# Creating indexes for world atlases at HarperCollins Publishers

Jim Irvine

*This paper describes the methods used in-house by HarperCollins Publishers to create indexes for some of the most prestigious world atlases available today. The 14-stage process described has been developed by HarperCollins, and uses a mix of non-specialist software packages and applications developed in-house.*

HarperCollins Publishers (HCP) General Reference division creates and publishes the Collins and Times ranges of world atlases and maps. Our flagship product is the *Times Comprehensive Atlas of the World*, 11th edition, published in October 2003. This is a huge book, weighing 5.5 kg and containing 56 pages of preliminary thematic content, 249 pages of full colour reference maps and 217 pages of index containing over 200,000 entries. This article focuses on the characteristics of this and other HCP atlas indexes, and how they are created.

## The characteristics of world atlas indexes

Most world atlas indexes share certain characteristics that distinguish them in terms of form and function from those found in other reference works.

- The primary function of an atlas index is to help the user locate places on the map pages. Each index entry therefore directs the user to a single map page which best shows the place in context, regardless of how many map pages the place actually appears on.
- A secondary function is to inform the user about places by providing information beyond that which can be gleaned from viewing it on a map. For example administrative qualifiers placing the feature into its political context, latitude/longitude coordinates, and geographical descriptors indicating the type of feature are often included. This leads to the index sections of many world atlases being more accurately described as index/gazetteers, reflecting this dual purpose. Indeed many users get the information they require directly from the index, without having to refer to the maps at all.
- Well-known English alternative place names are generally shown in brackets after the local name on world atlas maps, for example Torino (Turin). These alternative names appear as cross-references in the index. Some indexes also include cross-references for additional alternative names which do not appear on the map, but nevertheless assist the user to locate the correct place. These include alternative spellings, former/historic names and so on. World atlas indexes may therefore contain more names than the maps they refer to. The 11th edition of the *Times Comprehensive Atlas of the World* contains over 10,000 of these index-only cross-references.

The process by which HCP creates world atlas indexes reflects these characteristics, and is described below. It attempts to automate as much of the process as possible while at the same time maintaining a high degree of flexibility.

## Overview of technology used by HCP

Digital technology is used to create all maps and indexes. A suite of cartographic databases is maintained using specialist Geographic Information Systems (GIS) software, including a master place names database which underpins the index creation process. In the databases each town or physical feature is uniquely identified by an automatically generated ID number. The uniqueness enforced by this ID is key to the indexing process, as many names occur more than once in any world atlas. For example, there are 32 places in the *Times Comprehensive Atlas of the World* called San Pedro. To index each of these correctly we need to be able to distinguish effectively between them.

The master place names database includes several pieces of information required for indexing, including:

- separate map and index name forms for all features, allowing names to be permutated (such as 'Everest, Mount') in indexes
- alternative name forms
- administrative qualifiers, that is, the country, administrative division and if appropriate the island the place falls within
- a descriptor or type for all physical features: mountain, hill, river, salt flat and so on
- latitude/longitude coordinates.

Accented characters are represented by numeric codes attached to the character, the whole being enclosed in angled brackets. For example, code 03 represents a grave accent, so è becomes <03e> as in Gen<03e>ve for Genève.

Data is extracted from these cartographic databases and processed using complex map creation software to produce a huge variety of different types of map. Indexes to these maps are created and stored using standard database software from Oracle, supplemented at particular points of the process by software written in-house to accomplish specific tasks.

## The index creation process

The index creation process comprises a series of stages as described below.

*Stage 1: Set up Oracle table and screen*

A new product index table is created in Oracle for each atlas index, together with an interactive screen to allow individual index entries to be manually edited. This contains fields for each element that might be required in the final index, together with a number of non-printing fields called flags. These are used later in the process to control which elements are actually included in each index entry.

*Stage 2: Which places are on each atlas page*

To index an atlas we must know which names appear on each map. Map pages are held as Adobe Illustrator files. In-house software is run against these pages to extract a list of names with their unique ID numbers for each page. These ID numbers and the map names together with the relevant page numbers are then loaded into the product index table. Bracketed names are loaded separately with a flag set to indicate that these records should generate cross-references.

*Stage 3: Create full index entries*

The remaining fields in the product index table are then populated for each entry, including the index version of the name. Software is used to pull the information in from the master place names database, joining on the unique ID to ensure the correct links are made.

*Stage 4: Index-only cross-references*

If the atlas in question is to include index-only cross-references, these are then copied into the product index table from the master place names database. Generally a selection is made based on the categorization of each alternative name, and the appropriate entries are copied across, again joining on the unique ID. The flag that indicates that these are cross-reference entries is also set.

*Stage 5: Create list of alternative names under main entries*

Besides appearing as cross-references, a list of alternative names is also included under the main entry in many of our indexes. These are created automatically via software, with the list of alternative names held as a single string in the product index table, attached to the relevant main entry.

For example, the following two cross-references

**Den Haag** Neth. *see* **'s-Gravenhage**
**The Hague** Neth. *see* **'s-Gravenhage**

result in the following main entry, with both alternatives classified and listed below:

**59 F4 's-Gravenhage** Neth. 52° 04' N 4° 18' E
alt. **Den Haag**, conv. **The Hague**

*Stage 6: Create alphanumeric grid references*

Next alphanumeric grid references are created for each entry. The spatial referencing system on each map generally follows the graticule. This gives a curved grid, the exact shape depending on the map projection used. The letter and number labels referencing each feature to a particular cell can therefore be calculated arithmetically from the coordinates of the features themselves. This is achieved using in-house software which requires the maximum and minimum latitude and longitude and degree interval between graticule lines to be entered for each page.

*Stage 7: Check for missing elements and other data problems*

To ensure all relevant fields in the product index table are now complete, a series of checks are carried out. These are precautionary in nature, as the digital methods employed mean problems are unlikely to occur. They include:

- A simple check for missing data.
- A check for orphaned cross-references, that is, cross-references referring to main entries that do not exist.
- Duplicate index entries.
- A check for discrepancies between the map names copied from the final map pages and the index names copied in from the master place names database. The master database is constantly updated, and this traps any names that have changed during the time lag between initial map creation and indexing.

*Stage 8: Descriptors for physical features*

All physical features in the master place names database have a descriptor indicating their type. These are copied into the product index table during stage 3 above. Descriptors are not included where the type of the feature is implicit in its name, however. A software update based on a series of rules automatically sets a flag for each entry to indicate whether the descriptor should be included or not. For example the entry

102 N4 **Washington, Mount** mt. *NH* U.S.A. 44° 18' N 71° 18' W

unnecessarily includes the descriptor 'mt.'. In this instance the flag would be set to 'No', resulting in the following final index entry.

102 N4 **Washington, Mount** *NH* U.S.A. 44° 18' N 71° 18' W

Occasionally a physical feature that would otherwise have its descriptor suppressed shares a name with a town. In these cases the automatically generated flag must be overridden to re-include the descriptor in order to distinguish which is which. An example is Thunder Bay, where the final index entries look like this:

96 A3 **Thunder Bay** b. Ont. Can.
96 A3 **Thunder Bay** Ont. Can. 48° 26' N 89° 15' W

Software is run against the Oracle index table to list these instances, and the Include descriptor flag is set manually.

### *Stage 9: Include administrative divisions*

In a similar fashion, a flag is set in the product index table to indicate which entries should include an administrative division. Generally there are two cases where this occurs.

First there are normally some countries where first-order administrative divisions are included for all towns. In the *Times Comprehensive Atlas of the World* this includes Australia, Canada, China, India, the United Kingdom and the United States of America.

Second, administrative divisions may be used to distinguish between places of the same name and type in the same country. For example:

87 E5 **Fagwir** *Jonglei* Sudan 9° 07' N 30° 45' E
87 E5 **Fagwir** *Wahda* Sudan 9° 32' N 30° 24' E

In both cases the Include admin flag is set automatically using software.

### *Stage 10: Include coordinates.*

A flag is also used to indicate which entries should include latitude and longitude coordinates. In the *Times Comprehensive Atlas of the World* this includes all towns and certain types of physical feature. Again this flag is set automatically through software.

### *Stage 11: Overlapping*

Complete and final index entries now exist for all named features, with multiple entries for features that appear on more than one map page. The final yes/no flag is used to indicate which of these multiple entries should be included in the index. The general rule is that each feature should be indexed to the map page that shows it best in context. This is normally, though not necessarily, the largest-scale map it appears on. In order to get the best result our Cartographic Editors examine each page and generate a set of rules indicating which of the names appearing on that page should be indexed to it, and which should not. Example rules for say page 64 could be to include all names in India not already indexed to page 62 and all names in Pakistan south of 30 degrees north.

Setting the overlapping flag to implement these rules is one of the most complex tasks in index creation. Each page is done individually, with each rule applied via a software update against the product index table. Once the flag is set for all entries, two checks are carried out to ensure that each place is indexed once and once only. These take the form of two queries run against the product index table, using the Unique ID to identify multiple entries for the same place. The following are listed out.

- all places that are selected for inclusion in the index more than once
- all places that exist at least once in the index, but are never selected for inclusion.

The anomalies highlighted by these lists can then be fixed via software or by manually editing the overlap flag for individual entries. The checks are then run again to ensure the relevant fixes have been made. For a large atlas this may be an iterative process, with several checking and fixing stages carried out until all names on the maps are included in the index once and once only.

The deselected entries remain in the product index table as a record of exactly which names appear on each map.

### *Stage 12: Alpha sort*

For alpha sorting, a third version of each name is automatically created which strips out all non a–z characters, expands 'St' to saint and converts all letters to lower case. This also removes the codes used for accented characters. The following examples illustrate:

| **Index name** | **Sort version** |
|---|---|
| North Queensferry | northqueensferry |
| M<04u>nchen | munchen |
| St-Agnant | saintagnant |
| Ste-Alv<03e>re | saintealvere |

Several place names worldwide contain numbers. If required the sort version of these names can be edited manually to ensure they alpha sort correctly. For example the policy in the *Times Comprehensive Atlas of the World* is for numbers to sort before letters, so that names beginning with numbers have their own section at the top of the index, and the mountain K2 is the first entry under 'K'. Examples of the sort names given to these entries are:

| **Index name** | **Sort version** |
|---|---|
| 6 October City | aaaaaaoctobercity |
| K2 | K-aaaa |

The hyphen included in the sort name for K2 ensures that this entry comes before the entry for 'Ka', a settlement in the Democratic Republic of Congo.

The next task is to actually put the entries into the correct order. A software program sorts all entries by the sort name, then country, then descriptor, then administrative division and finally by the alphanumeric grid reference. The program updates a number field for each entry in the product index table indicating its position in the sort. This number field starts at 5 and increments by 5 each time, allowing for subsequent insertion of additional index entries into their correct position in the sorted index. This is however a rare occurrence.

Sections of the index are then checked to ensure that the alpha sort is correct. Generally the top and bottom 100 entries are scanned, together with a few sections from elsewhere.

### *Stage 13: Extract tagged text to flow into index pages*

The alpha sorted index is then combined with Quark XPress tags to control typesetting, and extracted to a text file. The various yes/no flags control what is included for each entry, and the cross-reference flag ensures that these entries are structured correctly. A search and replace is carried out on this file to substitute accent codes with real accented characters in the appropriate font. The final text file is then flowed into Quark XPress, resulting in a first-cut properly typeset index.

### *Stage 14: Final editing*

Final editing of the index is carried out in Quark XPress. This includes:

- Addition of running heads on each page.
- Addition of alpha breaks.
- Tops and tails of each column are checked and edited to ensure that individual entries do not span columns.
- Entries spanning more than one line are checked and edited to ensure that line breaks occur appropriately. Automatic inclusion of non-breaking spaces in multi-word place names and coordinate strings reduces the number of entries that need to be edited.

It is also possible at this stage to fine-tune the typesetting of the index to ensure it fits the available pages. Proofs of the completed index are then printed out and checked. The robust nature of the creation process together with the checks carried out throughout mean that we can be confident the content of the index is correct.

## Benefits of this approach

The method has been fine-tuned over several years, and brings many benefits, including:

- Increased efficiency. The high level of automation significantly reduces the total time required to create an index over traditional manual methods.
- Automatic checking. Many routine content checks can be performed automatically through software: for example searching for orphaned cross-references, overlap checks and ensuring all duplicate names in the same country include administrative divisions.
- Facilitates reuse. Maintaining all occurrences of all names in the product index table makes it easy to reuse this name set in the future. For example this might be to construct an index for a new atlas that reuses a selection of pages from existing atlases; or it might be to ensure that exactly the same name set appears on a completely new map or atlas page.
- Facilitates atlas revision. During creation of a new edition, the names held in the product index table can be compared against the names currently held in the master place names database to highlight names that have changed in the interim.
- Facilitates foreign language translation. As the index for an atlas is held digitally, all names can be extracted into popular word processing or spreadsheet applications for co-publishing partners in foreign markets to add translated name forms. The index file can then be automatically translated and re alpha sorted, and so on. We have also developed software in-house that automatically translates names on the maps in Illustrator, again based on the translated names file.

## Conclusion

Indexes are necessarily created towards the end of each project, with delivery deadlines looming. It is important therefore that the methods and technology used are robust. The approach described above has been used to create indexes for some of the most prestigious world atlases available today, passing all tests with flying colours. However, as technology develops and we continue to look for ways to improve our atlas production flowline, it is certain that the process described above will evolve and adapt in the future.

## References

*The Times Comprehensive Atlas of the World*, 11th edition, 2003, London: Times Books.

*Jim Irvine has worked for HarperCollins cartographic operation for 15 years, initially as part of the team responsible for the introduction of digital technology into the map creation process, and most recently as Head of Digital Resources and the Collins Newsroom. He is a geography graduate from the University of Edinburgh and also holds a postgraduate diploma in digital mapping from the University of Glasgow.*
*Throughout his time at HarperCollins Jim and his team have been directly involved in developing indexing processes for a variety of maps and atlases. Email:* jim.irvine@harpercollins.co.uk.