UNITED
NATIONS

Group of Experts on Geographical Names

Eleventh Session
Geneva, 15-23 October 1984

WORKING PAPER
No. 42

Item No. 6(d) of the agenda

THE COMPUTER PROCESSING OF ABBREVIATIONS, SPECIAL CHARACTERS, AND THE SORTING OF GEOGRAPHICAL NAMES

(submitted by the United Kingdom)[1]

1/ Copyright © Controller HMSO, London, 1984

Introduction

At the Fourth United Nations Conference on the Standardization of Geographical Names (Geneva 1982) it was apparent that the computer processing of geographical names was becoming increasingly important in many countries, and would continue to do so. The United Kingdom described how it had developed and introduced an automated production system for gazetteers, based on the use of micro-computers and printers, and other countries described similar systems. In the lasttwo years we have considerably developed this automated system, and have become increasingly aware of the many benefits resulting from this. We have also been made aware of the many problems associated with automation that were not apparent at the outset.

The Benefits Resulting From The Automation of Geographical Names Data

The detailed administrative and locational information associated with geographical names result in large amounts of textual information which is often required in a variety of standard formats. Gazetteers are the best examples of such formats, but atlas or map indexes, and lists of names sorted by administrative and/or geographical area are also commonly required. The application of computer processing to such data has greatly enhanced the ease with which such large amounts of detailed information can be manipulated. The following benefits resulting from automation are quickly achieved:

1. Ease of Amendment/Update:- Once data has been input to a names database it can be quickly and easily recalled for correction and amendment. New or additional information can similarly be added to the database. As this does not involve changes to printing processes it simplifies output procedures.

2. Reduced Likelihood of Typographical Error:- In a fully automated names processing system, once data has been input to the names database and verified, there is no requirement for it to be typed or input again. Simple transcription or typing errors should thus be eliminated.

3. Variety of Output From a Single Database:- An automated names database allows selection from the database of names that meet given criteria. Thus names can be selected by geographical area, by administrative area or by feature code. Once selection has been made the data can be sorted to meet user requirements. Thus names can be sorted by map sheet number, by ascending latitude and longitude, or in the lexical order of the country concerned. The use of computer type setting and changeable printer-heads allows a variety of type faces to be used without having to change the database.

4. Output on Demand:- Through the continuous update of a names database, output can genuinely reflect the latest state of information. This can be true of "hard copy" printed output as much as cf computer display output. The flexibility of printed output and ease of update facilitate regular output to meet user demands.

5. Exchange of Information:- The increasing use of automated databases in cartography, and the rising demand for data in digital form can both be met from computer processing of geographical names. This would further ensure consistency in the treatment of geographical names between users exchanging information, and by users in successive editions of their products.

Problems Associated With The Computer Processing of Geographical Names Data

The major problems we have found when processing geographical names by computer have fallen into three broad categories: The treatment of diacritical marks and special characters, the treatment of abbreviations, and the automated sorting of geographical names to the required lexical order. None of these problems is insoluble; they are to some extent inter-related, but they are important factors that should be considered in system design and hardware procurement.

1. The Treatment of Special Characters/Diacritical Marks:- Most computers offer the user a range of special characters over and above the local linguistic requirements of the region they are designed to be used in. Many of these additional characters can be input directly from the keyboard as normal letters, but often input is cumbersome, and sometimes impossible. It is not uncommon to find computer printers which can print characters that compatible keyboards cannot input, and visa versa. When characters can be input the resolution of the VDU screen and/or printhead is often such that similar characters become blurred or indistinguishable.

2. The Treatment of Abbreviations:- The major problem caused by abbreviations is not how to portray them, but how to ensure they are sorted into the correct lexical order on output. It is common practice to show the term Saint/Sainte as St./Ste. such that in the United Kingdom we always refer to St Albans and never to Saint Albans. It is also the agreed norm to position the abbreviated term in a gazetteer as if it had been spelt in full. Thus St. Albans will be found just before St. Andrews and before Sandwich, but not just before Stalham. This is not easily achieved using a computer generated sorting routine, and requires careful system design.

The abbreviation Mac (also spelt Mc, Mck and McC) poses no problem however. It is the convention here to alphabetise as the term is spelt, so that Macarthur comes before McArthur and both before Mckinney. This is easily done by computer.

3. The Automated Sorting of Geographical Names:- The simplest of computer generated alphabetical sort routines take the unique numerical code allocated to each character the computer recognises, and compares each character of a name with the equivalent position character of the next name. The smaller the code, the higher the priority given to the character in the lexical order. Unfortunately the need for a unique code for each character results in upper case letters being distinguished from lower case letters. The commonly used ASCII code gives all upper case letters priority over all lower case letters, and some punctuation marks priority over the capital letters. (This includes the space character). Thus in Annex A, Cha Yue Pai has been given priority over Chai Kek, while Annex B shows the correct lexical order.

In many languages certain letter and diacritic combinations have a very different lexical order from the base letter. Thus in Scandinavian languages the lexical order is A,B,C,.....Y,Z,Æ,Ä,Å etc. While in English we would not distinguish between A and A/Å, the local requirement would be to make this distinction. Neither of these needs is met by ASCII codes which give the special characters a priority below lower case letters. Similar examples can be found in other languages.

To achieve a suitable sorted output from an automated names database suitable for gazetteer production requires the user to be able to define his own lexical order that does not give priority to punctuation marks, that does not distinguish between upper and lower case variants of the same letter, and that allows diacritical marks and special characters to be allocated their correct local order.

Annex A. ASCII Computer Sorting Routine.

NAME	DESG	LATITUDE	LONGITUDE	GRID COORDS	SHEET
Boulder Point see: Pak Kok	PT	22°14′N	114°06′E	KV 026 625	15
Boulder Point see: Kau Lau Wan Tsui	PT	22°27′N	114°21′E	KV 281 864	8
Boundary Street	RD	22°19′N	114°10′E	KV 088 719	11
Sowring Camp	PPL	22°24′N	113°58′E	HQ 069 815	6
Brick Hill see: Nam Long Shan	HLL	22°14′N	114°10′E	KV 083 625	15
Bride's Pool	PND	22°30′N	114°14′E	KV 166 913	3
Bride's Pool Road	RD	22°29′N	114°14′E	KV 160 892	3
Bridge Hill see: Liu Fa Tseng Shan	HLL	22°14′N	114°13′E	KV 129 619	15
Brothers, The	ISLS	22°20′N	113°58′E	HQ 061 728	10
Brothers Point see: Tai Lam Kok	PT	22°21′N	114°01′E	JV 927 756	6
Buffalo Hill see: Shui Ngau Shan	HLL	22°22′N	114°13′E	KV 149 771	7
Buffalo Pass see: Tai Lo Au	PASS	22°22′N	114°14′E	KV 152 769	7
Burma Lines	CMPM	22°30′N	114°09′E	KV 073 915	3
Butler, Mount see: Pat Na Shan	HLL	22°16′N	114°12′E	KV 123 653	11
Butterfly Beach see: Wu Tip Wan	BAY	22°22′N	113°57′E	HQ 043 772	5
Butterfly Estate	PPL	22°22′N	113°57′E	HQ 050 775	5
Butterfly Valley see: Wu Tip Kuk	VAL	22°20′N	114°08′E	KV 055 737	11
Byewash Reservoir	RSV	22°21′N	114°08′E	KV 063 746	11
Cafeteria Beach	BCH	22°22′N	113°59′E	HQ 073 774	6
Calf's Head see: Fu Yung Pit	HLL,	22°22′N	114°14′E	KV 154 769	7
Cameron, Mount	HLL	22°15′N	114°10′E	KV 086 648	11
Camp Cove see: Pak Sha Tau Wan	BAY	22°32′N	114°17′E	KV 218 946	4
Cape Collinson Training Centre	BLDG	22°15′N	114°15′E	KV 166 633	11
Cape D'Aguilar Road	RD	22°13′N	114°14′E	KV 157 591	15
Care Villages	PPLS	22°12′N	114°01′E	KV 923 583	14
Caroline Hill	LCTY	22°16′N	114°11′E	KV 099 663	11
Casam Beach	BCH	22°22′N	114°04′E	JV 984 766	6
Cassino Lines	CMPM	22°29′N	114°04′E	JV 993 898	2
Castle Peak see: Tsing Shan	HLL	22°23′N	113°57′E	HQ 038 789	5
Castle Peak Bay see: Tsing Shan Wan	BAY	22°22′N	113°58′E	HQ 060 708	6
Castle Peak Beach	BCH	22°22′N	113°58′E	HQ 066 780	6
Castle Peak Firing Range	CMPM	22°24′N	113°56′E	HQ 030 810	5
Castle Peak Road	RD	22°25′N	113°59′E	HQ 075 833	6
Castle Rock see: Lo Chau Pak Pai	RKW	22°10′N	114°14′E	KV 148 550	15
Causeway Bay see: Tung Lo Wan	LCTY	22°16′N	114°10′E	KV 096 667	11
Causeway Bay Typhoon Shelter	HBR	22°17′N	114°11′E	KV 098 673	11
Cemetery Gap see: Po Leng Au	LCTY	22°29′N	114°07′E	KV 039 900	3
Central District see: Chung Wan	LCTY	22°17′N	114°09′E	KV 070 670	11
Centre Island see: A Chau	ISL	22°26′N	114°13′E	KV 138 843	7
Cha Hang see: Tai Hang	PPL	22°28′N	114°08′E	KV 064 877	3
Cha Kwo Chau	ISL	22°12′N	113°56′E	HQ 034 597	13
Cha Kwo Leng see: Cha Kwo Ling	PPL	22°18′N	114°13′E	KV 142 689	11
Cha Kwo Ling	PPL	22°18′N	114°13′E	KV 142 689	11
Cha Liu Au	PASS	22°20′N	114°13′E	KV 140 726	11
Cha Yue Pai	ISL	22°22′N	114°17′E	KV 216 775	8
Chai Kek	PPL	22°26′N	114°07′E	KV 042 847	7
Chai Wan	LCTY	22°16′N	114°14′E	KV 149 650	11
Chai Wan	BAY	22°16′N	114°14′E	KV 155 655	11
Chai Wan Estate	PPLX	22°15′N	114°14′E	KV 153 648	11
Chai Wan Kok	PPL	22°22′N	114°06′E	KV 018 775	6
Chai Wan Road	RD	22°16′N	114°13′E	KV 146 652	11
Cham Keng Chau	ISL	22°33′N	114°25′E	KV 349 960	4
Cham Pai	RF	22°29′N	114°21′E	KV 286 896	4
Cham Shan	HLL	22°23′N	114°09′E	KV 073 786	7
Cham Shan	HLL	22°31′N	114°07′E	KV 048 934	3
Cham Tau Chau	ISL	22°22′N	114°17′E	KV 205 775	8
Cham Tin Shan	HLL	22°21′N	114°13′E	KV 142 744	11

Annex B. Gazetteer Sorting Routine.

NAME	DESG	LATITUDE	LONGITUDE	GRID COORDS	SHEET
Boulder Point see: Pak Kok	PT	22°14'N	114°06'E	KV 026 625	15
Boulder Point see: Kau Lau Wan Tsui	PT	22°27'N	114°21'E	KV 281 864	8
Boundary Street	RD	22°19'N	114°10'E	KV 088 719	11
Bowring Camp	PPL	22°24'N	113°58'E	HQ 069 815	6
Brick Hill see: Nam Long Shan	HLL	22°14'N	114°10'E	KV 083 625	15
Bride's Pool	PND	22°30'N	114°14'E	KV 166 913	3
Bride's Pool Road	RD	22°29'N	114°14'E	KV 160 892	3
Bridge Hill see: Liu Fa Tseng Shan	HLL	22°14'N	114°13'E	KV 129 619	15
Brothers, The	ISLS	22°20'N	113°58'E	HQ 061 728	10
Brothers Point see: Tai Lam Kok	PT	22°21'N	114°01'E	JV 927 756	6
Buffalo Hill see: Shui Ngau Shan	HLL	22°22'N	114°13'E	KV 149 771	7
Buffalo Pass see: Tai Lo Au	PASS	22°22'N	114°14'E	KV 152 769	7
Burma Lines	CMPM	22°30'N	114°09'E	KV 073 915	3
Butler, Mount see: Pat Na Shan	HLL	22°16'N	114°12'E	KV 123 653	11
Butterfly Beach see: Wu Tip Wan	BAY	22°22'N	113°57'E	HQ 043 772	5
Butterfly Estate	PPL	22°22'N	113°57'E	HQ 050 775	5
Butterfly Valley see: Wu Tip Kuk	VAL	22°20'N	114°08'E	KV 055 737	11
Byewash Reservoir	RSV	22°21'N	114°08'E	KV 063 746	11
Cafeteria Beach	BCH	22°22'N	113°59'E	HQ 073 774	6
Calf's Head see: Fu Yung Pit	HLL	22°22'N	114°14'E	KV 154 769	7
Cameron, Mount	HLL	22°15'N	114°10'E	KV 086 648	11
Camp Cove see: Pak Sha Tau Wan	BAY	22°32'N	114°17'E	KV 218 946	4
Cape Collinson Training Centre	BLDG	22°15'N	114°15'E	KV 166 633	11
Cape D'Aguilar Road	RD	22°13'N	114°14'E	KV 157 591	15
Care Villages	PPLS	22°12'N	114°01'E	KV 923 583	14
Caroline Hill	LCTY	22°16'N	114°11'E	KV 099 663	11
Casam Beach	BCH	22°22'N	114°04'E	JV 984 766	6
Cassino Lines	CMPM	22°29'N	114°04'E	JV 993 898	2
Castle Peak see: Tsing Shan	HLL	22°23'N	113°57'E	HQ 038 789	5
Castle Peak Bay see: Tsing Shan Wan	BAY	22°22'N	113°58'E	HQ 060 708	6
Castle Peak Beach	BCH	22°22'N	113°58'E	HQ 066 780	6
Castle Peak Firing Range	CMPM	22°24'N	113°56'E	HQ 030 810	5
Castle Peak Road	RD	22°25'N	113°59'E	HQ 075 833	6
Castle Rock see: Lo Chau Pak Pai	RKW	22°10'N	114°14'E	KV 148 550	15
Causeway Bay see: Tung Lo Wan	LCTY	22°16'N	114°10'E	KV 096 667	11
Causeway Bay Typhoon Shelter	HBR	22°17'N	114°11'E	KV 098 673	11
Cemetery Gap see: Po Leng Au	LCTY	22°29'N	114°07'E	KV 039 900	3
Central District see: Chung Wan	LCTY	22°17'N	114°09'E	KV 070 670	11
Centre Island see: A Chau	ISL	22°26'N	114°13'E	KV 138 843	7
Cha Hang see: Tai Hang	PPL	22°28'N	114°08'E	KV 064 877	3
Chai Kek	PPL	22°26'N	114°07'E	KV 042 847	7
Chai Wan	LCTY	22°16'N	114°14'E	KV 149 650	11
Chai Wan	BAY	22°16'N	114°14'E	KV 155 655	11
Chai Wan Estate	PPLX	22°15'N	114°14'E	KV 153 648	11
Chai Wan Kok	PPL	22°22'N	114°06'E	KV 018 775	6
Chai Wan Road	RD	22°16'N	114°13'E	KV 146 652	11
Cha Kwo Chau	ISL	22°12'N	113°56'E	HQ 034 597	13
Cha Kwo Leng see: Cha Kwo Ling	PPL	22°18'N	114°13'E	KV 142 689	11
Cha Kwo Ling	PPL	22°18'N	114°13'E	KV 142 689	11
Cha Liu Au	PASS	22°20'N	114°13'E	KV 140 726	11
Cham Keng Chau	ISL	22°33'N	114°25'E	KV 349 960	4
Cham Pai	RF	22°29'N	114°21'E	KV 286 896	4
Cham Shan	HLL	22°23'N	114°09'E	KV 073 786	7
Cham Shan	HLL	22°31'N	114°07'E	KV 048 934	3
Cham Tau Chau	ISL	22°22'N	114°17'E	KV 205 775	8
Cham Tin Shan	HLL	22°21'N	114°13'E	KV 142 744	11
Channel Rock see: Cham Pai	RF	22°29'N	114°21'E	KV 286 896	4