A PRELIMINARY ANALYSIS OF THE CGA

Pontignano/Siena, July 2001

At the last SCAR Meeting, Tokyo, July 2000, R.Cervellati proposed to perform an accurate analysis of the CGA to be ready for a discussion at the next WG meeting in 2002. More exactly, it was proposed that the letter "A" of the CGA were inspected with the aim to comment each feature, the comment becoming a separate field to be added to the record, and describing what is the present situation of the given feature.
From this point on, the discussion should then move to the kind of action one could possibly undertake in order to approach a settlement of the multiple naming problem.
In our opinion, now that the structure of the Composite Gazetteer of Antarctica has been set up and the future of the CGA seems confined to a continuous, actually endless, updating of the database, it is time for the WG-GGI, or a part of it, to think ahead and to foresee what use should be done of the collected data.

In the months which followed the Tokyo Meeting some work on that subject has been undertaken by Cervellati and Ramorino. The present paper summarises the preliminary results.

The first thing that appeared quite clear was that it is not easy, perhaps it is meaningless, to speak of the letter "A" of the CGA. The SCAR CGA is the collection of several national gazetteers in different languages and alphabets: it is obvious that the content of the letter "A" in one gazetteer is different from the content of letter "A" in another gazetteer, and if all names of the CGA were listed according to the alphabetical order an entirely new set of the letter "A" names would arise.
What we are really interested in, however, are the features: we want to see which features have received a single name or multiple naming, and in case of multiple naming which features present an easy solution (i.e. an easy way out toward a single, internationally recognised, name) and which ones would require an exhaustive discussion.

Accordingly, we decided to subject to our analysis the first thousand features of the CGA and not the letter "A". As it is well known, each feature in the CGA has received a reference number. The reference number, although arbitrary, is a unique identifier of that feature. To take the first thousand features into consideration means exactly that we analyse the features with a reference number less than 1005 (four reference numbers have been suppressed during the previous updating of the CGA).
In the following we will refer to this set as the "1000-set" (or "our set").
The difference in the approach (one thousand features instead of letter "A") is however more conceptual than practical. As a matter of fact, the first thousand features correspond more or less to the names beginning with "A" in most gazetteers because the present numbering system still shows traces of the previous ordering criterion where the alphabetical order was preferred.

We deem that our limited set may give a rough idea of what is inside the overall CGA in its present form and size and possibly in the future too. Accordingly, our set deserves an accurate study and discussion.

How many features in the 1000-set have received a single name? This is one important question because from the answer one can put apart all those features for which there is no question about which name should be used for them. The answer is: 376. In other words 37.6% of the features appears beyond dispute. Let us add "at least" because the following analysis shows that the fraction of features beyond dispute is larger than that.

The remaining features, i.e. 624 features out of 1000, are present in more than one gazetteer and, accordingly, have received a name in each of the gazetteers. The names, two or more, can be identical or not. We have found it useful to classify the multiple naming into three levels.

One level, let us call it **level 1**, corresponds to the case where the applicable names are identical. By identical we mean that also a computer simply instructed to sort out names which are made of an identical string of characters would reach the same result as a man inspecting the list, feature after feature. This criterion is very strict. For example the names Atherton islas and Atherton Islas (feature no. 602) are not considered identical. The difference, is only in the letter I, the Argentinean gazetteer preferring the lower case, while the Chilean gazetteer prefers the upper case.
The features, present in two or more gazetteers, which also fall into the level 1 class are 378. This figure, summed up to the previous figure (376, features present in only one gazetteer) gives a total of 754. We have thus 754 out of 1000 features, i.e. 75.4% , which have received a single name. We will come later on this encouraging result.

There are many cases where one recognises a difference in naming but not a substantial disagreement.
We will put these cases into to a new class, of nearly identical names that we will call **level 2** class.
The previous example, referring to Atherton Island, belongs to the level 2 class**.** Other examples of this quasi-equal class are given in the following lines:

feature no.942: Barrios, Islote (CHL) and Barrios Rocks (GBR,USA)
feature no. 444: Apollo, glaciar (ARG) and Apollo Glacier (GBR,USA)
feature no. 793: Bakhallet (NOR) and Bakhallet Slope (USA)
feature no. 573:Orejas de Burro, islas(ARG)/Islas (CHL) and Asses Ears (GBR, USA)

Most cases of level 2 arise, as in features 942 and 444, from a difference in the generic part of the name which depends mostly upon the language used. So to speak, it is an apparent difference, because the specific name is common to all gazetteers. In other words, all those differences (or most of them) would disappear if the Countries would like to agree on specific names only, leaving the choice of the generic parts to the language in use.

Other cases of level 2 arise from a translation, more or less acceptable, of the specific name. Feature 793 is a typical and well known case of misunderstanding connected with many Norwegian names. Actually the Norwegian gazetteer very often makes a single word from the specific and the generic parts: here Bakhallet, when translated into English, turns out to mean "the back slope".

Feature 573 is another example of a translation. Here it would be difficult to recognise that the specific part of the names have the same meaning (unless you know both languages). However, also in this case, there is a substantial agreement about the origin of the name, despite the diverging final result.

We found it useful to count how many features, in our 1000-set, belong to level 2.
For those features it should be possible to discuss and to agree on the general rules which should eventually get the SCAR gazetteer near to a single naming system.
By the way, the features in the level 2 are 205, which means another 20.5% of the total.

The remaining features, not falling in the classes of level 1 or 2, exhibit incurable differences and are to be included in a new class: the **level 3** class. The features left are 41; i.e. only 4.1% of the total is at level 3.

Examples of level 3 are given in the following. Most of them are quite obvious.

feature no.287: Carminatti, bahia (ARG) and Ambush Bay (GBR, RUS, USA)
feature no.618: Larga, isla (ARG) and Atriceps Island (GBR, RUS, USA)
feature no. 440: Apéndice, isla (ARG), Apéndice Island (USA), Rivera, Isla (CHL),
Sterneck Island (GBR)

All features of level 3 require an analysis based on historical and other suitable elements in order to ascertain which name is to be eventually preferred. An easy way out, although provisional, could be to simply recognise that at present there are features to which more than one name is applicable. The situation is not unique of Antarctic, however. There are areas in the world where multiple naming is an accepted practice, e.g. in a region of the Northern Italy which has got itself two names, Alto Adige (Italian) and Süd Tyrol (German), most geographical features are officially indicated by two, sometimes three, different names.

It is an important result, in any case, to recognise that the features of level 3 are a relatively small fraction of the total.

Table 1 shows the distribution of the 1000 features taken into consideration according to the number of gazetteers which contain the given features (we may also say "according to the number of names received", if it is understood that two identical names are considered as two names). We see from table 1 that one feature has been given six names. It is:

feature no.431: Amberes, isla (ARG), Anvers, Ile (BEL), Anvers, Isla (CHL),
Anvers Island (GBR, RUS, USA)

No other feature has received six or more names.

For each row of table 1 the features corresponding to level 1, 2 or 3 have been counted. (see columns 2, 3,4). The total of columns 2, 3, 4, gives the numbers of features in level 1 (*754*), 2 (*205*) and 3 (*41*).

The same results are illustrated by the bar-graphs and pie-graphs in figures 1 to 3.

While the analysis of the1000-set required time-consuming procedures, the counting of multiplets throughout the CGA (17038 features, as per July 1st, 2001) is a simple computer facility. The table 2 and the pie-graph of figure 4 show the result. A comparison of figure 2 with figure 4 shows that the pie percentages for the full CGA and for the subset taken into consideration, which represents about 6% of the total, are quite similar, thus suggesting that a study on the 1000 feature sample is probably meaningful for the full CGA.

Rome, June 2001

R.Cervellati, M.C.Ramorino

Features with:	**Level 1**	**Level 2**	**Level 3**	**Total**	**%**
1 name	376			376	37.6
2 names	285	51	4	340	34.0
3 names	92	93	12	197	19.7
4 names	1	37	20	58	5.8
5 names		23	5	28	2.8
6 names		1		1	0.1
TOTAL	**754**	**205**	**41**	**1000**	100.0

Table 1 – Distribution of multiplets in the first thousand features of the CGA
(Level 1: identical names; level 2: nearly identical; level 3: substantially different)

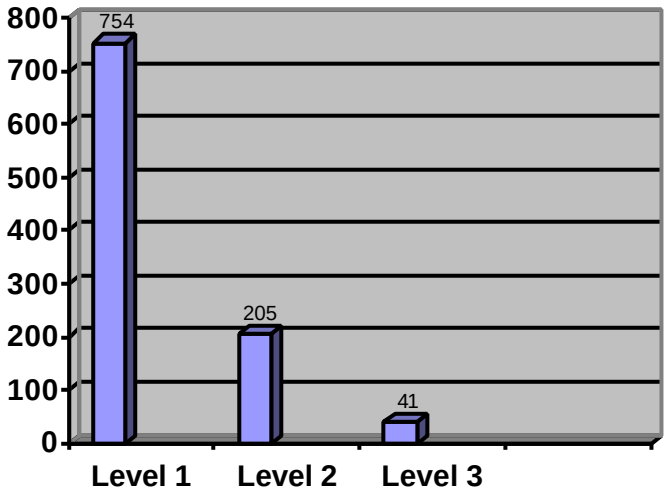

Fig. 1 – Distribution of names according to the level (first thousand features)

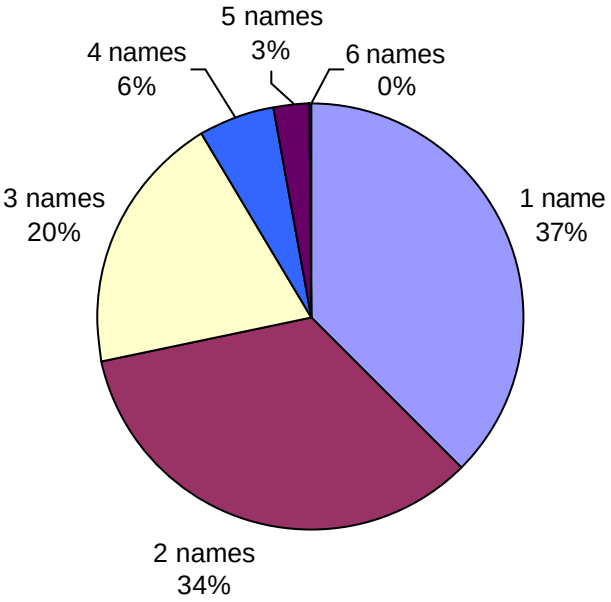

Fig 2 – Percentage of single or multiple naming (first thousand features)

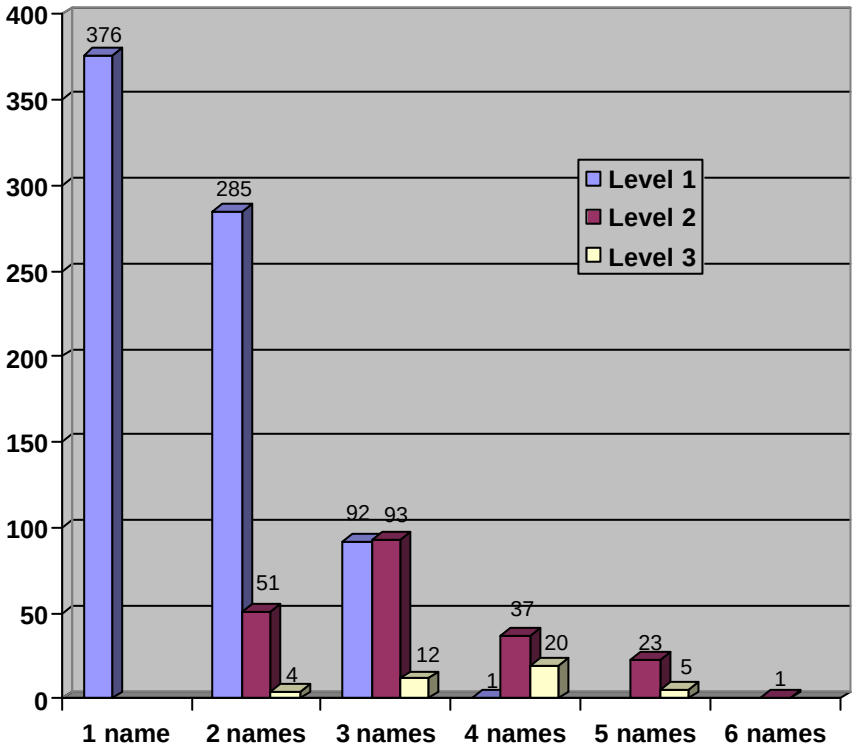

Fig. 3 – Distribution of multiplets of names (first thousand features)

Features with:	**Number**	**%**
1 name	6453	37.87
2 names	6088	35.73
3 names	3052	17.91
4 names	977	5.74
5 names	448	2.63
6 names	19	0.11
7 names	1	0.01
TOTAL	**17038**	**100.00**

Table 2 –Distribution of multiplets in the whole CGA

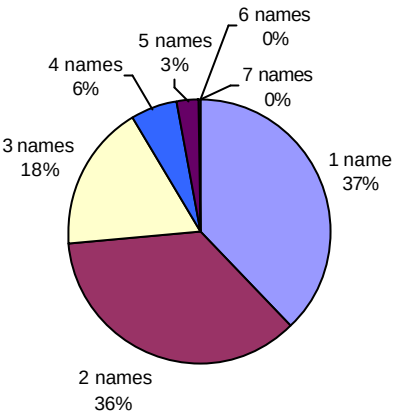

Fig. 4 – The pie distribution of table 2