Data Basis (*Base dos Dados*): Universalizing Access to High-Quality Data*

Ricardo Dahis[1,4], João Carabetta[2,4], Fernanda Scovino[2,4], Frederico Israel[3,4], and Diego Oliveira[2,4]

[1]PUC-Rio
[2]Rio de Janeiro City Hall
[3]Google
[4]Data Basis

July 5, 2022

Abstract

In this paper we explain how the Data Basis platform helps decisively solve the data access problem for different types of users. We describe its core products: a powerful search engine, a freely accessible data lake featuring a unified schema and hundreds of interoperable tables, and APIs in various programming languages. We exemplify the platform's utility with discussions of three datasets on labor markets, elections, and local public finances in Brazil. The project is extraordinarily cost-effective: dividing a measure of yearly benefits generated by a conservative estimate of yearly costs to run the organization yields a lower bound social return of 74. We conclude by laying out a roadmap to guide the organization's future steps.

JEL classification: C8, O12

Keywords: open data, big data, administrative data, search, data lake

*We thank all Data Basis' donors and collaborators who have helped turn an old shared dream into a reality. Corresponding author: Ricardo Dahis (rdahis@econ.puc-rio.br).

Contents

1 Introduction

Data is a critical input to a variety of activities. Governments use it to guide public policy. Researchers use it to describe reality and to test theories with statistical analyses. Companies use it to track markets, to inform business decisions, and to innovate over new products. Journalists use it for fact checking, to uncover new stories. Non-profits use it to design reform agendas. In summary, users (i.e. decision-makers, researchers, journalists) need data to answer questions (e.g. "how did the economy perform last year?", "what neighborhood should receive government transfers?", "what model can better predict the weather?") and to make decisions.

Yet, accessing data in a state ready for analysis is often tremendously costly, if not outright impossible: as illustrated in Figure 1, the distance between a question and an answer can be long. Searching for data online is not easy. Users often do not know exactly what dataset they need, what datasets exist, or who exactly provides them. Government websites have low-quality unreliable layouts that make finding data difficult. Online addresses change frequently and little or no useful metadata is provided. When metadata does exist, it is often incomplete, outdated, or wrong. Auxiliary files, such as dictionaries, tutorials, or reports, are often confusing, outdated, difficult to parse, or absent altogether.

Even conditional on finding the data one needs, other important challenges still remain. Downloading data can be difficult when files are large or when providers' servers are slow and unstable. Loading data locally can be costly when it comes divided in dozens or hundreds of files in archaic formats, or when its total size is larger than one's local RAM memory. Most importantly, users still need to clean the data, which can require months of work. Each user repeats the same process, each time with new bugs being introduced to the code.

Despite the progress observed in recent years, as we will discuss below, the data access scenario worldwide is still dire. And there are understandable reasons behind this situation: there is virtually no institution with both the incentives and the technical and organizational capabilities to provide high-quality data at scale to the public. First, data is a public good whose provision generates substantial positive externalities. As is well understood by economists, this implies that its market provision will be sub-optimal. Companies collecting and organizing it will charge for access; researchers have no career incentives to release and maintain data. Second, for the entities with incentives aligned to social welfare, such as governments and non-profits, there are enormous technical, organizational, and political challenges to high-quality data provision. Technological capabilities are low on average. Data is produced in vast amounts by a variety of agencies which often do not communicate

or are unwilling to share. Turnover of staff makes public data provision extremely unstable. For example, decisions of what data to collect and provide, how to name files, or even what platform to use, change every few years.

In this paper we introduce the Data Basis[1] project, which simultaneously solves all problems listed above and, thus, attempts to reduce the distance between users' questions and answers to zero. Data Basis is a non-profit organization founded in 2019 with the mission to universalize access to high-quality data.[2] We aim to promote economic development and social welfare by supporting evidence-based policy making, science, and rational discourse.

The organization exists at the intersection of four communities: modern open-source software development, data science, academic research, and policy-making. We develop three core products that help us take data access to an unprecedented level of quality and scope: a search engine, a data lake, and APIs in various programming languages. Our search engine allows users to quickly find the data they need. The unified schema and filters powering it make searching for datasets easy. Our data lake provides users access to hundreds of structured, normalized, and up-to-date tables. Data are readily available for joins, aggregations, selections, and analysis in general. Users can query data at scale for free without ever having to download files locally. Lastly, we provide APIs in Python, R, Stata, and Excel that make accessing data a matter of two or three lines of code. All our software is open-source on GitHub.[3]

As we detail in Section 5, the organization generates an extraordinarily high benefit-cost ratio of 74. That is, for every unit of currency invested in Data Basis, society receives back *at least* 74 times that, as measured by the value of time saved for our users. This happens because our impact is highly scalable: a well-cleaned dataset made available once can be accessed by an arbitrarily large number of users forever. In fact, our impact is growing rapidly: in 2021 our data lake received about 400,000 queries, and by the end of 2022 we project that number to grow to at least one million. The costs to run the organization, on the other hand, scale at a much lower rate.

Our platform is not the first or only project out there providing structured data online, but we believe we make unique contributions to the landscape of open data initiatives. First, we provide a powerful curated search engine of universal scope that returns meaningful results to users. While other initiatives offer a search engine covering datasets within their platforms, only Google Dataset Search has ambitions for universal scope. The tool relies on dataset

[1] In Portuguese it is called *Base dos Dados*. In Spanish it is called *Base de los Datos*.

[2] The organization's website in Portuguese is at `https://basedosdados.org`.

[3] See `https://github.com/basedosdados/`.

providers' providing metadata in accordance with the standards defined by the schema.org consortium. Results returned by its search, however, have little texture and are not very useful in practice. Our search engine benefits from internal curation and from a layout better attuned to user experience to yield them a high-quality service.

Second, we provide a novel SQL-powered normalized serverless database where tables share a unified schema and can be freely manipulated at scale. Other open data initiatives offer at most a subset of these features. For example, some initiatives provide a rich catalog of datasets with documentation and some degree of schema unification, but offer no platform for SQL data manipulation. This set includes IPUMS [Ruggles et al., 2021] and the Socioeconomic High-resolution Rural-Urban Geographic Platform for India (SHRUG) [Asher et al., 2021]. Another set of platforms provide structured repositories for individual data providers to upload their data. They offer general documentation and other features, but they impose no schema requirements on authors. This includes the World Bank Open Data platform, a variety of government data websites, the Dataverse, Zenodo, and Kaggle. While we mainly focus on providing structured data, one project that stands out instead by producing high-quality content based off data and research is Our World in Data (OWID). They also structure and provide datasets, but their focus is text entries and visualization, as exemplified by their entry on CO_2 emissions [Ritchie et al., 2020].

To the best of our knowledge, the platform that most approximates ours is called DataCommons.[4] They aggregates data from a wide range of sources into a unified database. Their universal schema allows for structuring data as a graph, which can then facilitate manipulation and visualization. While some aspects of DataCommons are currently more advanced than Data Basis', our differentials are providing our data through a SQL database, hosting a variety of large-scale administrative microdata, and more closely interacting with decision-makers in policy and other sectors in search of social impact.

The remainder of this paper proceeds as follows. In Section 2 we describe the organization's basic business model and organogram. In Section 3 we discuss our core products in more detail. In Section 4 we outline how we apply our technologies to three examples of datasets covering labor markets, elections, and local public finance in Brazil. In Section 5 we perform a benefit-cost analysis of the project. In Section 6 we conclude and lay out a roadmap to guide the organization's future steps.

[4]See the platform's website at `https://datacommons.org/`.

2 Our Business Model

Data Basis is a non-profit organization currently based in Brazil. Our goals are not to maximize profits, but to do the most good possible following our mission to "universalize access to high-quality data." We developed a business model that both aligns with this goal and promotes sustainability and growth over time.

Our efforts are split into development of core products, fundraising, the execution of projects, and the delivery of services to clients. These actions are complementary: better core products improve the quality of projects and services, and vice-versa. For example, producing online educational material enlarges our user base, which increases our fundraising capacity and helps improve our core products.

Each of our core products solves one or more components of the question-to-answer problem illustrated in Figure 1. Users can use our search engine to easily find data, and can use our data lake platform and APIs to directly download and merge data. All other steps, such as loading, cleaning, and processing are all taken care of by our methodologies described in Section 3 and by Google Cloud's BigQuery capacities.

Our core products and the data expertise generated within the organization then undergird our fundraising and service provision efforts. Our fundraising initiatives' purpose is to obtain core long-term funding to lengthen our survival horizon and to allow for more ambitious goals. Our service initiatives are divided into data engineering, data analytics, and consulting.

We structure the organization in teams following two dimensions: one function-oriented and another product-oriented. So far our function-oriented division has happened in six teams: data, infrastructure, communications, fundraising/projects, website, and administrative.

The data team's functions include populating and maintaining our core products with data and metadata; supporting our teams in fundraising/projects and communications with data analytics demands; interacting with infrastructure team to incorporate new tools into the pipeline; and providing support to the community answering data questions and eliminating bugs. The infrastructure team works on developing and maintaining our API packages in Python, R, Stata, and Excel; maintaining our orchestration infrastructure for data ingestion and other operations; and developing our website's back-end validation and APIs.

The communications team manages our interaction with the outside world, including managing our social medias; producing and publicizing content based on our core products, methodologies, partnerships, and services; communicating with donors and supporters; and

producing institutional material such as our annual reports. Our fundraising/projects team is responsible for finding new clients and scoping projects; managing ongoing projects along every dimension (scoping, choice of squads, executing, delivery of results); writing grant proposals; and communicating with current and potential donors. The website team manages all matters related to our website's front-end, which involves designing visual and textual content, writing code, and coordinating with other teams how content will be displayed online. Our administrative team is responsible for dealing with the organization's legal and admin issues, such as financial management, contracts, external audits, among others.

As we discuss in Section 5, our cost structure is divided into three broad categories: payroll, management, and infrastructure. Payroll accounts for the majority of expenditures on a monthly basis, while management accounts for occasional legal or related bills. Our infrastructure expenditures account for a small percentage of the total budget. They are concentrated in cloud services such as Google Cloud Platform (GCP) for virtual machines, storage, and processing of our data. Thanks to GCP's computational capacity and billing model where datasets can be made public and where users pay for their own data processing (with a free 1 terabyte per month-user), our costs scale only with our storage needs. We do not need to worry about setting up and maintaining large servers, which drives costs down significantly.

3 Providing High-Quality Data at Scale

In this Section we describe each of our core products and how they help us provide high-quality data at scale. We also describe the auxiliary processes and methodologies developed in-house that sustain each of the products.

3.1 Search

Searching for data online is not easy. Users often do not know exactly what dataset they need, what datasets exist, or who exactly provides them. Government websites have low-quality unreliable layouts that make finding data difficult. Addresses change frequently and little or no useful metadata is provided. When metadata exists, it is often incomplete, outdated, or wrong. Column metadata is usually absent so the only way to maybe understand what is available inside the dataset is to download it all and assess locally. Auxiliary files, such as dictionaries, tutorials, or reports, are often confusing, outdated, difficult to parse, or absent altogether.

We have developed the Data Basis search engine to solve all these problems at once. With a mix of expertise and curation, high-quality metadata, a search engine with filters, and universal coverage, our search allows users to quickly find the data they need with high confidence.

3.1.1 Metadata

Good search requires high-quality metadata and filters, and at Data Basis we take that seriously. We have developed a unique ontology and schema with complex fields that serve as basis for our ecosystem. The ontology is organized around *datasets*. Each dataset has its metadata (e.g. title, description, themes, tags) and may contain one or more *resources* such as *tables*, *original sources*, and *information requests*. Tables have their own metadata, such as spatial and temporal coverage, observation level, and update frequency. Tables contain columns, which have their own metadata (e.g. name, description, foreign key, measurement unit). Original sources' metadata include the download URL, the license type, whether access is free, and the spatial and temporal coverage. An information request's metadata includes the request URL, status updates, a link to the response and the data provided, among others.

Next, we discuss two examples of how complex metadata fields power our search engine. First, we keep track of resources' and columns' spatial coverage with a list of area ID strings. For example, the spatial coverage for a table covering Brazil as a whole is `["sa.br"]`. The field for a table covering the states of Minas Gerais and São Paulo is `["sa.br.mg", "sa.br.sp"]`. Together with a complete list of area IDs structured in a hierarchy as `"<continent_id>.<country_id>.<admin1_id>.<admin2_id>"`, the search engine then returns the datasets *covered by* a specific selection of keys. For instance, when a user filters for datasets covering the United States, the search engine will return results that cover the whole world, the North American continent, and the full United States. datasets only covering specific American states will not be displayed.

Second, we document a every resource's observation level. In the case of a table, this field lists what entities are represented by a given row in the data. For example, tables may have each row representing a municipality-year or a country. For each resource, we structure its observation level information in our metadata database as list of dictionaries. Each dictionary has three optional fields: `country`, `entity`, and `columns`. We maintain a list of allowed entities in our back-end to serve as validation. For example, this list includes entities such as `country`, `state`, `year`, `month`, `person`, and `company`. Users can then use our

search engine to filter datasets for exactly what entities they are interested in.

Or, together with our other fields and to take an example, users can search for datasets about education, covering Brazil as a whole, with data at the state level, covering the years of 2014 to 2019, and updated yearly. Searching for data has never been this easy, and the possibilities here are endless.

3.2 Data Lake

Even conditional on finding the data one needs, other important challenges still remain before a user can answer questions with it. Downloading the data can be difficult when files are large or when providers' servers are slow and unstable. Loading the data locally can be costly when files come in archaic formats, or when they are split into dozens or hundreds of separate files, or when files are larger than one's local RAM memory. Most importantly, users still need to clean the data, which can take months of labor. Each user repeats the same work, potentially introducing new bugs in the code each time. The end product is often of low quality, not well documented, and effectively useless to the rest of the community.

We have developed our data lake and auxiliary technologies to solve all these problems at once. We provide a structured, normalized and up-to-date data lake where users can quickly query high-quality data virtually for free. We structure data from a variety of sources under one common schema following a set of best practices and our own style manual. The end-result is a data base with hundreds of tables ready for joins, aggregations, selections, and analysis in general.

This data lake has a series of other advantages. Scale is virtually limitless given Google Cloud's serverless infrastructure. There are specific tables with about 300 gigabytes of data, and users can query terabytes at a time with running times of seconds or minutes. Moreover, users can process data directly on the cloud using BigQuery and can thus avoid downloading and processing massive datasets locally. Finally, our tables follow Hive partition guidelines and are optimized for reducing data processing costs when users only need information from say a specific state or year.

The data lake is organized around datasets and tables, but it has an underlying structure of entities. As described in Section 3.1.1, entities are ontological objects for which data refers to. Each entity in principle has a corresponding *directory* table in the data lake with a primary key. For example, a table called `municipality` located inside the Brazilian directory represents a Brazilian municipality. Its primary key is named `municipality_id`. The analogous is true for other entities such as `year` (with directory and primary key named

`year`) and `country` (with directory named `country` and primary key named `country_id`). And to ensure interoperability and to allow all tables to be easily joined with each other, we maintain consistent naming for foreign keys whenever possible across the whole data lake. In other words, users can effortlessly join any two tables that have a municipality as one of its entities on the column `municipality_id`.

Moreover, we maintain dataset-specific dictionaries whenever at least one of its tables' columns are categorical strings. Having dictionaries allows us to encode string columns to numbers in the main tables and thus to save storage space and processing cost. We compile and organize dictionaries from original data sources, which are often outdated and confusing, into one coherent structure. Each dictionary *key* is uniquely identified by a `table_id`, a `column_name`, a `temporal_coverage`, and a `key`. Its values may be null or any string.

Finally, we perform automatic *data checks* to help us catch errors and maintain high data quality. For instance, based on the identifying columns reported in the metadata subfield `observation_level.columns`, we check what percentage of rows are uniquely identified in the data. We also automatically compare primary keys to those in our directories and throw errors when there are discrepancies.

3.3 APIs

Our third core product is the set of APIs used for interacting with our platform. We maintain packages in most of the relevant data science languages: Python, R, Stata, and Excel. These packages contain functionality to allow users to search for datasets, to add and modify metadata, to upload and to download data, and to configure the required GCP authentications. All code is open-source on GitHub.

Our APIs provide users and our internal teams unparalleled simplicity when interacting with data. They can, for example, download large-scale high-quality tables with two lines of code: one for loading our package and one for calling the download function with specific `dataset_id` and `table_id` parameters. Downloading data generated by arbitrary queries takes only one extra line.

3.4 Automated Workflows

The work we perform at Data Basis benefits from automations across various strategic areas such as data engineering, data analytics, and website deployment. We believe automation is the only path to large-scale operations at a manageable cost.

And implementing automations is one of the main responsibilities of the infrastructure team. They are constantly searching for opportunities to automate workflows employing modern open-source technologies.

3.4.1 Development and Production Environments

Maintaining and developing a stable, large-scale, high-quality data and metadata platform would be challenging in the absence of quality-control mechanisms. Having all additions and modifications to our data and metadata become immediately public introduces risks of introducing mistakes, our team inadvertently deleting prior work, among others. To deal with such problems, as described by Figure 2 and analogously to the best practices in software development, we have structured our data work into two separate environments: one for development and another for production.

The two environments have mostly the same structure, except for a few publicity and permission configurations. The production environment holds the data and metadata we judge to be "ready for publication". It is public on our website and BigQuery, and is unmodifiable by non-admin users and team members. On the other hand, the development environment's purpose is exactly that of development and experimentation. It is within that environment that we perform quality checks in our new data, run test queries, and catch bugs. We have an internal development environment, but external users can replicate it to upload their own data to our platform in production.

The two environments interact via GitHub Actions as follows. First, a user performs our data upload steps to add new data to BigQuery and to open a pull request in our GitHub repository. Second, with the pull request open, our team then performs quality-control checks, both automatic and manual. For example, we search for bugs in the code and perform basic queries in the data to map empty columns or invalid primary keys. Finally, once the pull request is approved, we merge it into our main branch. An action with appropriate permissions is then triggered to copy the data (and other information such as metadata, raw files, and architectures) from development to production. The end-result goes public to the wider community as soon as the action finishes running.

3.4.2 Orchestration and Pipelines

Providing always-up-to-date data at scale is only remotely feasible with a robust system for orchestration of pipelines. At Data Basis we employ an orchestration system based on modern open-source technologies such as GitHub, Prefect, Kubernetes, and dbt.

The system not only is robust, but it is also scalable and flexible. We use Prefect to schedule *flows*, each running a given list of *tasks*. Tasks can be about virtually anything, from full Extract, Transform and Load (ETL) pipelines to publishing treated tables on our production environment or posting analyses on social media.

4 Data Use Examples

In this Section we discuss a few use examples of our core products. We briefly discuss each dataset's history, how our work significantly improved its data quality and make access to it easier, and illustrate powerful analyses performed with them.

4.1 Labor Markets

The *Relação Anual de Informações Sociais* (RAIS) dataset is one of Brazil's public data gems. The dataset has been maintained by the Ministry of Labor for decades, and contains the valuable information on formal labor markets undergirds official annual government reports. Because of its broad coverage (i.e. data starts in 1985) and detailed information on workers, firms, and employment spells, it has also been widely exploited for research purposes in Economics, both in Brazil and abroad.

The dataset has both a public and private version (with identifiers). But accessing even the public version has always been difficult. First, the dataset itself is massive: each year since 1985 list dozens of millions of observations with a variety of columns. The microdata table alone has a total size of more than 300 gigabytes. Second, the Ministry of Labor only provides the data via an FTP server restricted to Brazilian IPs. Third, the data is split into hundreds of files, with no standard naming standards or common schema. Finding the data is not easy; downloading it all is not easy; loading it requires a powerful server; cleaning it all can take months.

We have employed the set of steps and practices described in Sections 2 and 3 and have made the full public RAIS available in our platform. Running queries on these hundreds of gigabytes now takes only a few seconds. As an example, we calculate the college premium for Brazil between 2006 and 2020. The query calculates the ratio between average wages for college-educated and high-school-educated workers. It processed 331.97 GB in 203 seconds on BigQuery. The result in Figure 3 documents that the college premium in the Brazilian formal labor market was about 200% until 2012, fell to about 175% in 2018, and grew again until 2020.

4.2 Elections

Brazil has fantastic elections data centrally provided by the *Tribunal Superior Eleitoral* (TSE). The data cover the universe of results and candidate's characteristics in federal, state, and municipality elections since 1998, and various other aspects (such as campaign revenues and expenditures, characteristics of the electorate, and number of legislative seats) with varying degrees of quality and coverage, dating back to 1945. Similar to RAIS, Brazilian elections data have been widely used in research, journalism, and other applications in Brazil and internationally.

Despite its many qualities, the data provided by TSE still have limitations bounding usage to its full potential. For example, accessing the data is made difficult by users having to download hundreds of separate files with with little standard naming standards or common schema. Moreover, appropriately cleaning the full dataset requires significant prior knowledge of the mechanics of elections and can also take months. To make matters worse, TSE uses its own municipality identifier that is different from the standard provided by IBGE. Finally, the candidate identifiers TSE provides have problems and do not serve as primary keys: there are name misspellings, wrong CPF numbers, and duplicate observations everywhere. The more trustworthy identifier in the data, named `sequencial_candidato`, only identifies candidates *within* elections, but not across years.

As with RAIS, we have employed the set of steps and practices described in Sections 2 and 3 and have made the majority of TSE elections data available in our platform. The version we provide is of significantly higher quality than the original. First, our tables are normalized, structured with a common underlying schema, and stack all years of available data. Second, we have employed our municipality directory described in Section 3.2 and merged the standard `municipality_id` column into all tables. Lastly, we have developed a novel candidate identifier allowing users to effortlessly track 99.51% of candidates over time and space since 1998.

Our Brazilian elections dataset allows for an unpredictably large set of possible applications and research. To give some simple examples of its potential, we have calculated two important statistics. First, in Figure 4 we display a gender representation index for local legislative chambers in Brazil. The index is defined as the share of legislators being female divided by the population share of females. The Figure documents that very few Brazilian municipalities display proportional gender representation (measured by 100%). The index average across years is 24.3%, while 17% of municipalities in 2020 had zero elected female legislators. Second, with data for the 2020 elections, we find that 37.5% of municipalities

had two or less candidates for mayor. And 2% had only one.

4.3 Public Finance

Data about municipality public finance in Brazil are organized and provided by the *Tesouro Nacional* (TN). Ranging two different systems called *Finanças Brasileiras* (FINBRA) and the *Sistema de Informações Contábeis e Fiscais do Setor Público Brasileiro* (SICONFI), there exists data on municipalities' revenues, expenditures, and balance sheets since at least 1989. These data yield a rich picture of Brazilian local governments, which has been frequently studied and reported on.

The current way TN makes the data available also has many of the standard data access problems described in Section 3. The FINBRA data for 1989-2012 is available in an obscure web address, and comes only in Microsoft Access format. The SICONFI data from 2013 onwards has to be downloaded one table-year at a time. There is no common schema between years, and extracting and cleaning the relevant data can take months of work. Additionally, there is enormous variation in how accounts are structured over time. The hierarchy of accounts changes, their names and spellings change, and there are no account identifiers.

As with the previous examples, we have generated and provide a version of SICONFI that deals with all the problems mentioned above. We stack all years of data into few normalized tables, we deflate currencies to Brazilian real, and we generate unique identifiers for accounts down to the third level in the hierarchy. This allows users to track individual municipality-account pairs over time since 1989.

As an example, in Figure 5 we document a negative relationship between expenditures as a share of GDP and log GDP per capita for Brazilian municipalities in 2019. To generate this binned scatter, we have incorporated two other important datasets, municipality population and GDP. This Figure shows that richer municipalities have more fiscal slack in Brazil.

5 Benefit-Cost Analysis

The work we do at Data Basis has an extraordinarily high benefit-cost ratio. In this Section we calculate this number based on estimates of the benefits we generate and the costs of running the project. We then discuss how our effectiveness compares to other non-profits and philanthropic projects.

5.1 Benefits

Data Basis generates benefits across a range of outcomes, only a part of which is measurable. Among the less measurable benefits, our work nurtures a culture where individuals use data to answer important questions. It supports evidence-based policy-making, particularly to small and low-capacity governments. It increases the quantity and quality of science and journalism, with less time being spent on cleaning data and less bugs surviving down the analysis pipeline, and with there being more trustworthy data to base research on. It empowers data-driven education in schools and universities worldwide with students learning data science based on real-world high-quality data. Our core products serve as an example to society on what is the gold standard for data access, and helps clarify a path for governments and other organizations to get there. Our work ultimately strengthens democracy by promoting transparency and accountability to government actions.

One measurable benefit of our work is the value we generate to users in terms of time saved in cleaning data. This benefit B can be approximated as

$$B = Q \times A \times T \times W$$

where Q represents the total number of queries run on our data lake, A represents a test-query adjustment factor, T represents the number of hours saved per query (e.g. in data pre-processing), and W represents the hourly wage earned by our users.

Measuring Q is easy based on data lake usage records. In 2021 we received about 400,000 queries. For 2022 we expect receiving at least one million queries: between January and June alone we have received about 580,000 queries. Therefore we set $Q = 1,000,000$.

We estimate A and T based on some assumptions. First, because users often run test queries before running a "final" query, we incorporate a "test-query adjustment factor" of 2 to 1, or set $A = 0.5$. Second, we assume that a "final" query to our data lake saves our users on average 10 hours of data pre-processing time. In other words, we set $T = 10$.

Finally, we estimate W based on our knowledge of our user base and current wage estimates in the data science labor market. Our user base is composed of various groups: government officials, senior researchers, data scientist in various capacities, graduate and undergraduate students, and others. Currently about 94% of website users reside in Brazil, with 1% being in the U.S., 1.5% being in Portugal, and the remaining spread around the world. The current age profile is 27.5% between 18-24, 33.5% between 25-34, 15.5% between 35-44, and the remaining above 45 years old.

We base our wage estimates off our knowledge of what our average user, namely data scientists and economists, earns in the labor market. Monthly wages for data scientists and developers in Brazil today can be as high as 6,000 BRL for recent graduates, and 12,000 BRL for more senior analysts. Taking into account that a share of our users are younger students, we make a conservative assumption that our average user earns 4,000 BRL and works 180 hours per month. In other words, we set $W = 4000/180 = 22.2$.

In sum, for the year 2022 alone, we estimate $B = 1{,}000{,}000 \times 0.5 \times 10 \times 22.2$ BRL $= 111{,}000{,}000$ BRL, or about 22,200,000 USD. As discussed above, given that we cannot properly measure many of longer-term and more abstract benefits of our work, this number is simply a lower bound to the true total benefit we create.

5.2 Costs

Our total economic costs are divided into what is effectively paid for and what the organization receives for free (e.g. volunteer work and cloud computing credits). As we will detail below, measuring the former is easy based on our financial accounting, whereas measuring the latter is harder. Consequently, the cost we estimate below is a lower bound on the yearly cost to run the organization.

We project the total long-term yearly cost to have our operations and teams be fully professionalized and paid for, at our current scope described in Section 2, based on our observed costs in 2021 and 2022. In 2021 we spent about 127,000 BRL. A significant percentage of the work hours invested into the project were still unpaid for, as volunteer work done by cofounders and other volunteers. In 2022 we expect to spend a total of 550,000 BRL. Not only has our team significantly grown since 2021, with new hires across the board, but we have taken measures to more fully account for costs we previously received as donations or volunteer work. Today we spend about 82% of our total costs on payroll, 10% on infrastructure variable costs, and 8% on legal and administrative matters.[5]

Assuming Data Basis remains at its current scope and size, we conservatively project our long-term yearly costs to be around $C = 1{,}500{,}000$ BRL, or 300,000 USD.[6] This already incorporates, for example, larger and better-paid teams, full-time employees at key managerial roles, and higher usage of infrastructure and services. This number ignores, however, any

[5]Importantly, a large share of our potential infrastructure cost (i.e. for maintaining large scale servers which can support hundreds of thousands of queries to the data lake) is currently borne by Google. Naturally, if GCP went bankrupt or changed their business model, we would need to reassess our estimates of infrastructure costs.

[6]Assuming an exchange rate of 5 BRL to 1 USD, according to the current economic scenario.

potential change to the scope of our activities, such as developing new products or making the project international.

5.3 Ratio

Given the estimates provided above, we arrive at a benefit-cost ratio of $B/C = 111{,}000{,}000$ BRL / 1,500,000 BRL = 74. In other words, for every 1 USD or BRL spent in the project, society gets an extraordinarily high return of at least 74.

A return this high is explained by at least two reasons. First, our core products are basic, low-level public goods with enormous positive externalities. Demand for high-quality public data is widespread across sectors like academia, government, journalism, education, private enterprise, and civil society more broadly. Second, our impact is highly scalable: access to data happens over the internet and is based on powerful cloud infrastructure that is cheap and reliable. Once a dataset has been made available in our data lake, it can be instantly accessed by millions of users worldwide at no extra cost to the organization.

6 Roadmap

This paper introduces and discusses various aspects of the Data Basis project. We have outlined the project at its current form, with details about its business model, its technologies and core products, examples of datasets, and a benefit-cost analysis. In this final Section we discuss our vision for what lies ahead in our path to universalizing access to high-quality data.

First, we will continue to develop and maintain our core products. Our search engine, data lake, and APIs are the bedrocks that power every other action we promote and therefore will always be a priority to the organization. We will scale data ingestion and maintenance by automating as many ETL pipelines as possible. We will continue to expand our catalog of datasets with new additions in every relevant theme. We intend to develop a more robust schema and data graph to help us connect all datasets to each other and develop more powerful visualization capacity. We will also continue to incorporate new technologies to our platform. These include a new space for users where they can participate in forums, follow specific datasets, receive recommendations, and access data directly from our website.

Second, we will work to broaden our user base, which is a key metric for success of the organization's mission. We will develop more educational material, both online and offline, to serve our various user types in government, academia, journalism, education, and others.

We will develop more content in the form of dashboards, blog posts, and articles. We will strive to have our data be used and cited in content outlets, such as the press, think-tank reports, and academic research.

Third, we will make Data Basis international. We will translate all our core products and materials to English and other languages in order to facilitate international usage of Brazilian data. But, even more importantly, we will start adding data from other countries to the platform. We envision a federated system where teams local to each country use our technology to upload and maintain their datasets. Our unique schema and quality-control practices will then guarantee all metadata and data are compatible, which will allow easy search and analysis at scale merging data from multiple countries.

Fourth, we plan to export our infrastructure-as-a-service (IaaS) to governments, companies, and other organizations. In 2022 we already had one important proof-of-concept with the Rio de Janeiro's city government adopting our technology and workflows as basis to their Data Office activities. As our infrastructure matures and is better documented, we expect more institutions to adopt it as basis to their data workflows.

Fifth, we intend to incorporate Data Basis into the scientific research pipeline. First, researchers can base their analyses on our public datasets and make promptly replicable code available to the scientific community. Second, journals could make publication conditional on authors publishing their datasets on our platform. This would enforce a common schema and interoperability for all datasets used in research. This would empower true transparency and expand possibilities for replication, as opposed to today's practices where authors publish isolated datasets and code that is not externally valid.

References

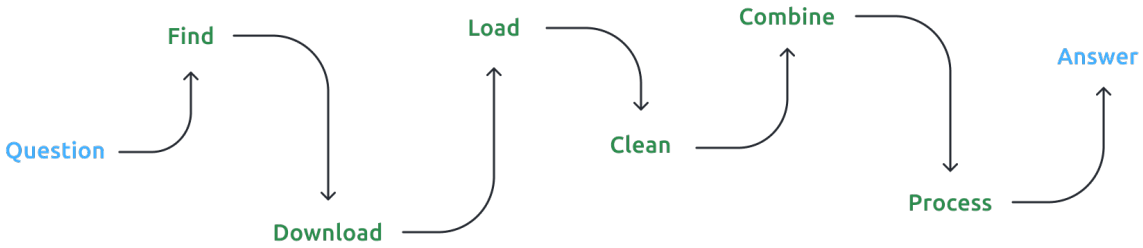
Sam Asher, Tobias Lunt, Ryu Matsuura, and Paul Novosad. Development Research at High Geographic Resolution: An Analysis of Night-Lights, Firms, and Poverty in India Using the SHRUG Open Data Platform. *World Bank Economic Review*, 35(4):845–871, 2021. ISSN 0258-6770. doi: 10.1093/wber/lhab003.

Hannah Ritchie, Max Roser, and Pablo Rosado. CO2 and Greenhouse Gas Emissions. *Our World in Data*, 2020. URL `https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions`.

Steven Ruggles, Sarah Flood, Sophia Foster, Ronald Goeken, Jose Pacas, Megan Schouweiler, and Matthew Sobek. IPUMS USA: Version 11.0 [dataset]. Technical report, IPUMS, Minneapolis, MN, 2021. URL `https://doi.org/10.18128/D010.V11.0`.

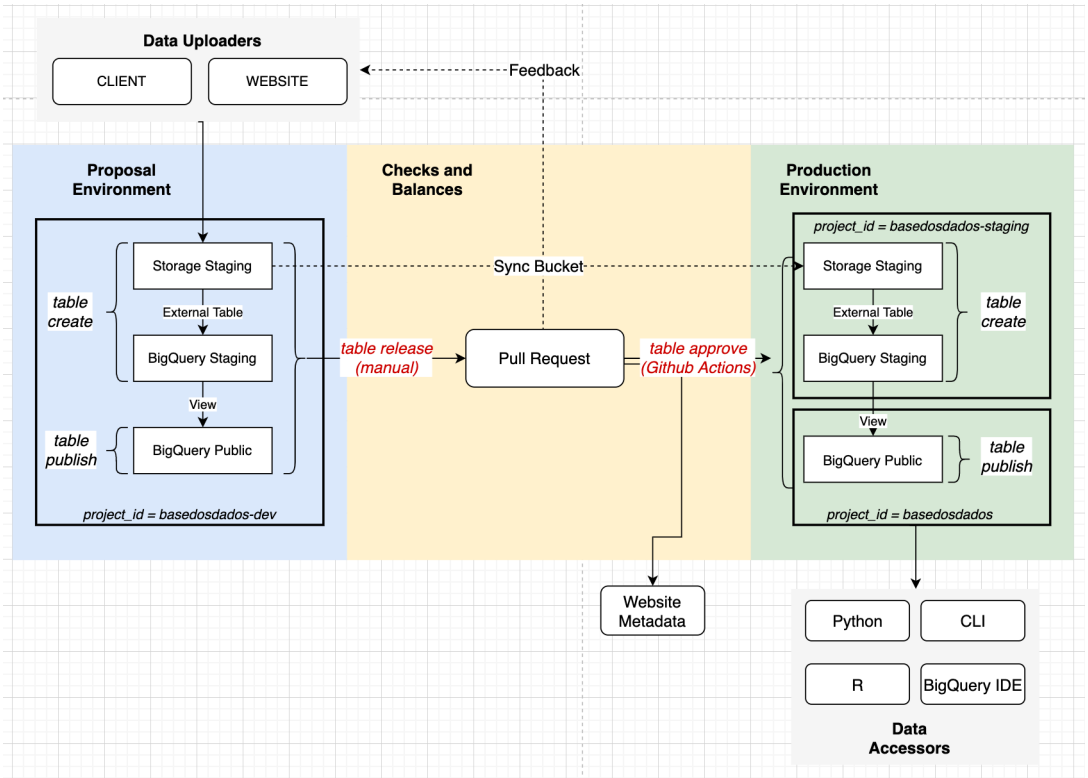
A Figures and Tables

Figure 1: The Distance Between Questions and Data-Based Answers



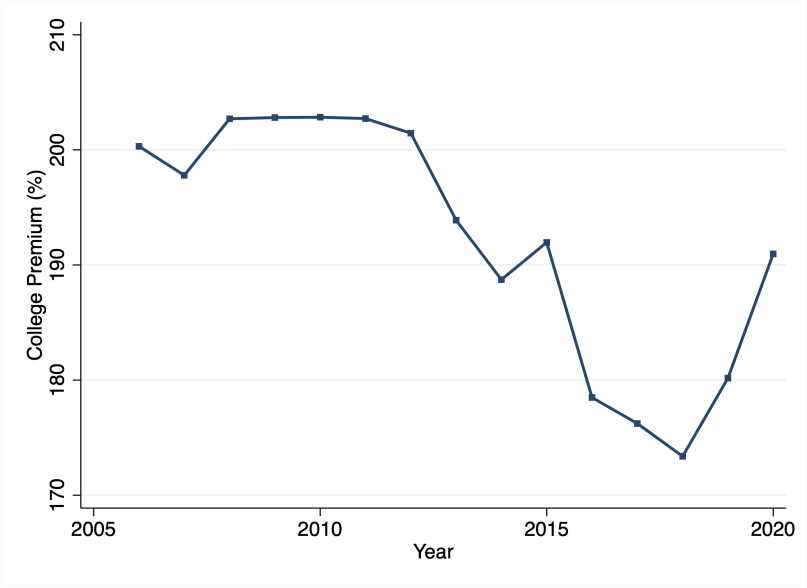
Notes: This Figure illustrates the typical distance between a user's question and its data-based answer.

Figure 2: The Data Basis Infrastructure Diagram



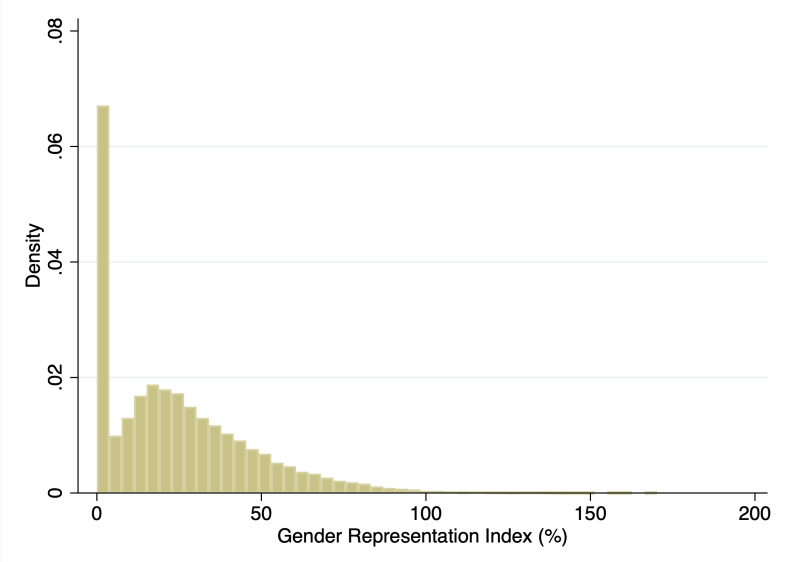
Notes: This diagram illustrates the Data Basis' infrastructure as described in Section 3.

Figure 3: College Premium in Brazil between 2006 and 2020



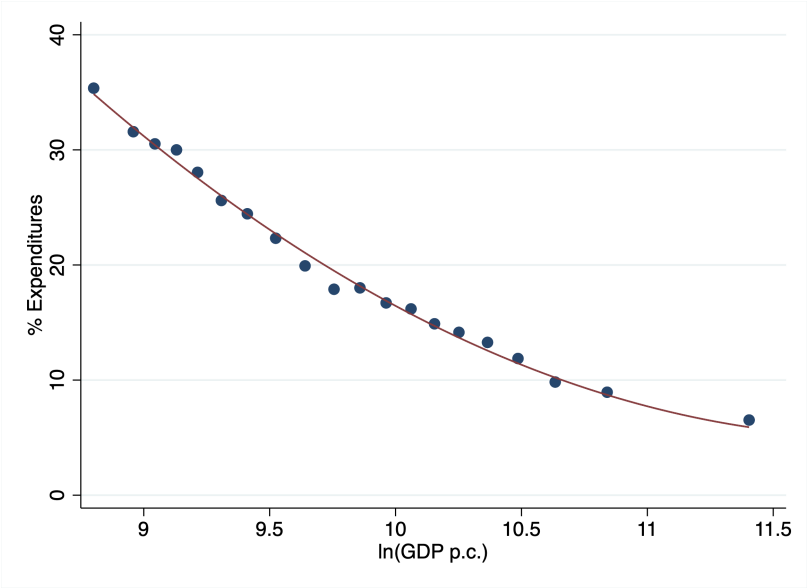
Notes: This figure plots the college wage premium for Brazil between 2006 and 2020. The premium is calculated as the ratio between average wages for college-educated workers and high-school-educated workers.

Figure 4: Gender Representation Index in Brazil between 2000 and 2020



Notes: This histogram plots the gender representation index' distribution across Brazilian municipalities between 2000 and 2020. The index is defined as % female elected legislators / % female population, where population is measured by the last Census (in 2000 and 2010).

Figure 5: Local Government Expenditures vs GDP per capita in Brazil in 2019



Notes: This binned scatter plot documents a negative relationship between local expenditures as a share of GDP and log GDP per capita across Brazilian municipalities in 2019.