A Frictionless Approach for Statistics and SDG Indicators

Rufus Pollock, Annabel van Daalen & Paul Walsh

Datopian

Notes on the Author

Datopian are experts in data management. They work with organizations of all sizes to design, develop and scale solutions to unlock their data's potential. They have worked on data strategy and built data platforms for numerous clients, including the US Government, Fortune 500 companies, and some of the world's largest NGOs. Led by CKAN creator and Open Knowledge Foundation founder Rufus Pollock, Datopian are one of the creators and curators of the Frictionless framework.

List of Interviewees

DANE, Colombia: Alexander González Coca (Head, Prospective and Data Analysis Unit); Andrés Holguín (Head, Systems Office); Camilo Andres Mendez Coronado (Head, International Affairs Office); Julieth Alejandra Solano Villa (Technical Director, Regulation, Planning, Standardization, and Normalization); Luis Fernando Romero Ardila (Interoperability Architect, Systems Office); Marly Esther de Moya Amaris (Head, Information Management Process, Systems Office).

Statistics Sierra Leone: Bintu Baindu Kandeh (Head, Open Data Unit); Clementina I. Akran (Head, Open Data/ Data Revolution and Visualization Section, Data Science Division); Mohamed Keita Samura (Head, Visualization Unit); Sallieu Mansaray (Focal Person, SDGs).

World Bank: Hiroko Maeda (Statistician, Development Data Group); Timothy Herzog (Senior Data Scientist, Development Data Group); Tony Fujs (Data Scientist, Development Data Group); Umar Serajuddin (Manager, Development Data Group).

Disclaimer

The findings, interpretations, and conclusions expressed in this work do not necessarily reflect the views of The World Bank Group, the National Administrative Department of Statistics Colombia (DANE), Statistics Sierra Leone, or the governments they represent.

This report was produced with support from the Trust Fund for Statistical Capacity Building and is made available under the Creative Commons Attribution 3.0 IGO license (CC BY 3.0 IGO) http://creativecommons.org/licenses/by/3.0/igo.

Executive summary

This report considers how the Frictionless framework (https://frictionlessdata.io/) could contribute to data management within statistical organizations such as National Statistical Agencies (NSOs) and the World Bank. It bases its findings on interviews with key staff from the Bank and two of its partner NSOs, Statistics Sierra Leone (Stats SL) and the Colombian National Administrative Department for Statistics (DANE), as well as the documentation they provided.

Worldwide, data-oriented organizations find themselves at a critical juncture when it comes to data management as they respond to the rapid growth in the variety, volume and velocity of supply *and* demand for data. Statistical organizations are at the forefront of these changes. However, they may differ from other data-oriented organizations in that they deal less with 'big data', with its focus on data size and volume, and more with large numbers of smaller but more diverse datasets. Their data may include financial data in numerous currencies, complex geospatial data, and sensitive data that may need high levels of security and governance. Statistical agencies also face the added challenge of needing to handle data coming from external partners, over which they often have limited control.

Reflective of these trends has been the rapid growth in 'data science' in the last decade, matched, more recently, by the rise in 'data engineering'. It is now widely recognized that the data engineering work to 'manage' data and create robust, evolvable data pipelines is an essential prerequisite for good data science. More generally, it has become clear that the work to 'produce' data (data engineering) is as crucial as the work to 'consume' data, that is to analyse and apply it (data science etc). This is leading many organizations to re-examine their data strategy and place greater emphasis on data management and data engineering.

Our research showed that statistical organizations face many of the same challenges. They all have to find ways to alleviate the difficulties that come with receiving large amounts of data from diverse sources in a variety of formats. Their systems for processing the incoming data are predominantly large, complex and bespoke, often without a common framework. This can create a variety of challenges: in terms of data reliability, in terms of a lack of transparency into the data process, in terms of ongoing maintenance, in terms of adapting or scaling to meet new needs, or in terms of reliance on key staff. Ultimately, the challenges faced by these organizations across the areas of data ingestion, cleaning and publishing affect the quality of the data produced, and by extension the impact that that data can have.

Frictionless is an open-source data management framework developed over the last decade that offers a solution to many of these problems. Developed in response to real-world use cases, it has a strong emphasis on simplicity whilst retaining the power to scale with complexity. Frictionless offers both core specifications as well as a rich suite of tools that can be adopted in incremental steps and are designed to work with existing workflows and systems. This makes it a very attractive solution for the organizations examined in this report, all of whom have substantial human and capital investments in existing tooling.

Based on the analysis of this report, statistical organizations would benefit from using the Frictionless specifications and tools. Frictionless' 'atomic', modular approach allows organizations to build up data infrastructure in an incremental and agile fashion – starting with a minimum viable product and scaling as they grow. Frictionless is open-source, which makes it accessible both financially and technically, and its focus on tooling popular among data scientists and engineers (like Python and R) allows for reduced dependency on specialised staff. Moreover, Frictionless also provides simple, reliable solutions to the problems of data validation and metadata standardization, two key problem areas across the three organizations.

Context

High-quality, statistical data is key to solving some of the greatest challenges facing the world today, from poverty to climate change. One of the primary statistical data collections influencing crucial decisions and policy making processes in these areas is the UN's collection of Sustainable Development (SDG) Indicators, large datasets that help various stakeholders to track the progress of the Sustainable Development Goals (SDGs). Worldwide, various international organizations, among them the World Bank and National Statistical Agencies such as DANE and Stats SL Leone, collect, process and publish SDG and other development data with the aim to transform countries across the globe into dynamic, innovative economies with modern, citizen-centered public sectors.

The World Bank is widely considered a world leader in open data, not least due to its work on some of the world's pioneering open data initiatives. In recent years, its Development Data Group (DECDG) has worked in collaboration with other World Bank units and external partners to support the implementation of open data projects and best practices in multiple countries and across the world. Several of these projects focused specifically on increasing the capacity of National Statistics Offices (NSOs) to support open data.

To coordinate their work program, a working group of NSOs met regularly between 2013-2017, during which time they identified several key takeaways of which two are relevant here. The first is that standards are essential to ensure the interoperability of published data, as the current lack of interoperability in the realm of SDG indicators has slowed the pipeline for organizations to share data efficiently and transparently, whilst also increasing transaction costs. The second is that openness must be a design consideration throughout the entire data production and dissemination cycle. So far, implementing openness has been made difficult by legacy data systems that are often complex, opaque, and difficult to integrate with more modern approaches. As such, there remains a need for specific guidance and tools that NSOs and other statistical organizations can use to implement both open data best practices and the data production systems that support open data.

However, there are not many templates for opening and modernizing the production of statistical products in an iterative fashion alongside a legacy production environment. It is for this reason that the Bank, Stats SL and DANE are considering Frictionless as a promising approach for statistical organizations.

The State of Play

Overview

This section provides an overview of current data practices within the statistical community, in particular the ways in which technology is applied to data management in producing SDGs, as well as broader organizational objectives. By analysing the current state of play in this way, Datopian sought to uncover how a Frictionless approach to data management can help address a range of challenges across statistical organizations, and improve collaboration and data exchange among them.

The World Bank

The Bank is a major global data provider. Its primary data collection, the World Development Indicators (WDI), housed within the DECDG, receives over a million hits per month and provides access to around 1,400 key global development indicators. The Bank also works alongside NSOs and international agencies to produce indicators for the Sustainable Development Goals, which it submits to the UN Statistics Divisions' Global SDG database. Its contribution here is crucial to global efforts to tackle some of the most pressing challenges facing humanity today, from poverty to climate change.

While the Bank is certainly a leader in data production, this does not mean to say that they are not trying to improve their services even further. In particular, teams at the Bank want to better data access and discoverability for end users. To this end, they are focusing their attention on improving observability of critical processes, modernizing their systems and approaches, and making their data processes more efficient. To understand why these areas could improve data for the end users, we first need to have an overview of the Bank's current processes and systems.

Much of the data management in DECDG centers around the World Development Indicators (WDI), specifically two systems called the WDI Working and the WDI Final. These two WDI systems are subsets of an SQL database called the Data Collection System (DCS).

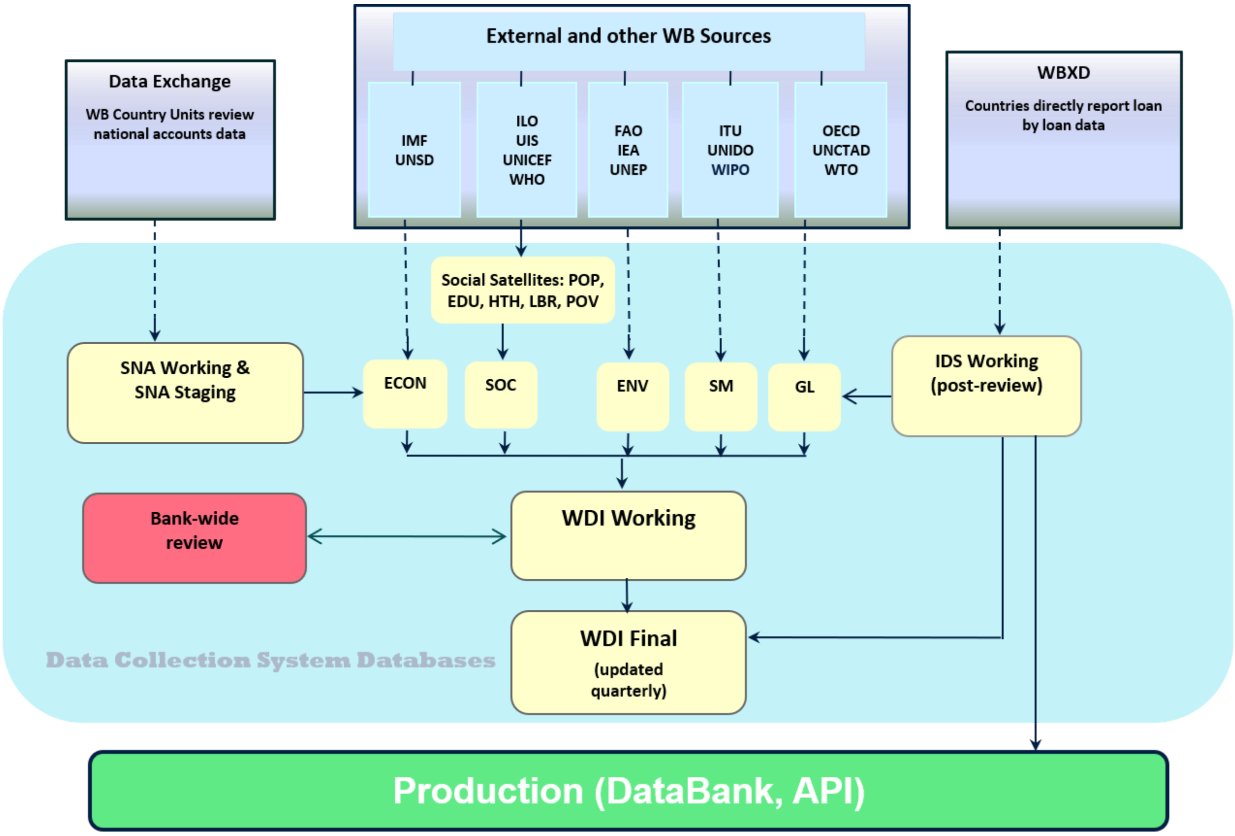
External and other WB Sources
IMF
UNSD
ILO
UIS
UNICEF
WHO
FAO
IEA
UNEP
ITU
UNIDO
WIPO
OECD
UNCTAD
WTO
Data Exchange
WB Country Units review national accounts data
WBXD
Countries directly report loan by loan data
Social Satellites: POP, EDU, HTH, LBR, POV
SNA Working & SNA Staging
ECON
SOC
ENV
SM
GL
IDS Working (post-review)
Bank-wide review
WDI Working
WDI Final (updated quarterly)
Data Collection System Databases
Production (DataBank, API)

Fig 1. WDI data flow (flow chart).

The Bank collects data from a number of internal and external sources. External sources consist of multiple large, international organizations such as the United Nations (UN), the International Monetary Fund (IMF) and the World Trade Organization (WTO). Internal sources include the World Bank External Debt system (WBXD), into which member countries directly report loan-by-loan data, and the Data Exchange, into which Country Management Units (CMUs) review national accounts data. Data reaches the Bank in a number of formats, including Excel spreadsheets, SDMX, PDFs and Web APIs.

Data then enters the Satellite DCS, where it is aggregated into five domains: Economy (ECON), Environment (ENV), Global Links (GL), State & Market (SM) and Social (SOC). Data from external data providers is pushed directly into the Satellite DCS, while data from the Data Exchange and WBXD first travels through the System of National Accounts (SNA) and the International Debt Statistics system (IDS) respectively. Here the data undergoes initial outlier detection by the relevant indicator specialist.

Data is then pushed from the Satellite DCS to the WDI working, where it undergoes further aggregation. For example, here GDP data from the Data Exchange is combined with population data from the UN Population Division and NSOs to calculate GDP per capita. As a final step, data is then pushed into the WDI Final, where it is refactored into either the DataBank or sent via API to the open data website for publication. This website offers a more modern UI for exploring WDI data than the DataBank.

National Statistical Agencies (NSOs)

National Statistical Offices (NSOs) are responsible for collecting, processing and disseminating official statistics for their country. While there may be many ministries responsible for official statistics in one country, NSOs typically serve as the coordinating entity and record the social, economic and environmental condition of the country. The ability of NSOs to produce high-quality data is essential for each country to set priorities and make informed choices.

DANE, Colombia

DANE is very active in the sphere of open data in Colombia and works closely with government officials on a number of policies. Their current open data work is set against a background of increasing recognition in Colombia of the importance of open data. Colombia's 2018 National Policy document, for example, set out a framework for increasing the use of public data and ensuring that it is managed as an asset capable of generating social and economic value. DANE wants to see their data continue to increase transparency and social accountability, improve products and services, and generate knowledge for decision making (especially in the current context of COVID-19).

To this end, DANE has invested heavily in implementing a bespoke, best-practice data management system with a sophisticated approach to collecting, processing and publishing statistical indicators. DANE's technical capability is reflected in the complexity and sophistication of their data collection methods. Organizations providing SDG data deliver data to DANE in Excel files with a specific format designed by DANE, which they themselves upload to DANE's database using web services like SDMX. Larger providers send data using Microsoft SQL server or or inject the data directly into Oracle databases.

DANE uses various tools to clean data, from Excel to custom scrapes designed in tools like R, SPSS, and SAS. They also have their own custom application cleaning tools that upload data into repositories ready for consumption. Once the data is in repositories, DANE uses various tools to publish data. They use R, SPSS, and Excel to create new products like statistical reports and use different tools to create new data with geographical components. Data publishing happens through a software called NADA that uses DDI to publish information. They also publish Excel files on their online portal, as well as a geoportal that allows users to download data in geo format, and offer applications through which users can query the data warehouse.

While DANE's approach is effective for data collection and processing in Colombia, it is unclear whether the approach would replicate as successfully in different contexts for a number of key reasons. Firstly, DANE has unusually strong technical capacity, which affords the expertise to develop and sustain a highly customized and sophisticated data platform. Second, they have been able to take a relatively greenfield approach unencumbered by legacy systems or approaches.[1] Third, their solution relies on closed-source software that is not readily available to other organizations, especially those with more restricted funding. Fourth, and importantly, government-regulated publishing requirements allow DANE to impose strict data serialisation requirements on other government publishers (based on SDMX). This ability to impose strict requirements on upstream providers of data is not something the World Bank (nor many others) can do.

Statistics Sierra Leone (Stats SL)

Stats SL is Sierra Leone's central authority for the collection, processing, analysis and dissemination of high quality statistical information, with a vision to create a viable National Statistical System (NSS) that can support evidence-based decision making. They are major contributors to the Sierra Leone Open Data Portal and were recognized by the Disclosure of Information Compliance Award as the only MDA to proactively disclose all information as stipulated in the Right to Access Information Act of 2013. Currently, as part of the government of Sierra Leone's National COVID-19 Emergency Operations Center, Stats SL are engaged in a collaboration with various international partners to produce crucial datasets and tools under an open, non-commercial licence to support the country's COVID-19 response.

[1] A Greenfield project lacks constraints imposed by other work, ie. starts with a 'clean slate'.

With regards to data collection, Stats SL find themselves in a period of transition. Up until recently, they collected data using Paper Assisted Personal Interview (PAPI). In 2011, the team was introduced to Computer-Assisted Personal Interview (CAPI) by means of hand-held devices called Personal Digital Assistants (PDAs), which will be used to collect data for Sierra Leone's first Mid-Term Population and Housing Census in 2021. However, certain issues surrounding data conversion to user-friendly formats still stand in the way of improving their ability to fully support open data initiatives.

Stats SL simply cannot collect certain types of data needed for indicators in the SDG framework by themselves. This is either because the data cannot be collected through surveys due to its nature, like climate data, or because they do not have the financial resources to collect the data. This leaves the team reliant on secondary data, which they often consider unreliable due to technical concerns. Secondly, the team struggles to collect data for SDG indicators that fall within Tier III (indicators that have no clear standardization or definition), as many categories, such as 'poverty', mean different things in different countries of the world. In response to this issue, Stats SL have domesticated certain indicators to form the Sierra Leone Specific SDG indicators.

The team at Stats SL hopes that the recent introduction of a data accreditation committee will help to increase the credibility of secondary data for official statistics. They also hope to help to reduce friction in the collection process by providing training to data standardization training to Ministries Departments and Agencies (MDAs) and Local Councils. To this end, they are taking part in a project funded by the UN to foster collaboration between local councils and research institutes. As part of this process, the team at Statistics Sierra Leone have asked Local Councils to highlight the data that they already collect, plus the data they have the capacity to start producing. This allows Statistics Sierra Leone to provide them with the relevant data standards.

Drivers of Change

Datopian has observed a number of trends in the wider data space that are influencing the ways in which many organizations, the World Bank, DANE and Statistics Sierra Leone included, are working with data. In particular, these shifts have catalyzed calls for modernization in terms of data infrastructure and approaches to data management. In this section, we explore how these external trends are mirrored in internal shifts at statistical organizations.

Global data growth and diversification

Worldwide, data is growing: in amount, in demand, in variety, in velocity, and simply in its importance and the degree of data-driven activity. That the Bank's open data website has only been around for ten years is testament to this. Today, even small organizations perform an increasing amount of data collection and data-driven analysis. Diversification of the *supply* of data has been met with diversification of *demand* in the form of users and use cases, and tools like Google spreadsheets have democratized and expanded the range of users.

Shift from statistics to data science

Over the past decade there has been significant change in the typical profile of staff working with data. While true of both DANE and Stats SL, it is especially noticeable at the Bank, where formerly staff were mainly trained as statisticians, working on tasks such as computing indicators. They had a toolbox of finished, bespoke tools and were trained specifically in how to use these tools to move data through the pipelines. Knowledge of a more general suite of data science tools was not common. In recent years, the Bank has been hiring more data scientists, who have brought with them experience in tools such as Pandas and R Studio. These provide an alternative way to build sophisticated data management systems, which previously required expensive, inflexible tools.

Need for data engineering and data management

As a direct result of a maturation of data science over the past decade, Datopian has observed an increased need for data engineering and data management. Data science started off focused on new potential for data analysis through a combination of coding and statistics. However, Datopian recognized that data preparation and processing steps absorbed a large proportion of data scientists' time and that managing data and creating and maintaining robust, evolvable data pipelines was a job in itself with specific skills. In short, it became apparent that data engineering was an essential complement and prerequisite for good data science. This newfound understanding of the relationship between the two disciplines has led to a demand for the development of robust, maintainable and evolvable data management systems and pipelines.

The Problem Space

In this section, we will explore some of the current challenges facing the World Bank and the two NSOs in the collection, processing and publishing of statistical data. Our analysis will interrogate the various systems in each organization, as well as the teams' general approaches to data management and any associated gaps in knowledge or expertise.

Lack of autonomy and agility slows development and responsiveness

All three organizations face problems of dependency, particularly when it comes to data exchange. At the Bank, the workflows of the Development Data Group (DECDG) depend greatly on a suite of large, legacy systems such as the Data Collection System (DCS). In the case of the NSOs, DANE's workflows rely on their providers sending data in a particular format, while Stats SL relies on external technical expertise to assist with data processing of complex surveys and censuses. A more open-source and agile approach can empower data teams to manage their own toolchain.

Limited observability reduces the quality of data governance

In some cases, there was evidence of a lack of observability into key systems and processes, which impedes reliability and retardes the pace at which problems are diagnosed and fixed. There may be central 'blackbox' systems which perform key data aggregation and transformation functions but that are not transparent to all stakeholders. While the teams can be certain that the datapoint was derived based on a set of assumptions, without observability into the transformation process, they cannot tell what these assumptions were. This affects those teams' ability to control and understand data provenance, including what happens to data in the journey from source to publishing. The lack of observability also negatively impacts data curation teams, who 'don't know what they don't know' about the data.

Insufficient metadata results in poor data quality

All organizations struggle to collect sufficient amounts of metadata. As explored above, gaps around data provenance within World Bank systems mean that the end user often misses out on the important contextual information usually provided by provenance metadata. For example, missing information such as when databases and statistical series were last updated can cause the user to question the validity of certain datasets. As a large institution working with countries from around the world, missing currency metadata can also cause headaches for the Bank. This is a problem also familiar to DANE, who often receive entire data backups without the accompanying data dictionary.

Challenges in data cleaning and normalization

Data cleaning and normalization is a key problem area for statistical organizations. Stats SL in particular highlighted cleaning as one of the most problematic processes and expressed a need for capacity building in the use of the data cleaning tools SPSS and Stata, especially when dealing with large amounts of data (such as in censuses).

Data normalization is also often a problem across the organizations due to the diversity of the data being processed. For example, while a statistical organization can rely on well-defined country codes at national level, identifiers are less well defined—or are missing entirely—at subnational level.

Manual data quality validation is tedious and time consuming

A common challenge for all organizations we spoke to was ensuring the quality of data, especially published data. All of the organizations were seeking better and easier ways to (automatedly) check the quality of data. In the case of the Bank, there is use of an outlier detection algorithm. But even here experts have to manually check whether the outlier is genuine, ie. whether it actually is an incorrect value. These checks could be avoided or made much more efficient if key contextual metadata were present (at present, the information is either documented separately from the data or not documented at all). Statistics Sierra Leone have no automated data quality validation checks and rely on manual data re-inputting to see whether the error margin is over 20%.

Key tools require restrictive formats reduces agility

Datopian works with many clients whose reliance on proprietary software leaves them inflexible when it comes to data formats. At the Bank, the Data Collection System (DCS) only accepts excel files, meaning that staff have to manually extract datasets and convert them to excel before they can be ingested into the DCS. Not only is this time consuming, but Excel is not designed for data exchange in data management systems like this and lacks many features one would want for a base data format (e.g. the ability to store associated metadata).

Similarly, Stats SL takes data in the format in which it is provided, though they do this manually. In Colombia, DANE has taken a different approach and sought regulatory pressure to enforce data formatting requirements on providers. Of the 27 organizations currently providing SDG data to DANE, 15 provide data through SDMX, an open standard specifically designed for data and metadata exchange. Soon, this will become a legal requirement for all statistics organizations in Colombia.

Concerns around data security slow down data work

Data security concerns are a common factor when it comes to the exchange of public data. Colombia is no exception, with some of DANE's data providers concerned about sharing their data via web services. This results in an infrequent data delivery process, whereby DANE receives very large volumes of data at once via SSH File Transfer Protocol (SFTP). To address this bottleneck, DANE is to start encouraging providers to send smaller volumes of data on a more frequent basis by introducing a more secure data exchange software called Xroad. Stats SL are also looking at different approaches to data security with file transfer protocols.

Desired Outcomes

On the whole, there is agreement that efficiency, quality, and agility gains are to be made by adopting a more modern, best-practice approach to data management for statistics and indicators. During our conversations a number of common desired outcomes emerged:

- The development of a set of principles towards data management modernisation, explicitly covering increasing transparency around data processing and practices.
- Any changes to the current tooling and methods of data management should be adoptable for the medium to long-term, i.e., 5-10+ years.
- A move towards a best-practice approach to data management should accommodate for staff capacity building.

- Specifications and tooling that are simple and lightweight, to the extent that staff can run the primary toolchains on their personal computers.
- More transparency into the data processing pipeline by allowing staff to read and modify code that consolidates, normalizes and transforms data.
- Better (more accessible, more flexible) tooling for validating data quality, so that errors and inconsistencies surface earlier in the data production process.
- Data exchange processes should be explicitly secure to assuage security concerns currently held by multiple data providers.
- Incremental adoption both in terms of architecture and tooling - metaphorically speaking, one wants to renovate the house, not build a new one.

The Frictionless Framework and Toolkit

Overview

The Frictionless project holds promise as a possible approach for statistical organizations to modernize and open their data production pipelines. Designed around open-source, scalable standards for data and metadata, and intended to be implemented in incremental steps, Frictionless is already being used by open data programs in Great Britain, France and the US as well as in many other organizations. Its core principles provide a way for organizations to build on existing workflows and platforms rather than replace them outright at the beginning. For NSOs and other statistical agencies that have substantial capital and human investments in existing systems, this incremental approach could provide a realistic blueprint for modernizing their processes and moving towards greater interoperability, efficiency and openness.

Context: why data management matters

If Frictionless helps organizations to manage data, then to truly grasp its significance it is important to understand what data management actually means, and why it is important in the first place.

Data management is the job of collecting, organizing, curating and integrating data so that is readily available to downstream users and applications. A key part of data management is ‘data integration’: bringing diverse data together, cleaning it up, knitting it together and pushing it into downstream applications, analytics or warehouses – and doing this reliably, repeatedly and automatedly.

Data management is essential to realizing value from data. It’s much like baking a cake, in the sense that many people want the delicious cake (ie. the insight) but they need to do a lot of preparation and processing to get there. In the end, the time you spend eating the cake is negligible in comparison to the time you spend getting it ready. Frictionless Data aims to maximize the time we can spend extracting the value from data by shortening the path from data to insight.

Understanding atomicity as Frictionless’ underlying principle

Frictionless is an open framework for building data infrastructure that is both powerful and simple. It is designed to cater to the full range of data producer needs, whether it’s building the simplest of pipelines or powering sophisticated workflows in combination with modern tooling and supporting libraries. The reason why Frictionless is able to scale up and down in this way is because its framework follows an ‘atomic’ approach, which we will explain briefly here.

Atomicity describes the process of breaking something down to its smallest components or fundamental building blocks (‘atoms’) and combining them together piece-by-piece to create more complex structures (‘molecules’). The Frictionless framework follows an atomic approach in the sense that tools and specifications are broken down to their minimum viable components, thereby allowing users to build data infrastructure in incremental steps.

It’s about more than just a toolkit: the significance of having a *framework*

Frictionless is more than simply a suite of tools. Through combining tooling with a set of specifications, Frictionless provides an overall framework for describing and organizing data. To understand what this means we need to briefly explain what we mean by specifications. Specifications are a bit like a set of instructions that tell a tool how to interact with your data. In this sense, tools and specifications are synergistic, meaning they can have more impact when used together than they can by themselves. By using tools and specifications in combination, users can create ecosystems of different infrastructure in which data flows fluidly between teams and tools.

The Frictionless specifications have been refined over more than a decade to a zen-like simplicity. As a result, they can be picked up in minutes and immediately integrated with other libraries or existing projects. The specifications can be used by themselves or in combination with others. They include:

- *Data Package* for datasets
- *Data Resource* for files
- *Table Schema* for tables

Let’s take the example of the *Data Package* specification. *Data Package* is a simple container format used to describe and package a collection of data (a dataset). If we wanted to, we could create a Data Package specialized for a specific type of data, eg. we could create a Tabular Data Package for tabular data. To do this, we could combine Data Package (to describe the dataset) with Table Schema (to describe the table structure). Datopian refers to this approach as ‘small pieces, loosely joined’.

On top of these core specifications, Frictionless then offers a rich toolkit ranging from core SDKs for most major languages (Python, R, Javascript, Java, Ruby, Julia etc) to low-level tooling for validation, data inspection, conversion, metadata creators and editors and finally high level platforms such as DataHub and GoodTables and integrations with tooling like CKAN, Zenodo, GitHub etc.

Further reading

- https://frictionlessdata.io/
- https://frictionlessdata.io/guide/
- Frictionless Data: standards and tooling (Youtube video)

Recommendations

Based on our interviews with the World Bank, DANE, and Stats SL, we have a number of recommendations to create more efficient, sustainable and reliable data systems for statistics and indicator data. Taken together, these recommendations offer a clear path for adoption and training as part of data capacity building, and also enable the transfer of skills and knowledge across boundaries between statistical data work, and other fields of data science and data engineering.

Keep data systems as simple as possible

As was established in the 'Drivers of change' section of this report, data management needs will continue to grow in complexity, variety and volume. The single best way to ensure that our solutions are agile, maintainable and reliable enough to apprend this change is to keep them simple. A great way to do this is using a componentized, atomic approach, as this allows systems to scale in a way that limits complexity. It also makes systems more approachable by allowing developers to run ket data processing steps on their PC or laptop with the same toolchain as that deployed for production. Ultimately, keeping systems simple results in greater autonomy for data teams, reduced burden on individuals with certain skill sets, increased observability into data transformations, and reduced infrastructure costs.

How Frictionless helps: Frictionless is designed around the principle that you should start with what you need and add in tools and specifications as you grow. Stream-based, open-source modular tooling means that even workloads with tens of gigabytes of data can be run by personal computers with ease. Common tooling and data formats that do not require specialized computer infrastructure or know-how to run, and which are compatible with many existing tools and platforms, also make Frictionless simple to use.

Leverage open-source solutions

There are many large open source communities developing tooling and approaches to data work. Our work in Frictionless Data is part of this broader ecosystem of communities. Participation in such communities will ensure that the statistical organizations avoid vendor lock-in and knowledge silos.

How Frictionless helps: Frictionless is an existing open-source framework with connection into the wider ecosystem. Its core specifications provide a robust, open foundation for integration of additional tooling whether open-source or proprietary.

Prioritize tooling that is widely used by data scientists and data engineers

Data ecosystems in Python, R and Javascript programming languages are the most widely used set of tools used for all kinds of data work. Adopting tooling and principle based on and/or compatibles with these ecosystems ensures that statistical organizations and stakeholders use and have access to both best-of-breed technical solutions as well as a large talent pool.

How Frictionless helps: Frictionless is designed by data scientists and engineers for data scientists and engineers. It is even built with the tooling they use the most.

Change existing end-user experience as little as possible

Working with existing tooling and workflows is integral to the success of any change process in relation to data management in an organization. Favor approaches, tools and standards that are interoperable with existing end-user tooling and workflows such as MS Excel, as well as professional databases and analysis platforms. Avoid requiring immediate and disruptive changes.

How Frictionless helps: Frictionless specifications already have integrations with many standard end-user tools as well as databases and other platforms. For example, Frictionless' default tabular format is based on CSV, which is supported and used by almost every common tool, from Excel to relational databases. In addition, the simple nature of Frictionless specifications and the wide range of existing Frictionless tooling makes integrating with new systems quick and easy.

Use standardized metadata

If publishers and consumers have mechanisms to view metadata related to data, a range of assumptions about the data can be codified and communicated. In turn, this codification of metadata helps reduce questions about data integrity and consistency. In addition, good structured metadata is foundational for pipeline automation and automated testing. In essence, standardized metadata provides critical upstream information to downstream consumers of data, eg. the currency of financial data. It also provides critical context for data reuse, eg. the data's source and the date of its creation.

At the same time, it is important to avoid 'metadata overload', whereby data producers are burdened with supplying large amounts of complex metadata. Metadata overload results either in the provision of no metadata, or, potentially worse, incorrect and inconsistent metadata. One wants to find the “goldilocks” level of metadata: not too much and not too little.

How Frictionless helps: Frictionless has a small core yet supports the publication of almost any metadata, capturing everything from general notes to specific type information for columns of data. One also has the option to extend Frictionless for custom metadata or to integrate existing metadata standards. Moreover, a rich set of existing tooling supports the creation, validation and display of Frictionless metadata.

Test and validate your data

In software, (automated) testing is standard – in fact, software without tests is often considered unusable. With data, testing and validation is usually manual and hence costly, error-prone and irregularly used. Validating data is a key issue, both at the curation stage (working with source data from data producers), and the publication stage (presenting data for public use). Dedicated, automated data validation flows that detect not only outlier detection, but structural and schematic consistency, will greatly increase the quality and efficiency of data work. This is especially true if the validations are customized to handle the types of problems that commonly occur.

How Frictionless helps: Frictionless (specifically the Table Schema specification) allows you to describe data in a way that makes it easier for tools to validate it. Frictionless also provides Table Schema-based tooling in various programming languages that can run automated tests and report errors to users. For example, “GoodTables” is a command line and GUI tool for doing data validation that works with simple text files like CSV, and can also leverage the Frictionless Table Schema format in its validation runs.

Data processing observability

Observability into data processing routines is critical for teams to trust their data outputs. Data processing pipelines should be viewable and modifiable by those responsible for running them. This reduces reputational risk, which can occur when data is processed in ways that a team does not fully understand or have control over. It also empowers team members to make changes and communicate to colleagues increasing efficiency and satisfaction.

How Frictionless helps: Due to Frictionless' modular approach to data pipelining, teams can compose pipelines of data processors, and modify them with use, using standard Python code over data in standard text files. In addition, the Frictionless specifications provided a standardized exchange language between processors and larger components making it easier to log, rerun and debug pipelines.

Proposed Next Steps

The recommendations above provide the groundwork for a deeper analysis of key use cases for Frictionless in each organization. In order to maximize the impact of Frictionless on each individual organization, Datopian suggests that it would be productive to focus on developing demonstration projects or pilots. This would involve defining the key Frictionless components (metadata or other standards), designing and developing processing or validation code, and identifying integration points.

For example:

- The World Bank might pilot a Frictionless approach to document the complicated and decentralized processes of assessing inputs to its World Development Indicators, national accounts, or external debt products.
- Stats SL might test out Frictionless for enhancing and simplifying their manual data validation process and metadata extraction.

- DANE might trial Frictionless as a means of providing external data providers with lightweight tools for publishing in the formats required by DANE.

Glossary

For a glossary of technical terms used in this report, please refer to datopian.com/glossary.