argendataR: la cocina de los datos de Argendata

Anónimo

Palabras clave: data - argentina - etl - desarrollo

Argendata procura ser un sitio de referencia con datos sobre Argentina y sumar conocimiento basado en evidencia al debate público. Para ello el punto de partida de Argendata es el conocimiento de dominio de múltiples investigadores. Desde allí se busca generar un producto comunicable de calidad basado en datos reproducibles y actualizables. **argendataR** es un paquete de R pensado como herramienta para el proceso de generación de los datos que sirven de insumo de Argendata.

Arquitectura del proyecto

La construcción de Argendata involucró equipos de investigación, de diseño y de desarrollo, que participaron en cuatro grandes etapas:

(I) una primera etapa de propuestas temáticas y definiciones curatoriales, para la que se convocó a especialistas de cada uno de los temas;

(II) la generación de contenidos de base de los diversos tópicos por parte de investigadores;

(III) la armonización de los procesos de generación de datos y preparación de recursos para su publicación;

(IV) la generación de productos gráficos y narrativos para la comunicación.

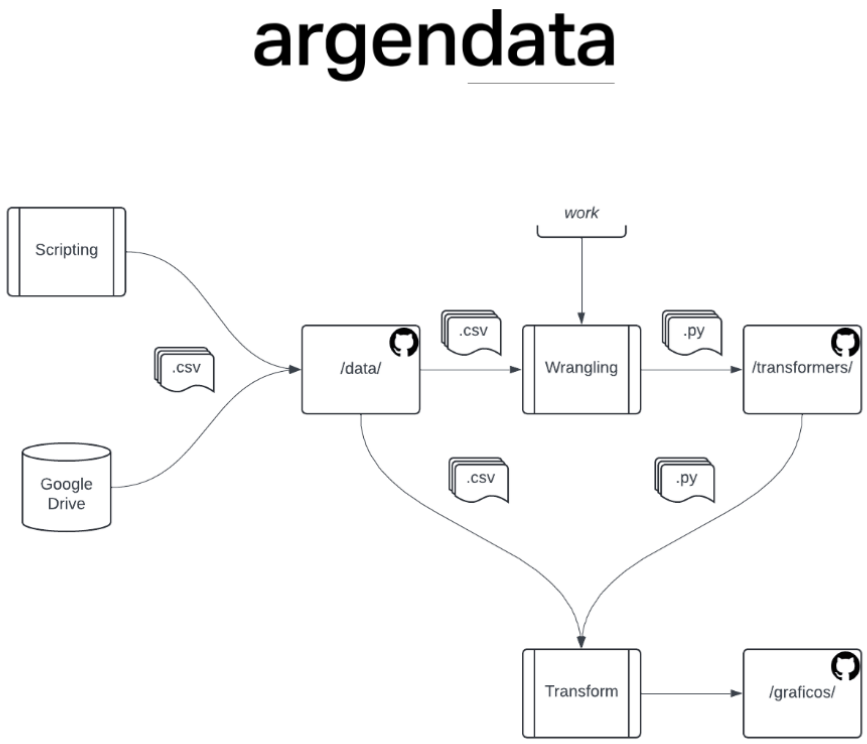
El punto de partida del proyecto prioriza el conocimiento de dominio de los investigadores involucrados, dados ciertos lineamientos y estándares generales para la elaboración de contenidos. Se puede decir que la etapa inicial de generación de contenido -etapa (ii)- está "atomizada". Esto supone un desafío para las etapas subsiguientes, que deben lidiar con diferentes técnicas y formatos. En la etapa de armonización se desarrolló código en Python y R para realizar lidiar con esto pensando principalmente en las dimensiones de:

Administración de recursos: que implicó el desarrollo y mantenimiento de estándares y programas para facilitar la gestión, control y posprocesamiento de datos elaborados por investigadores con el objetivo de minimizar la distancia entre procesos de investigación individuales y la necesidad de contar con información armonizada y coherente.

Control de calidad y reportes: la generación de programas que permiten realizar consultas sistemáticas sobre la calidad de los datos, de acuerdo a los estándares del proyecto.

Armonización del código fuente: gestión de programas y flujos de trabajo para administrar de manera armónica el proceso de generación de datos, con el objetivo de procurar la reproductibilidad y actualización en el tiempo.

El esquema del flujo de datos de Argendata puede ilustrarse así:



Se utilizó `Python` para el desarrollo de programas de gestión de recursos (interactuando con APIs de `Google` y `Github`, por ejemplo), para los módulos de aseguramiento de calidad (*QA* por sus siglas en inglés) y el de reportes para analistas (en combinación con `markdown` y `pandoc`).

Para la armonización del código fuente (*scripts* para el manejo de fuentes y generación de *outputs* asociado a cada gráfico del sitio), se definió un flujo de trabajo basado en un proyecto de `R`.

Finalmente la generación de gráficos interactivos publicados en el sitio web se basó en la librería *amcharts* de `JavaScript` (en conjunto con `html5`, y `css3`)

Características principales de argendataR

El paquete argendataR se desarrolló para dar una soluciones unificadas a problemas y necesidades recurrentes de los usuarios encargados de realizar la armonización de los procesos de generación de datos -etapa (iii)-. Estas soluciones implican:

- Gestionar la consulta, creación y acutalización de archivos desde el entorno de R a recursos en Google Drive que es la herramienta usadas por investigadores no programadores para trabajar colaborativamente.
- Facilitar la comparación entre datos nuevos y datos anteriores mediante métricas estadísticas comunes
- Facilitar la documentación de los datos exportados.
- Fijar el estándar de escritura de archivos csv del proyecto.

El objetivo de esta presentación es describir el *pipeline* de generación de datos en general, dando cuenta del flujo de trabajo "poliglota" antes descritpo. Y, en particular, poniendo el foco sobre el proceso de *extracción, transformación y carga* (*ETL*, por sus siglas en inglés) centrado en `R`, utilizando para ello `argendataR` - un paquete de `R` para facilitar las taras de esta etapa.