

Limpiado de fechas en OpenRefine con llamado a Canadensys Date Parsing

BREVE INTRODUCCIÓN

Uno de los campos sobre el que podemos corroborar la calidad de los datos es el campo de fecha: eventDate.

Recordemos primero la **definición de eventDate en el estándar Darwin Core**

<http://rs.tdwg.org/dwc/terms/index.htm#eventDate>:

The date-time or interval during which an Event occurred. For occurrences, this is the date-time when the event was recorded. Not suitable for a time in a geological context. Recommended best practice is to use an encoding scheme, such as ISO 8601:2004(E).

Si pensamos en un ejemplar de museo, eventDate refiere a cuándo fue colectado el ejemplar. Si pensamos en una observación, eventDate refiere a cuándo fue realizada esa observación.

Darwin Core sugiere que se utilice para capturar la información de fecha el estándar **ISO 8601:2004(E)** (https://en.wikipedia.org/wiki/ISO_8601). Para fechas únicas, este estándar tiene el siguiente formato:

AAAA-MM-DDTHH:mmX

Donde:

AAAA: año, con cuatro dígitos.

MM: mes, con dos dígitos. E.g.: mayo sería 05.

DD: día, con dos dígitos. E.g.: segundo día de un mes sería 02.

T: indica que lo que viene a continuación es la hora.

HH: horas, con dos dígitos, en formato de 24 hs.

mm: minutos, con dos dígitos.

X: indica la zona horaria. La zona horaria se determina tomando como base UTC (Coordinated Universal Time). Si uno está justo sobre la zona horaria UTC, X se reemplaza por "Z". Si uno está en otra zona horaria, debe reemplazarse X por la diferencia horaria correspondiente.

Por ejemplo, Argentina es UTC-3, o sea, 03horas00minutos al oeste (-) de UTC, por lo cual X debe reemplazarse por “-0300”.

NOTAS:

- De este formato, uno puede utilizar tanto el formato completo (incluyendo la hora) como sólo la primera parte, AAAA-MM-DD.

- Este formato también puede utilizarse para expresar rangos de fecha de manera estandarizada. Para ello, se usa el mismo formato y se separan las fechas con barras “/”, ver ejemplos abajo.

EJEMPLOS:

FECHA ORIGINAL	FECHA ESTANDARIZADA
12 Feb 1809	1809-02-12
12/02/1809	1809-02-12
Jun 1906	1906-06
1971	1971
20 Feb 2009 8:40am UTC	2009-02-20T08:40Z
8 Mar 1963 2:07pm, en la zona horaria 6 horas más temprano que UTC	1963-03-08T14:07-0600
13-15 Nov 2007	2007-11-13/15
1 Mar 2007 1pm UTC – 11 May 2008 3:30pm UTC	2007-03-01T13:00:00Z/2008-05- 11T15:30:00Z

LIMPIEZA DE FECHA USANDO OPENREFINE Y CANADENSYS DATE PARSING

Muchas veces, a pesar de lo que indica el estándar Darwin Core, encontramos en el campo eventDate fechas que no siguen el formato sugerido. Para limpiarlas, podemos hacer uso de la herramienta que ofrece Canadensys: Date Parsing (<http://data.canadensys.net/tools/dates>).

Esta herramienta permite interpretar fechas, devolviéndolas en formato estándar. Ejemplos de los tipos de valores que puede interpretar son:

Jun 13, 2008
15 Jan 2011

2009 IV 02
2 VII 1986

Algunas fechas, sin embargo no las interpreta, veamos el siguiente ejemplo:

Date parsing results

original	year	month	day	ISO 8601
2-4-1980				
2/4/1980				
2/13/1980	1980	2	13	1980-02-13
13/2/1980	1980	2	13	1980-02-13

En las dos líneas inferiores, "13" sólo puede referir a días, pues no hay un mes "13".

En las dos líneas superiores, en cambio, "2" y "4" pueden ambos referir a mes y día. Como en distintas partes del mundo se utilizan sistemas distintos (primero se pone día y luego mes, o viceversa), la herramienta no puede determinar inequívocamente cuál es cuál, y por ende no hace la interpretación.

Debemos tener esto en cuenta cuando utilicemos la herramienta para limpiar nuestros datos.

Ahora sí, invoquemos Date Parsing desde OpenRefine. Para ello:

1. Sobre la columna eventDate: Edit column → Add column by fetching URL...
Lo que haremos con esto es crear una nueva columna con los resultados que indique Canadensys.
2. En la ventana que aparece, nombrar la nueva columna y pegar en el cuadro de texto la siguiente expresión:

```
"http://data.canadensys.net/tools/dates.json?data="+escape(cells["eventDate"].value,"url")"
```

Lo que hace esta expresión es pedirle a herramienta que evalúe los valores del campo eventDate y que nos envíe los resultados en formato JSON.

Add column by fetching URLs based on column eventDate

New column name Throttle delay milliseconds

On error set to blank store error

Formulate the URLs to fetch:

Expression Language No syntax error.

Preview History Starred Help

row	value	"http://data.canadensys.net/tools/dates.json?data="+escape(cells["eventDate"].value,"url")"
1.	9/16/1967	http://data.canadensys.net/tools/dates.json?data=9%2F16%2F1967
2.	6/4/1968	http://data.canadensys.net/tools/dates.json?data=6%2F4%2F1968
3.	5/27/1968	http://data.canadensys.net/tools/dates.json?data=5%2F27%2F1968
4.	2/9/1973	http://data.canadensys.net/tools/dates.json?data=2%2F9%2F1973
5.	2/1/1996	http://data.canadensys.net/tools/dates.json?data=2%2F1%2F1996
6.	2/1/1996	http://data.canadensys.net/tools/dates.json?data=2%2F1%2F1996

OK Cancel

3. La limpieza puede tomar bastante tiempo, incluso horas, sea paciente... váyase a almorzar, o incluso a dormir y lo revisa al día siguiente... Cuando vuelva, encontrará el nuevo campo con los valores estandarizados! En formato JSON...

eventDate	Canadensys_eventDate
9/16/1967	<pre>{"data":{"results":[{"originalValue":"9/16/1967","year":1967,"month":9,"day":16,"iso8601":"1967-09-16","partial":false}]}}</pre>
6/4/1968	<pre>{"data":{"results":[{"originalValue":"6/4/1968","error":"The date [6-4-1968] could not be precisely determined.","partial":true}]}}</pre>

Fíjese que en el primer caso de la figura, Canadensys ha podido resolver la fecha, mientras que en el segundo caso no ha podido, dado que no puede interpretar inequívocamente “6” y “4” como día y mes o viceversa (como se explica más arriba).

4. Ahora que tenemos el JSON, queremos extraer de allí los valores de interés. Podríamos extraer sólo la fecha en formato ISO, o también año, mes y día en campos separados. Para ello, a partir de la columna que tiene el JSON, creamos nuevas columnas: Edit column → Add column based on this column.

Para extraer sólo la fecha en formato ISO, en la ventana nombramos la nueva columna y en el cuadro de texto pegamos la siguiente expresión:

```
forEach(value.parseJson().get("data").get("results"),v,v.get("iso8601"))[0])
```

Add column based on column Canadensys_eventDate

New column name:

On error: set to blank store error copy value from original column

Expression: No syntax error.

Language:

Preview History Starred Help

row	value	forEach(value.parseJson().get("data").get("results"),v,v.get("iso8601"))
1.	{\"data\":{\"results\": [{\"originalValue\":\"9/16/1967\",\"year\":1967,\"month\":\"09-16\",\"partial\":false}]}}	1967-09-16
2.	{\"data\":{\"results\": [{\"originalValue\":\"6/4/1968\",\"error\":\"The date [6-4-1968] could not be precisely determined.\",\"partial\":true}]}}	null
3.	{\"data\":{\"results\": [\"1968-05-27\"]}}	1968-05-27

OK Cancel

Para extraer el año, mes o día, pegamos en cambio una de las siguientes expresiones:

Año:

```
forEach(value.parseJson().get("data").get("results"),v,v.get("year"))[0])
```

Mes:

```
forEach(value.parseJson().get("data").get("results"),v,v.get("month"))[0])
```

Día:

```
forEach(value.parseJson().get("data").get("results"),v,v.get("day"))[0])
```

Veremos que algunos de los resultados serán nulos, estos corresponden a los casos que Canadensys no ha podido resolver (como se explica más arriba).

eventDate	Canadensys_eventDate	ISO_eventDate
9/16/1967	<pre>{ "data": { "results": [{ "originalValue": "9/16/1967", "year": 1967, "month": 9, "day": 16, "iso8601": "1967-09-16", "partial": false }] } }</pre>	1967-09-16
6/4/1968	<pre>{ "data": { "results": [{ "originalValue": "6/4/1968", "error": "The date [6-4-1968] could not be precisely determined.", "partial": true }] } }</pre>	

5. Para terminar de limpiar las fechas, entonces, tendremos que revisar los valores que no hayan sido estandarizados por la herramienta. Para ello, sobre el campo ISO_eventDate podemos armar una faceta y seleccionar el valor "blank". Luego, armamos una faceta sobre el campo eventDate (el que tenía los valores originales) y si estos son pocos, podemos hacer un chequeo manual y completar el campo ISO_eventDate.