John Snow Labs and Open Knowledge International Partner to Donate 200+ High Quality Datasets to the Open Data Community

The data philanthropy commitment yields clean, current, and enriched datasets in eight categories published in the Frictionless Data standard on DataHub.io

John Snow Labs and Open Knowledge International have recently announced a partnership around data philanthropy. This partnership will yield over 200 clean, fresh, updated datasets for the open data community. In addition, John Snow Labs and Open Knowledge International have committed to jointly maintain 11 datasets that will be part of DataHub Core Datasets. These datasets are hosted on DataHub.io website. DataHub.io is one of the oldest and most widely used online data portals, hosting and providing a wide range of high quality datasets maintained by a community of data professionals and experts. John Snow Labs datasets are in compliance with FrictionlessData.io specifications and are easily shareable over multiple platforms.

- John Snow Labs has partnered with Open Knowledge International to make data freely accessible and available through Datahub.io.
- John Snow labs will work with Open Knowledge International to maintain core data sets.
- There are challenges to making data open for uploading to Dathub.io, including such problems as messy data and missing values.
- John Snow Labs has found solutions to these challenges and are able to provide clean and frictionless data in the form of data packages.
- There are several benefits to open data for both government and non-government organizations.

What is Open knowledge and what are frictionless datasets?

Open Knowledge International is a global organization that has been promoting open data initiatives since 2004. It educates, builds and promotes the concept that data should be free and easily accessible so that society can benefit from its application. To this end, Open Knowledge International provides tools and works to advocate for open data.

This is an area that John Snow Labs feels strongly about and recognizes that data should be used for the benefit of all. There are many benefits to open data, including increasing credibility and promoting progress and innovation. Such access to data also enables decisions to be made by policy makers. Data can also be used for education and community development, which leads to progress.

The use of open data has dramatically increased over the years, yet there are still many challenges that must be overcome to make it more useable. There are standards that have to be met before data can be considered open.In fact, one such format is the data package standard which is suggested for frictionless data. This standard is simple and web friendly. Adopting standards is therefore important in achieving open data that is easy to access.

Frictionless data is data that can easily be transported between systems and used with ease. Open Knowledge initiated the Frictionless Data standard and Datahub (Datahub.io). Datahub is committed to making data easily accessible, which John Snow Labs is also passionate about. Frictionless Data focuses on the logistics of data with the aim of cutting costs to allow for the mass automation of data cleansing by making it efficient, cheap and fast.

What were the initial challenges?

Gaining insight from data requires collecting, cleaning and maintaining data - each being a very time-consuming process. Challenges can include the legalities of accessing data, interoperability and the financial costs involved when it comes to data access. Data is messy, incomplete, and may come in different formats. It can be very difficult to combine datasets because the data formats and their schema could differ significantly. These are all challenges which have been recognized and are constantly being addressed at John Snow Labs.

The idea behind Open Access is to upload datasets onto Datahub for free access and use. However, there were some significant technical challenges that were involved in the process of making data available and accessible. Such challenges included the compliance of all datasets to the Frictionless Data specification and the problems with datasets being messy or difficult to access.

What solutions were implemented?

John Snow Labs has successfully delivered over 200 datasets to Datahub. These datasets are now available for everyone to use free of charge. These datasets comply with frictionless data specifications which enables using various analytical tools for data ingestion. Frictionless data effectively "containerizes" the data to improve interoperability between systems. John Snow Labs uses data packages that implement the frictionless data specifications such as, for instance, well-formed datapackage.json files and data.csv files.

Decreased data preparation time is crucial to many organizations. Standardized packages allow for the use of tools such as data validation methods, storing and searching of data. Such packages can also facilitate the efficient and automatic import or export of data. Ultimately, by providing frictionless data John Snow Labs aims to reduce the time required to get from raw data to insights.

At present, John Snow Labs has successfully integrated Datahub into its workflow. This entails joint maintenance of 11 open datasets and over 200 free datasets available for the community. These clean and fresh datasets are curated regularly by domain experts at John Snow Labs, and go through three levels of quality reviews - two manual reviews and then a set of over 60 automated validations checks. The extensive automated and manual curation of data ensures a level of quality control that is unrivaled in the industry and is a hallmark of John Snow Labs.

These datasets contain clean data and have been validated against frictionless data specifications. The partnership between John Snow Labs and Open Knowledge International is a long term commitment on both sides - to provide free access to high quality, curated and always up-to-date datasets that will benefit the entire open data community.