



El futuro digital  
es de todos

MinTIC

# OPEN REFINE



## REGLAS DEL TALLER

**Horario**  
**10:00 a.m.**  
**11:00 a.m.**

**No abandonar  
la reunión**

**Presencia  
plena**

**Preguntas por  
medio del link  
o al final de la  
presentación**



# INTRODUCCIÓN

La falta de calidad de los datos es uno de los principales inconvenientes a los que se enfrentan los responsables de sistemas de información y las empresas en general, pues representa claramente uno de los problemas “ocultos” más graves y persistentes en cualquier organización.

En efecto, una buena calidad de datos es el activo corporativo más potente, ya que permite acelerar el crecimiento y administrar de mejor manera los costos y las iniciativas para obtener mejores rentabilidades.



# ¿Qué es la iniciativa de software libre?



El futuro digital  
es de todos

MinTIC

# Objetivos de la iniciativa de Software libre



## ¿Qué es OPEN REFINE?

Es una herramienta que nos permite perfilar y limpiar conjuntos de datos.



# CARACTERÍSTICAS

Es una aplicación de escritorio que se puede descargar e instalar desde la web : <http://openrefine.org/>  
Está disponible para **Windows, Mac y Linux.**

Utiliza lenguaje GREL en el que podemos usar expresiones regulares, concretamente en sus funciones: **replace, match, partition, repartition y split.**

Permite transformar archivos de diferentes extensiones como XLS, CSV, JSON, XML, TSV y las **spreadsheets** de Google.

Es compatible con cualquier navegador.



## CONSEJOS DE ELECCIÓN DE UN LENGUAJE DE PROGRAMACIÓN Y RAZONES PARA APRENDER OPEN REFINE

**Ámbito  
de aplicación.**

**Costo.** El factor económico es también crucial en nuestra decisión

**Características**  
propias del lenguaje o entorno de programación.



# PROCEDIMIENTO RÁPIDO PARA LA DESCARGA E INSTALACIÓN DE LA HERRAMIENTA

Entramos a la página oficial de OPEN REFINE  
y seleccionamos la opción **DOWNLOAD - DESCARGAR**

<https://openrefine.org/download.html>

## OpenRefine 3.4.1

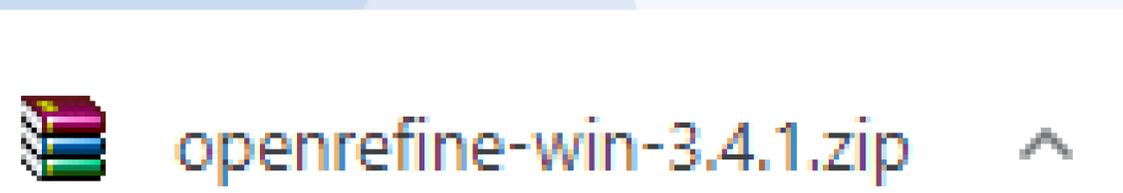
La última versión estable de OpenRefine 3.4.1, publicada el 24 de septiembre de 2020. Haga una copia de seguridad del directorio de su espacio de trabajo antes de instalar e informe cualquier problema que encuentre. Se proporciona un registro de cambios en la [página de lanzamiento](#).

- **Kit de Windows**, este requiere que Java esté instalado en su computadora. Descargue, descomprima y haga doble clic en *openrefine.exe* o *refine.bat* si el primero no funciona.
- **Kit de Windows con Java integrado**, incluye OpenJDK Java, disponible bajo la licencia GPLv2 + CE. Descargue, descomprima y haga doble clic en *openrefine.exe* o *refine.bat* si el primero no funciona.
- **Kit de Mac**, Descargue, abra, arrastre el icono a la carpeta de Aplicaciones y haga doble clic en él. No es necesario instalar Java por separado.
- **Kit de Linux**, descargue, extraiga, luego escriba *./refine* para comenzar. Esto requiere que Java esté instalado en su computadora.

**SELECCIONAR**

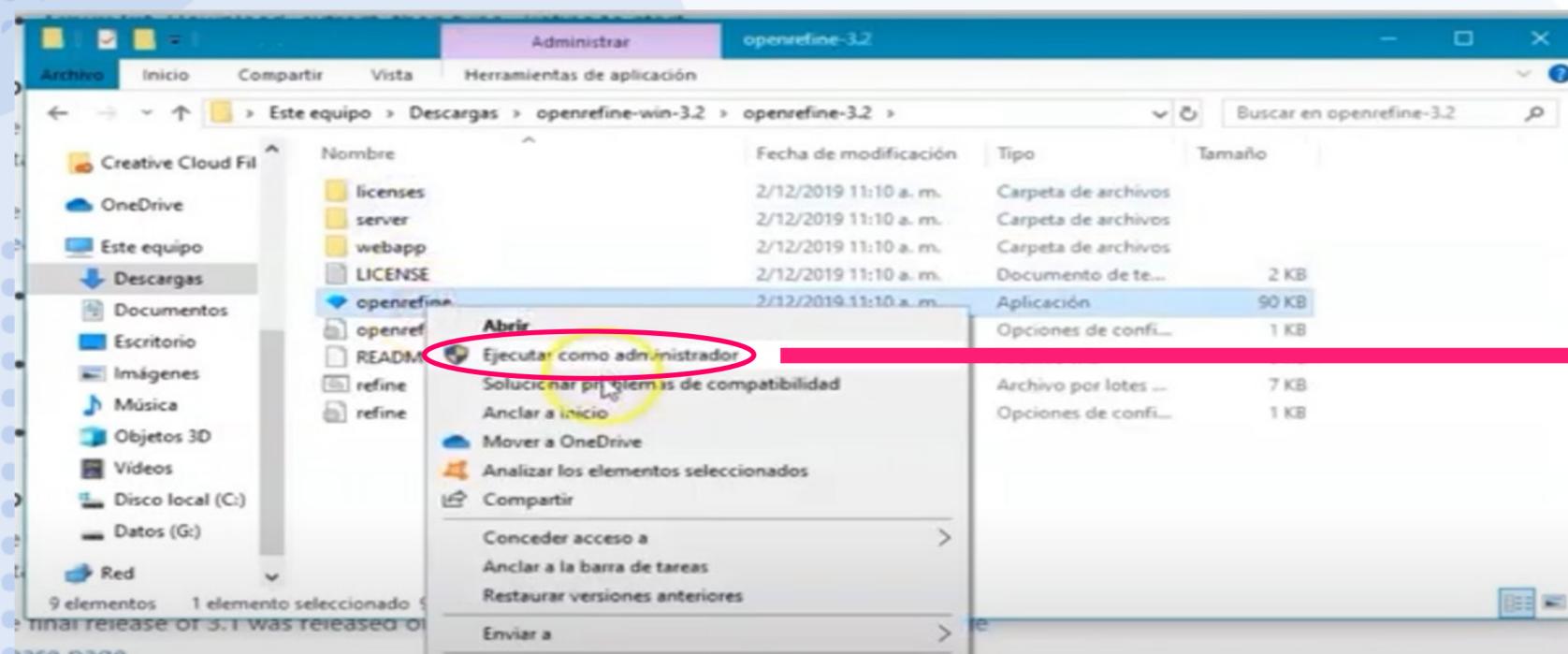


# Nos descarga una carpeta **.ZIP** para descomprimir



Vamos a la carpeta de  
descarga, seleccionamos  
la herramienta

**OPEN REFINE**, y damos  
clic derecho a la opción  
**EJECUTAR COMO  
ADMINISTRADOR**





## MIENTRAS DESCARGA...

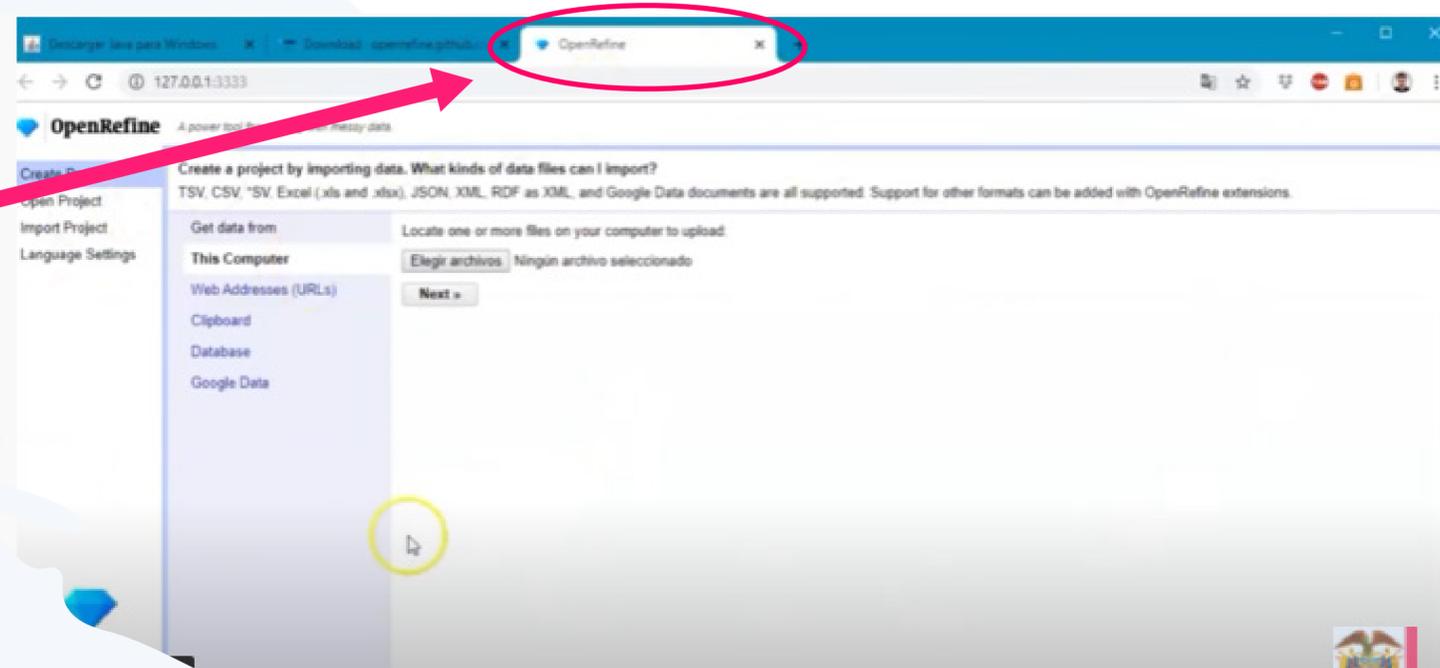
Nos muestra esta interfaz donde ya se está ejecutando el sistema

```
Open Refine
C:\Users\Rentadvisor\Downloads\openrefine-win-3.2\openrefine.exe
[refine_server] Starting Server bound to '127.0.0.1:3333' (0ms)
[refine_server] Initializing context: '/' from 'C:\Users\Rentadvisor\Downloads\openrefine-win-3.2\openrefine-3.2\webapp' (11ms)
[refine_server] class path contains multiple SLF4J bindings.
[refine_server] binding in [jar:file:C:/Users/Rentadvisor/Downloads/openrefine-win-3.2/openrefine-3.2/server/target/lib/slf4j-api-1.7.18.jar!/org/slf4j/impl/StaticLoggerBinder.class]
[refine_server] binding in [jar:file:C:/Users/Rentadvisor/Downloads/openrefine-win-3.2/openrefine-3.2/webapp/WEB-INF/lib/slf4j-api-1.7.18.jar!/org/slf4j/impl/StaticLoggerBinder.class]
[refine_server] See https://www.slf4j.org/codes.html#multiple_bindings for an explanation.
[refine_server] Using the selected logger factory class org.slf4j.impl.Log4jLoggerFactory
[refine_server] Starting OpenRefine 3.2 [55c921b]... (1157ms)
[refine_server] initializing FileManager with dir (1ms)
[refine_server] C:\Users\Rentadvisor\AppData\Local\OpenRefine (0ms)
```

Automáticamente se abrirá una pestaña nueva en el navegador

**CON LA SIGUIENTE DIRECCIÓN IP**

<http://127.0.0.1:3333/>





El futuro digital  
es de todos

MinTIC

# EJEMPLO DE USO



OpenRefine *A power tool for working with messy data.*

Create Project  
Open Project  
Import Project  
Language Settings

Version 3.4.1 [437dc4d]  
Preferences  
Help  
About

Create a project by importing data. What kinds of data files can I import?  
TSV, CSV, \*SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for other formats can be added with OpenRefine extensions.

Get data from

**This Computer**  Ningún archivo seleccionado

Web Addresses (URLs)  
Clipboard  
Database  
Google Data

**ENTRAMOS EN LA INTERFAZ DE OPEN REFINE**



## Seleccionamos

**Elegir archivos**

lo que nos lleva a la carpeta  
donde elegiremos el documento  
y le damos clic en **ABRIR**

Create a project by importing data. What kinds of data files can I import?  
TSV, CSV, \*SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for other formats can be added with OpenRefine extensions.

Get data from: **This Computer** | Locate one or more files on your computer to upload:

**Elegir archivos** | Directorio E...0307 | **Next »**

Abrir

Este equipo > Descargas > | Buscar en Descargas

Organizar | Nueva carpeta

OneDrive | Este equipo | Descargas | Documentos | Escritorio | Galaxy A7 (2017) | Imágenes | Música | Objetos 3D | Videos | Windows (C:)

Hoy (4) | Directorio Empresas Sociales del Estado 03... | DataSync-1.9.6 (1) | DESCARGAS | openrefine-win-3.4.1

Ayer (6) | Al principio de esta semana (1) | La semana pasada (25) | Al principio de este mes (12)

Nombre de archivo: Directorio Empresas Sociales del E | Todos los archivos

**Abrir** | Cancelar



OpenRefine A power tool for working with messy data.

« Start Over Configure Parsing Options Project name Directorio Empresas Sociales del E Tags Create Project »

	GOBERNACION DEL VALLE DEL CAUCA	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10	Column 11
1.	SECRETARIA DEPARTAMENTAL DE SALUD										
2.	DIRECTORIO DE EMPRESAS SOCIALE DEL ESTADO E.S.E DEL VALLE DEL CAUCA										
3.											
4.	No.	Codigo del Departamento	Nombre del Departamento	Codigo del Municipio	Nombre del Municipio	Codigo de Habilitación	NIT	Razon Social Organización	Gerente	Dirección	Teléfono
5.		1	76 Valle del Cauca	020	ALCALÁ	7602001657	891900438	EMPRESA SOCIAL DEL ESTADO HOSPITAL SAN VICENTE DE PAUL	CARLOS ARBEY PELAEZ GOMEZ	BARRIO VILLA FERNANDA	092-2004120-3113182460

Parse data as

- Excel files
- JSON files
- Line-based text files
- CSV / TSV / separator-based files
- Fixed-width field text files
- PC-Axis text files
- MARC files
- JSON-LD files
- RDF/N3 files
- RDF/N-Triples files

Worksheets to Import

Select All Unselect All

- Directorio Empresas Sociales del Estado 03072020.xlsx#ESE 61 rows

Ignore first 0 line(s) at beginning of file

Parse next 1 line(s) as column headers

Discard initial 0 row(s) of data

Load at most 0 row(s) of data

Store blank rows

Store blank cells as nulls

Store file source (file names, URLs) in each row

Update Preview

Version 3.4.1 [437dc4d]

Preferences Help About

YA TENEMOS EL ARCHIVO EN NUESTRA INTERFAZ PARA TRABAJAR SOBRE ELLA



OpenRefine *Una herramienta poderosa para trabajar con datos desordenados.*

Crear proyecto "Comenzar de nuevo" Configurar opciones de análisis Nombre del proyecto **EJERCICIO WEBINAR** Etiquetas

Proyecto abierto

Importar proyecto

Configuraciones de idioma

	GOBERNACION DEL VALLE DEL CAUCA	Columna 2	Columna 3	Columna 4	Columna 5	Columna 6	Columna 7	Columna 8	Columna 9	Columna 10	Columna 11
1.	SECRETARIA DEPARTAMENTAL DE SALUD										
2.	DIRECTORIO DE EMPRESAS SOCIALES DEL ESTADO ESE DEL VALLE DEL CAUCA										
3.											
4.	No.	Codigo del Departamento	Nombre del Departamento	Codigo del Municipio	Nombre del Municipio	Codigo de Habilitación	LIENDRE	Razon Social Organización	Gerente	Dirección	Teléfono
5.		1	76 Valle del Cauca	020	ALCALÁ	76020016	891900438	HOSPITAL EMPRESA SOCIAL DEL ESTADO SAN VICENTE DE PAUL	CARLOS ARBEY PELAEZ GOMEZ	BARRIO VILLA FERNANDA	092-2004120-3113182460
6.		2	76 Valle del	036	ANDALUCÍA	76036017	891900390	ESE HOSPITAL	PAIII A	CARRERA 4 No	092-2235100

Analizar datos como

Archivos de Excel

Archivos JSON

Archivos de texto basados en líneas

Archivos basados en CSV / TSV / separadores

Archivos de texto de campo de ancho fijo

Archivos de texto de PC-Axis

Archivos MARC

Archivos JSON-LD

Archivos RDF / N3

Hojas de trabajo para importar

Directorio Empresas Sociales del Estado 61 03072020.xlsx # ESE 61 filas

Ignorar primero 0 línea (s) al comienzo del archivo

Analizar siguiente 1 línea (s) como encabezados de columna

Descartar inicial 0 fila (s) de datos

Cargar como máximo 0 fila (s) de datos

Almacenar filas en blanco

Almacenar celdas en blanco como nulos

Almacene la fuente del archivo (nombres de archivo, URL) en cada fila

CREAR PROYECTO

Personalizamos el nombre del proyecto



OpenRefine: Una herramienta poderosa para trabajar con datos desordenados.

Crear proyecto: "Comenzar de nuevo" | Configurar opciones de análisis | Nombre del proyecto: EJERCICIO WEBINAR | Etiquetas: | Crear proyecto »

	GOBERNACION DEL VALLE DEL CAUCA	Columna 2	Columna 3	Columna 4	Columna 5	Columna 6	Columna 7	Columna 8	Columna 9	Columna 10	Columna 11
1.	SECRETARIA DEPARTAMENTAL DE SALUD										
2.	DIRECTORIO DE EMPRESAS SOCIALES DEL ESTADO ESE DEL VALLE DEL CAUCA										
3.											
4.	No.	Codigo del Departamento	Nombre del Departamento	Codigo del Municipio	Nombre del Municipio	Codigo de Habilitación	LIENDRE	Razon Social Organización	Gerente	Dirección	Teléfono
5.	1	76	Valle del Cauca	020	ALCALÁ	7602001657	891900438	HOSPITAL EMPRESA SOCIAL DEL ESTADO SAN VICENTE DE PAUL	CARLOS ARBEY PELAEZ GOMEZ	BARRIO VILLA FERNANDA	092-2004120-3113182460
6.	2	76	Valle del	036	ANDALUCÍA	7603601707	891900390	ESE HOSPITAL	PAIIA	CARRERA 4 No	092-2235100

Analizar datos como:

- Archivos de Excel
- Archivos JSON
- Archivos de texto basados en líneas
- Archivos basados en CSV / TSV / separadores
- Archivos de texto de campo de ancho fijo
- Archivos de texto de PC-Axis
- Archivos MARC
- Archivos JSON-LD

Hojas de trabajo para importar:  Directorio Empresas Sociales del Estado 61 filas (03072020.xlsx # ESE)

Opciones de configuración:

- Ignorar primero: 0 línea(s) al comienzo del archivo
- Analizar siguiente: 1 línea(s) como encabezados de columna
- Descartar inicial: 0 fila(s) de datos
- Cargar como máximo: 0 fila(s) de datos
- Almacenar filas en blanco
- Almacenar celdas en blanco como nulos
- Almacene la fuente del archivo (nombres de archivo, URL) en cada fila

Actualización de Vista previa

**Trabajando con funciones de GREL** en Open Refine nos permite crear nuestras propias facetas ad hoc usando GREL, Jython o Clojure. GREL (Google Refine Expression Language) es un lenguaje de programación con un gran número de funciones que permite realizar tareas de depuración avanzadas. Veamos algún caso sencillo. Vamos a modificar las abreviaturas de los nombres de los viales a su denominación extendida. En el submenú de operaciones de la columna 'vial' seleccionaremos **Edits cells -> Transform** e ingresamos la siguiente expresión GREL: `value.replace('Avdav', 'Avenida')`



Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started? Watch these screencasts

	All	GOBERNACION	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10	Column 11
1.	SECRETARIA DEPARTAMENTAL DE SALUD											
2.	DIRECTORIO DE EMPRESAS SOCIALES DEL ESTADO E.S.E DEL VALLE DEL CAUCA											
3.												
4.	No.	Codigo del Departamento	Nombre del Departamento	Codigo del Municipio	Nombre del Municipio	Codigo de Habilitación	NIT					
5.		1	76	Valle del Cauca	020	ALCALÁ	7602001657	891900438				
6.		2	76	Valle del Cauca	036	ANDALUCÍA	7603601707	891900390				
7.		3	76	Valle del Cauca	041	ANSERMANUEVO	7604104065	891900446				
8.		4	76	Valle del Cauca	054	ARGELIA	7605401652	891901101				
9.		5	76	Valle del	100	ROI ÍVAR	7610002524	891900414				

- Transform...
- Common transforms

- Facet
- Text filter
- Edit cells
- Edit column
- Transpose
- Sort...
- View
- Reconcile

Seleccionamos la opción

- EDIT CELL
- TRANSFORM



### Transformación de texto personalizado en la columna Columna 10

Expresión

Idioma

`value.replace('CRA', 'CARRERA')`

Sin error de sintaxis.

Avance

Historia

Sembrado de estrellas

Ayudar

3.	nulo	Error: el reemplazo espera 3 cadenas o 1 cadena, 1 expresión regular y 1 cadena
4.	Dirección	Dirección
5.	BARRIO VILLA FERNANDA	BARRIO VILLA FERNANDA
6.	CRA 4 No 13-31	CARRERA 4 No 13-31
7.	CARRERA 8 No 8-57	CARRERA 8 No 8-57
8.	CRA 4 No 1-65	CARRERA 4 No 1-65
9.	CALLE 8 No 5-09	CALLE 8 No 5-09
10	CRA 17 No 2-08	CARRERA 17 No 2-08

Por error

- mantener original
- establecer en blanco
- error de tienda

Vuelva a transformar hasta  veces hasta que no haya cambios

OK

Cancelar



Las expresiones las podemos encadenar complicándolas tanto como queramos. El ejemplo de la imagen siguiente realiza múltiples reemplazos de una sola vez.

### Custom text transform on column vial

Expression Language Google Refine Expression Language (GREL) ▾

```
value
.replace('Avenida','Avenida de')
.replace('Avda','Avenida de')
.replace('Calle','Calle de')
.replace('C/','Calle de ')
.replace('Pza','Plaza de')
.replace('Plaza','Plaza de')
.replace('C.C.','Centro Comercial')
```

No syntax error.

**Preview** [History](#) [Starred](#) [Help](#)

row	value	value .replace('Avenida','Avenida de') .replace('Avda','Avenida de') .replace('Calle','Calle de') .replace('C/','Calle de ') .replace('Pza','Plaza de') .replace('Plaza','Plaza de') .replace('C.C.','Centro Comercial')
1.	Calle José María Cossío	Calle de José María Cossío
2.	Avda General Dávila	Avenida de General Dávila
3.	Calle Jimenez Díaz	Calle de Jimenez Díaz

On error  keep original  Re-transform up to  times until no change  
 set to blank  
 store error



### Transformación de texto personalizado en la columna Columna 10

Expresión

value

Idioma

- Lenguaje de expresión general refinado (GREL) ▼
- Lenguaje de expresión general refinado (GREL)
- Python / Jython**
- Clojure

Sin error de sintaxis.

Avance

Historia

Sembrado de estrellas

Ayudar

hilera	valor	valor
1.	nulo	nulo
2.	nulo	nulo
3.	nulo	nulo
4.	Dirección	Dirección
5.	BARRIO VILLA FERNANDA	BARRIO VILLA FERNANDA
6.	CRA 4 No 13-31	CRA 4 No 13-31
7.	CALLE 8 N 8-57	CALLE 8 N 8-57

Por error

- mantener original
- establecer en blanco
- error de tienda

Vuelva a transformar hasta  veces hasta que no haya cambios

OK

Cancelar

Es una herramienta de código abierto



# Kit de herramientas que pueden encontrar

## <https://gobiernodigital.mintic.gov.co/portal/Iniciativas/Software-libre/>

### MySQL Workbench

- Instalación de MySQL Workbench.
- Entorno y funcionalidades de MySQL Workbench.
- Ejercicio práctico en MySQL Workbench.

### Open Refine

- Descargar Open Refine.
- Instalación Open Refine.
- Funcionalidades Open Refine.
- Datos Georreferenciados con Open Refine

### Python:

Instalación de Python con pandas y Jupyter:

Comandos e Instrucciones Video1.txt.

- Funcionalidades y conceptos básicos:
- Comandos e Instrucciones Video2.txt.
- Entrada y visualización de datos desde datos.gov.co:  
Comandos e Instrucciones Video3.txt..

### Tock Chatbot:

Instalación de Tock:

Tock\_Video1\_Código.txt.

- Funcionalidades de Tock - Parte 1.
- Funcionalidades de Tock - Parte 2.

### X-ROAD:

Conferencia sobre la implementación de X-ROAD, herramienta de interoperabilidad.

- Integración de servicios web a X-Road.
- Video 1 Generalidades X-Road.
- Video 2 Arquitectura y tecnología X-Road.
- Video 3 Implementación X Road.



## Conclusiones

Hemos demostrado cómo resolver problemas recurrentes tales como duplicados e inconsistencias ortográficas de forma automatizada con la ayuda de Open Refine, todos los datos están sucios, pero tú puedes hacer algo para remediarlo

No dudes en experimentar con las funciones de limpieza, ya que estás realizando estos pasos en una copia de tu conjunto de datos, y Open Refine te permite rastrear todos tus pasos (y volver atrás) en el caso de que hayas cometido un error.



El futuro digital  
es de todos

MinTIC

**En este momento daremos unos minutos para llenar el formulario de diagnóstico**



El futuro digital  
es de todos

MinTIC

# PREGUNTAS