

UNITED
NATIONS



Economic and Social Council

Distr.
GENERAL

E/CN.4/INF/56
20 August 1987

ENGLISH ONLY

Fifth United Nations Conference
on the Standardization of
Geographical Names
Montreal, 18-21 August 1987
Item 6(c), of the provisional agenda*
11(b)

TOPONYMIC DATA FILES: AUTOMATED DATA PROCESSING (ADP) SYSTEMS
WRITING SYSTEMS AND GUIDES TO PRONUNCIATION
CONVERSION INTO NON-ROMAN WRITING SYSTEMS

A computerized toponymic transliteration/transcription/translation system

Paper submitted by Israel *

* Prepared by Naftali Kadmon, Geography Department, The Hebrew University of Jerusalem, Israel.

A COMPUTERIZED
TOPONYMIC TRANSLITERATION/TRANSCRIPTION/TRANSLATION SYSTEM

At the 12th Session of the United Nations Group of Experts on Geographical Names in Geneva, September 1986, the Delegate from Israel reported on an automated system for the transliteration, transcription and translation of geographical names in general and world atlases in particular (see Working Paper No. 57).

The present paper, presented to the 5th United Nations Conference on the Standardization of Geographical Names, although based on the Working Paper mentioned above, reports on progress made towards the production stage achieved during the past year.

Perhaps no written product requires more individual investment in verbal transfer methodology than a world atlas. Producing a map of a single country with a linguistic/toponymic base other than that of the receiver language may require a single transliteration or transcription system only. The task is made more difficult if the scripts of the two countries are different, too, and even more so if they belong to different generic types such as alphabetic, syllabic and logographic. In a world atlas produced in a given receiver language all names, including those which in their original language are unwritten, must be transferred to a single writing and pronunciation system. This problem is well known to every atlas editor. But only in few cases are all names indeed transferred directly from the donor language to that of the receiver. In most instances "standardized" intermediate atlases are used for the production of a new atlas, and the more thorough the work which has been invested in the former, the higher can be the toponymic quality of the latter. If the number of published general atlases may be regarded as an indicator, it seems that the majority of work has been done on atlases in Roman script. This fact alone would justify the insistence of the UN Conferences on the

Standardization of Geographical Names, as well as of the UN Group of Experts on Geographical Names, on the development of a single official romanization system for all non-Roman scripts. Although the English language does not appear to the present writer to be the single most efficient medium for phonetic transmission of sounds of different languages (Armenian, for example, has a much wider range of consonants), no language group seems to have produced more good, well-founded world atlases than the English-speaking countries.

Some years ago the first computerized bi-scriptual gazetteer of Israeli place names, carrying Hebrew and official romanized forms, was developed (Kadmon, document E/CONF.74/L.24, 4th UN Conference on the Standardization of Geographical Names, 1982). The work by "Carta" of Jerusalem, a well-known Israeli publishing house with a tradition of map and atlas production, and described below, carries automated toponymy a very great step forward. Some years ago this firm acquired the translation rights of a major world atlas in the English language, the Times Atlas of the World. A magnetic tape of the index of over 200,000 names has been supplied, and this serves as the starting point for the computer-assisted production of the new atlas which is being printed in Hebrew. This language has an alphabetic script in which consonants are written in full, but most vowels are usually omitted, a practice followed also, e.g., by Arabic. However, foreign words, and in particular geographical names, require at least partial vocalization in order to avoid ambiguity.

A team of linguists and toponymists, mostly from among the faculty of the University of Jerusalem and the Academy of the Hebrew Language, was assembled with experts representing each of the linguistic blocks or regions, resembling to a certain degree the divisions of the UN Conferences on the Standardization of Geographical Names. Various methods of transfer apply, of course, to different languages and scripts in relation to Hebrew. Thus, Arabic, another Semitic language, has many similar attributes and Arabic script can be referenced via a single system of transliteration.

For Roman alphabets a phonetic transfer system has to be used, and one of the first acts required is setting up a basic table of equivalence. Since names in non-Roman scripts are ingested into the system through Roman characters with certain diacritics, this table is used also for toponyms in non-Roman scripts. A statistical analysis is then made, resulting in a basic "universal" transliteration system into Hebrew. On this the specific transliteration and transcription rules for various languages are then superimposed. This set of instructions is then incorporated in the processing software. Also included are the rules of stress for the various languages.

A computer-generated printout is now produced. Each record, i.e., each name, carries all relevant information pertaining to the name, including location, typeface and size, and, of course, the automatically transformed name. Transformation, in the present context, means either transliteration or transcription as applicable, as well as translation of generic terms (see below). All records of names relating to a given map, or country, or language block - as the case may be - are checked by the expert involved and corrected if necessary. On the checking document the latter can also add secondary names and decide whether these should appear only in the atlas index or in the maps, too. If word order of a name in the index has to be changed in relation to that in the maps, e.g. in complex names involving generic and specific terms (see N. Kadmon, Working Paper No. 3, UNGEGN, 19th Session) an appropriate code is inserted. Figure 1 shows a typical editing sheet of the computer printout, with an amendment by the "human" editor.

Geographical terms forming part of toponyms in all languages, some 2,000 in all, were translated and inserted into a terminology table. The software recognizes these strings of characters in the donor language and automatically translates them into Hebrew. However, the human element the linguistic-toponymic expert - retains overriding control in all stages and of all elements, both linguistic such as spelling, and cartographic, such as typeface and size.

After the corrections are inserted in the computer file and after checking, the magnetic tape is run on an optical typesetting system. This produces strip film which is duplicated as necessary for names appearing on more than one map. Names placement is carried out manually, in conformity with the views expressed in the Report of the 11th Session of UNCEGN, 1984 (document ESA/RT/C/GN/9, para. 49).

Sample map sheets in final printed form are already available.

The expertise gained enables the firm concerned to carry out map and atlas translation work in practically any language. The system is especially useful for maps and atlases using non-Roman script.

<p>13 01 119-B-008 01265 07852w</p> <p>ECUADOR</p> <p>MT</p> <p>Chimborazo</p> <p>016-00098-1</p> <p>12</p> <p>11</p> <p>D</p> <p>שְׁמֵי הַמְּדִינָה</p>	<p>13 01 119-B-008 01365 07930w</p> <p>ECUADOR</p> <p>Guacanda</p> <p>016-00090-1</p> <p>12</p> <p>11</p> <p>D</p> <p>שְׁמֵי הַמְּדִינָה</p>
<p>13 06 119-B-008 01265 07824w</p> <p>ECUADOR</p> <p>Baños</p> <p>016-00095-1</p> <p>12</p> <p>11</p> <p>D</p> <p>שְׁמֵי הַמְּדִינָה</p>	<p>13 02 119-B-008 01365 07929w</p> <p>ECUADOR</p> <p>Catarama</p> <p>016-00091-1</p> <p>12</p> <p>11</p> <p>D</p> <p>שְׁמֵי הַמְּדִינָה</p>
<p>13 07 119-B-008 01265 07954w</p> <p>ECUADOR</p> <p>Baños</p> <p>016-00096-1</p> <p>12</p> <p>11</p> <p>D</p> <p>שְׁמֵי הַמְּדִינָה</p>	<p>13 03 119-B-008 01365 07932w</p> <p>ECUADOR</p> <p>Pueblo Viejo</p> <p>016-00092-1</p> <p>12</p> <p>11</p> <p>D</p> <p>שְׁמֵי הַמְּדִינָה</p>
<p>13 06 119-B-008 01265 07830w</p> <p>ECUADOR</p> <p>Pelileo</p> <p>016-00097-1</p> <p>12</p> <p>11</p> <p>D</p> <p>שְׁמֵי הַמְּדִינָה</p>	<p>13 04 119-B-008 01365 07826w</p> <p>ECUADOR</p> <p>Tungurahua Vol</p> <p>016-00093-1</p> <p>12</p> <p>11</p> <p>D</p> <p>שְׁמֵי הַמְּדִינָה</p>

Figure 1: Sample page of the preliminary computer printout, with an amendment by the editor. The transformed name (in Hebrew) appears in the lower right corner of each "box".