Natural History forums as source of valuable biodiversity data

De Felici S.[1,2,3], Mazzei P.[4,5], Dinolfo T.[4], Gioli F.[4], Labriola C.[4], Sbordoni V.[3,6], Cesaroni D.[3]

[1]LifeWatch-ITA, Collections Thematic Center; [2]National Research Council of Italy - CNR-IBAF; [3]Department of Biology, University of Rome "Tor Vergata"; [4]Natura Mediterraneo Forum; [5]Forum Entomologi Italiani; [6] Italian National Academy of Sciences, called of the XL

The unaware Citizen Scientists

Among the scientific activities so far less investigated there are the Natural History forums, where users exchange knowledge and opinions by sending photos and messages.

Forums usually do not represent or implement any projects, including those of Citizen Science and yet their enthusiastic users produce scientific data.

How do the Natural History forums work?

1. A user opens a topic by publishing one or more photos.

2. If the user is an expert, proposes or provides an identification, comments on the observation in many ways and other users leave other comments.

3. If the user is a not-expert asks something like: "*What's this*?"

How do the Natural History forums work?

2. In the latter case, one or more other users answer the question or attempt to do so.

3. In most cases a moderator or an expert user close the topic with a final identification of the subject.

Forums like biodiversity repositories

Over the years, the forums have accumulated large amounts of data on primary biodiversity.

We have estimated that only in three of the most well-known Italian forums: "Forum Entomologi Italiani" (FEI) and "Forum Natura Mediterraneo" (FNM) and "Acta Plantarum" are currently record more than **400.000 threads** (about 480.000 observations) on Plantae, Fungi, Lichens and Metazoa, but…

Are these data really suitable for scientific use?

What we test

We selected randomly topics from Italy in which at least one photo was available to identify the specimens and annotated the all the data and the final identification; the dataset consists of **3.382 threads** (1.479 from FNM and 1.903 from FEI) with **4.029 observations**.

To evaluate the reliability of the final identifications we set up a "Standardized Statement Grid" in which we grouped the statement used by the identifiers in qualifying their own identifications:

Statement type	Reliability level and p (assigned) values
Plain identification; **"The species is…"**	R= Reliable identification (p>= 95%)
"I'm almost sure is…"; **"Very probably it is…"**	AR= Almost reliable identification (p between 94% and 75%)
"Could be…"; **"I think it is…",**	LR=Low reliable identification (p between 74% and 50%)
"I think it is … but I'm not experienced with the genus"; **"I'm not sure";** **"Could be …** **"I try to guess…"**	UR=Unreliable identification (p< 50%)
Genus sp. **No answer for the identification request**	NA=Not Available

What we test

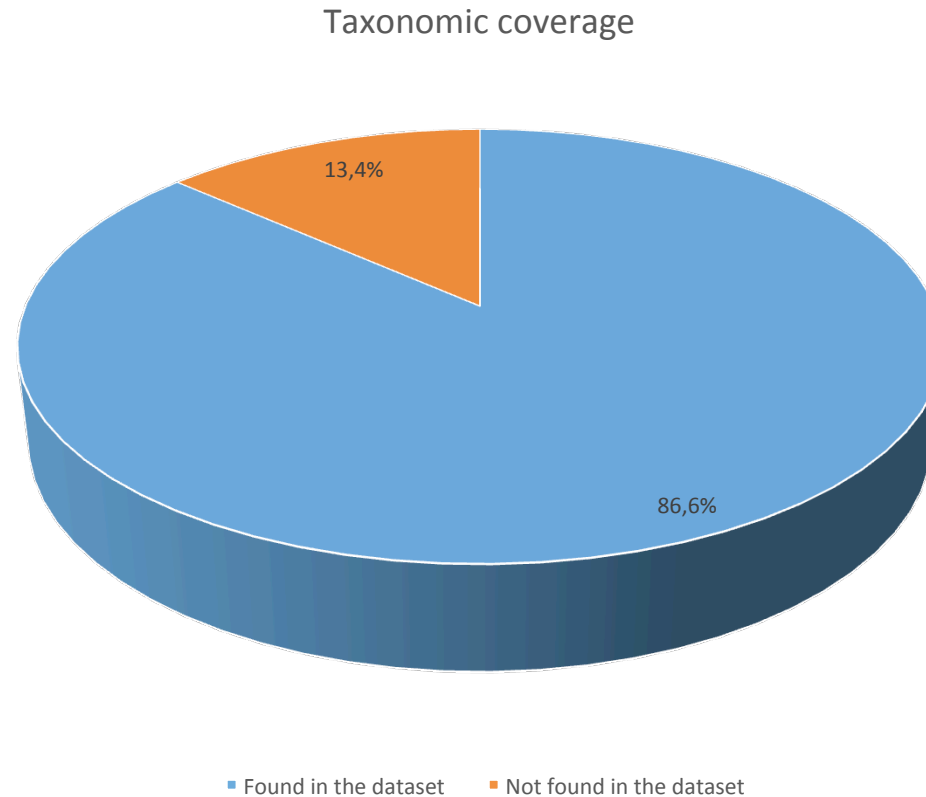The same standardized statement grid was used by ourselves to qualify our identifications.

In addition to the taxonomic reliability, the site descriptions were assessed for spatial accuracy on the base of the method outlined in this conference by De Felici et al.

Statement type	Reliability level and p (assigned) values
Plain identification; **"The species is…"**	R= Reliable identification (p>= 95%)
"I'm almost sure is…"; **"Very probably it is…"**	AR= Almost reliable identification (p between 94% and 75%)
"Could be…"; **"I think it is…",**	LR=Low reliable identification (p between 74% and 50%)
"I think it is … but I'm not experienced with the genus"; **"I'm not sure";** **"Could be …** **"I try to guess…"**	UR=Unreliable identification (p< 50%)
Genus sp. **No answer for the identification request**	Not Available

Results

3.810 observations refer to adult stage and 219 refer to other stages (eggs, larvae and pupae).

1. Taxonomic coverage: **251 species on 290 of the italian checklist (86.5%) has been identified by the expert.**

Taxonomic coverage

13,4%

86,6%

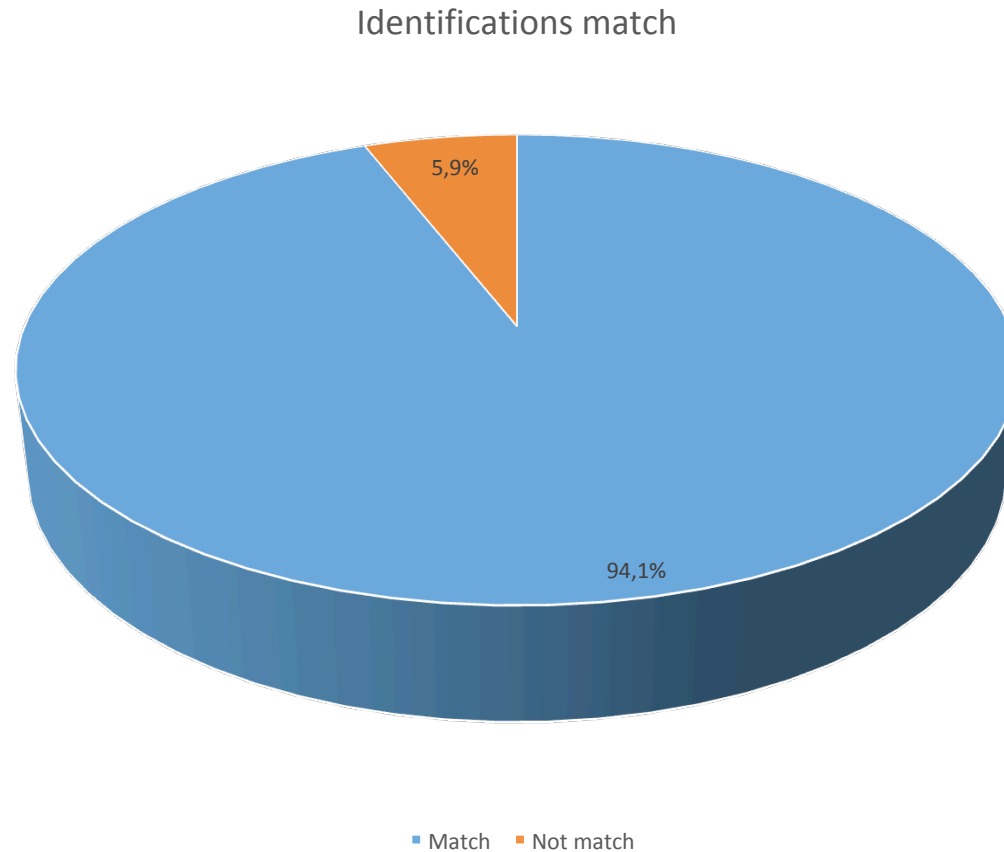■ Found in the dataset   ■ Not found in the dataset

Results

3.810 observations refer to adult stage and 219 refer to other stages (eggs, larvae and pupae).

2. Identifications match: **3.585 on 3.810 adult identifications in the forums match with expert identifications (94.0%)**

Identifications match

5,9%

94,1%

■ Match   ■ Not match

Identifications reliability

With the aim of asses the overall reliability of identifications we have compared separately the reliability level for matching and nonmatching identifications for adult stage (3.585).
On the main diagonal forum and experts identifications have the same reliability: **3286 (91.6%)**

MATCHING IDENTIFICATIONS N=3585		EXPERT				
		R	AR	LR	UR	NA
FORUM	R	3027	30	15	6	
	AR	57	66	3		
	LR	133	39	41	7	
	UR	7		2	1	
	NA					151

In the cells above the main diagonal, the reliability ratings in the forums is higher than that assigned by the experts: **54**
In the cells below the main diagonal the reliability ratings in the forum is lower than that assigned by the experts: **238**
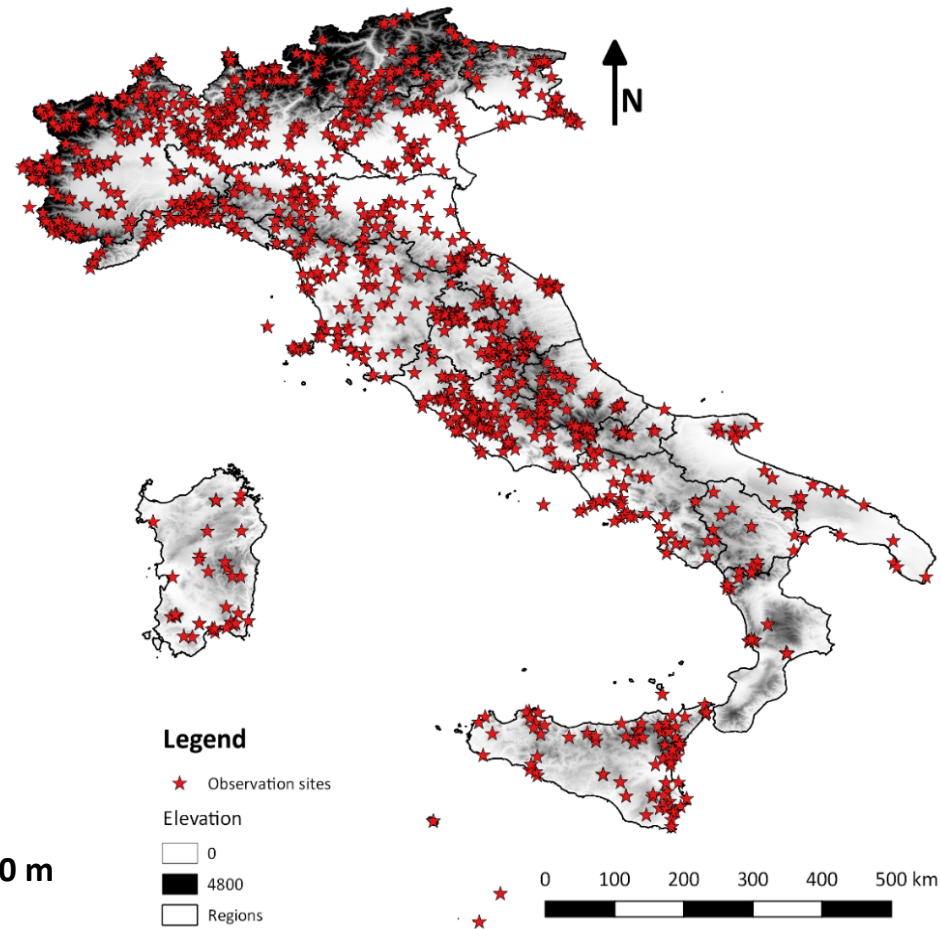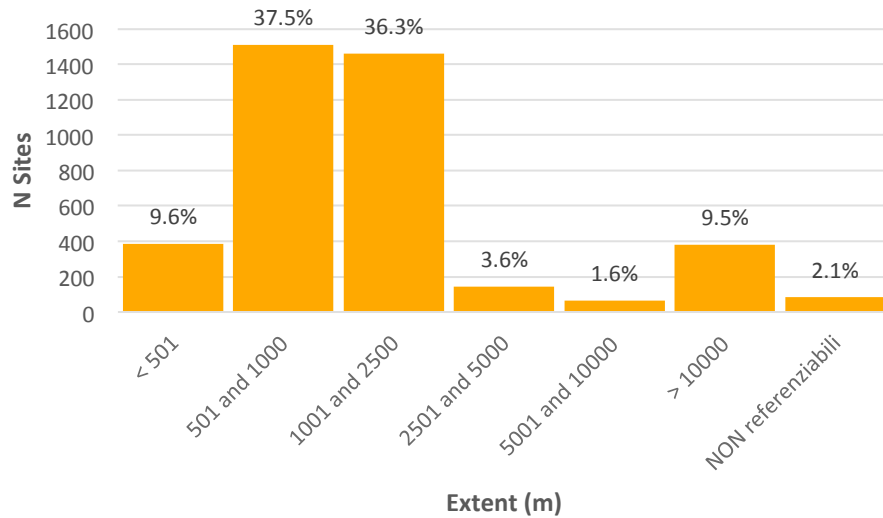
Identifications reliability

With the aim of asses the overall reliability of identifications we have compared separately the reliability level for matching and nonmatching identifications for adult stage (3.585).

NON MATCHING IDENTIFICATIONS N=225		EXPERT				
		R	AR	LR	UR	NA
	R	**24**	**4**	1	1	4
	AR	**5**	**3**		1	22
FORUM	LR	17	4	5	1	30
	UR		2		1	7
	NA	55	12	12	2	12

Cells with red numbers contains identifications, reliable or almost reliable according to the forums that mismatch with reliable or almost reliable identification by the experts: **36** (1% on adult identifications).

Geographical accuracy from sites descriptions

3. More than 80% of the georeferenced sites have extent < 2.500 m

Yes, data from tested Natural History forums are really suitable for scientific use!

Looking forward

1. Other taxa must be tested.

2. Other Forums must be tested.

3. Reliable Forums must be supported, both for ethical question than to facilitate data exchange.

4. A software (or a service) with Natural Language Processing for automatic interpretation of thread contents is needed to recover large amount of data

5. A software (or a service) is needed for the geo-referencing of large amounts of geographical data.

Thank you for your time and attention!