TITULO {arcenso}: primeros pasos desarrollando un paquete en comunidad con rOpenSci

**ENFOQUE**

## parte 1: rOpenSci

## parte 2: el armado del paquete {arcenso}

\_\_\_\_\_\_\_\_\_\_\_\_

# parte 1: rOpenSci

## Slide 1: hola

Hola, BIENVENIDOS A LA CHARLA de arcenso: primeros pasos desarrollando un paquete en comunidad con rOpenSci.

## Slide 2:

`arcenso` es un proyecto apoyado por el programa de Campeones y Campeonas de rOpenSci! cohorte 2023-2024, con Andrea Gómez Vargas como autora principal, Emanuel Ciardullo como co-autor y Luis D. Verde Arregoitia como mentor.

## Slide 3:

¿Que es el rOpenSci Champions Program o en español el programa de Campeones y Campeonas de rOpenSci?

## Slide 4:

Es un programa diseñado para identificar, reconocer y premiar a miembros apasionados de las comunidades. Desde rOpenSci reconocen la falta de diversidad en el ecosistema de las comunidades de software de investigación y de código abierto, y destacan que las comunidades de práctica pueden impulsar un cambio significativo.

Por esto, la importancia de apoyar a una cohorte diversa de personas, que colaboran en todos los aspectos de la creación de software de investigación y que son líderes de comunidades.

El desarrollo de este programa se de con una duración de 12 meses que incluye formación basada en cohortes, desarrollo de un proyecto y mentorías individuales.

El proyecto que presente fue la creación de un nuevo paquete,

## Slide 5:

El programa está enfocado y pensando para que participen personas que pertenecen a grupos histórica y sistemáticamente excluidos de las comunidades de software abierto y de software de investigación y que están interesadas en contribuir a rOpenSci y al ecosistema más amplio de comunidades de software de código abierto y de investigación.

## Slide 6: ¿Cuál fue(es) mi proyecto?

Dentro de las opciones e proyecto propuestas

Generar un paquete que permita disponer de los datos oficiales de los censos nacionales de población en Argentina provenientes del INDEC desde 1970 hasta 2022, homogeneizados, ordenados y listos para usar.

Contar con un paquete de información censal permitirá al sector público y privado, a los ciudadanos y a otros actores de la sociedad acceder a información actual e histórica sobre la población, los hogares y las viviendas de Argentina de una manera más accesible.

## Slide 7: ¿Por qué?

Actualmente los resultados históricos censales están disponibles en distintos formatos a través de libros físicos, PDFs, archivos en formato excel o en REDATAM, sin contar con un sistema o formato unificado que permita trabajar con los datos de estos seis periodos censales como base de datos.

Además, la presentación de los datos no está homogeneizada entre periodos dificultando la comparación histórica o en serie de la información disponible.

## 

## Slide 10: proceso de trabajo

## Slide 11: ¿Cómo empezar?

Como trabajar con datos elaborados de seis decadas distintas? cada año censal es un proceso de trabajo distinto y con el pasar de decada, hay un aumento exponencial de la cantidad de información publica

¿Cómo empezar y no morir en el intento?

les vamos a contar un poco

## Slide 12: trainings

Parte del programa consistió en una serie de capacitaciones virtuales de **seis semanas sobre: a) planificar y facilitar talleres atractivos e inclusivos para apoyar el éxito de quienes participan, b) conocer los diversos canales a través de los cuales se puede participar y contribuir a los proyectos de rOpenSci y de R, y c) conocimientos técnicos en desarrollo y revisión de software.**

## Slide 13: trainings

**Además, conte con la posibilidad de un seguimiento personalizado del proyecto con mi mentor, participantes del programa (tanto champions y mentores de este y la anterior cohorte) y community manager de rOpenSci y otros integrantes de ropensci para compartir avances, experiencias y sugerencias colectivamente. Y tuve la fortuna de conocer a algunos en persona en la posit::conf de este año.**

## Slide 14: Antes de las funciones: diseño conceptual

¿Cómo empecé? para darle forma al paquete tenia que estructurar la información que queria presentar y llegue al armado de esta clasificación censal, por año, alcance geografico, unidades de relevamiento y temáticas

## Slide 15: manos a la obra - EMA

Bueno, como llevamos toda esta idea a la práctica, a algo tangible y al paquete de R, usamos dos herramientas básicas, python para hacer web scraping en la página web del INDEC, para descargar los cuadros censales publicados para cada censo.

python no solo ayudó a la descarga de la información, sino tambien ayudar a listar la información descargada en algunos casos un mismo archivo de excel tenia, dos o tres cuadros censales, a veces con un índice en primer hoja otros no, que de manera automática se pudiera leer cada uno de los archivos y al menos a registrar el título de los cuadros que contenía ayudo mucho a la posterior organización del trabajo.

una vez descargados los archivos, con R se cambiaron los formatos originales que eran mas de presentación de la información a una tabla tidy que pueda ser usada para hacer cualquier tipo de análisis y procesamiento. Para que quede bien registrado que transformación se uso, para que pueda ser replicada o mejorada en el futuro.

## Slide 16: problemas EMA

Con que problemas nos encontramos en el camino?

A medida que avanzaban las décadas, los archivos excels se empezaban a contar de a cientos, en formatos que no estaban estandarizados, es decir, no podíamos convertir uno y repetir el mismo proceso para todos los demás porque cambiaba la organización del cuadro: como presentaba la información, filas, columnas, etc

## Slide 17: ¿Cómo nos organizamos? EMA

para enfrentar esta colosal tarea,

planteamos un proceso por etapas, donde vamos a ir avanzado censo por censo, desde 1970 en adelante, una vez finalizado uno avanzar con els siguiente, y lo que se plante es:

primero descargar cuadros, los clasificamos por temática y unidad de relevamiento, hacemos una subseleccion de cuadros a formatear, los pasamos a formato tidy uno a uno, hacemos una revision cruzada con una shiny, actualizamos el paquete en github y repetimos este proceso.

## Slide 18: etapas censales EMA

estas son las etapas de trabajamos que definimos para avanzar en el proyecto, actualmente estamos en la etapa 1, para que se hagan una idea, en 1970 tuvimos que convertir alrededor de 300 cuadros y para 1980 asciende a mas de 1100 cuadros, este aumento exponencial de cuadros se repite de un censo a otro

## Slide 19: andre

## slide 20 andre

## slide 21 andre - finalmente llegamos a una primera version de arcenso.

## slide 22-23- 24 EMA

## slide 25 shiny para usuarios

## {ARcenso} un paquete con rOpenSci (Gomez Vargas & Ciardullo)