

Data quality checks

EURISCO training workshop, 19th to 21st May 2015, Tirana, Albania

Helmut Knüpffer & Stephan Weise
19 May 2015



COMPLETENESS OF INFORMATION

Completeness of information

- Often, only limited information about certain accessions
- Some descriptors only sparsely populated
- In many cases, information available in the respective information systems, but not in EURISCO

Completeness per descriptor (all NI)

Descriptor	Completeness [%]
SPECIES	91,98%
SPAUTHOR	87,66%
STORAGE	81,92%
ORIGCTY	78,74%
SAMPSTAT	72,29%
CROPNAME	68,50%
ACQDATE	68,20%
ACCENAME	62,80%
MLSSTAT	48,59%
COLLSRC	43,70%
DONORDESCR	38,12%
ACCEURL	36,63%
OTHERNUMB	35,09%
COLLNUMB	31,63%
COLLSITE	31,19%
COLLCODE	31,15%

Descriptor	Completeness [%]
SUBTAXA	31,05%
COLLDESCR	30,43%
COLLDATE	28,07%
DONORCODE	26,98%
DONORNUMB	20,89%
AEGISSTAT	20,28%
SUBTAUTHOR	19,60%
LATITUDE	15,69%
LONGITUDE	15,65%
ELEVATION	14,19%
DUPLDESCR	11,87%
DUPLSITE	11,20%
BREDDESCR	8,47%
ANCEST	6,91%
BREDCODE	6,32%
REMARKS	5,83%

as of 2015-05-11

Completeness per NI

- 36 descriptors (4 mandatory) → 32 descriptors considered
- $\text{Completeness}_{\text{NI}} = \frac{\text{sum of all actual descriptor values}_{\text{NI}}}{\text{sum of all possible descriptor values}_{\text{NI}}}$

Completeness per NI

NICODE	Completeness [%]
MNE	60,00%
NGB	54,00%
SVK	54,00%
ALB	50,00%
NLD	48,00%
CYP	46,00%
ESP	46,00%
DEU	46,00%
UKR	46,00%
LTU	45,00%
TUR	45,00%
GEO	43,00%
IRL	43,00%

NICODE	Completeness [%]
AZE	42,00%
ISR	42,00%
MKD	40,00%
EST	40,00%
CZE	39,00%
SVN	38,00%
LVA	38,00%
CHE	38,00%
BIH	38,00%
HRV	37,00%
GRC	37,00%
ROU	37,00%
SRB	36,00%

NICODE	Completeness [%]
HUN	35,00%
MDA	34,00%
BLR	34,00%
AUT	32,00%
POL	31,00%
GBR	30,00%
ARM	29,00%
RUS	26,00%
FRA	25,00%
BEL	24,00%
PRT	24,00%
BGR	24,00%
ITA	19,00%

Completeness_{total} = 37,00%

as of 2015-05-11

SCIENTIFIC PLANT NAMES

Scientific plant names

- Inconsistencies

Quick Search

The screenshot shows a search interface with tabs for 'Taxonomy', 'Accession', and 'Status'. The 'Taxonomy' tab is active. A search box contains 'R.da' and a dropdown menu is set to 'Contains'. Below the search box, the results are displayed under the heading 'Genus containing "R.damascena"'. Two results are listed: '?? ?????.?????R.damascena' (circled in red) and 'R.damascena x R.gallica'.

- Synonyms
 - E.g. *Solanum lycopersicum* L. vs. *Lycopersicon esculentum* Mill.
- Typos

Check/improve scientific names

- Use available information systems, e.g.
 - GRIN taxonomy
(<http://www.ars-grin.gov/cgi-bin/npgs/html/index.pl>)
 - Mansfeld catalogue
(<http://mansfeld.ipk-gatersleben.de/>)
 - Catalogue of Life
(<http://www.catalogueoflife.org/>)

Catalogue of Life

- List Matching Service
 - CoL → Tools → List Matching Service
 - Bulk comparison
 - Accepted names and synonyms

List Matching Service



Column delimiter:

First row contains headers:

Only accepted names:

Input type: File upload Text area

No file selected.

Catalogue of Life

- Prepare list of scientific names

	A	B	C
1	genus	species	spauthor
2	Triticum	aestivum	L.
3	Hordeum	vulgare	L.
4	Malus	domestica	Borkh.
5	Pisum	sativum	L.
6	Avena	sativa	L.
7	Zea	z.mays	L.
8	Dactylis	glomerata	L.
9	Solanum	tuberosum	L.
10	Glycine	max	(L.) Merr.
11	Hordeum	vulgare	
12	Lolium	perenne	L.
13	Pyrus	communis	L.
14	Triticum	sp.	
15	Triticum	durum	Desf.
16	Vicia	fabas	L.
17	Capsicum	annuum	L.
18	Rhododendron		
19	Secale	cereale	L.
20	Triticum	aestivum	(L.)Thell.
21	Trifolium	pratense	L.
22	Lycopersicon	esculentum	Mill.
23	Triticum	turgidum	L.
24	Medicago	sativa	L.
25	Solanum	lycopersicum	L.
26	Hordeum	sp.	
27	Lens	culinaris	Medik.
28	Astragalus	szovitsii	Fisch. Et C. A. Mey.
29	Helichrysum	callichrysum	Dc.
30	Lappula	consanguaneae	Fisch.Et C.A.Mey.
31	Cerasus	mahaleb	(L.)Mill.
32	??	??	
33	? Festulolium	loliaceum	(Huds.) P. Fourn.
34	?Egilops	????? ?????? aegilops	
35	Ocimum	x africanum	Lour.
36	Populus	? beroliunensis	Dipp.

Catalogue of Life

- Upload/paste it to List Matching Service

List Matching Service



Column delimiter:

First row contains headers:

Only accepted names:

Input type: File upload Text area

```
genus species spauthor
Triticum aestivum L.
Hordeum vulgare L.
Malus domestica Borkh.
Pisum sativum L.
Avena sativa L.
Zea mays L.
Dactylis glomerata L.
Solanum tuberosum L.
Glycine max(L.) Merr.
Hordeum vulgare
Lolium perenne L.
Pyrus communis L.
Triticum sp.
Triticum durum Desf.
Vicia faba L.
Capsicum annum L.
Rhododendron
Secale cereale L.
Triticum aestivum (L.)Thell.
Trifolium pratense L.
Lycopersicon esculentum Mill.
Triticum turgidum L.
Medicago sativa L.
Solanum lycopersicum L.
Hordeum sp.
Lens culinaris Medik.
Astragalus szovitsii Fisch. Et C. A. Mey.
Helichrysum callichrysum Dc.
Lappula consanguaneae Fisch. Et C.A.Mey.
Cerasus mahaleb (L.)Mill.
?? ??
? Festulolium loliaceum (Huds.) P. Fourn.
? Egilops ?????? aegilops
Ocimum x africanum Lour.
Populus? beroliunensis Dipp.
```

Catalogue of Life

- Matching names

List Matching Service



Matched Species Download to file

Your Data	Scientific Name	Status
Triticum aestivum L.	<i>Triticum aestivum</i> L.	accepted name
Hordeum vulgare L.	<i>Hordeum vulgare</i> L.	accepted name
Malus domestica Borkh.	<i>Malus domestica</i> Borkh.	synonym
Pisum sativum L.	<i>Pisum sativum</i> L.	accepted name
Avena sativa L.	<i>Avena sativa</i> L.	accepted name
Dactylis glomerata L.	<i>Dactylis glomerata</i> L.	accepted name
Solanum tuberosum L.	<i>Solanum tuberosum</i> L.	accepted name
Solanum tuberosum L.	<i>Solanum tuberosum</i> Bertero ex Walp., nomen nudum	synonym
Glycine max (L.) Merr.	<i>Glycine max</i> (L.) Merr.	accepted name
Hordeum vulgare	<i>Hordeum vulgare</i> L.	accepted name
Lolium perenne L.	<i>Lolium perenne</i> L.	accepted name
Pyrus communis L.	<i>Pyrus communis</i>	accepted name
Pyrus communis L.	<i>Pyrus communis</i> (var.) tomentosa Koch	ambiguous synonym
Pyrus communis L.	<i>Pyrus communis</i> Thunb.	ambiguous synonym
Triticum durum Desf.	<i>Triticum durum</i> Desf.	synonym
Vicia faba L.	<i>Vicia faba</i> L.	accepted name
Capsicum annuum L.	<i>Capsicum annuum</i> L.	accepted name
Secale cereale L.	<i>Secale cereale</i> L.	accepted name
Triticum aestivum (L.)Thell.	<i>Triticum aestivum</i> L.	accepted name

Catalogue of Life

- Non-matching names

<i>Asragalus szovitsii</i> Fisch. Et C. A. Mey.	<i>Asragalus szovitsii</i> Fisch. & C.A.Mey.	accepted name
<i>Helichrysum callichrysum</i> Dc.	<i>Helichrysum callichrysum</i> DC.	provisionally accepted name
<i>Cerasus mahaleb</i> (L.) Mill.	<i>Cerasus mahaleb</i> (L.) Miller	synonym

Unmatched Species

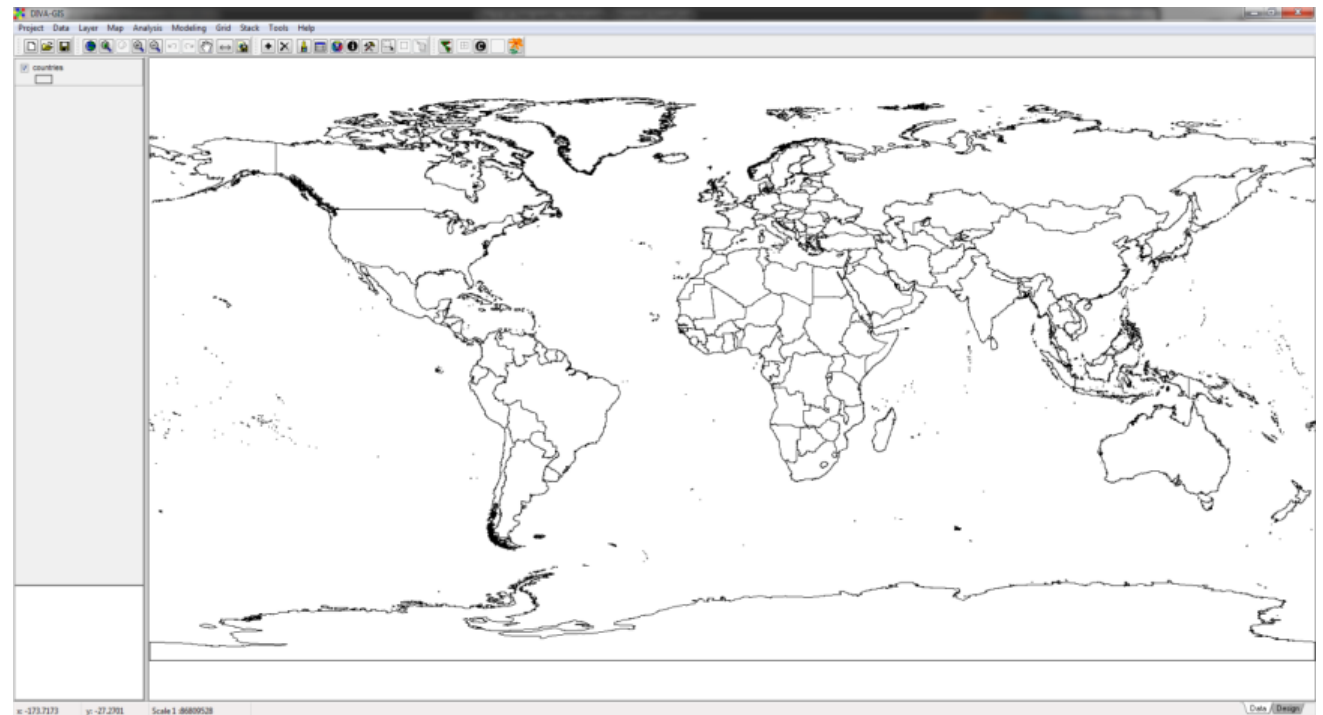
Zea z.mays L.
Triticum sp.
Rhododendron
Hordeum sp.
Lappula consanguaneae Fisch.Et C.A.Mey.
?? ??
? Festulolium loliaceum (Huds.) P. Fourn.
?Egilops ????? ???? aegilops
Ocimum × africanum Lour.
Populus ? beroliunensis Dipp.

GEOGRAPHICAL DATA

Check/improve geographical data

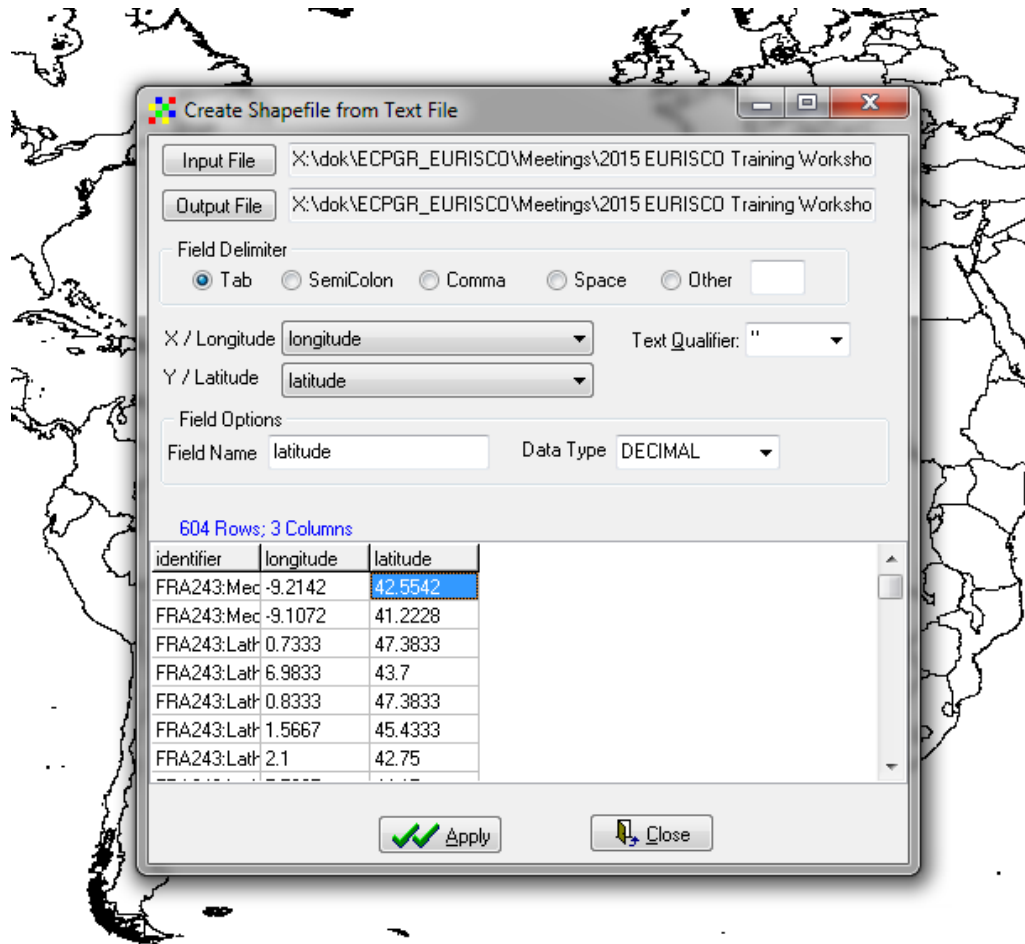
- E.g. by DIVA-GIS
 - Prepare collecting coordinates, e.g. by NICODE or ORIGCTY
 - Load country shapes

```
0 10 20 30 40 4
1 identifier longitude latitude
2 FRA243:Medicago:1548 -9.2142 42.5542
3 FRA243:Medicago:1608 -9.1072 41.2228
4 FRA243:Lathyrus:2048 0.7333 47.3833
5 FRA243:Lathyrus:2049 6.9833 43.7
6 FRA243:Lathyrus:2051 0.8333 47.3833
7 FRA243:Lathyrus:2052 1.5667 45.4333
8 FRA243:Lathyrus:2053 2.1 42.75
9 FRA243:Lathyrus:2054 7.5667 44.15
10 FRA243:Lathyrus:2055 6.9833 43.7
11 FRA272:Zea:FRA0410502 2.1333 43.7
12 FRA272:Zea:FRA0410650 -0.3667 43.4667
13 FRA272:Zea:FRA0410651 -0.3667 43.4667
14 FRA272:Zea:FRA0410652 -0.2333 43.4833
15 FRA272:Zea:FRA0410653 -1.2167 43.15
16 FRA272:Zea:FRA0410666 2.3167 43.8833
17 FRA272:Zea:FRA0410822 -0.55 43.2833
18 FRA272:Zea:FRA0411017 5.1167 46.6
19 FRA272:Zea:FRA0410736 1.5 42.7667
20 FRA272:Zea:FRA0410737 1.1833 42.75
21 FRA272:Zea:FRA0410738 0.25 45.9167
22 FRA272:Zea:FRA0410818 -0.6333 43.3
23 FRA272:Zea:FRA0410668 2.2667 43.3833
24 FRA272:Zea:FRA0411043 -0.25 42.9
25 FRA272:Zea:FRA0411044 -0.0167 42.9333
26 FRA272:Zea:FRA0411045 0.1667 43.05
27 FRA272:Zea:FRA0411046 -0.25 42.9
28 FRA272:Zea:FRA0411047 -0.8167 42.9833
29 FRA272:Zea:FRA0411048 -0.25 43.0833
30 FRA272:Zea:FRA0411015 5.1667 46.65
31 FRA272:Zea:FRA0411040 -0.55 43.2833
32 FRA272:Zea:FRA0411050 0.1667 43.05
33 FRA272:Zea:FRA0411053 0.5333 42.9667
34 FRA272:Zea:FRA0411054 -0.55 43.2833
35 FRA272:Zea:FRA0411057 -0.2 43.1167
36 FRA272:Zea:FRA0411058 -0.6 43.15
37 FRA272:Zea:FRA0411049 -0.25 43.0833
38 FRA272:Zea:FRA0411030 0.1667 45.95
39 FRA272:Zea:FRA0411018 5.4167 47.3
40 FRA272:Zea:FRA0411019 5.4167 47.3
41 FRA272:Zea:FRA0411020 5.4167 47.3
42 FRA272:Zea:FRA0411042 -0.25 43.3333
43 FRA272:Zea:FRA0411041 0.0167 43
44 FRA272:Zea:FRA0410001 -0.9 43.4667
45 FRA272:Zea:FRA0411036 -0.0167 43.05
46 FRA272:Zea:FRA0411038 -0.0167 43.05
47 FRA272:Zea:FRA0411039 0.4 42.95
48 FRA272:Zea:FRA0410647 -0.1833 43.2167
49 FRA272:Zea:FRA0410607 -0.2667 43.2333
50 FRA272:Zea:FRA0410614 2.5167 44.6167
```

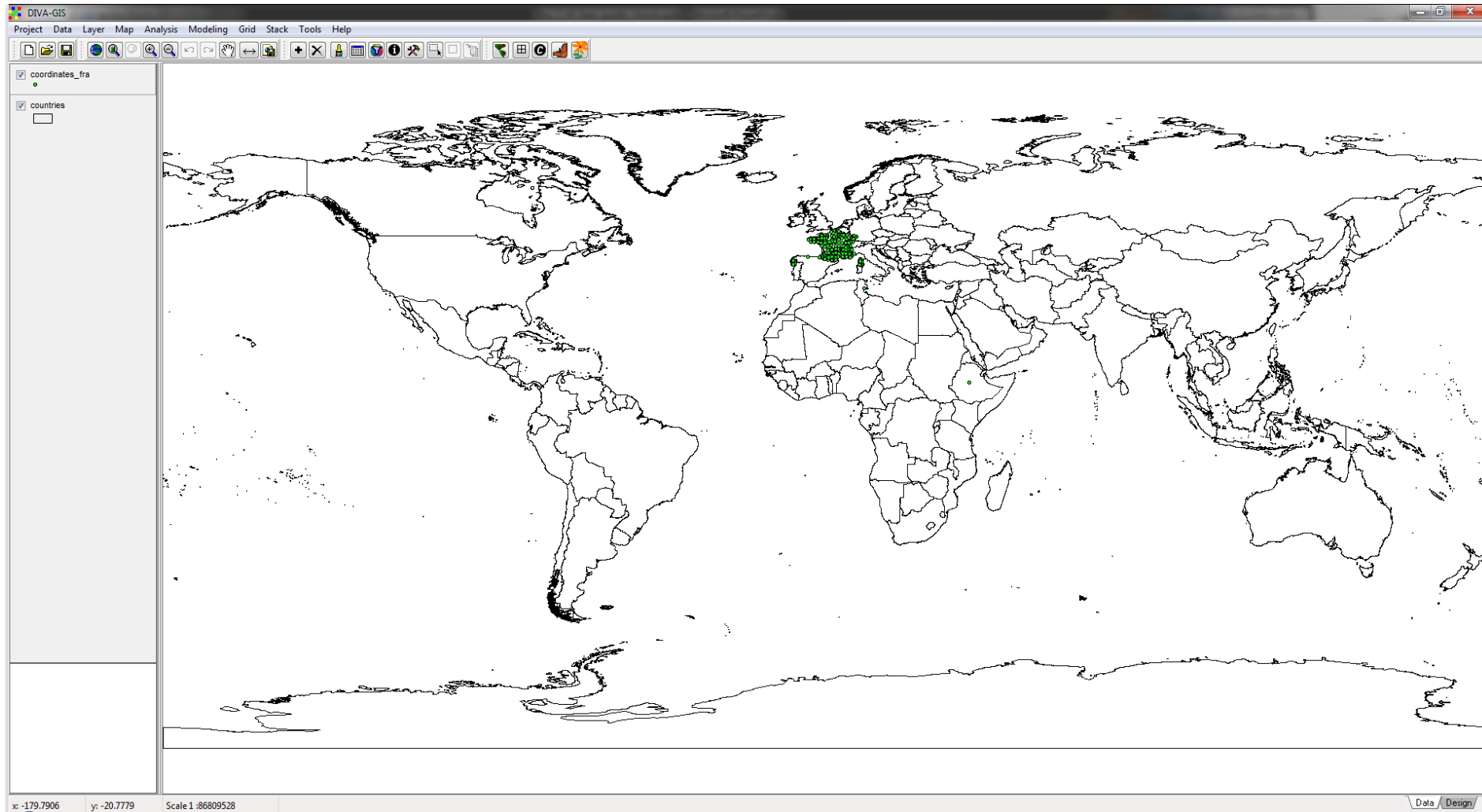


Check/improve geographical data

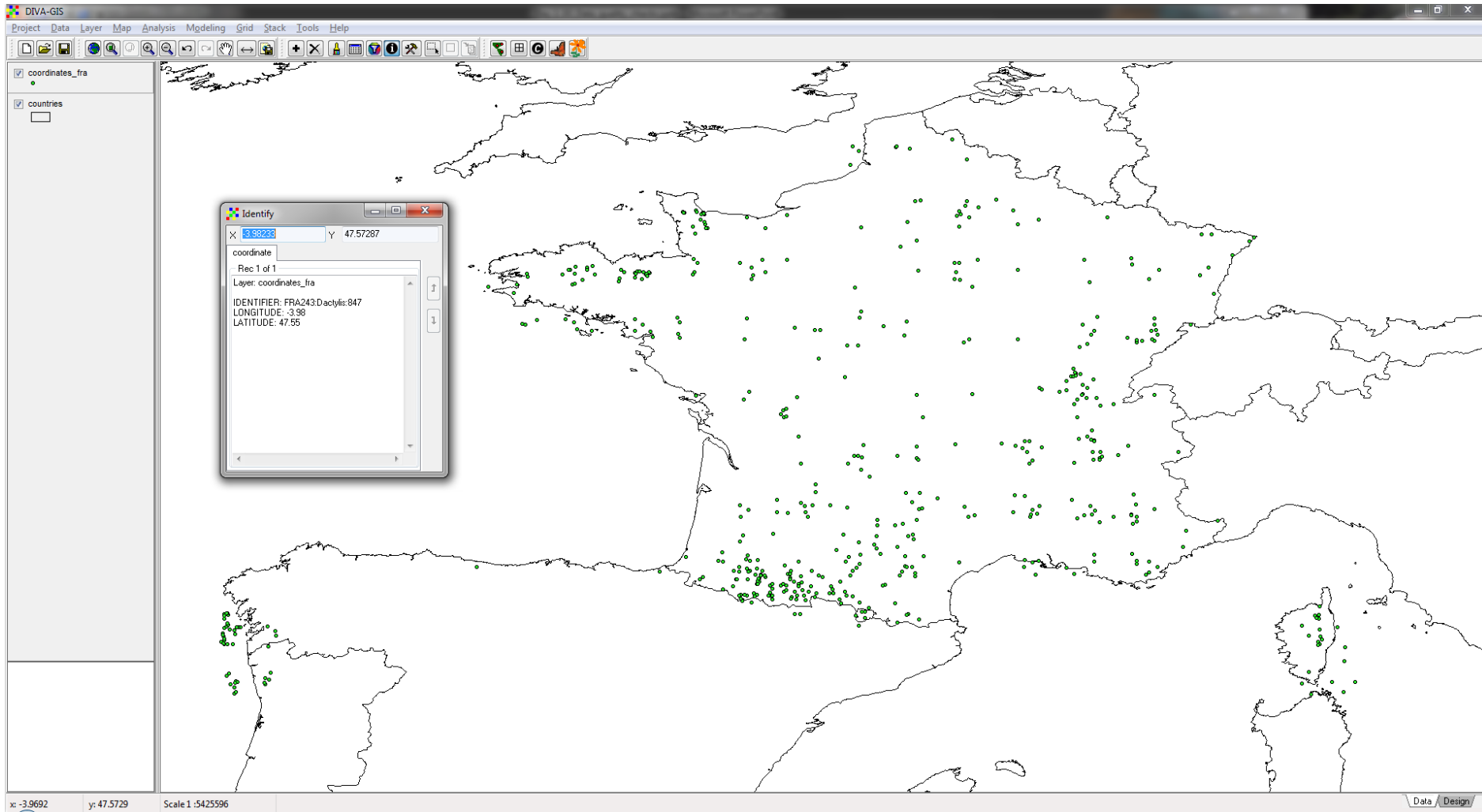
- Load and map coordinates



Check/improve geographical data

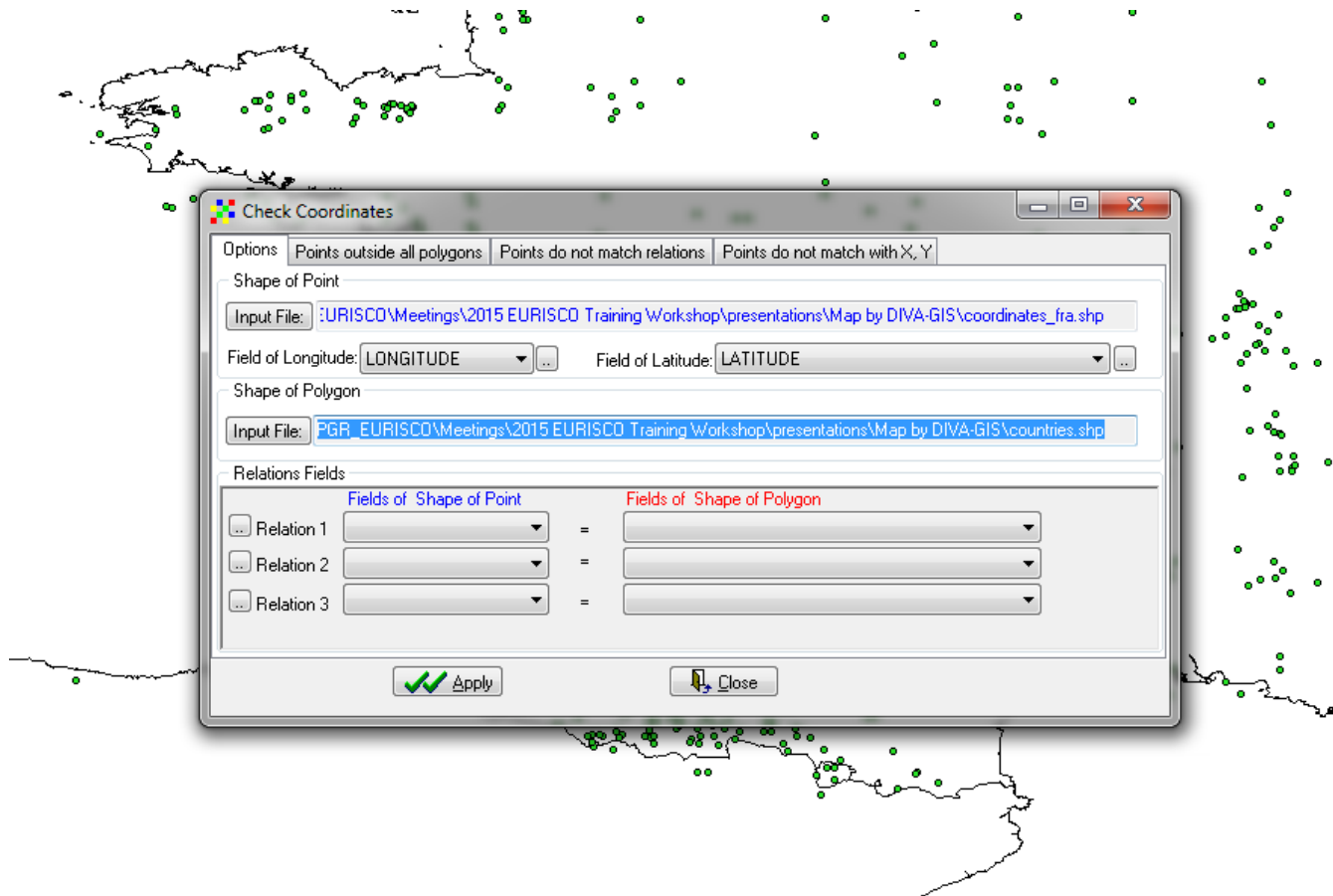


Check/improve geographical data



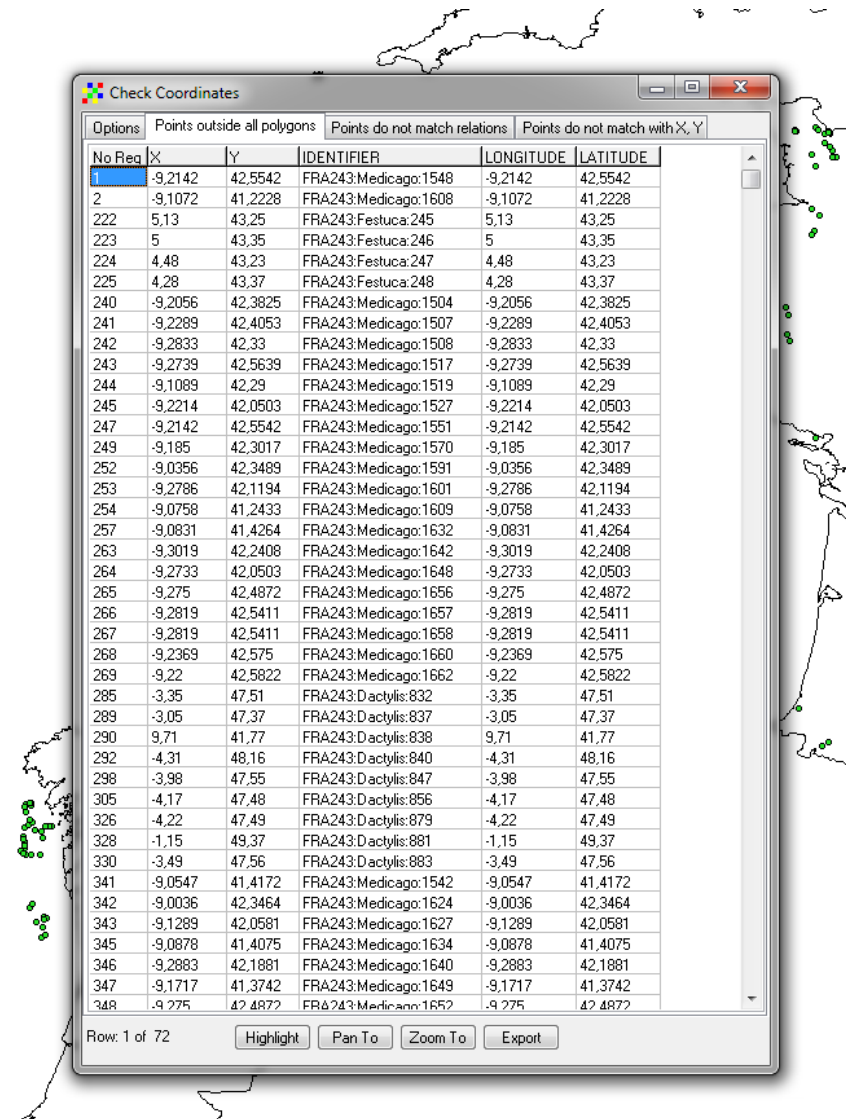
Check/improve geographical data

- Check coordinates



Check/improve geographical data

- Check results
 - 72 coordinates outside of county borders



Check/improve geographical data

