INTRODUCTION

On 17 November 2009, the UK government set out their vision for the role of public data and information in the delivery of Smarter Government. This included proposals to release certain datasets collected and maintained by the public sector (local and central Government), for free, with no restrictions on re-use. Initially the Ordnance Survey 1:25 000 scale and 1:50 000 scale raster products were proposed to be included in the open data portfolio (Gov.uk 2009). These are familiar to the general public in the paper mapping products Explorer and Landranger. But it emerged from a public consultation that these raster products would not meet the requirements. The reason for releasing these datasets was to create an economic stimulus, as developers begin to combine the information in new and innovative ways, creating fresh on-line business opportunities. One of the intentions of the Making Public Data Public initiative was to provide contextual backdrop maps allowing people to overlay and analyse 3rd party information. The current specifications were never designed with this in mind – the Ordnance Survey 1:25 000 and 1:50 000 scale mapping has been specifically tailored for the outdoor leisure market. Hence there is a lot of detail, cluttering these maps, making it difficult for the user to overlay data and carry out visual analysis in a GIS.

Ordnance Survey therefore decided in January 2010 to create a new product, OS VectorMap District, designed to work as a backdrop map, with just under four months to design and create the product. This paper describes the main technical challenges involved in creating this product, mostly automatically.

This paper starts with describing the new product family to which OS VectorMap District belongs, and how the product was designed. The following section describes how the system to create the product was designed and implemented, to meet the very tight time constraint. This is followed by a technical section that describes the most challenging tasks that had to be automated. This leads to the section presenting the results obtained in the first (alpha) release of the products, and the improvements brought to it for the second (beta release), followed by some information about how and where the product is published. Finally the conclusion highlights what we have learned during this project and the opportunities it brings to create products in the future.

OS VECTORMAP

The Making Public Data Public initiative brought the opportunity to create an entirely new backdrop vector mapping product to support district-level analysis in the 1:15 000 to 1:35 000 scale range - OS VectorMap District. This product is designed to be flexible and customisable, and is specifically tailored for use on the web.

Figure 1. OS VectorMap Local shown using three different display styles

OS VectorMap District is part of the VectorMap product family, which began in 2009 with the launch of OS VectorMap Local, see Figure 1 (Ordnance Survey 2009a). Being the second product in the series, it builds on the experience gained from creating the first. Ordnance Survey had already conducted an internal proof-of-concept project on district-level vector mapping in 2009. In addition Ordnance Survey Research have been investigating derived mapping for a decade and this knowledge was an essential element of the 2009 proof-of-concept and the implementation of the OS VectorMap District product.

During the creation of OS VectorMap District, it was necessary to produce a wide selection of maps demonstrating different possible content themes and styling. A "plot-folio" was assembled, containing paper plots of various alternative specifications in five representative sample areas (urban, rural, coastal,
mountainous and wooded). Like a wallpaper book, this allowed the specification to be openly discussed, iteratively improved and finally agreed through extensive consultation with Government.

OS VectorMap District is supplied in 10 x 10km tiles, with no features overlapping tile boundaries (features are chopped at the tile edge). The whole country is covered by 2860 of these OS VectorMap District tiles.

**AUTOMATION APPROACH**

Due to the time constraint, it was decided to build a tile based production system. The country would be divided into 10x10km tiles. This is much simpler than designing a system able to handle a continuous database. Updates would be made by full tile replacement. This meant that the system needed to be almost fully automatic, as manual finishing would have to be repeated at each refresh. Also the automation had to be reasonably efficient, to allow the whole country coverage to be refreshed with a few weeks of batch processing.

Some of the data themes included in OS VectorMap District already existed in Ordnance Survey, and could be included using simple attribute queries, reclassification, minor geometry changes and splitting into 10km tiles. This paper focuses on those transformations that were particularly challenging, namely buildings, railways, coastlines and cartographic text. These demonstrate the techniques of automated generalisation, automatic text placement and spatial data conflation.

Generalisation is the process of deriving smaller scale, less detailed, maps from larger scale maps. It involves selecting and representing information in a way that adapts to the display resolution of the map. The website of the ICA Commission on Generalisation and Multiple Representation, provides a good online resource on this topic, including downloadable workshop papers and a comprehensive bibliography (ICA Gen&MR 2010).

Conflation is a process carried out to integrate spatial datasets. It generally involves several steps: first translating the datasets into a common environment for comparison, secondly finding corresponding features between the datasets (data matching) and finally integrating the datasets such as correction of differences, geometry/attribute transfer, removal of duplicates and geometry alignment.

**BUILDING GENERALISATION**

The large scale vector data for buildings is very detailed and is unsuitable for district-level mapping. There was no building dataset in vector format available at Ordnance Survey that could be used to create OS VectorMap district. It has therefore been decided to generalise the large scale buildings. Figure 2 shows how the buildings generated for the alpha version compare with buildings available in other products.

![Figure 2. From left to right: OS MasterMap buildings, 1:10 000 scale raster, 1:25: 000 scale raster, OS VectorMap District alpha and 1:50 000 scale raster](image)

*Ordinance Survey © Crown Copyright. All rights reserved.*

To generalise the buildings, we have reused and extended the “growing tide” algorithm, initially implemented by Ordnance Survey Research between 2004 and 2005 in a project that created a prototype 1:50 000 scale vector map of the Glasgow area. The overall results of the project are described in (Revell et. al. 2007), while the detail of the building generalisation algorithms is presented in (Regnauld and Revell 2007).

To create OS VectorMap District, the parameters for this building generalisation process were adjusted to retain more details. The building generalisation algorithm starts by casting the “shadow” of groups of
buildings onto the surrounding roads. This is used to create a fixed-width buffer which follows the roads and encloses just those buildings in proximity to that road (Figure 4a). Buildings outside the initial buffer are subsequently included by amalgamation (Figure 4b).

The process was enhanced to allow it to call two different amalgamation algorithms. The original was a rural building "squared" amalgamation algorithm (Revell 2004). The algorithm is generally more applicable in rural areas. The second algorithm is an adaptation of the more recent work of (Dammen and van Kreveld 2008), which is based on Minkowski sums and differences. The algorithm works by applying a morphological dilate and erode to the buildings, using a square buffer oriented to the most common orientation in the building group (Figure 3).

![Figure 3. Large scale buildings (left) amalgamated using Minkowski (right)
Ordinance Survey © Crown Copyright. All rights reserved.](image)

After amalgamation has completed these building “amalgams” require simplification but only to be applied specifically to the angular parts (not on the side shared by the road symbol, materialised by a yellow tint on Figure 4a.). To achieve this, an existing building simplification algorithm, inspired by the algorithm described in (Lichtner 1979), was broken into its component parts: the primitives are called “hat”, “stairs”, “simp” (simplifying short edges) and “line” (removing collinear points). Instead of using a static sequence of these primitives, we have used an optimisation approach for determining the best combination to apply to each building. The optimisation engine used is based on ideas first conceived during the AGENT project (Agent 2000), and now available in Radius Clarity (1Spatial). The system converges towards the solution by evaluating a set of constraints, that control different aspects of the result (granularity, angles, size, convexity).

Finally holes get filtered based on their size and narrow corridors get either enlarged or removed. The final result is shown on Figure 4d and 4e (without and with the initial buildings on top).
It is impossible to process all the buildings at once, therefore it became necessary to partition the space. This was achieved using a combination of roads and urban extents, to create partitions of the space, which can be processed independently. The building generalisation was by far the most computationally intensive process.

For the alpha version the process would have taken 54 days on a single computer. To speed up the processing, tiles were processed in parallel on a cluster of 6 workstations.

Before the beta version, a number of improvements were made. The agent based generalisation has been optimised, which reduced the processing time of the simplification by approximately 50%. New algorithms have been added to tidy up the results, like cleaning narrow gaps and corridors, removing spikes, and simplifying the outlines of holes. The building generalisation process for the beta version (results shown in Figure 4) takes as long as the alpha process, but produces a much better result.

**RAILWAYS**

The source topographic line data for the railways is separated into three classes: multi-track, single-track and narrow gauge. The multi-track and narrow gauge railways could be included in OS VectorMap District directly without generalisation. However the single-track features were very cluttered in some areas, so it was decided to leave the sidings out of the product. The problem was to identify them, as they were classified as single track.
Some years ago Ordnance Survey created a detailed rail network dataset containing a detailed classification of rail types, including information about sidings, but the dataset has not been maintained, so is quite out-of-date.

Figure 5. Multi-track railways (black), single-track railways (red), deduced gaps (turquoise) and the eliminated sidings (green).

Ordnance Survey © Crown Copyright. All rights reserved.

The best solution was to transfer the attributes if this old dataset to the up-to-date topographic railway data. Initially open-source road network matching tools were tested (Revell and Antoine 2009), but these were found to be unsuitable as the structure of the railway data was fundamentally different from that of road networks. Consequently a bespoke algorithm was developed to match the two datasets. This algorithm uses the Hausdorff distance for testing potential matches, and only retains matches of high confidence.

After matching, the rails were still not completely classified. It was necessary to build a complete network by deducing gaps that occurred when the railways passed through tunnels or under roads. The classification from known sections of the network could then be propagated through to the unclassified sections. As a final step, everything not classified as single-track was removed, thus erasing the sidings and leaving the uncluttered single-track railways behind (Figure 5).

Since the data volume was relatively small, all this processing took place on the national set of railways, with the results being split into tiles as a post process. The processing of the railways took less than one hour.

GENERALISATION OF COASTAL AREAS

In the large-scale OS coastal data, linear features define the Mean High and Mean Low Water (MH&LW) tidal marks. The foreshore region between the MH&LW lines is represented by inter-tidal polygons. The region below the MLW is represented by permanent tidal water polygons. In VectorMap District alpha, the MH&LW lines were not generalised and inter-tidal areas were omitted from the map. For VectorMap District beta, some significant improvements have been made, using generalisation processes that employ explicit topology to ensure that shared-edge relationships between coastal features are preserved.

The Weighted Effective Area algorithm (Zhou & Jones 2005) was found to give the best results when simplifying MH&LW lines. When applying such modifications, it is important to maintain shared-edge relationships between features and ensure that the topology is not broken. If a small self-intersecting loop is created, then the line is locally modified to remove the loop. If modifying a line would cause it to self-intersect with another line, the problem section is locally held in place, while all other parts of the line are simplified. In rare cases when the modification still fails, the simplification tolerance is temporarily reduced, but this can cause a long section of coastline to remain too detailed. The resulting lines can be
rather angular, so Gaussian smoothing is applied to the result, again ensuring that all modifications are topologically valid. Results are shown in Figure 6.

Figure 6. Simplification and smoothing of MHW lines, original lines in red. 
Ordnance Survey © Crown Copyright. All rights reserved.

The next step is to generalise the inter-tidal and tidal polygons. The same process is applied to both: first adjacent (touching) polygons are merged together, then holes smaller than a threshold are eliminated, along with any line and area features inside those holes. As a final step, the tidal water polygons are extended to additionally cover inter-tidal areas. This means that the user can choose to turn off the MLW and the inter-tidal areas (brown) and still have a complete map, as shown on the right of Figure 7. This will also support future VectorMap products at smaller scales. Land polygons (cream) are used to represent all non-tidal and non-inter-tidal regions of the map.

Figure 7. Coastal area displayed with and without MLW and inter-tidal areas 
Ordnance Survey © Crown Copyright. All rights reserved.

NAMED EXTENDS AND CARTOGRAPHIC TEXT

Ordnance Survey has been investing in an integrated data storage and maintenance system for the National Geographic Database. This contains the large scale vector data supplied in the OS MasterMap product. This new database has continuous coverage of the entire country at 1:1250 scale in urban areas, 1:2500 in rural areas and 1:10 000 scale in mountain and moorland areas. A considerable amount of data enhancement takes place on initial load of this database, making the data much more suitable for deriving new products.
One of the major breakthroughs has been the use of cartographic text to improve the structuring and attribution of the data. Descriptive text is used to populate the form and function attributes of topographic features. Distinctive names are separated into point features for named buildings and structures, termed Functional Sites, and area features are created for fuzzy geographic features such as hills, valleys and named woodlands, termed Named Extents. Since the source scale is so detailed (ie. very zoomed in), features are often named multiple times in the cartographic text. A Named Extent geometry can then be approximated by using a rounded convex hull surrounding these cartographic text points, as shown in Figure 8.

Historically, skilled Ordnance Survey cartographers have placed the text by hand at each individual scale of mapping. Maintaining separate representations of the name in separate products causes data maintenance and consistency problems. There is now a project underway at Ordnance Survey to create a single Named Places database covering the entirety of Britain. So far this project has captured accurate extent geometries for Settlements, Islands, Sea Areas, Estuaries and Large Geographic Areas.

Having an extent geometry makes it straightforward to automatically decide at which scales to show the name, the appropriate text size for the name, in which tiles to show the name, and the suitable positions for the name. Text placement software can then use the named extents to give an optimal text placement solution.

By default it was decided to use the new Named Places database in preference to the Named Extents, since the Named Places have been captured and verified manually, and are therefore of greater accuracy. However some features existed only in the Named Extents and some features were duplicated between the two datasets. Conflation was required to identify and remove the duplicated features from the Named Extents, allowing both data sources to be used in combination.
<table>
<thead>
<tr>
<th>1:50 000 Scale Name</th>
<th>Nearest Large Scale Name</th>
<th>Match Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>St Loy’s Cove</td>
<td>St Loy’s Cove</td>
<td>0. Exact match</td>
</tr>
<tr>
<td>Porth Ty-mawr</td>
<td>Porth Ty-mawr</td>
<td>1. Accent difference</td>
</tr>
<tr>
<td>Porth Wen Bach</td>
<td>Porth Wen Bach</td>
<td>1. Accent difference</td>
</tr>
<tr>
<td>Nant Las</td>
<td>Nant Las</td>
<td>1. Accent difference</td>
</tr>
<tr>
<td>Ashby de la Zouch Canal</td>
<td>Ashby De La Zouch Canal</td>
<td>2. Punctuation difference</td>
</tr>
<tr>
<td>Caernaff-y-Ganedd</td>
<td>Caernaff-y-Ganedd</td>
<td>2. Punctuation difference</td>
</tr>
<tr>
<td>Ducks’ Pool</td>
<td>Ducks’ Pool</td>
<td>2. Punctuation difference</td>
</tr>
<tr>
<td>Allt Fionn a’ Ghlinne</td>
<td>Allt Fionn a’ Ghlinne</td>
<td>3. Case difference</td>
</tr>
<tr>
<td>Loch Na Doire Duinne</td>
<td>Loch Na Doire Duinne</td>
<td>3. Case difference</td>
</tr>
<tr>
<td>Loch Mhaoilech-coire</td>
<td>Loch Mhaoilech-Coire</td>
<td>3. Case difference</td>
</tr>
<tr>
<td>Fresh Water West</td>
<td>Freshwater West</td>
<td>4. Spacing difference</td>
</tr>
<tr>
<td>Loch a’Mhadaidh</td>
<td>Loch a’Mhadaidh</td>
<td>4. Spacing difference</td>
</tr>
<tr>
<td>Sour Milk Gill</td>
<td>Sour Milk Gill</td>
<td>4. Spacing difference</td>
</tr>
<tr>
<td>Caemeron Bay</td>
<td>Caemeron Bay</td>
<td>5. Likely match</td>
</tr>
<tr>
<td>Chang Burn</td>
<td>Chang Burn</td>
<td>5. Likely match</td>
</tr>
<tr>
<td>Sandy Sike</td>
<td>Sandy Sike</td>
<td>5. Likely match</td>
</tr>
<tr>
<td>Ffynnon Eldd</td>
<td>Afon Eldd</td>
<td>6. Unlikely match: common part</td>
</tr>
<tr>
<td>Llyn Ty’n-y-mynydd Reservoir</td>
<td>Llyn Ty’n-y-mynydd Reservoir</td>
<td>6. Unlikely match: common part</td>
</tr>
<tr>
<td>Ffynnon Caserg</td>
<td>Cwm Caserg</td>
<td>6. Unlikely match: common part</td>
</tr>
<tr>
<td>South Channel</td>
<td>South Cliff</td>
<td>7. Unlikely match</td>
</tr>
<tr>
<td>Dale Pond</td>
<td>Dale Wood</td>
<td>7. Unlikely match</td>
</tr>
<tr>
<td>Alt, Orinig</td>
<td>Alt, Orinig</td>
<td>7. Unlikely match</td>
</tr>
<tr>
<td>Car Fonds</td>
<td>Car Plantation</td>
<td>7. Unlikely match</td>
</tr>
</tbody>
</table>

Figure 9. Automated matching of 1:50 000 scale gazetteer against Named Extents
Ordnance Survey © Crown Copyright. All rights reserved.

A data matching algorithm was used to identify and remove the duplicates between the two sources. For each Named Place, the algorithm searches in radius for all the Named Extents features. A search is first made for an exact match, or a match with minor differences such as capitalisation, punctuation, accents or spacing. If no match is found, the Levenstein distance (the so-called "edit distance") is then used to decide on the best possible match. The Levenstein distance is an integer expressing the minimum number of single character edits to go from one name to another. Some sample results from this can be seen in Figure 9. Further details of this work are given in (Ordnance Survey 2009b).

After the matching we have just the required Named Extents - specifically hills, valleys, woodland, open landcover, rivers, canals, lakes, water inlets, localities, named railways, coastal names, hollows, passes and caves. Some of these are actually linear features (valleys, rivers, canals and railways) and an oriented name is required for the map. For these, new geometries, termed Name Alignments, were created, to allow the text placement software to orient the text along a line.

For the beta release, a new process was developed to create named river and canal alignments, by snapping large-scale text positions to topographic single line streams and the central lines of collapsed topographic water polygons (Figure 10). This improved the text placement by allowing the curved text to more closely follow water features. Name alignments for railways and valleys were created by using a minimum-spanning tree to connect together large-scale text positions representing the same name.
The data volume of the names data was relatively small, so all the processing could take place on the national set of names. The last step was to chop the names into tiles. This process takes account of very small portions of a name geometry overlapping a tile and can suppress such cases. The results of cutting the names into tiles are shown in Figure 11.

There was also a requirement to name isolated buildings in rural areas. The Functional Sites proved a useful source for these. Firstly the named Settlement areas were used to erase all the Functional Sites inside or close proximity to Settlements. This left behind just the isolated rural building names. However there was still some clutter - i.e names too close together, requiring typification. This was achieved by first creating partitions from the main road network. In each a proximity graph was built of the Functional Site locations and split based on distance. The names in each sub-graph were then analysed and the most representative one selected, leaving the rest to be deleted as shown in Figure 12.
RESULTS AND IMPROVEMENTS MADE BETWEEN VERSION ALPHA AND BETA

Following the release of OS VectorMap District alpha, improvements were made to the product, to integrate in the next release (beta). These improvements were decided in response to three main drivers:

• Better meet the user requirements. We have collected user feedback gained from the alpha release and adjusted the product accordingly. One of the main complaints was that the raster was still too busy to work as a good backdrop map. The beta version of the raster product therefore uses more faded colours, and includes less local amenity symbols (for school, hospital etc), although these are still part of the vector product.

• Fix problems that we had identified ourselves with the alpha version, either using the knowledge of the developers of the automatic processes, or by doing visual quality checks of the product.

• Minimise dependencies between products. Many feature types in the alpha version were coming from other existing products. This was the case for coastal lines and areas, wooded areas, roads, the hydrology, administrative boundaries and custom landform (hachures, rocks, etc). For the beta version, we have
updated the process so that most of these features get derived from our large scale database. The only data that still come from another product are the administrative boundaries and the custom landform.

Figure 14 shows the large scale data used to generate the beta version of the product, shown in Figure 15. The automatic derivation of all the topographic features shown in OS VectorMap District for the whole of great Britain took 63 days of processing using a server with a processor Intel® Xeon® X5550 (2.67GHz), with 4GB of RAM and running 1Spatial Radius Clarity 2.7 under Windows XP Professional. We used three identical servers running in parallel to take the processing time down to 21 days.
On the 1st April 2010 ten Ordnance Survey products were released through the OS OpenData website (http://www.ordnancesurvey.co.uk/opendata). These products can be viewed and downloaded for free, or ordered on EcoDisc for a small charge. Some have also been made available in linked data formats through http://data.ordnancesurvey.co.uk/.

One month later, OS VectorMap District was released through the OS OpenData website in ESRI shape file format. This format was chosen to maximise the usability of the product to the general public, who generally do not have access to GML translators.

On 7th May a raster version of the product was released with default Ordnance Survey styling. The pastel colours in the raster version have been deliberately chosen to facilitate the overlay of 3rd party data. The raster version of the product can also be viewed using the OS OpenData viewer (Figure 16): http://www.ordnancesurvey.co.uk/oswebsite/opendata/viewer/.

At the time of writing, the alpha version of OS Vector Map District (both raster and vector) is available online, but this should be replaced by the beta version by the time this paper is published.

Figure 16. OS OpenData Viewer showing an extract of OS VectorMap District

CONCLUSION

Automatic generalisation has proved to offer a cost-effective solution for creating OS VectorMap District. The project has also proved the value of investing in enhancements to the National Geographic Database, without which the creation of OS VectorMap District would have been impossible. The project has highlighted the importance of conducting long-term research, to prepare the organisation for creating new cartographic products on demand. Ordnance Survey Research continues to investigate technologies to support the creation of new products and replacement of legacy production systems with modern automated vector systems.

The development of a single geographic names database has been significantly progressed as a result of the OS VectorMap District project. This database will continue to be enhanced, and as it improves, so too will the quality of the cartographic text placement in automatically derived products. Long-term this will replace the disparate names datasets maintained by Ordnance Survey, reducing data maintenance costs and the new product creation costs.

Conflation is a relatively recent research area for Ordnance Survey, but conflation techniques have already been used in production for matching names datasets and matching rail datasets. Such tools are likely to be used more and more in future, both internally as we seek to reduce duplication between databases, and externally as developers combine and integrate data from disparate public sources and have to resolve issues of duplication and fuzzy matching between datasets.
The experience gained during the creation of OS VectorMap District has been invaluable in exploring a complete automated map production system from start to finish. This knowledge will help accelerate the modernisation of Ordnance Survey’s cartographic production systems, driving down costs and increasing organisational efficiency.

REFERENCES
Revell, P and Antoine, B. (2009) “Automated matching of building features of differing levels of detail”, International Cartographic Conference, Santiago, Chile, November 15-21