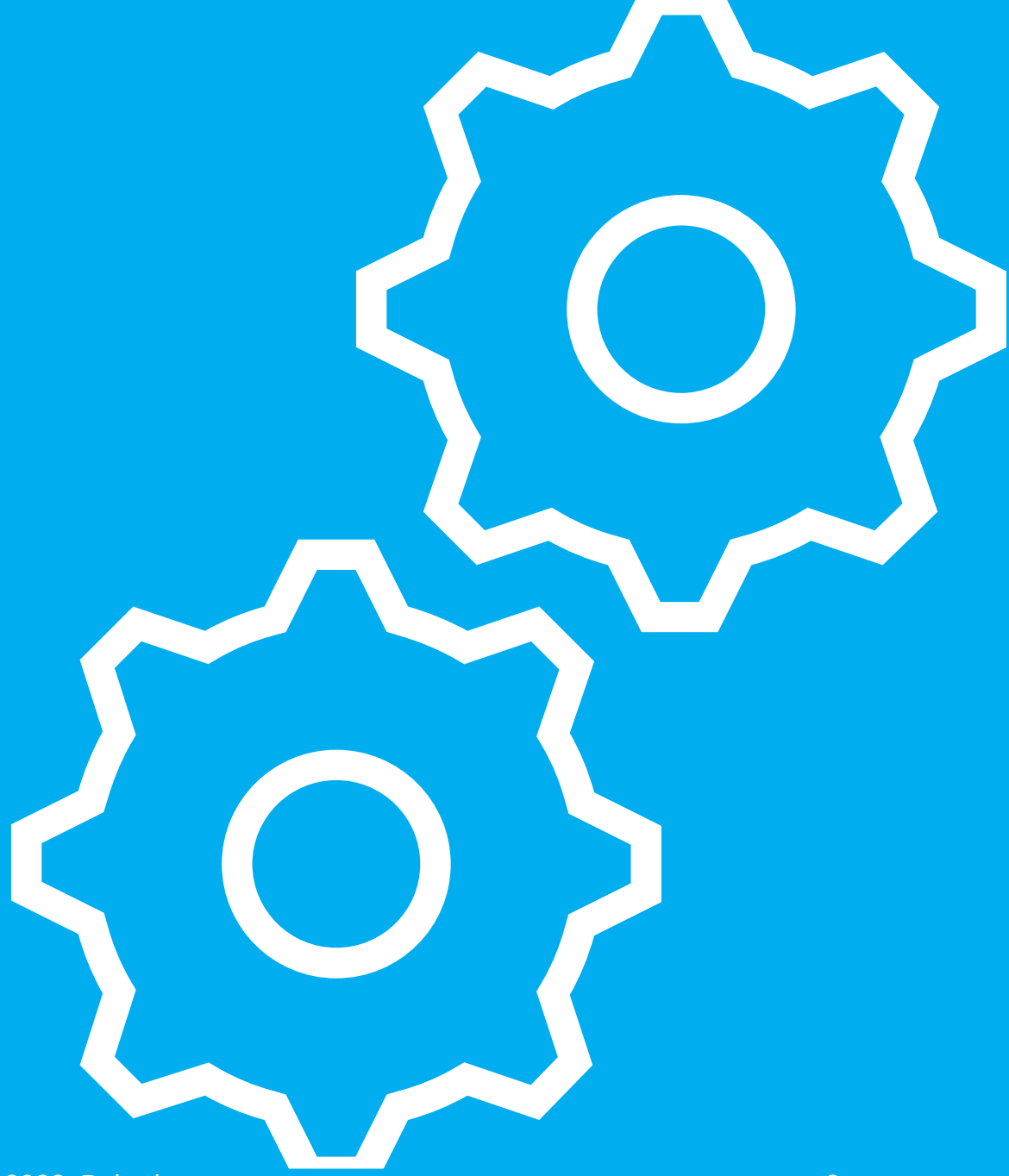
What the user sees



- The user types a question
- The application finds the most suitable indicator and geographic area matching the user's request.
- The application returns a value and a few related suggestions that can help refine the query or further explore their subject area of interest.

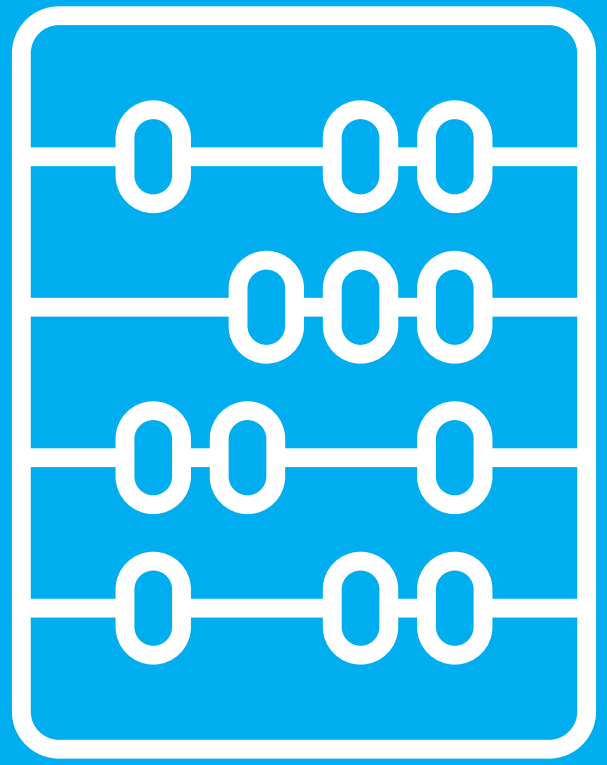
The data implementation

- We created a Solr index containing all the data points as “documents”.
- Each data point is augmented with the metadata contained in our SDMX registry, plus additional metadata stored in other repositories (we have a system dedicated to reference data and reference metadata).
- The metadata enrichment helps frame the “meaning” of the question.
- OpenNLP is used to find the most suitable (and most related) data points.



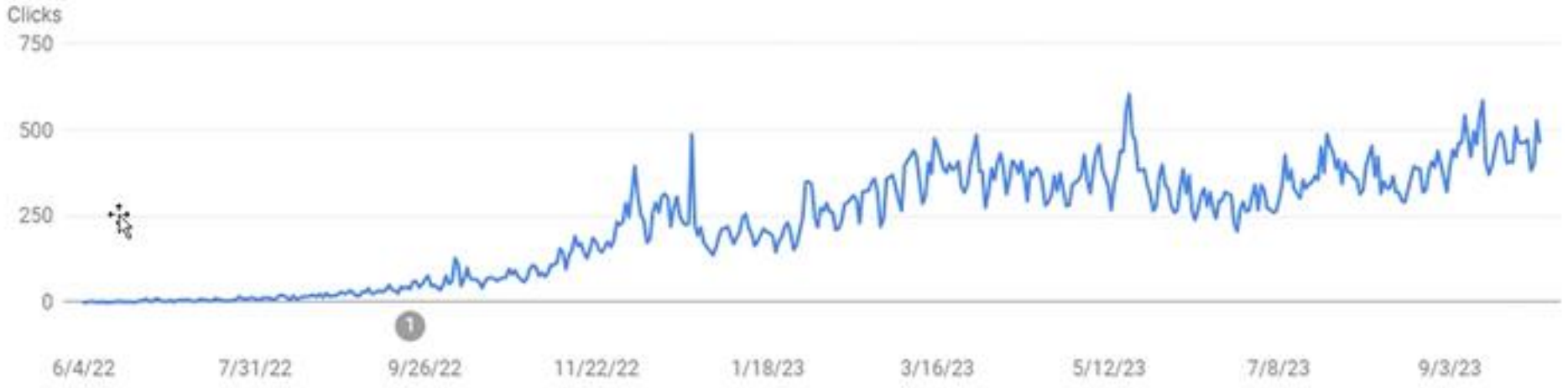
The way it executes

- The NLP algorithm can understand the synonyms and find the most suitable indicator.
- As an example, a common Google query such as “**How many kids are fat in America**” will be matched with the indicator: **“Weight-for-height >+2 SD (overweight)”** for the USA, most recent observation.
- As you can see, the algorithm frequently captures informal “slang” words and can also give good guesses at geographical entities and translate them (e.g., **America = United States of America**).

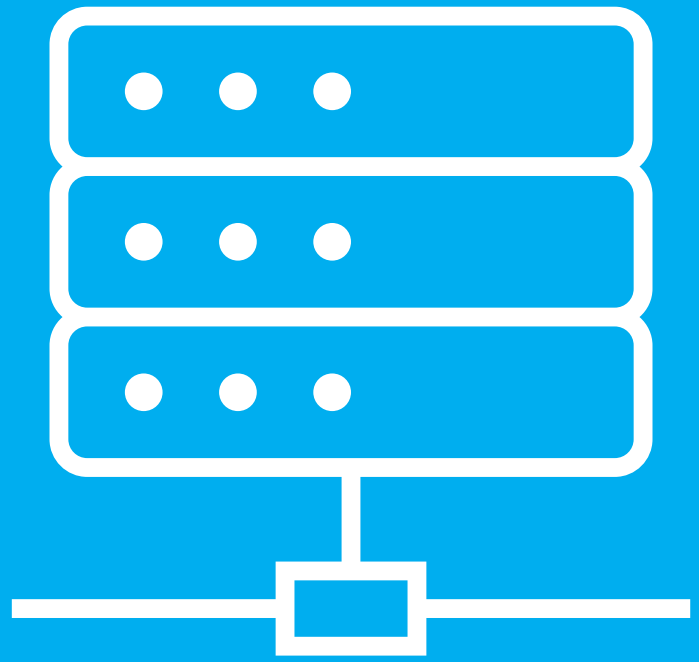


‘How Many’ tool

- Has delivered over 110,000 visitors to the website
- Now brings in ~6% of all SEO traffic to data.unicef.org!
- Now ranks #1 for a number of terms, including:
  - ‘How many boys are there in the world?’
  - ‘How many girls are there in the world?’
  - ‘How many children have been vaccinated against polio?’
  - ‘How many women have had the HPV vaccine?’



Next steps…



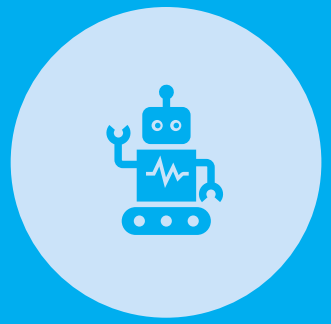
- No AI techniques are used (yet…) but this is our obvious next step when we get greenlit for such approaches!

Summary

We can now answer questions expressed in “users’ language”

We’re driving WAY more traffic to our data.unicef.org site

We are still native SDMX with strong data management baked into our work.



We are well positioned to slot in an AI large-language model when our agency is comfortable.