MORE DATA

1. Scenario

Based on the flooding scenario [1] we identify the following requirements:

(1) In case of severe weather conditions, or/and in case these conditions might affect transportation, we need to notify Royal Mail and provide alternative ways for delivering the mail in the affected areas. For example, if Dingwall is flooded an is inaccessible by car, we need to search for alternative ways to reach the city such as aeroplane or ferry (and vice versa).

(2) In case of severe weather conditions, or in case of a disaster, we need to notify the city's Public Authorities and the public for possible vulnerable public access transport nodes that might be or currently are in risk. That means that we need to be aware of public transport nodes in the affected area.

(3) We need to be aware of hospitals, emergency departments, or other units that can provide help in case of emergency. This requirement applies to both the preparation level (preparing for an emergency situation), but also during the disaster (we need to provided immediately information to the person in need).

For these scenarios we need to have access to transportation related datasets. Also we need datasets that contain information about emergency departments. The National Public Transport Access Nodes (hence NPTAN) and NHS provide such datasets.

We will present some results based on the DFO (flood dataset) entry:
country: UK; detailed Locations: Scotland, Highlands, Dingwall, Evanton, Alness, Tain, Caithness, Sutherland, Easter Ross, Moray Firth; began: 26-Oct-2006; ended: 28-Oct-2006; centroid x: -4.18; centroid y: 57,7.

Royal Mail Scenario The NPTAN dataset contains the following labels [2]: atco code, lata code, name, creation date time, modification date time, revision number, modification. The ferry related information is described by the labels: atco code, ferry code, name, name lang, grid type, easting, northing, creation date time, modification date time, revision number, modification.

There are no exact match on the data labels. However, we observe that all the data values under name (NPTAN) are of the form: *Town_name + airport*. For example Bristol Airport. The same format applies for most of the ferries entries: *Town_name+ferry terminal*.

However, we did not find any match on the areas that we are interested in (which is correct as there are not airports or ferry terminals in these areas). Another way to search

[1] In October 2006 3 days heavy rain caused floods in Scotland. It stopped raining in 28 October, and several places in the area of Highlands (Scotland - Highlands area - Dingwall, Evanton, Alness, Tain, Caithness, Sutherland, Easter Ross, Moray Firth) were affected (2299,58 km in total) and one person died

[2] As data labels we define the name of a column. A label defines the type of the entries above it.

for the areas is using the easting and northing of the area. In this case we are facing two problems. First of all we only have the latitude and the longitude of the area and not the easting and the northing. The second problem is that even if we had the easting and northing, we are searching for several areas which cannot be represented by a single pair of coordinates.

Transport Nodes: Coaches We search for coach terminals that might be at risk due to the flood. The coach dataset is described by the following labels: atco code, operator ref, national coach code, name, long name, long name lang, grid type, easting, northing, creation date time, moditfication date time, revision number, modification.

Again, there is no exact match on the labels. However, the data values of label "name" provides similar information as the data values of label detailed locations. We found matches for Evanton, Alness and Tain.

Transport Nodes: Rail Stations The rail dataset is described by the following fields: atco code, triploc code, crs code, station name, station name lang, grid type, easting, northing, creation date time, modification date time, revision number, modification.

The same pattern as in the airport dataset and in the ferry dataset applies here as well: data values of name are of the form *Town_name rail station*. We found matches for Alness, Tain, Dingwall.

Emergency Departments The NHS dataset is described by the following fields: *NHS Board* (the general area that the department belongs to, e.g. Highlands), *Site Type* (whether it offers 24h consultant or not), *Location Names* (hospital name –it could be more than one), *Location Address* (street, city, post code), *File Type* and *Comments*. This dataset contains two kind of emergency departments: those that provide a 24h medicine consultant, usually hospitals, and those that do not provide 24h consultant and usually are minor injury units, small hospitals, health centres *etc*. One problem is the format of this dataset (pdf), however it is possible to copy the data and convert it into the desirable format.

We identify mismatches on the labels between this set and the DFO set. However, we observe that the data values under the labels NHS Board, location name and location address are relevant. NHS Board represents the county of the hospital. In location name the name of the unit is specified, which usually contains information about the location of the unit. For example, Glasgow Royal Infirmary. But this applies only to some of the data entries. Finally, address can be the most representative field of the unit's location. The format that applies to all addresses is *street, city, postcode*, whereas city matches with the data values of detailed locations. We found matches for Dingwall and Caithness.

2. Mismatches: Summary

We have identified, that during an emergency response scenario, failing to obtain information is not only due to mismatches between the labels of the datasets, but also between their values. For example, a query which fails because of the mismatches in the labels is the following:

Fema	declared county/area
SPA	site name
DFO	country, other, nations, detailed locations
SEPA	river basin district, sub basin district
NHS	NHS board, location name, location address
Transport	X

TABLE 1. Different representations of location-related labels

disaster(Location,Disaster_type),
whilst the labels in our dataset are disaster(Area,Type). An example of a query failing to obtain results because of mismatches in the data values, is the following:

disaster(scotland, Disaster_type).
while in our dataset we have disaster(united kingdom, flood). Below we present the mismatches that we have encountered so far, regarding both levels of data,.

Location As shown in table 1, the ways the location-based labels differ are:

- Datasets do not use the same number of columns, i.e. location is broken down into sub-bales, each representing a different abstraction of a location. For example declared county/area Vs. country, other, nations, detailed location.
- Some labels are more domain specific than the the others. For example, site name Vs river basin district.

Furthermore, the mismatches that occur in the data-values fall into one of the following categories:

- More than one area Vs One area. For example Scotland-Highlands, Dingwall,.. and Dingwall
- Part of relation. For example, Scotland and United Kingdom, Caithness Lochs and Caithness, 84 Castle Street Glasgow G405F and Glasgow.
- Belongs to relation. For example Glasgow Royal Infirmary and Glasgow.
- One label for different meanings. For example, under the label declared county/area other entries contain cities, while other countries.
- Different structure. For example, all the declared county/area entries are of the form Name(County), e.g. Scotland(County).

Date In table 2 we present the mismatches on the date-related labels. As we can see, both date labes (FEMA, DFO) are really close semantically. Apart form the label mismatch, we further categorise the date-related data value mismatches into:

- Different representation of the same meaning. For example May and 05, '12 and 2012.
- Different structure. For example, 12/05/1953 and 05/12/1953.

Coordinates In table 3 we present the different labels the datasets use to describe coordinates. The mismatch between them is either that they use different representation for the same meaning, or that they use different coordination system. However, from

Fema	incident begin date, incident end date, disaster close out date
SPA	X
DFO	began, ended
SEPA	X
NHS	X
Transport	X

TABLE 2. Different representations of date-related labels

Fema	X
SPA	X
DFO	centroid-x, centroid-y
SEPA	lat, long
NHS	X
Transport	easting, northing

TABLE 3. Different representations of coordinate-related labels

the data-value level, more problems arose: the possibility of finding an exact match of coordinates is very low, as is it is very rare two point to share the same coordinates (unless these points are the same).

Missing labels/data Some datasets lack of labels that others provide. For example, the transport dataset does not provide any location-related information. However, there exist some patterns that can provide the missing information. For example, consider name(aberdeen airport) (transport dataset). By splitting the argument, we have name(aberdeen, airport). If we generalise it, name(Name, airport) then name can match with the detailed location (DFO) data value, declared county/area etc.