BIEN 3 - Bug #950

fix view_full_occurrence_individual_view rows with is_geovalid NULL

09/05/2014 04:51 PM - Aaron Marcuse-Kubitza

Status:	Resolved	**Start date:**	09/05/2014
Priority:	High	**Due date:**	
Assignee:	Aaron Marcuse-Kubitza	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:			

Description

test case

eg. this happens for CVS rows:

datasource	country	state_province	county	latitude	longitude	is_geovalid
CVS	United States	TENNESSEE	Sevier	35.654350967	-83.444906936	<NULL>

```
SET enable_seqscan = off;
SET enable_mergejoin = off;
SELECT * FROM view_full_occurrence_individual_view WHERE datasource = 'CVS' LIMIT 1;
```

and FIA rows:

datasource	country	state_province	county	latitude	longitude	is_geovalid
FIA	United States	Alabama	Covington	31.39	-86.36	<NULL>
FIA	United States	Alabama	Escambia	31.17	-86.72	<NULL>

```
SET enable_seqscan = off;
SET enable_mergejoin = off;
SELECT
DISTINCT ON (country, state_province, county)
*
FROM (SELECT * FROM view_full_occurrence_individual_view WHERE datasource = 'FIA' LIMIT 1000) s
;
```

info

since is_geovalid is NULL, and state_province is falling back to the unscrubbed value, this indicates it is unable to find a geoscrub.geoscrub_output row to join to

however, there is a matching row in the geoscrub DB's result table:

decimallatitude	decimallongitude	country	stateprovince	county	countrystd	stateprovincestd	countystd	latlonvalidity	countryvalidity	stateprovincevalidity	countyvalidity
35.654350967	-83.444906936	United States	TENNESSEE	Sevier	United States	Tennessee	Sevier	1	3	3	3

```
SELECT * FROM geoscrub WHERE
    country = 'United States'
AND    stateprovince = 'TENNESSEE'
AND    county = 'Sevier'
AND    decimallatitude = 35.654350967
AND    decimallongitude = -83.444906936
LIMIT 1
```

```
;
```

this suggests that the problem is in the transfer from the geoscrub DB to vegbien

```
grep -E '^35\.654350967,-83\.444906936,United States,TENNESSEE,Sevier.*$' inputs/.geoscrub/geoscrub_output/geoscrub.csv # returns no rows
```

because this row is not present in the extract, the problem is in the export from the geoscrub DB

then determine if decimal truncation is causing the problem:

```
grep -E '^35\.65.*,-83\.44.*,United States,TENNESSEE,Sevier.*$' inputs/.geoscrub/geoscrub_output/geoscrub.csv # returns no rows
```

but this doesn't appear to be the case

trying just the first coordinate:

```
grep -E '^35\.65.*$' inputs/.geoscrub/geoscrub_output/geoscrub.csv # returns many rows
grep -E ^35.654350967 inputs/.geoscrub/geoscrub_output/geoscrub.csv # returns 1 row:
---
35.654350967,-83.444906936,United States,Tennessee,Sevier,United States,Tennessee,Sevier,1,3,3,3
---
```

the returned row has the wrong value for stateprovince (the scrubbed value instead of the input value), so the problem is in the export of the placenames, not the coordinates

regenerating the extract:

```
ssh -t vegbiendev.nceas.ucsb.edu exec sudo -u aaronmk -i
rm=1 inputs/.geoscrub/geoscrub_output/geoscrub.csv.run export_
grep -E ^35.654350967 inputs/.geoscrub/geoscrub_output/geoscrub.csv # returns 1 row:
---
35.654350967,-83.444906936,United States,TENNESSEE,Sevier,United States,Tennessee,Sevier,1,3,3,3
---
```

the returned row is now correct, so there was likely some problem in the exporting of the geoscrub DB results

the next step will be to figure out if the other unmatched rows are also fixed by the reload

---

this appears to be fixed by the extra_float_digits fix for the 1st bug of #955. CVS, which has numerous rows with the maximum # of floating-point digits, now has is_geovalid populated for these rows.

History

#1 - 09/05/2014 06:15 PM - Aaron Marcuse-Kubitza

- Description updated

#2 - 09/18/2014 02:33 PM - Aaron Marcuse-Kubitza

- Description updated

#3 - 09/18/2014 02:49 PM - Aaron Marcuse-Kubitza

- Description updated

#4 - 09/18/2014 03:06 PM - Aaron Marcuse-Kubitza

- Description updated

#5 - 09/18/2014 03:12 PM - Aaron Marcuse-Kubitza

- Description updated

#6 - 09/18/2014 03:54 PM - Aaron Marcuse-Kubitza

- Description updated

#7 - 09/18/2014 03:54 PM - Aaron Marcuse-Kubitza

- Description updated

#8 - 09/18/2014 05:37 PM - Aaron Marcuse-Kubitza

- Description updated

- % Done changed from 0 to 50

#9 - 10/17/2014 11:26 AM - Aaron Marcuse-Kubitza

- % Done changed from 50 to 100

- Description updated

- Status changed from New to Resolved

#10 - 10/17/2014 11:27 AM - Aaron Marcuse-Kubitza

- Description updated