B03.3

'Spatial indicators' and their value in data cleaning

Tim Jones, Gazetteers and Service Manager, blue8 technologies ltd

'Spatial Indicators and their value in Data Cleaning'

Anyone who has ever been involved in attempting to reconcile a set of addresses from a legacy system with a public-domain reference set such as PAF or AddressPoint will know how difficult this can be if the user has not adhered to a postal format when they entered the address.

Without going into this in too much detail, the structure of the record in AddressPoint has its origins in PAF, and refers to the town where the mail is sorted, rather than its true location.

And, actually, even when you try and compare against something like the National Street Gazetteer and/or the National Land and Property Gazetteer, you could still come up with issues where the perception of the person who entered the data in the legacy system, does not match the structure of the reference record as it exists in the NSG or NLPG

This real-life example demonstrates the problem. It appears in PAF/AddressPoint as….

3, Cromwell Close, Penyffordd, Chester, Cheshire CH4 0GD

But Penyffordd (as its name suggests) isn't in England, it's in North Wales, in the county of Flintshire (which came back into being in the Local Government re-organisations of 1998). It appears like this because the postal sorting town for Penyffordd is Chester, and Chester is undeniably in Cheshire, so now Penyffordd appears to be in Cheshire as well.

Things are complicated by the fact that there are 2 villages with the same name in the county of Flintshire. Postally, these are represented as

A) Penyffordd, Chester

B) Penyffordd, Holywell

Of course, this is not restricted to Wales. Any county will have addresses near its border that will be aligned to postal towns in an adjacent county and vice versa.

Again, using real-data, here are examples of how these locations appear in a legacy data-set.

1) Penyffordd Nr Hope

2) Penyffordd Nr Ffynongroew

The legacy data set does not contain postcodes, and it does not contain co-ordinates. As a result, only local knowledge or reference to a relatively detailed map will reveal whether
1) = A) or 1) = B) and whether 2) = A) or 2) = B)
or whether these are not the same records at all.

Similarly, an automated data-cleaning process will find it almost impossible to distinguish between them and, as a consequence, the usual outcome is that they would be referred back to a user for manual intervention.

Some further examples come from a legacy Command and Control System in Greater Manchester Police.

Legacy = Rugby Street, Lower Broughton, Salford

PAF = Rugby Street,{no locality}, Manchester

And

Legacy = Ridyard Street, Walkden, Salford

PAF = Ridyard Street, Little Hulton, Manchester

Many of us will have heard of 'soundex', 'phonetic' and 'fuzzy' matching, but when neither the Locality nor the Town matches, an automated process or even a human with no local knowledge will find it very difficult to recognise that these are, in fact, one and the same.

Of course, this is not an issue that is restricted to the Emergency Services…it's a universal one which is probably familiar to most Local Authorities, where the address in the Housing system is not the same (even though it was intended to be) as that for the same property in the Benefits system etc etc.

And again, without going into too much detail, the origin of the problem lies in the fact that the legacy addresses were entered without reference to a centralised standard datastore, so people used their local knowledge to a greater or lesser degree (without even realising that they were doing so).

Although people could configure aliases or synonyms at the town and locality level, this requires a large degree of local knowledge and input from staff.

It's also true that products such as QAS-Pro include what are called NPRLs (or **N**on **P**ostally **R**equired **L**ocalities) which can fill in a lot of the gaps but not all of them, whereas this paper describes a different technique, one which assumes nothing and tries to make the most use of the existing legacy data.

The basic premise is that, hidden away in a lot of legacy data, there are what we've christened 'Spatial Indicators', hence the name of this paper, 'Spatial Indicators and their value in Data Cleaning'. These 'Spatial Indicators' are pieces of operational data that only mean something to the owner of the system but, nonetheless, help to position the item in question within the landscape of the organisation and, in so doing, help an automatic process to match it to a public-domain equivalent that may be very different in form.

Let's return to the first example once more. What else is known about the 2 records whose legacy form is

1) Penyffordd	Nr Hope	E03	{other user data}
2) Penyffordd	Nr Ffynongroew	C13	{other user data}

The Legacy data tells us that record 1) falls under the jurisdiction of the Fire Station whose call-sign is E03, whilst record 2) falls under the jurisdiction of the Fire Station whose call-sign is C13.

So what?

This is the crux of the technique, which uses spatial comparison as well as conventional pattern matching, including phonetic encoding akin to Soundex.

If we can understand where these ‘reference objects’ are, and understand their proximity to both the legacy locations and the postal locations, we can resolve automatically the paradox of local area names and postal sorting centres. In other words, the locality cited in the legacy data may never appear in any form in the postal address, and yet an automated process can still identify that the records are one and the same.

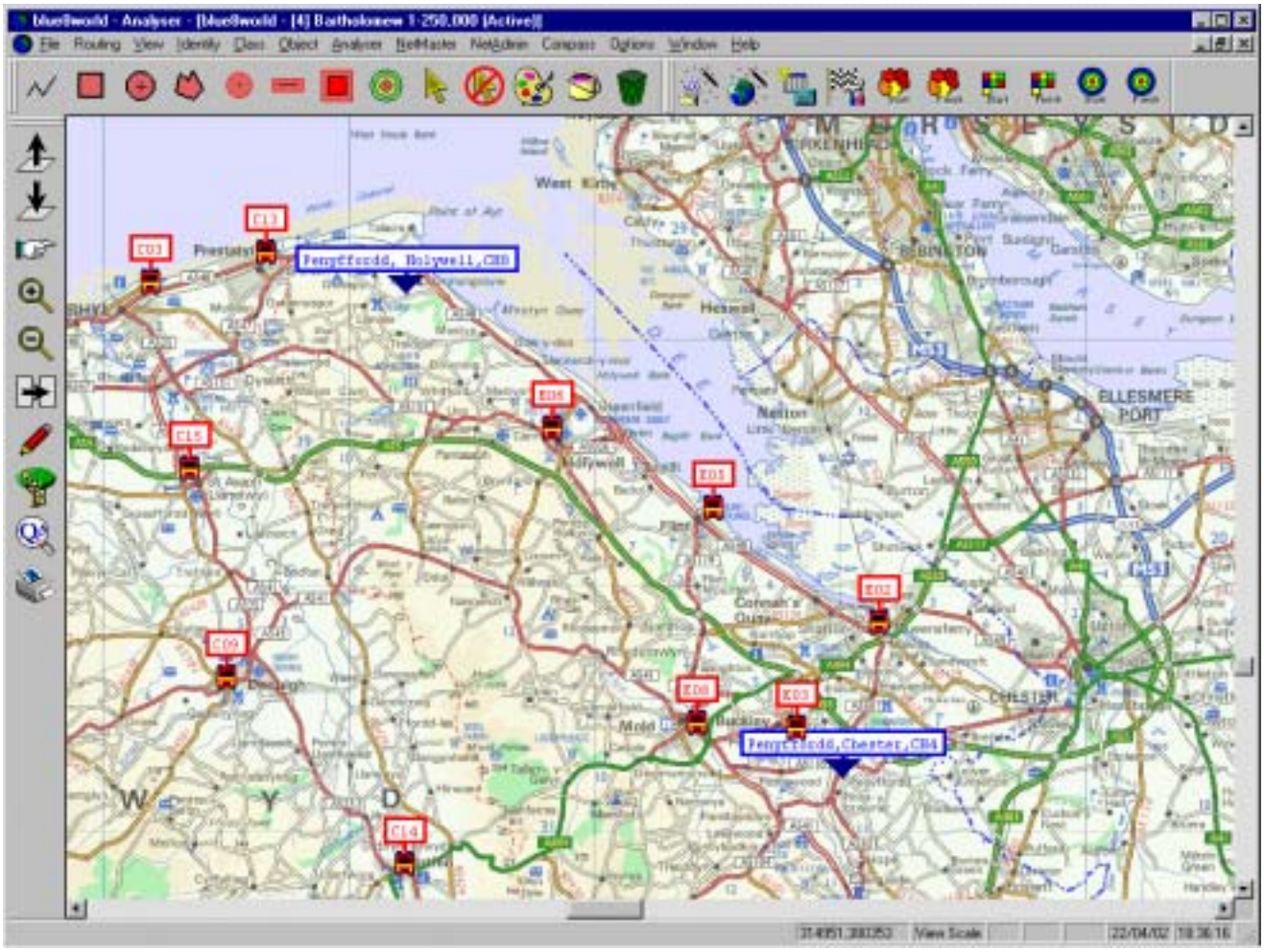
Geographically, the 2 postal records are distributed thus…

If one studies the map closely enough, it is possible to deduce that Penyffordd, Holywell is ‘near’ to Ffynongroyw (note a spelling difference to the Ffynongroew cited in the legacy data) and that Penyffordd, Chester is ‘near’ Hope.

So, for the first time, it is apparent that A = 1 and B = 2.

But this would not be apparent to an automated process, unless it takes the ‘spatial indicators’ into consideration.

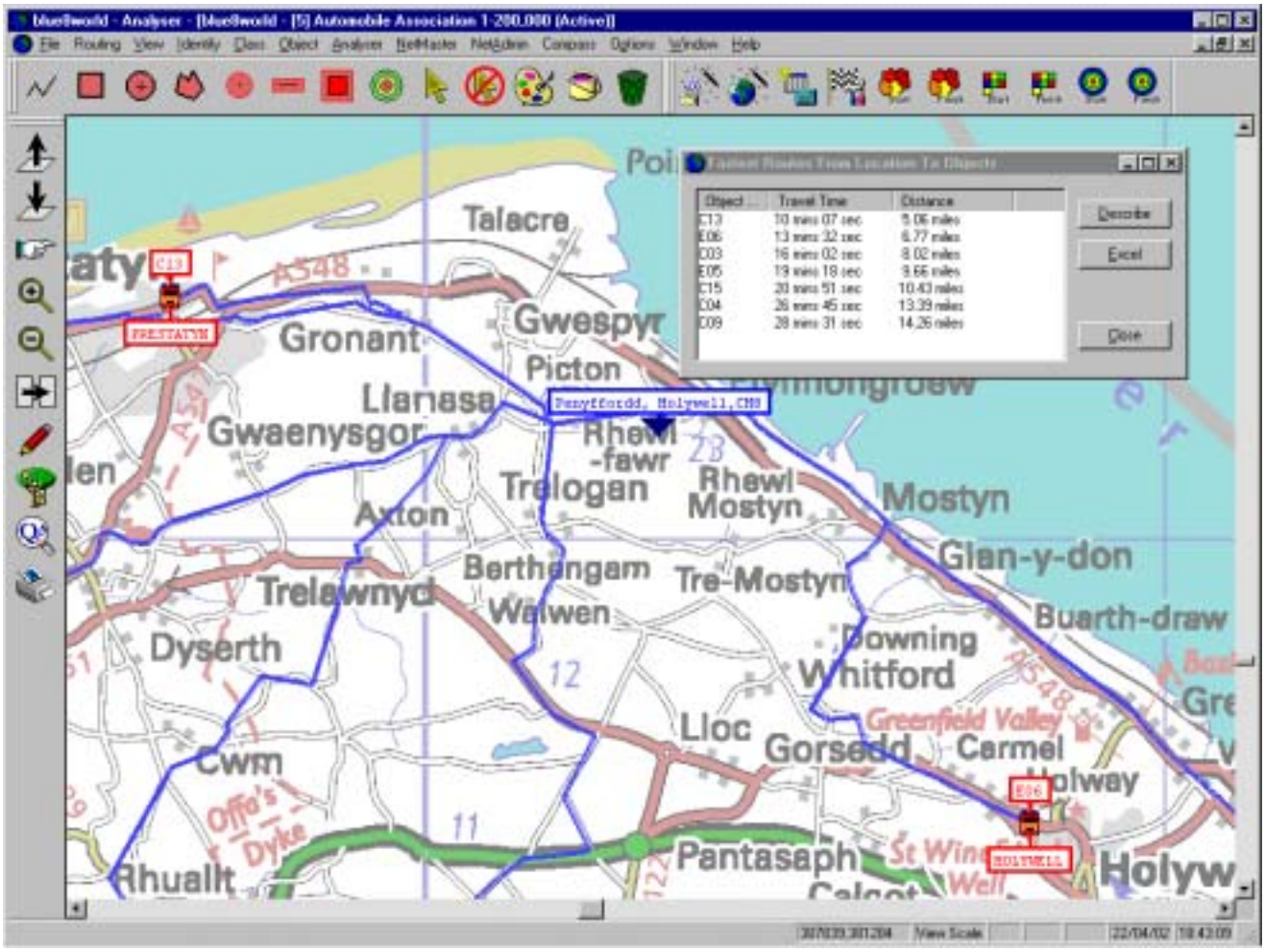
Looking at the map again, but with an overlay of point data representing Fire Stations…



The human eye tells us that the Penyffordd, Holywell record is ‘near’ the Fire Station whose call-sign is C13, whilst the Penyffordd, Chester record is ‘near’ the Fire Station whose call-sign is E03.

This is exactly what the legacy data was telling us, if we understood it properly. The legacy data told us that the record named Penyffordd Nr Hope fell under the jurisdiction of the Fire Station whose call-sign was E03. From the positions of the only 2 public-domain candidates, only one has a position that places it in the correct place in the landscape of the client.

An automated process can attempt to confirm this visual impression by routing from each candidate record to all of the reference objects (Fire Stations) and comparing the nearest to the ‘spatial indicator’ in the legacy data.



And all is revealed. The challenge of the legacy data was to reconcile Penyffordd, Holywell with either Penyffordd Nr Ffynnongroew or Penyffordd Nr Chester, with no local knowledge or manual intervention. The legacy data 'betrayed' the position of Penyffordd, Holywell through reference to an unambiguous object (C13) whose position can be referenced and, in so doing, a join between legacy and public data can be established and maintained.

This used OSCAR to compare position using predicted drive-times (a technique that is particularly useful for a Fire Service client, because it follows the model that they adopted when they created the data, probably at a time before GIS when staff physically drove vehicles along 'timed-runs'). It is particularly useful because it doesn't require the client to have prepared any geographical overlays.

However, where these do exist, then they can be used in a slightly different technique, and to equally good effect.

Returning to the second set of examples, which came from Greater Manchester Police (GMP), the legacy record that needs to be matched is

1) Ridyard Street, Walkden, Salford, Greater Manchester

whilst PAF/AddressPoint contains the following candidates

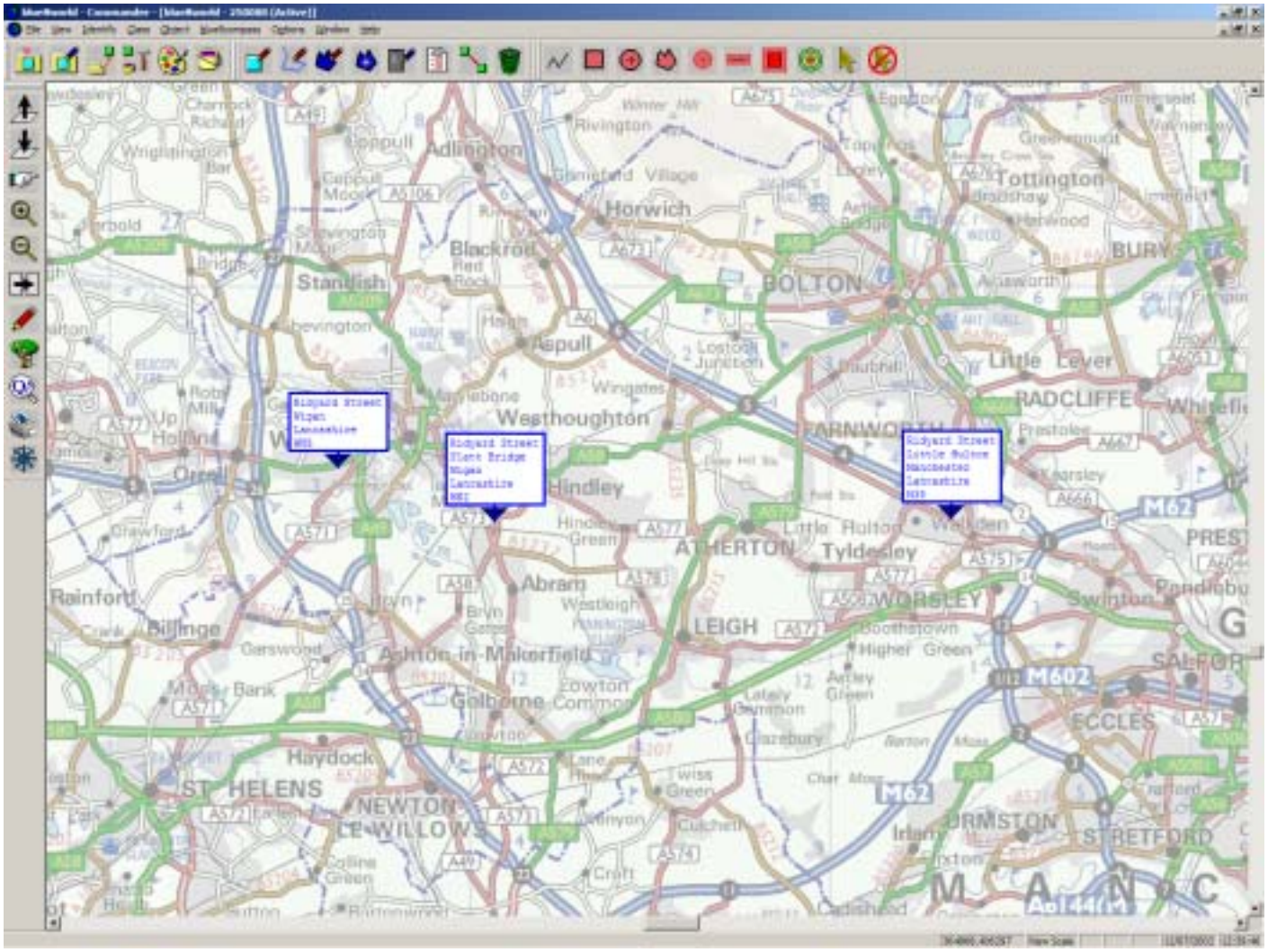
A) Ridyard Street, Little Hulton, Manchester, Lancashire

B) Ridyard Street, Platt Bridge, Wigan, Lancashire

C) Ridyard Street, {no locality}, Wigan, Lancashire

None of the Locality, the Town or the County can be validated and, without local knowledge, it is just not apparent which is the equivalent record.

Once again, the human eye can deduce that, in fact, A) is the most likely candidate, because the map shows us that there is a locality named Walkden in the immediate vicinity, and this is on the Salford side of Manchester . . .



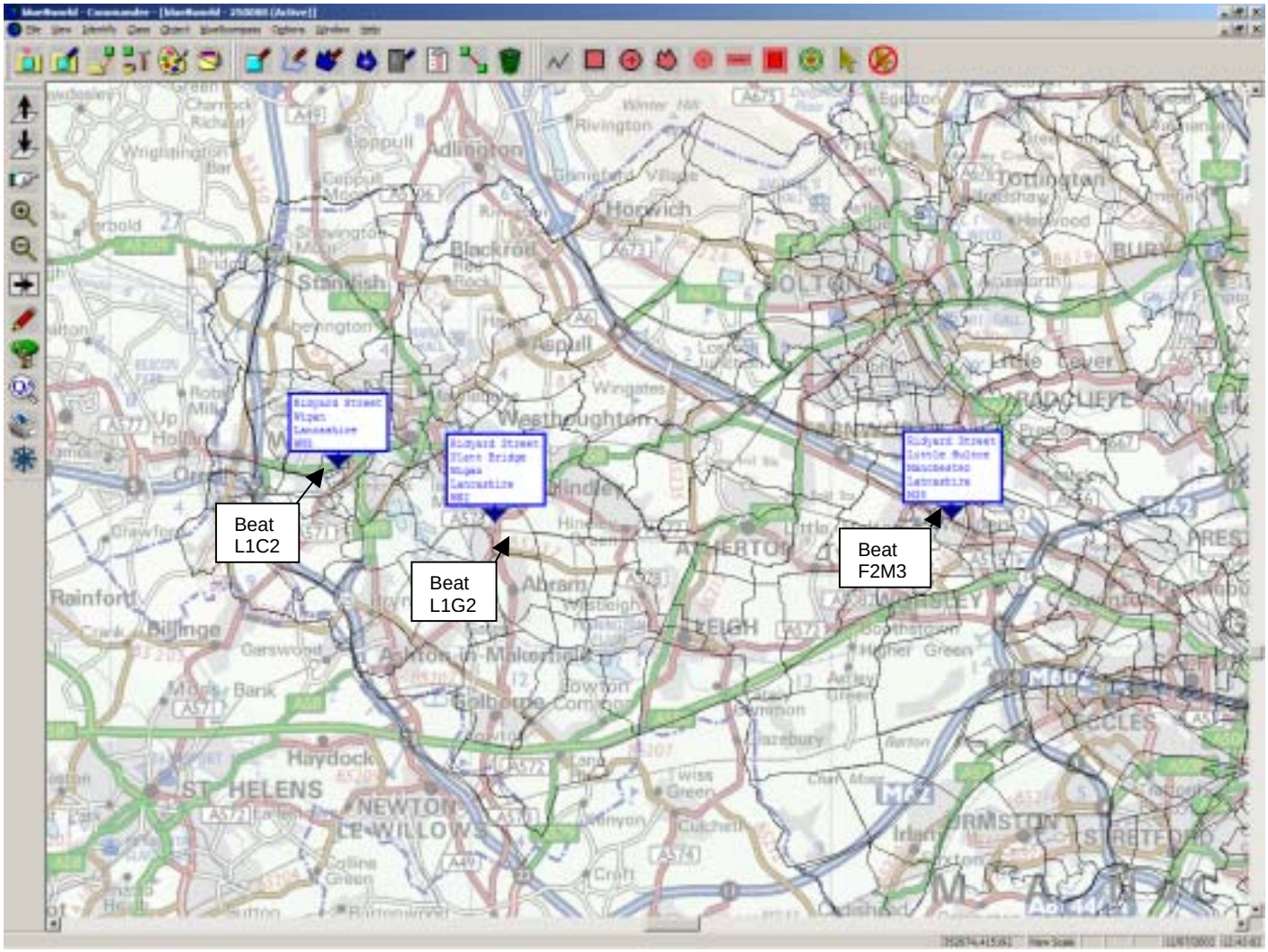
. . . but this would not be apparent to an automated process.

Let's turn to 'Spatial Indicators' once more . . . what pieces of data already exist in the legacy data that can be used as a 'reference object'?

In this example, the record that is

1) Ridyard Street, Walkden, Salford, Greater Manchester

has an attribute that says that it falls within the operational Police 'Beat' whose call-sign is F2M3.



If we activate the polygon dataset that represents the operational 'Beat' areas, then only one of the candidate records has a position that places it within the polygon that represents the Beat area whose call-sign is F2M3…..this is a positive spatial indicator…..what is the likelihood that these 2, and only these 2 records share a critical operational attribute and they are not one and the same?

And, as a result, an automated process can ascertain that

1) Ridyard Street, Walkden, Salford, Greater Manchester

is represented in public-domain datasets as

A) Ridyard Street, Little Hulton, Manchester, Lancashire

…with no local knowledge, no time-consuming (and therefore expensive) configuration of local synonyms and aliases, and no manual intervention.

This technique has been used to excellent effect on a number of recent exercises and has yielded automated hit-rates that were hitherto impossible, given the structure of the data.

Like everything in Location Management, nothing is black and white, and there is no right and no wrong. Why did a client choose to refer to Penyffordd as being 'Near Hope' when Hope is actually much smaller than Penyffordd. Wouldn't it have made more sense to refer to it as being 'near Buckley' (ie the nearest location of a size larger than itself)?

Perhaps it does, perhaps it doesn't, but it's undeniable that it didn't to the client and it is undeniable that data of this type exists throughout the UK, from Scotland to the Isle of Wight and from Anglesey to the Fens.

By understanding more of what the data is trying to tell us, and by gaining spatial intelligence, techniques such as the ones described here can lead to a new generation of spatial-data cleaning products.

The author

Tim Jones is Gazetteer Products and Services Manager for blue 8 technologies ltd, a leading supplier of Location Based Decision Support solutions. Tim qualified from Sheffield University with a B.Sc (Hons) degree in Geography in 1980 and has since worked in real-time Command and Control and GIS applications, with a specific focus on Location Management.

Acknowledgments

Many thanks to North Wales Fire Service and to Greater Manchester Police for the use of their data in the preparation of this paper.