The Role of Geography in Automated Generalisation

Mackaness, W.A.[1], Gould, N.M.[2]

[1]School of GeoSciences, The University of Edinburgh
Email: william.mackaness@ed.ac.uk

[2]Geography and Environmental Management, Manchester Metropolitan University
Email: nickgould@live.co.uk

Generalisation of existing topographic data is a well understood problem. Here we explore the role of generalisation in simple thematic queries, as a way of exploring the role of geography in governing design decisions. We argue that the modelling of geographic relationships is fundamental to higher levels of automation and machine reasoning.

1 Generalisation in a thematic context

In general(!), generalisation research has focused on topographic mapping. The data are a given, the ambition is reduction of content over smaller scales. But what if we explore generalisation methodologies from a thematic focus, starting with a blank canvas and assume we have access to distributed open source data? Furthermore, let us imagine that the user wishes to create a map that shows the location of the 10 longest rivers in the world – a map that enables their comparison. There are some delicious solutions to such questions – in some cases much of the geography has been removed in order to greatly facilitate and focus upon, a comparison of length (Figure 1 but also Thomson 1832). Figure 2 provides a more conventional solution and one certainly more easy to automatically produce than Figure 1!

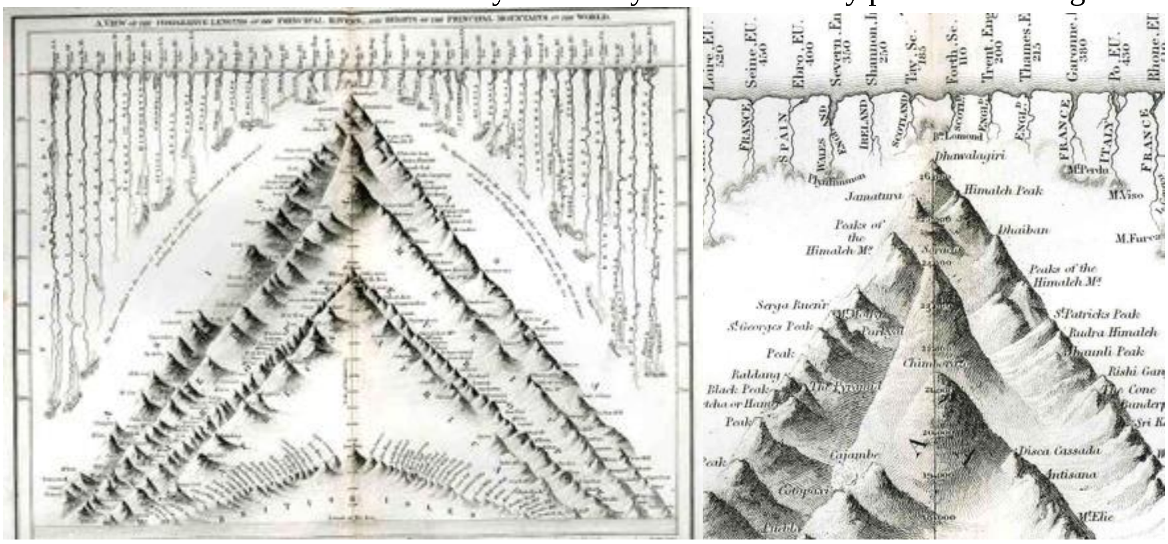

Figure 1: “A View of the Comparative Lengths of the Principal Rivers and Heights of the Principal Mountains in the World”, published by Orr & Smith in London (1836)

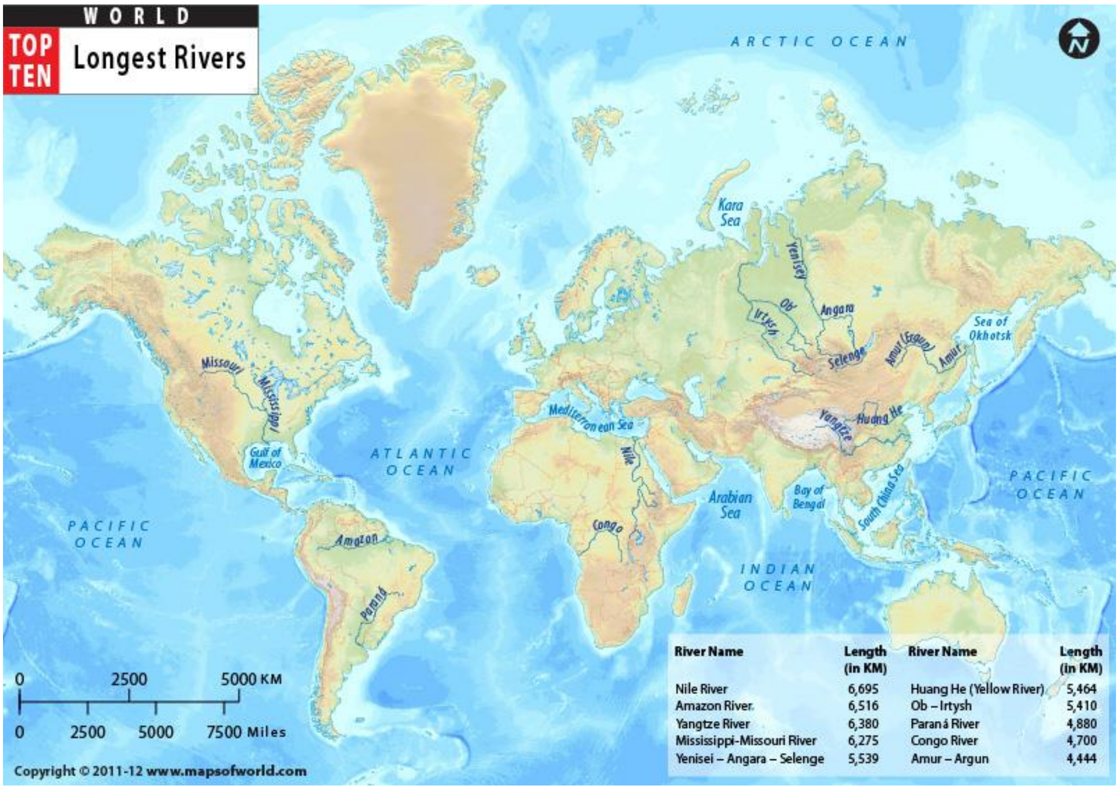

Figure 2: Where are the 10 longest rivers of the world? (http://www.mapsofworld.com/)

2 Some Reverse Engineering

We can learn a lot about how the geography governs generalisation by examining Figure 2. The salient information is the rivers themselves. Both Figure 1 and Figure 2 reveal the close linkage between topography and river course. If we look at Figure 2, we see that the contextualising information includes: outline of the continents, bathymetric data of the oceans, and topographic data revealing the major mountain regions of the world. The oceans are labelled, as are the rivers. The projection is probably Mercator, and basic marginalia reveals scale, the direction of north and a table showing river lengths. Close inspection reveals inclusion of political boundaries.

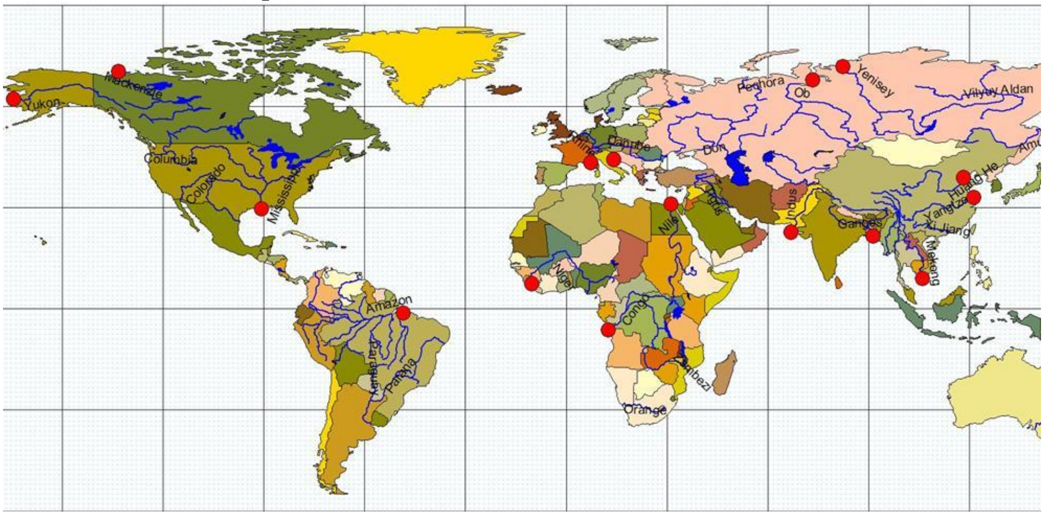

Figure 3: Major deltas of the world overlaid on a political map of the world.

It is worth noting what is not included: the continents are not labelled, no political regions are labelled, and no deltas are shown. It is important to acknowledge that a range of solutions exist, some of which might indeed include additional data such as deltas, or the same information presented in a different manner (Figure 3). We might argue that unlabelled political regions are of limited value, yet if we assume geographical knowledge on the part of the user, most of us can ascertain which countries these rivers border or flow through. We note that the scale/level of detail prevents inclusion of both relief AND political boundaries.

2.1 Scale

From this simple bit of 'reverse engineering' we can start to unpick the cartographic reasoning that governed these designs. From a geographical perspective, we see that such a question (where are the 10 longest rivers of the world?) requires us to view the solution at the global scale. Because the salient data is comprised 10 objects, we can probably get the map on a piece of A5 paper, so the scale is:1:150,000,000. We know that many factors govern the choice of scale, *yet it would seem there are no models that enable the selection of scale based on the type of query*.

The scale will now govern the level of detail for each feature mapped – perhaps salient content deserving of more detail than contextual data. *Despite the importance of this differentiation, we know of no generalisation algorithms in which the degree of application is controlled by whether the data are salient or geographically contextual in nature*.

2.2 Content Measures

It is worth noting that if the task was to try and show the 100 longest rivers of the world, that A5 would not suffice. Collapsing rivers to points, numbered and linked to a list in a table, or colour coding continents according to how many rivers per continent might be possible solutions – the large solution space is what makes automation so challenging. Whichever way you look at it, a method of measuring content is critical to governing the degree of generalisation required, *yet absolute (and even relative) content measuring tools are poorly developed*.

2.3 Salient and Contextualising Information

We know too that topography (and the underlying geology) govern the paths of such rivers, and that rivers drain into the sea. Many of the largest rivers drain into the sea via deltas. Geographical context is fundamental to meaning. Human cartographers know how landscapes, oceans and rivers interact, and can thus select the datasets that will provide context to such thematic questions, *yet in automated cartography, there are no models that enable contextualising content to be selected based on the choice of salient information.* If we had some formal description of the relationships between landscapes, oceans and rivers (Kavouras and Kokla 2008), we could use this as a basis for selecting the content for this task (Table 1).

Salient:

Dataset of the rivers for the whole world with attributes: length, name;
Dataset of major lakes and 'lake systems', with attributes density, size, name;

Contextualising:

Outline of the continents, name;
relief map of the world, height information;
Outline of the seas, name and bathymetric data.

Table 1: A list of content according to whether it was salient or contextual.

The binary of what is salient and what is contextualising is not always distinct. Major lake systems are surely a key aspect of what makes river systems so large. For this reason we might argue that major lakes are salient rather than just contextualising. In any event, we observe that the list of contextualising information has a clear hierarchy in terms of its relative importance. For example, in the contextualising information, we would argue that showing the major deltas is more of a complimenting/contextualising element than showing the bathymetry of the sea. This is because there is a weak relationship between river mouth and bathymetry; only at the large scale do we see such a link (Figure 4). Such an observation reveals 1) the importance of understanding the geographical dimensions of the task, and 2) how the scale/level of detail can affect the relative importance of contextualising information. From this we deduce that we need to explicitly model the effect that scale has on the relative importance of contextualising (and even salient) information, *yet we have neither models that enable us to hierarchically prioritise content, nor methods by which those rankings are adjusted according to changes in scale/ the level of detail.*

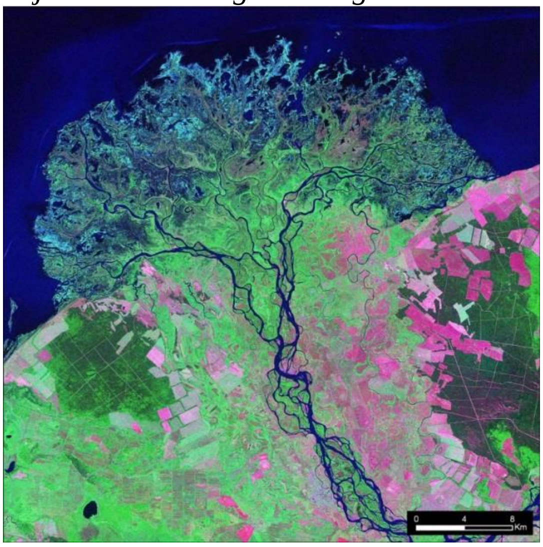

Figure 4: River sediment discharge and its effect on bathymetry: Selenga River delta on the southeast shore of Lake Baikal, Russia. http://soundwaves.usgs.gov/2008/01/SelengaDESLG.jpg

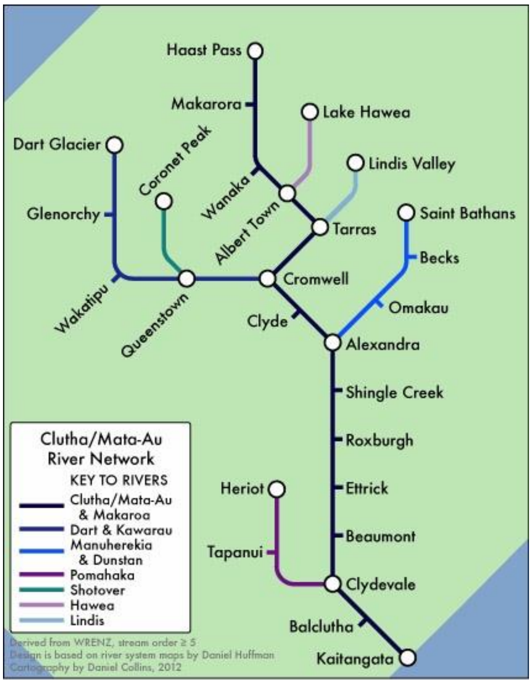

Figure 5: Rivers imagined as a transportation network (in the style of Beck): The rivers of the Clutha network and associated major towns, New Zealand. http://sciblogs.co.nz/

2.4 The Generalisation of this Geographic Information

The scale of the map largely governs the degree of generalisation – in this case the coastline is highly generalised, whilst the rivers mouths remain 'connected' to the coastal boundary. It would not be meaningful if the rivers did not connect directly into the sea (rivers flow after all!). So this is a simple example that illustrates how the relationship between rivers and coastlines imposes constraints on a particular generalisation process. *Some methods do exist that explicitly model such constraints (Gaffuri et al. 2008, for example) but broader, more integrated modelling is required.*

A subset of all rivers has first to be selected (top 10 longest). The rivers have been heavily pruned, then labelled. Applying a Horton ordering (Horton, 1945) is a simple way of identifying branches that need to be pruned such that rivers of order 1 are only shown. The reason for such intense pruning of the network lies in the nature of the question. Interpreting the question in order to set the level of pruning is very difficult to model. This selected ordering is then simplified. Both pruning and simplification require us to understand the geography of the feature. For example, applying Horton ordering to a cyclic network is meaningless; simplifying rivers to straight lines is counter intuitive (though it has to be said, that some lovely solutions exist that do just that – Figure 5). *Though algorithms for pruning and simplification do exist (Stanislawski 2009; Stanislawski and Buttenfield 2011), general models are required that link degree of application to thematic focus and geographic context.*

Major lakes sometimes lie at the head of a river system, and these have been included (Figure 2). As important as they are, there is insufficient room for them to be labelled. Being contextual, perhaps this is a fair compromise. Neither is their room to label the continents, though there is room to label the oceans. Some lake systems sit in isolation, but where a lake is clearly part of a major river, then the river needs to remain connected to that lake during the map generalisation process; a constraint that needs to be modelled.

Deltas (and estuaries) lie 'between' the river and the sea. *There are no generalisation algorithms for handling the generalisation of deltas or estuaries.* One can readily imagine some generalisation solutions to this problem – in any event, during their generalisation, the river still needs to make it to the sea. Humans have a conceptual understanding of what a delta is – a discrete labyrinthine structure connecting a river to the sea. This conceptual understanding of what a delta is, is something the cartographer has taken advantage of when designing Figure 3 – in which the deltas are represented as simple discrete dots. The cartographer has trusted the reader to make the connection between sea/river and delta. A similar dimensional reduction, or collapse, is typically applied to cities at very small scales. *There are no generalisation algorithms able to reduce dimensionality of geographic features in this manner, though 'functional generalisation' attempts some redress (Mackaness and Chaudhry 2011; Chaudhry et al. 2009).*

The contextualising bathymetric information (Figure 2) is somewhat of a 'luxury' but lends some aesthetic to the map; once again there is space enough to accommodate this information without interfering in any way with the river information. This is an interesting observation and it is an interesting question to know how such knowledge could be explicitly modelled such that a system could reason about the availability of space in this manner. In other words, altering any of the ocean information would have no effect on the clarity with which the rivers was conveyed (whereas showing relief does have a direct impact on the visibility of the river data). *We argue that there are no algorithms able to assess whether the generalisation or removal of one feature will increase/decrease the clarity of another.*

The relief is the dominant agency in governing the course of rivers and is therefore very important to the contextualising process, though not as important as the outlines of the continent. In this instance, a simple hypsometric colour tint is sufficient to convey the major mountain regions of the world, though these remain unlabelled and unquantified due to space constraints. *No generalisation algorithms exist that model relief such that it can be represented according to geographic context or thematic intent.*

For consistency the labelling of continents and mountain ranges remain unlabelled even where there is space to do so. The location of marginalia avoids being co-located with river information. *No generalisation algorithms exist that are capable of determining suitable locations for marginalia according to location of salient information.*

2.5 Missing Jigsaw Pieces

During 50 years of research in the field of map generalisation, we have focused on geometric solutions that fail to take into account the geography of the features being generalised. Instead of seeing the map as being made of networks, surfaces and rigid anthropogenic features, we have seen the map as a set of lines, points and areas. In the absence of consideration of the geography, it is not surprising that the algorithms we have created require so much human intervention, and operate over tiny changes in scale. Nyerges (1991) talks of *internal* and *external* meaning. Internal meaning describes properties specific to that entity; external meaning refers to the relationships among other phenomenon. Our focus on primitive representations of geographic phenomenon has led to a focus on the *internal* meaning; our solutions are impoverished because they fail to consider the *external*.

All of these observations point to the urgent and pressing requirement that we model the *geographic* not the *cartographic*. Geometrical and topological modelling is well developed, but the geographic knowledge underpinning a capacity to reason about cartographic design is almost completely absent. There is nothing new in this observation; as Mark (1989; 1991, p104) observed: 'progress in cartographic generalisation will be achieved by attempting to model and generalise real world objects rather than their cartographic representation'. The choice of solutions (even to this simple example) illustrate that 'cartographic generalisation is a particular instance of a semi structured problem that will not reduce to a solution by a lock-step set of deterministic rules or a single algorithm'. So wrote Armstrong in 1991 (p86). From an epistemological perspective, map generalisation should be about revealing different patterns and relationships among geographic phenomena. This then, is to see the problem as being both the generalisation of real world objects, AND the generalisation of the many relationships that exist between such phenomenon (to echo the idea of Nyerges 1991).

3 Modelling the Geography

Ontological modelling (a debased and overworked term) is an interesting path to take since its ambition is to make explicit both the entities AND their relationship with other entities (Kavouras & Kokla 2008; Smith 2003). Through several iterations, we tried to summarise the entities and their relationships in the form of a mind map (Figure 6). It was an interesting process; it was apparent that scale was a determinant in defining various partonomic, topological, and taxonomic relationships. Is a delta part of the sea, or part of the river network? What is the relationship between a town at a confluence and a river network, and how is it modelled? When a river acts as a basis for transporting goods (Figure 5) is this a property of the river network, a river, or a transportation network? Each question reveals the importance of context, map task and use (another area for which no automated solutions exist). In essence, the solution represented in Figure 6 was derived by reverse engineering Figure 2. Figure 6 takes a geographic perspective – making explicit the geographic meaning of things. It seems that what we require are ways of selecting subsets of these entities according to the task, and observing the relationships between these entities and using this information to define algorithmic constraints. In other words, we start with the geographic, and then move to the cartographic.

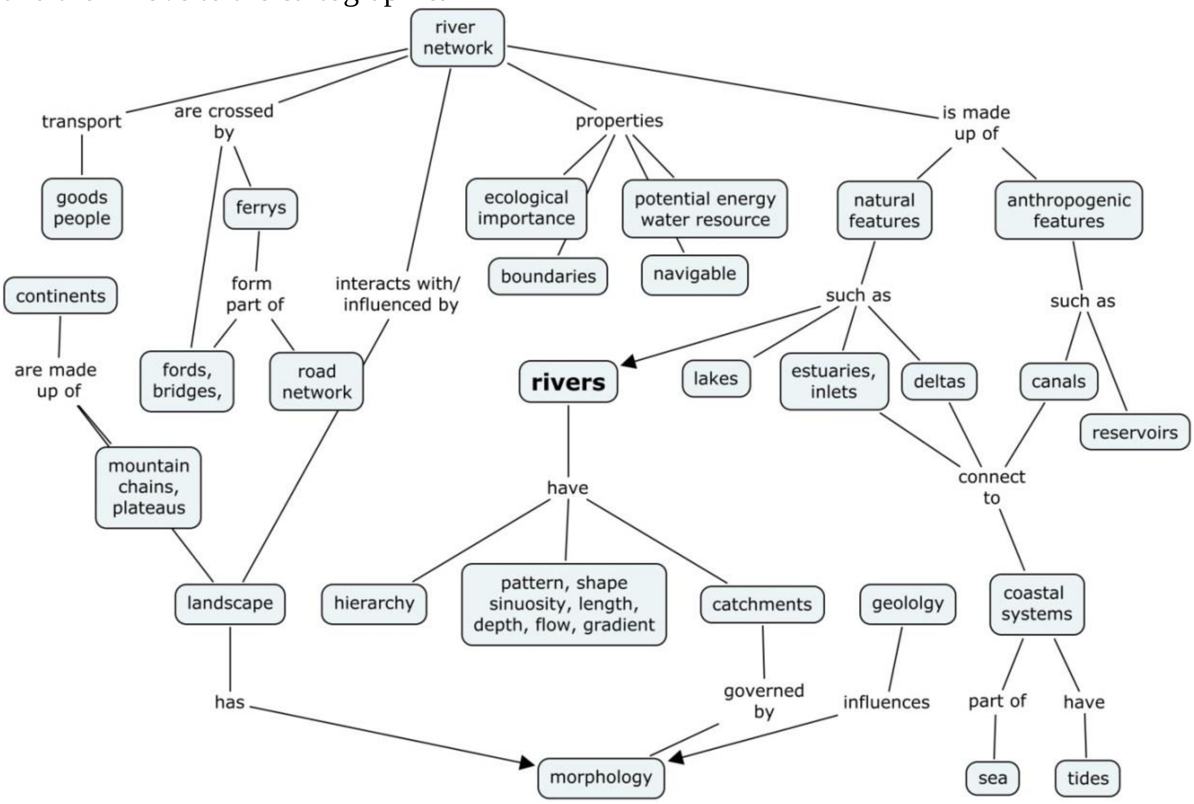

Figure 6: A schematic of the relationship between land, sea, rivers, lakes, continent. Which of these entities and which of these relationships do you wish to cartographically portray?

Thus from Figure 6, in seeking to answer the question set, we need to use scale (room available) to constrain the set of entities that provide context (lakes but not deltas; continents but not roads), and the properties of rivers (curving/ natural) and their connectedness to other entities as constraints – unbroken connection to the sea, and lines that remain 'natural'.

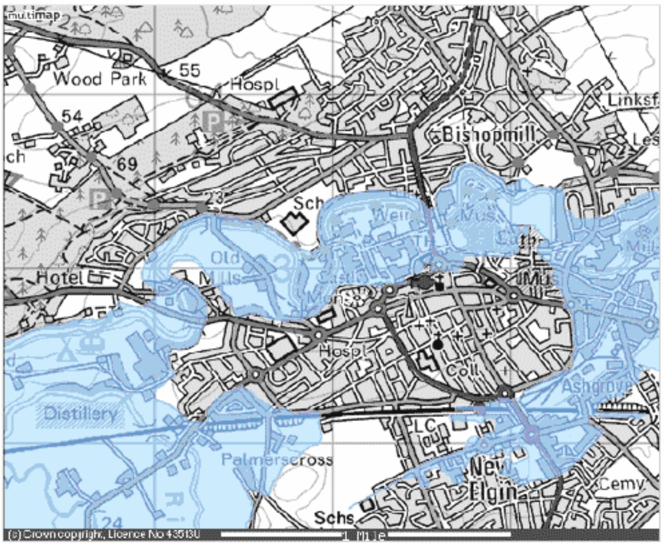

Figure 7: Flood risk map of the town of Elgin, Scotland

We then began to explore a different thematic question, though still related to rivers. This time, we started to think about how we might convey the relationship between rivers and the risk of flooding in towns (Figure 7). Now my mind map took on a very different form – we found ourselves describing a different set of entities, and a different set of relationships (Figure 8).

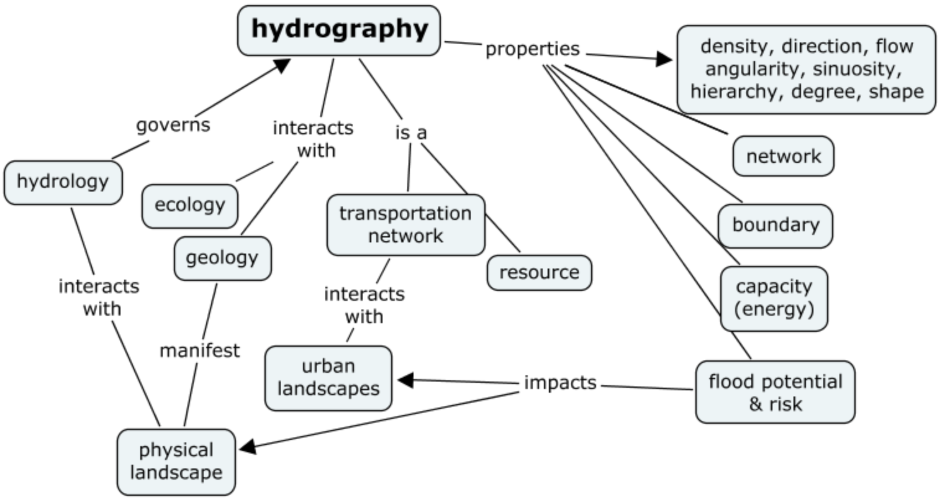

Figure 8: A schematic of the relationship between hydrography, urban and physical landscapes

Looking at these two mind maps (Figure 6 and 8), made me realise that just as in cartography, these semantic models have ‘cusps’ in which groups of entities ‘collapse’ together to form new ‘higher order’ entities. This in turn reveals a different set of external relationships. This idea of modelling the geographic, and thinking about internal and external relations really has purchase!

4 Conclusion

It feels as if our ideas have developed very little since Buttenfield and McMaster (1991). Re-reading this text brought to mind the story of Pooh Bear and Piglet following the path of the Heffalump (Milne, 1926); unknowingly they were going round in circles following their own tracks (Figure 9); I would argue that research in automated generalisation has failed to build upon the ideas of geographic modelling, and has instead opted for simplistic solutions that do not thematically scale.

Figure 9: Pooh Bear and Piglet on the trail of the Heffalump.

A reasoning system is only as effective as the geographical knowledge that informs it. If we are to achieve higher levels of automation and a capacity to reason about design then we must model the geographic. We must ensure that that geography centrally underpins all that we do in automated cartography.

References

ARMSTRONG, M. P. 1991. Knowledge classification and organization. *In*: BUTTENFIELD, B. & MCMASTER, R. B. (eds.) *Map Generalization: Making rules for knowledge representation*. New York: Longman Scientific & Technical.

BUTTENFIELD, B. & MCMASTER, R. B. (eds.) 1991 *Map Generalization: Making rules for knowledge representation*. New York: Longman Scientific & Technical.

CHAUDHRY, O. Z, MACKANESS, W.A. & REGNAULD, N., (2009) A Functional perspective on Map Generalisation, *Computer Environment and Urban Systems: Special Issue on Geo-Information Generalisation and Multiple Representation*, 33 (5): 349-362

GAFFURI, J., DUCHENE, C. & RUAS, A. 2008. Object-field relationships modelling in an agent-based generalisation model. *11th Workshop of the ICA commission on Generalisation and Multiple Representation*. 20-21 June Montpellier, France.

HORTON, R. 1945. Erosional development of streams and their drainage basins; hydrophysical approach to quantitative morphology. *Geological Society of America Bulletin*, 56, pp275-370.

KAVOURAS, M. & KOKLA, M. 2008. *Theories of geographic concepts : ontological approaches to semantic integration,* Boca Raton, CRC Press

MACKANESS, W.A. CHAUDHRY, O.Z 2011 Automated Classification of Retail Spaces from a Large Scale Topographic Database *Transactions in GIS* 15(3): 291-307.

MARK, D. 1989. Conceptual basis for geographic line generalization. *Proceedings of AUTO-CARTO 9, 9th International Symposium on Computer-Assisted Cartography*. Baltimore, Maryland.

MARK, D. 1991. Object Modelling and phenomenon-based generalization. *In*: BUTTENFIELD, B. & MCMASTER, R. B. (eds.) *Map Generalization: Making rules for knowledge representation*. New York: Longman Scientific & Technical.

MILNE, A. A. 1926. *Winnie the Pooh*, Penguin.

NYERGES, T. L. 1991. Representing geographical meaning. *In*: BUTTENFIELD, B. & MCMASTER, R. B. (eds.) *Map Generalization: Making rules for knowledge representation*. New York: Longman Scientific & Technical

SMITH B, 2003 Ontology, in L. Floridi (ed.), Blackwell Guide to the Philosophy of Computing and Information, Oxford: Blackwell: 155–166.

STANISLAWSKI LV (2009) Feature Pruning by Upstream Drainage Area to Support Automated Generalization of the United States National Hydrography Dataset. *Computers, Environment and Urban Systems*, 33: 325-333

STANISLAWSKI LV, BUTTENFIELD BP (2011) Hydrographic Generalization Tailored to Dry Mountainous Regions. *Cartography and Geographic Information Systems* 38(2):117-125

THOMSON, J 1832 Comparison of River Length and Mountain size, Atlas of Scotland.