Data quality checks

EURISCO training workshop, 12th to 14th October 2016, Angers, France

Helmut Knüpffer & Stephan Weise
13 October 2016

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)

The four mandatory descriptors NICODE, INSTCODE, ACCENUMB and GENUS are not contained in these numbers.

Descriptor	Completeness [%]
SPECIES	97,43 %
SPAUTHOR	91,66 %
STORAGE	91,52 %
ACQDATE	80,55 %
CROPNAME	79,16 %
SAMPSTAT	70,90 %
MLSSTAT	68,60 %
ACCEURL	59,86 %
COLLSRC	57,36 %
ORIGCTY	52,11 %
AEGISSTAT	51,48 %
DONORNUMB	49,74 %
DONORDESCR	48,58 %
DUPLSITE	44,10 %
ACCENAME	41,70 %
BREDDESCR	40,36 %

Descriptor	Completeness [%]
BREDCODE	40,18 %
DUPLDESCR	38,16 %
ANCEST	32,38 %
DONORCODE	27,09 %
OTHERNUMB	21,75 %
SUBTAXA	20,57 %
COLLNUMB	19,62 %
COLLSITE	19,21 %
COLLDATE	18,48 %
COLLCODE	18,25 %
SUBTAUTHOR	12,67 %
COLLDESCR	12,33 %
ELEVATION	11,66 %
LONGITUDE	10,50 %
LATITUDE	10,27 %
REMARKS	3,04 %

as of 2016-10-07

Completeness per NI

- 36 descriptors (4 mandatory) → 32 descriptors considered

- $\text{Completeness}_{NI} = \frac{\text{sum of all actual descriptor values}_{NI}}{\text{sum of all possible descriptor values}_{NI}}$

Completeness per NI

The four mandatory descriptors NICODE, INSTCODE, ACCENUMB and GENUS are not contained in these numbers.

NICODE	Completeness [%]
MNE	60,00%
NLD	50,00%
ALB	50,00%
GBR	48,00%
NGB	47,00%
CYP	46,00%
UKR	46,00%
LTU	45,00%
SVK	45,00%
AZE	45,00%
TUR	45,00%
ESP	44,00%
IRL	43,00%

NICODE	Completeness [%]
GEO	43,00%
DEU	43,00%
ISR	42,00%
FRA	42,00%
MKD	40,00%
CZE	39,00%
LVA	38,00%
BIH	38,00%
SVN	38,00%
CHE	38,00%
ROU	37,00%
GRC	37,00%
HRV	36,00%

NICODE	Completeness [%]
EST	36,00%
SRB	36,00%
HUN	35,00%
MDA	34,00%
BLR	34,00%
AUT	32,00%
POL	32,00%
ARM	30,00%
ITA	29,00%
RUS	26,00%
BGR	24,00%
PRT	24,00%
BEL	24,00%

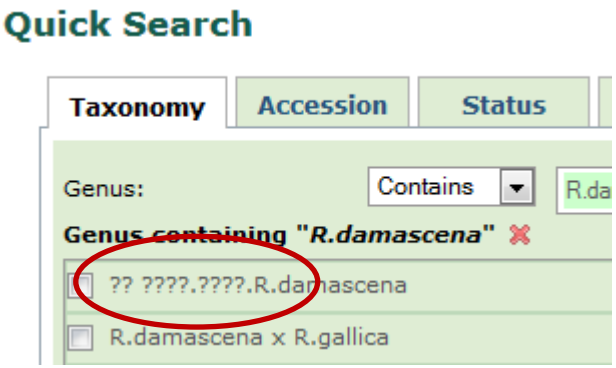
$Completeness_{total}$ = 42,00% (last year‘s workshop: 37,00%)

as of 2016-10-07

SCIENTIFIC PLANT NAMES

Scientific plant names

- Inconsistencies



- 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

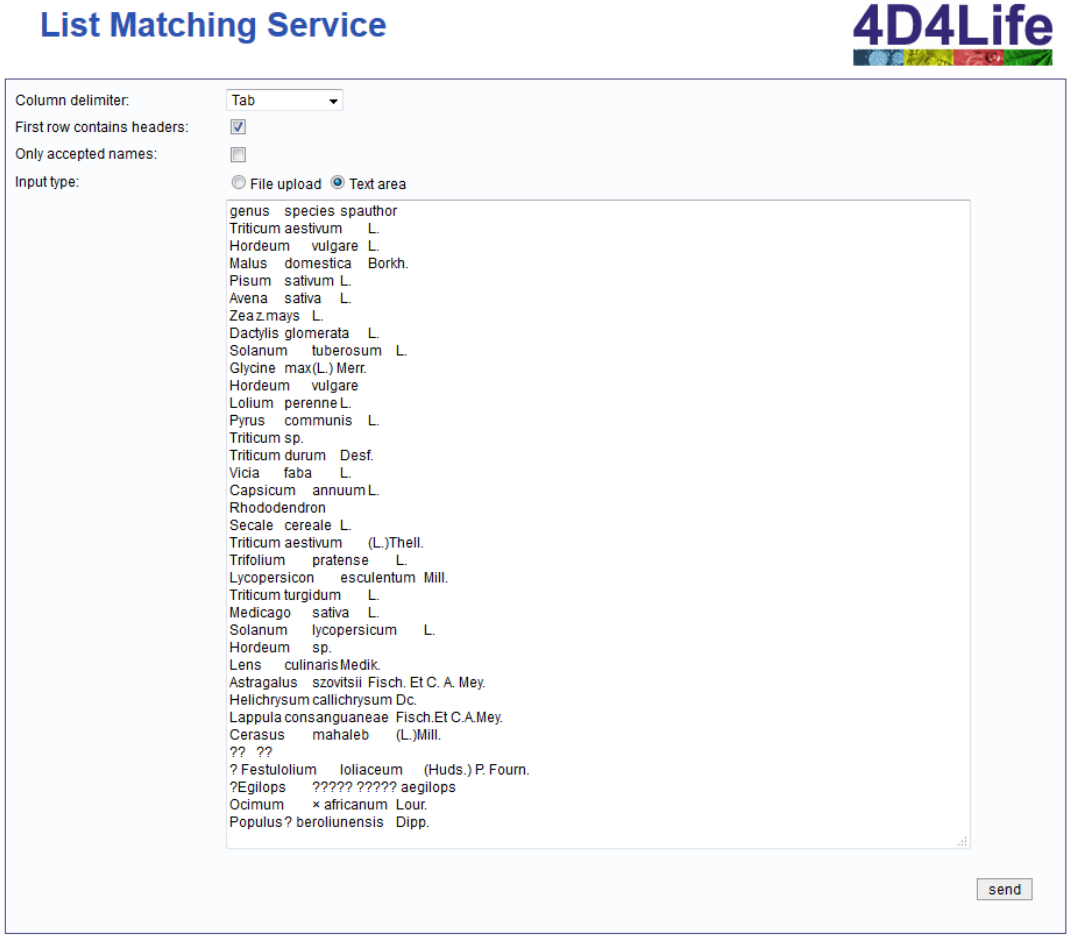
Catalogue of Life

- 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	faba	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	× africanum	Lour.
36	Populus	? beroliunensis	Dipp.

Catalogue of Life

- Upload/paste it to List Matching Service



Catalogue of Life

- Matching names

List Matching Service

4D4Life

Matched Species Download to file

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

Catalogue of Life

- Non-matching names

Astragalus szovitsii Fisch. Et C. A. Mey.	*Astragalus szovitsii* Fisch. & C.A.Mey.	accepted name
Helichrysum callichrysum Dc.	*Helichrysum callichrysum* DC.	provisionally accepted name
Cerasus mahaleb (L.)Mill.	*Cerasus mahaleb* (L.) Miller	synonym

Unmatched Species

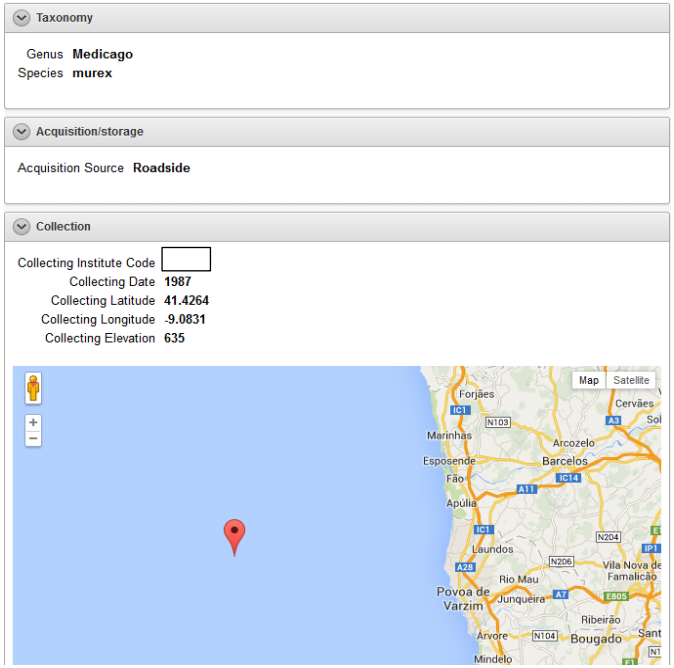
Zea z.mays L.
Triticum sp.
Rhododendron
Hordeum sp.
Lappula consanguaneae Fisch.Et C.A.Mey.
?? ??
? Festulolium loliaceum (Huds.) P. Fourn.
?Egilops ????? ????? aegilops
Ocimum × africanum Lour.
Populus ? beroliunensis Dipp.

GEOGRAPHICAL DATA

Geographical data

- Sparse information

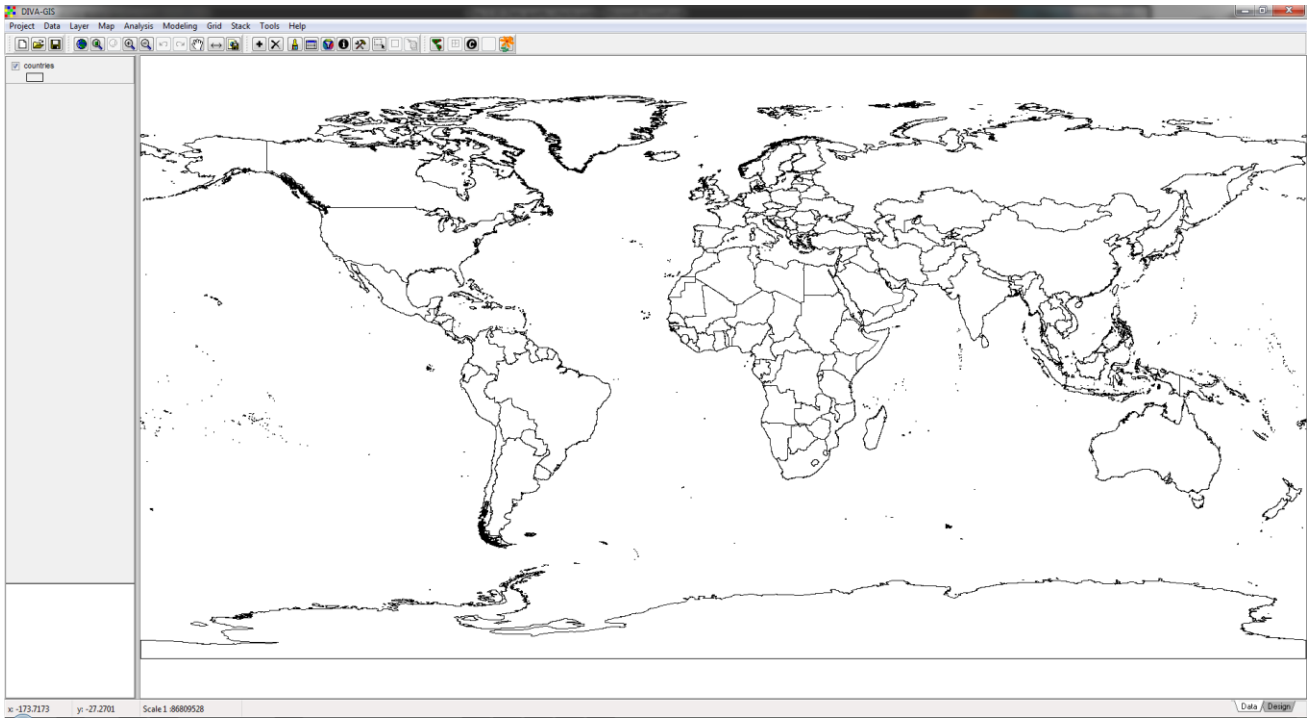
- Mistakes in geo referencing



Check/improve geographical data

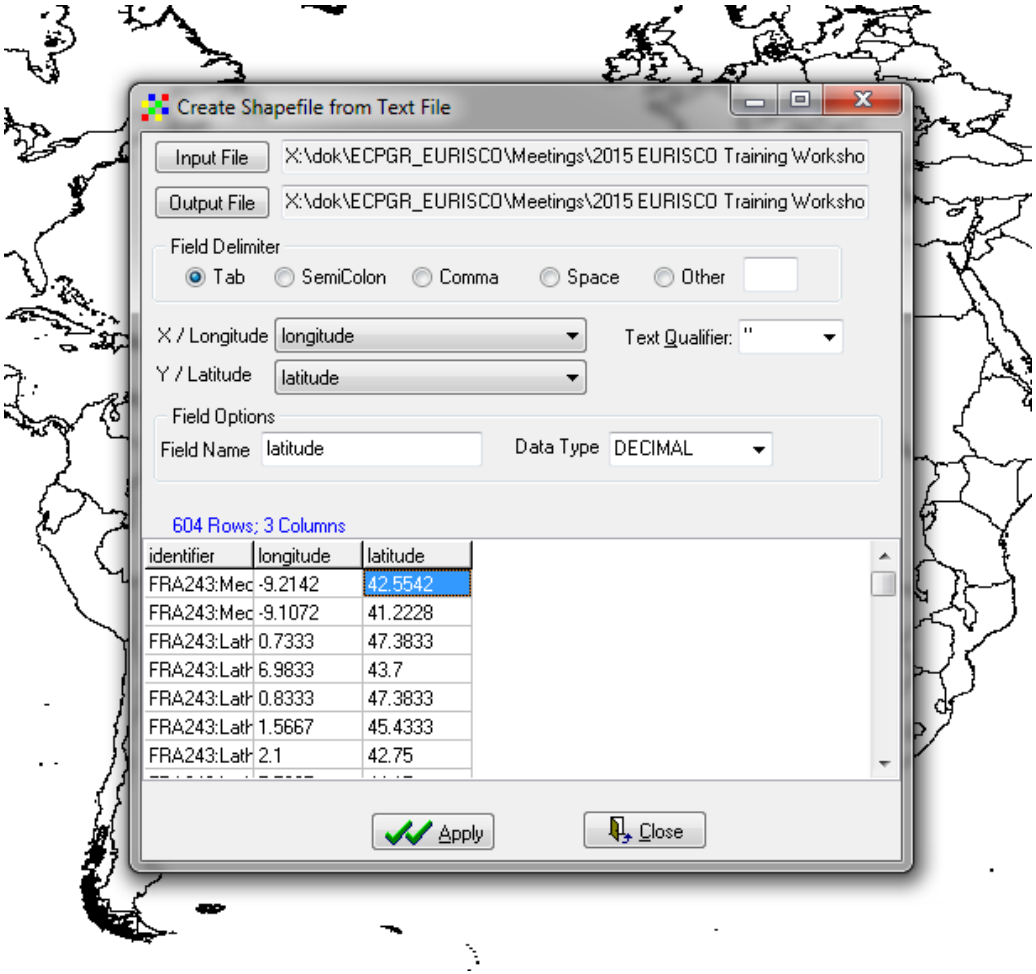
- E.g. by DIVA-GIS
  - Prepare collecting coordinates, e.g. by NICODE or ORIGCTY
  - Load country shapes

	identifier	longitude	latitude
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:FRA0411052	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

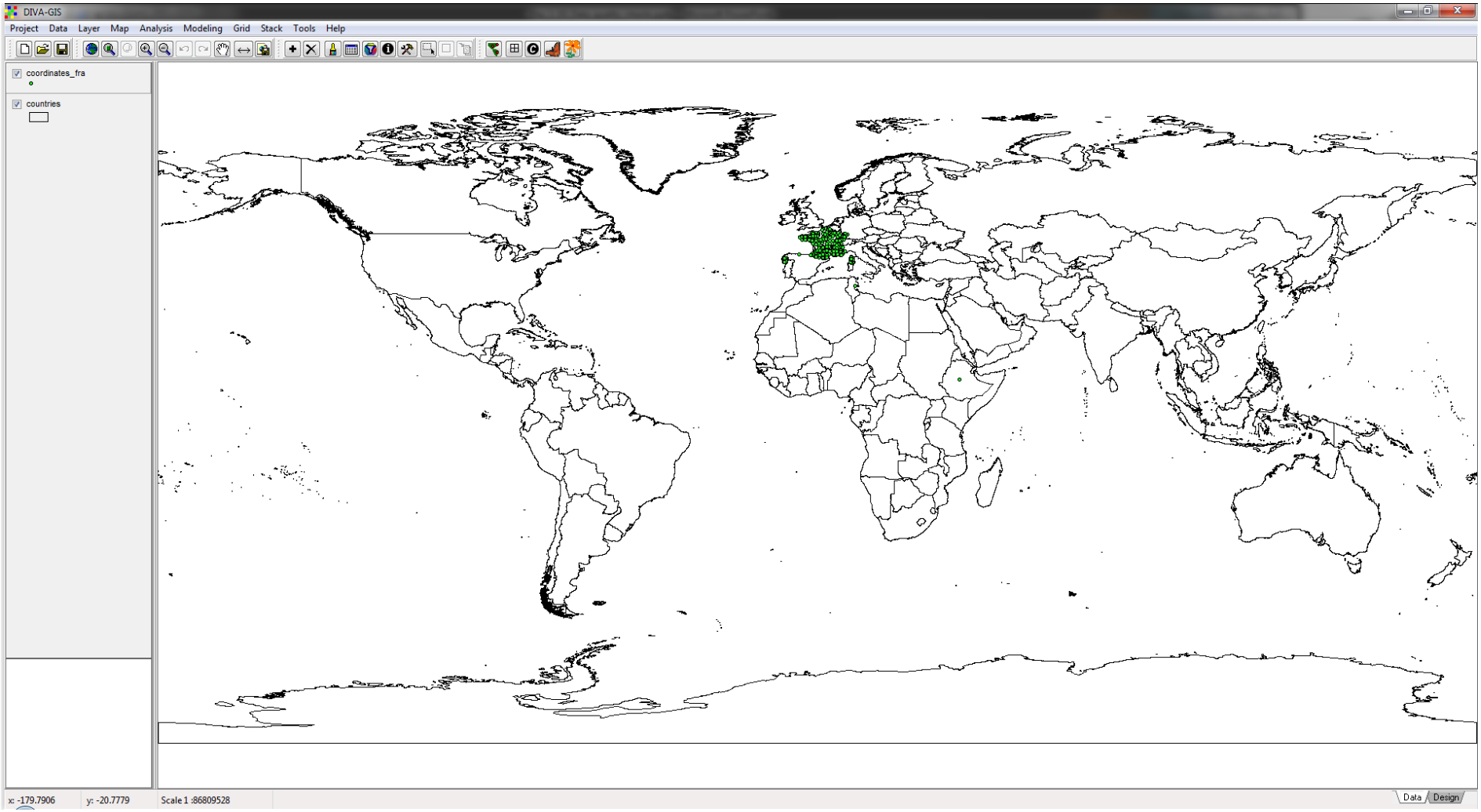


Check/improve geographical data

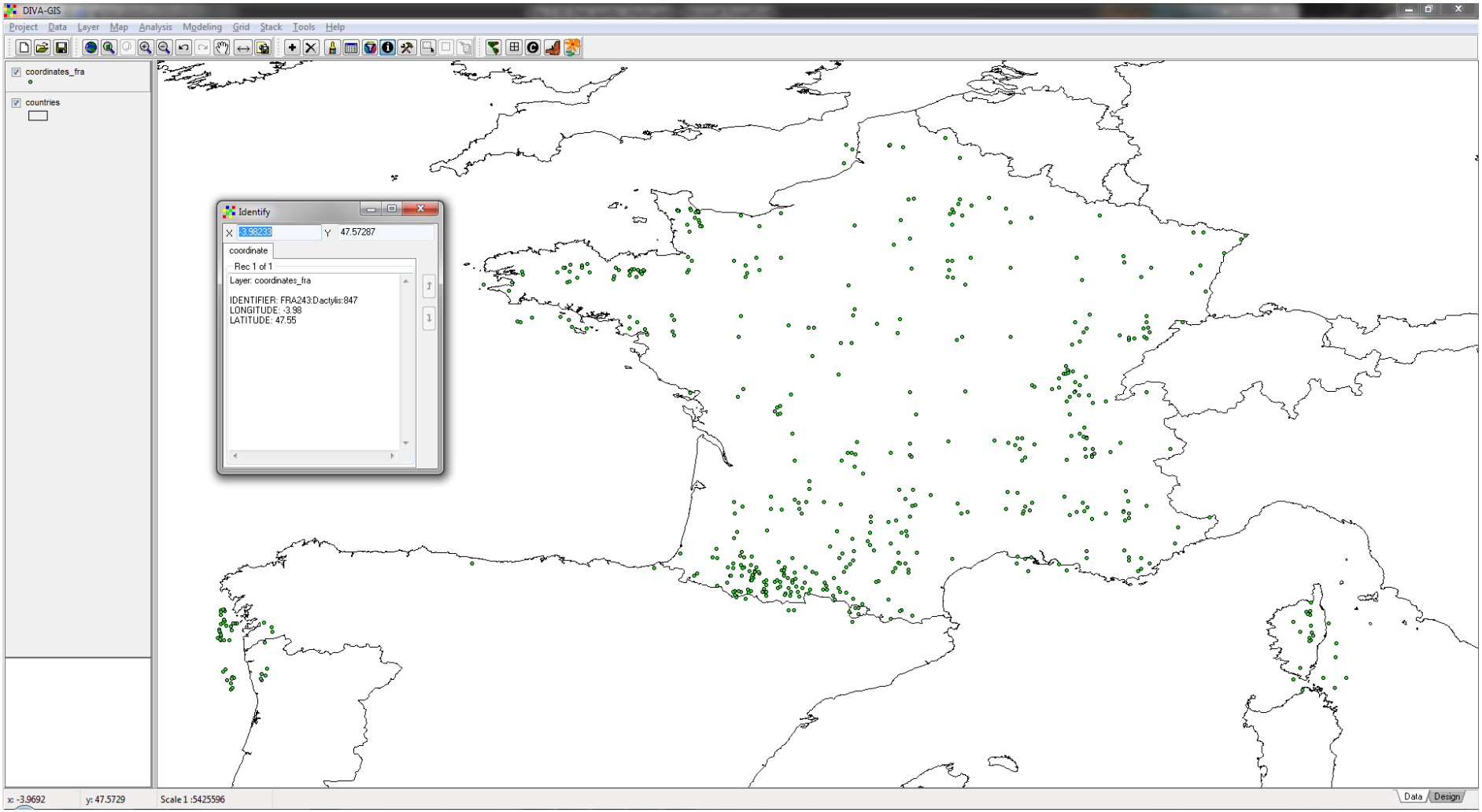
- Load and map coordinates



Check/improve geographical data

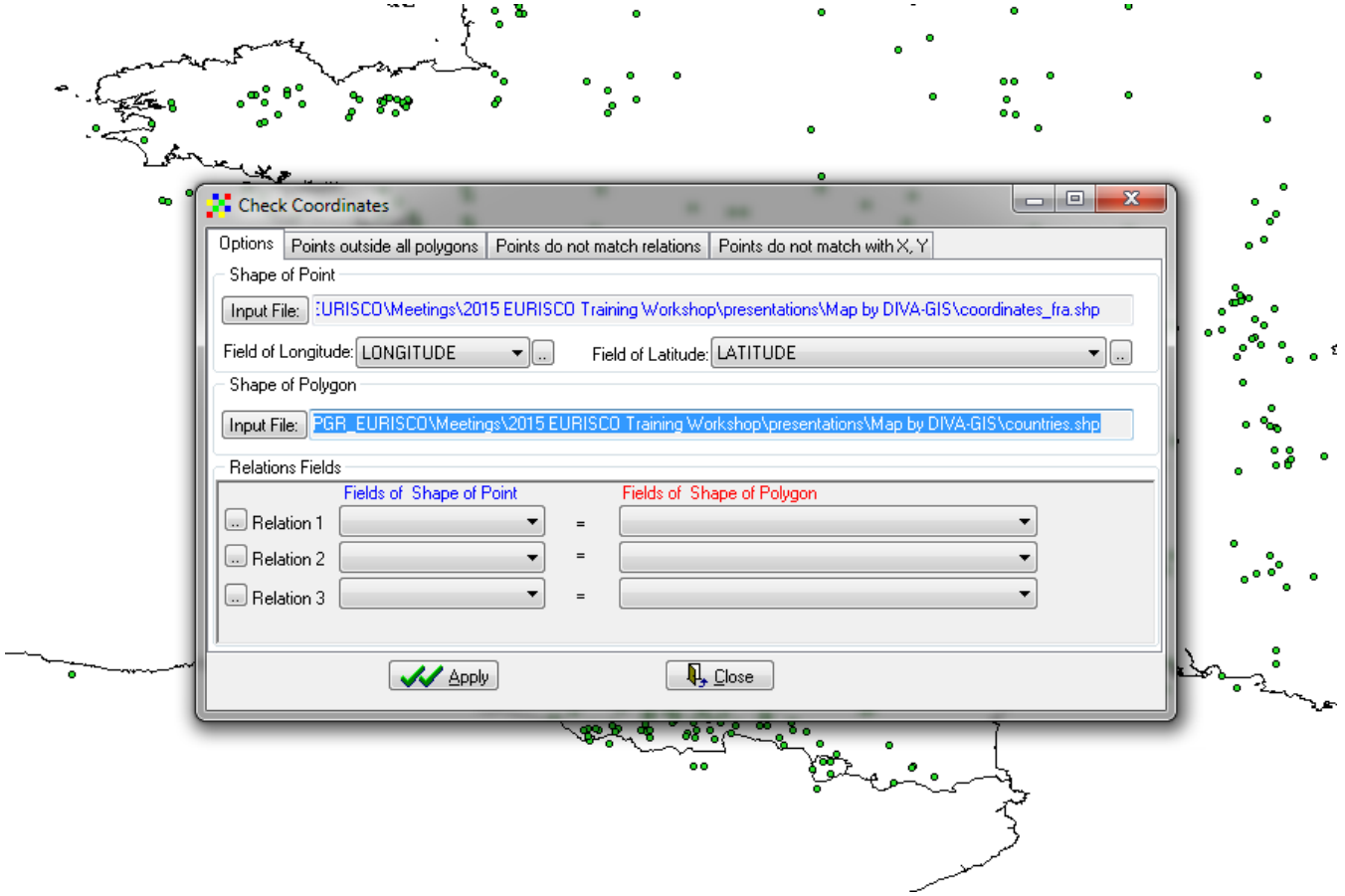


Check/improve geographical data



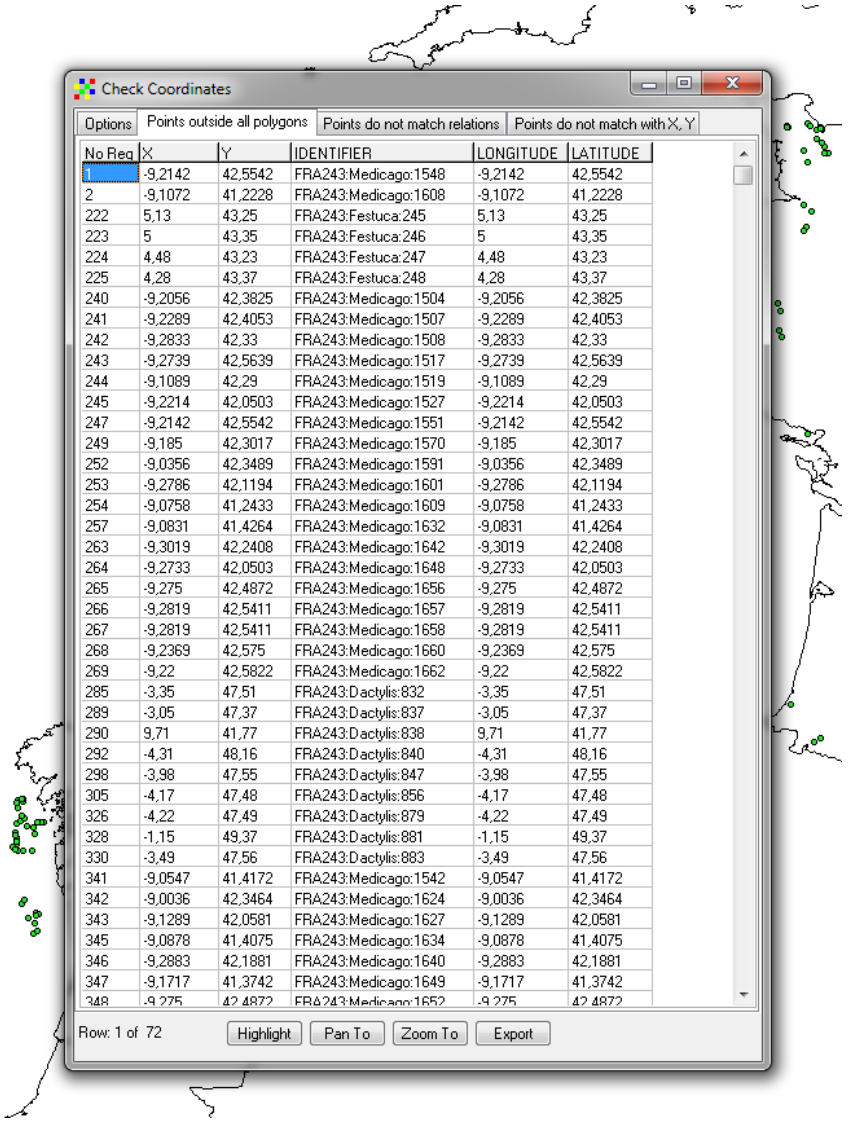
Check/improve geographical data

- Check coordinates



Check/improve geographical data

- Check results
  - 72 coordinates outside of county borders



No Reg	X	Y	IDENTIFIER	LONGITUDE	LATITUDE
1	-9,2142	42,5542	FRA243:Medicago:1548	-9,2142	42,5542
2	-9,1072	41,2228	FRA243:Medicago:1608	-9,1072	41,2228
222	5,13	43,25	FRA243:Festuca:245	5,13	43,25
223	5	43,35	FRA243:Festuca:246	5	43,35
224	4,48	43,23	FRA243:Festuca:247	4,48	43,23
225	4,28	43,37	FRA243:Festuca:248	4,28	43,37
240	-9,2056	42,3825	FRA243:Medicago:1504	-9,2056	42,3825
241	-9,2289	42,4053	FRA243:Medicago:1507	-9,2289	42,4053
242	-9,2833	42,33	FRA243:Medicago:1508	-9,2833	42,33
243	-9,2739	42,5639	FRA243:Medicago:1517	-9,2739	42,5639
244	-9,1089	42,29	FRA243:Medicago:1519	-9,1089	42,29
245	-9,2214	42,0503	FRA243:Medicago:1527	-9,2214	42,0503
247	-9,2142	42,5542	FRA243:Medicago:1551	-9,2142	42,5542
249	-9,185	42,3017	FRA243:Medicago:1570	-9,185	42,3017
252	-9,0356	42,3489	FRA243:Medicago:1591	-9,0356	42,3489
253	-9,2786	42,1194	FRA243:Medicago:1601	-9,2786	42,1194
254	-9,0758	41,2433	FRA243:Medicago:1609	-9,0758	41,2433
257	-9,0831	41,4264	FRA243:Medicago:1632	-9,0831	41,4264
263	-9,3019	42,2408	FRA243:Medicago:1642	-9,3019	42,2408
264	-9,2733	42,0503	FRA243:Medicago:1648	-9,2733	42,0503
265	-9,275	42,4872	FRA243:Medicago:1656	-9,275	42,4872
266	-9,2819	42,5411	FRA243:Medicago:1657	-9,2819	42,5411
267	-9,2819	42,5411	FRA243:Medicago:1658	-9,2819	42,5411
268	-9,2369	42,575	FRA243:Medicago:1660	-9,2369	42,575
269	-9,22	42,5822	FRA243:Medicago:1662	-9,22	42,5822
285	-3,35	47,51	FRA243:Dactylis:832	-3,35	47,51
289	-3,05	47,37	FRA243:Dactylis:837	-3,05	47,37
290	9,71	41,77	FRA243:Dactylis:838	9,71	41,77
292	-4,31	48,16	FRA243:Dactylis:840	-4,31	48,16
298	-3,98	47,55	FRA243:Dactylis:847	-3,98	47,55
305	-4,17	47,48	FRA243:Dactylis:856	-4,17	47,48
326	-4,22	47,49	FRA243:Dactylis:879	-4,22	47,49
328	-1,15	49,37	FRA243:Dactylis:881	-1,15	49,37
330	-3,49	47,56	FRA243:Dactylis:883	-3,49	47,56
341	-9,0547	41,4172	FRA243:Medicago:1542	-9,0547	41,4172
342	-9,0036	42,3464	FRA243:Medicago:1624	-9,0036	42,3464
343	-9,1289	42,0581	FRA243:Medicago:1627	-9,1289	42,0581
345	-9,0878	41,4075	FRA243:Medicago:1634	-9,0878	41,4075
346	-9,2883	42,1881	FRA243:Medicago:1640	-9,2883	42,1881
347	-9,1717	41,3742	FRA243:Medicago:1649	-9,1717	41,3742
348	-9,275	42,4872	FRA243:Medicago:1652	-9,275	42,4872

Check/improve geographical data