Geographic Reference Datasets Memo

TIN 3.165 Big Data Experimentation to Improve Understanding and Decision Making

THE INVESTIGATION WHICH IS THE SUBJECT OF THIS REPORT WAS INITIATED BY THE PROGRAMME AND DELIVERY DIRECTORATE DSTL AND WAS CARRIED OUT UNDER TERMS OF CONTRACT NO. DSTLX-1000095527.

Reference..... O-DHCSTC_I395527_I_T3_165/006

Version..... 2

Date ..... 27 April 2015

Customer Information	
[redacted]	
Domain or Dept:	Dstl
Milestone Number:	O-DHCSTC_I395527_I_T3_165/006
Contract Number:	DSTLX-1000095527
Project Start Date:	21/01/2015
Project End Date:	31/03/2015
Due Date:	30/03/2015

Authors	Organisation	Tel	email
[redacted]			

Approvals	Organisation	Email	Date
[redacted]			

Amendments			
Version	Amendment date	Amended by	Remarks
1			Original issue
[redacted]			

Distribution		
Name	Role	Organisation
[redacted]		

Contents

Tables

Acronyms & Abbreviations

Acronym	Definition
ASCII	American Standard Code For Information Interchange
CIA	Central Intelligence Agency
CUDE	CSV Universal Data Exchange for Geospatial Data
DBMS	Database Management System
DHCSTC	Defence Human Capability Science & Technology Centre
ESRI	Environmental Systems Research Institute
FBI	Federal Bureau of Investigation
GIS	Geographic Information System
GNIS	Geographic Names Information System
GPS	Global Positioning System
GTOPO30	Global 30 Arc-Second Elevation
HTTP	Hypertext Transfer Protocol
NASA	National Aeronautics and Space Administration
NGA	National Geospatial-Intelligence Agency
NOAA	National Oceanic and Atmospheric Administration
NRCan	Natural Resources Canada
OS	Ordnance Survey
OSM	OpenStreetMap
POI	Points of Interest
TIGER	Topologically Integrated Geographic Encoding and Referencing
UN	United Nations
WGS84	World Geodetic System 1984
WOEID	Where On Earth Identifier

1 Introduction

Table 1-1 : DHCSTC Tin 3.165 Report Deliverables

Tin 3.165 Report Deliverable List
Geographic Datasets Memo – This Document

1.1 Geocoding Overview

Geocoding is the process of converting addresses or locations into geographic coordinates (e.g. *latitude 50.999707 and longitude -1.519818*) from spatial reference data (i.e. building polygons, land parcels, street addresses, postal codes), which can be used to place markers or position a map. Geocoding facilitates spatial analysis using Geographic Information Systems (GIS), allowing the data to be indexed and stored, which can be used to perform geospatial queries.

2 Available Reference Datasets

This section provides a brief overview of the most accessible reference datasets for use in geocoding. The following table (Table 2-1) summarises the reference datasets that are described in more details in the subsequent sections.

Table 2-1: Summary of Reference Datasets

Name	Licence	Coverage	Type
Geopostcodes (*section 2.1*)	Commercial	Worldwide	Offline, CSV file
Geonames (*section 2.2*)	Open Source	Worldwide	Offline, CSV file
OpenStreetMap (*section 2.3*)	Open Source	Worldwide	Offline, CSV file
OpenCage (*section 2.4*)	Commercial	Worldwide	Offline, CSV file
OpenData (*section 2.5*)	Commercial	UK	Offline, GML/CSV file
ESRI World Geocoding (*section 2.6*)	Commercial	Worldwide	Online web service
Yahoo! GeoPlanet (*section 2.7*)	Open Source	Worldwide	Online web Service [1]
OpenGeocode (*section 0*)	Open Source	Worldwide	Offline, CSV file

2.1 Geopostcodes

Geopostcodes (4) provides a common dataset structure for all countries, containing all localities, postal codes, administrative divisions, statistical units, reference codes, time zones, elevations and, for selected countries, neighbourhoods, suburbs and streets. All data are georeferenced and available in local language, transliterated English and non-accented ASCII versions. In addition, each record within the dataset is georeferenced using standard World Geodetic System 1984 (WGS84) datum.

The data is provided in a consistent structure in raw Comma Separated Values (CSV) and GIS formats that are can be imported into many database management systems (DBMS).

Geopostcodes has worldwide coverage, which is summarised as follows:

- 250+ countries & territories.
- 150000+ regions.
- 5000000+ places and postcodes.
- 12500000+ streets.

Coverage is most comprehensive in Europe, and areas of North America, Africa and South East Asia. Central Asia, Oceania and South America have limited coverage.

2.2 Geonames

The GeoNames (5) database contains over 10,000,000 geographical names corresponding to over 7,500,000 unique features. A key concept used within Geonames is "Feature Codes", which categorise locations into a series of distinct classifications. All features are categorised into one out of nine feature classes and further subcategorised into one out of 645 feature codes. Beyond names of places in various languages, data stored include latitude, longitude, elevation, population, administrative subdivision and postal codes. All data is georeferenced using the WGS84 datum.

Geonames utilises data from a variety of different sources. The main data sources are identified as follows:

- National Geospatial-Intelligence Agency's (NGA) and the U.S. Board on Geographic Names (6).
- U.S. Geological Survey Geographic Names Information System (GNIS) (7).

[1] GeoPlanet provides downloadable data, but hasn't been updated since 1 June 2012 [26]

- Ordnance Survey OpenData (8).
  - Gazetteer: Contains Ordnance Survey data and public sector information (9).
  - Postal codes: Contains Royal Mail data.
- GeoBase (10).
- Global 30 Arc-Second Elevation (GTOPO30) (11).

The Feature Codes functionality has the potential to improve geospatial searching by providing an extra level of granularity to filter likely locations from a potential list of results. For example, allowing the user to specify the location is an Airport or Railway Station.

2.3 OpenStreetMap

OpenStreetMap (OSM) (12) is a collaborative project built by a community of mappers that contribute and maintain data. OSM emphasises local knowledge with contributors using aerial imagery, GPS devices, and field maps to verify that OSM is accurate and up to date. All data is georeferenced using the WGS84 datum.

OSM data quality and coverage differs between regions. For example, many European cities are covered to a high level of detail. Due to the collaborative nature, OSM data tends to benefit from locations being georeferenced before alternative systems. For example, OSM is usually the first to have a new housing development or a new motorway exit mapped. However in some, mostly rural areas, there may be limited data in the database.

In addition to obtaining data from the user community, some government agencies have released official data on appropriate licenses. The following summarises the key government data contributions to OSM:

- Landsat 7 satellite imagery from National Aeronautics and Space Administration (NASA) (13).
- Prototype Global Shorelines Prototype Global Shoreline (14) from National Oceanic and Atmospheric Administration (NOAA).
- Topologically Integrated Geographic Encoding and Referencing (TIGER) (15) from the US Census Bureau.
- Ordnance Survey OpenData (8).
- CanVec (16) vector data from Natural Resources Canada (NRCan).
- GeoBase (10) provides land-cover and streets.

2.4 OpenCage

OpenCage (17) provides a commercial offering to OSM by offering bespoke georeferenced data and formats. OpenCage extracts required data from OSM and formats it as required. OpenCage also provides an interface that combines multiple geocoding systems in the background, each optimised for different parts of the world and types of requests.

2.5 OpenData

Ordnance Survey (OS) OpenData (18) provides detailed gazetteer data, covering the United Kingdom. It is commonly combined with other open datasets available from a variety of sources and is used as a data source in several georeferenced datasets, for example, OSM (12) and Geonames (5). OpenData is composed of several products and gazetteers, the most relevant products are described below:

- OS Locator: A fully searchable point-based national gazetteer of road names. Specific locations can be found and identified by a number of criteria, including locality, settlement, local authority and county.
- 1:50 000 Scale Gazetteer: The 1:50 000 Scale Gazetteer is a reference tool or location finder, similar to the index in a road atlas. It can be used as a simple list to discover relevant coordinates and six-figure grid references for a town or area.
- OS Open Rivers: A connected river network for Great Britain which has been derived from OS large scale data. It shows the flow and the locations of rivers, streams, lakes and canals, across Great

Britain, providing a structured and attributed network for sharing information and simple analysis of the river network.

- OS Terrain 50: OS Terrain 50 is a new height product for Great Britain. It is supplied both as a set of 50m gridded digital terrain model (OS Terrain 50 grid) and 10m contours and spot heights (OS Terrain 50 contours).

OpenData sources its data from a combination of OS street level data and external data from other government and non-government agencies. Some of the key data sources are listed below:

- Land Registry Open Data (19).
- Environment Agency Datashare (20).
- Office for National Statistics Open Geography (21) .
- Guardian Datastore (22).

2.6 Environmental Systems Research Institute (ESRI) World Geocoding

The World Geocoding Service (23) finds addresses and places in all supported countries around the world. The service can find point locations of addresses, cities, landmarks, business names, and other places. The output points can be visualised on a map, inserted as stops for a route, or loaded as input for a spatial analysis. World Geocoding provides extensive world-wide coverage for addresses, cities, landmarks, business names, and many other points of interest (POI).

The service is available as both a geosearch and geocoding service:

- Geosearch Services: Locate a feature or POI and then have the map zoom to that location. The result might be displayed on the map, but the result is not stored in any way for later use.
- Geocoding Services: Convert an address to an x,y coordinate and append the result to an existing record in a database.

World Geocoding is provided as an online web service. The Geosearch service can be accessed free of charge, however, the Geocoding service requires an annual subscription. Due to the nature in which this data is provided, World Geocoding is not suitable for use in closed systems without external Internet access.

2.7 Yahoo! GeoPlanet

Yahoo! GeoPlanet (24) is a platform for coordinating geographic information. It provides a complete geolocation infrastructure for search engines, portals and Web sites. An integral part of GeoPlanet is the "Where On Earth Identifier" (WOEID) (25), which is a unique 32-bit reference identifier, that identifies any feature on Earth. In addition to the strict numerical WOEID, GeoPlanet also has a hierarchical structure that allows accessing surrounding locations, and zooming up and down administrative divisions.

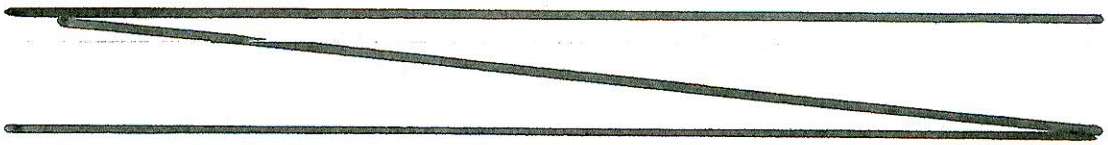
GeoPlanet provides information for approximately six million named places globally. Coverage varies from country to country, but includes several hundred thousand unique administrative areas with half a million variant names; several thousand historical administrative areas; over two million unique settlements and suburbs, and millions of unique postal codes covering approximately 150 countries. It also includes a significant number of POIs, Colloquial Regions, Airports, Area Codes, and Time Zones.

Yahoo! released GeoPlanet's WOEID data to the public for external download (26), however, the last downloadable release was on 1st June 2012. The WOEID data is now hosted online via the Yahoo! GeoPlanet web service. For this reason, GeoPlanet may not be suitable for use in closed systems without external Internet access. However, the downloadable data may be useful when combined with other data sources to enhance the georeferenced datasets.

2.8 OpenGeoCode

OpenGeoCode (27) is an open-data project providing datasets free-of-charge to the developer community. All datasets are compiled and aggregated from public domain sources from a wide variety of agencies, including; the United Nations (UN), National Oceanic and Atmospheric Administration (NOAA), U.S. Department of Agriculture (USDA), U.S. Geological Survey (USGS), Federal Bureau of Investigation (FBI) and Central Intelligence Agency (CIA). The datasets can be obtained from a central CSV Universal Data Exchange for Geospatial Data (CUDE) (28), which is updated regularly. All data is georeferenced using the WGS84 datum.

3 Bibliography



4. Geopostcodes. [Online] http://www.geopostcodes.com.

5. Geonames. [Online] http://www.geonames.org.

6. National Geospatial-Intelligence Agency. [Online] http://geonames.nga.mil/gns/html.

7. US Geological Survey. [Online] http://geonames.usgs.gov/index.html.

8. Ordnance Survey. [Online] http://www.ordnancesurvey.co.uk/business-and-government/products/opendata-products.html.

9. National Archives. [Online] http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/.

10. Geobase. [Online] http://www.geobase.ca/.

11. GTOPO30. [Online] https://lta.cr.usgs.gov/GTOPO30.

12. OpenStreetMap. [Online] http://www.openstreetmap.org.

13. Landsat 7. [Online] http://geo.arc.nasa.gov/sge/landsat/l7.html.

14. Prototype Global Shoreline Data. [Online] http://shoreline.noaa.gov/data/datasheets/pgs.html.

15. TIGER. [Online] https://www.census.gov/geo/maps-data/data/tiger.html.

16. CanVec. [Online] http://geogratis.gc.ca/api/en/nrcan-rncan/ess-sst/-/%28urn:iso:series%29canvec.

17. OpenCage. [Online] http://geocoder.opencagedata.com.

18. OpenData. [Online] http://www.ordnancesurvey.co.uk/business-and-government/products/opendata-products.html.

19. Land Registry. [Online] http://landregistry.data.gov.uk.

20. Environment Agency. [Online] http://www.geostore.com/environment-agency.

21. ONS. [Online] https://geoportal.statistics.gov.uk/geoportal/catalog/main/home.page.

22. Guardian Datastore. [Online] http://www.theguardian.com/data.

23. ESRI. [Online] http://www.esri.com.

24. Yahoo! GeoPlanet. [Online] https://developer.yahoo.com/geo/geoplanet/data.

25. WOEID. [Online] https://developer.yahoo.com/geo/geoplanet/guide/concepts.html.

26. GeoPlanet Data. [Online] https://developer.yahoo.com/geo/geoplanet/data.

27. OpenGeocode. [Online] http://opengeocode.org.

28. CUDE Dataset. [Online] http://opengeocode.org/cude1.2.

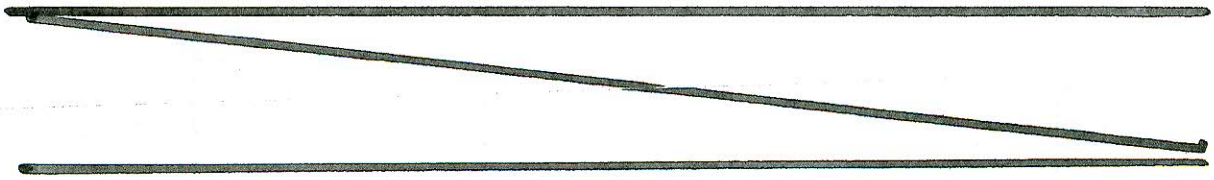
-End of Document-

Subject to Protective Marking, National Caveats and announcement or availability restrictions, the information provided in this form will be published outside UK government and should, where possible, be of a commercially non-sensitive nature.

To aid completion notes can be found at the end of this form.

1a.	**Report number:**	**O-DHCSTC_I395527_I_T3_165/006**	1b.	**Version number:2**
2.	**Date of publication:** 27/04/2015		3.	**Number of pages:iv+6**
4a.	**Report UK protective marking:**	OFFICIAL		
4b.	**Report national caveats: N/A**	N/A		
4c.	**Report descriptor: N/A**	N/A		
5a.	**Title:**	**DHCSTC Tin 3.136 Big Data Experimentation to Improve Understanding and Decision Making Geographic Reference Datasets Memo**		
5b.	**Title UK protective marking:**	OFFICIAL		
5c.	**Title national caveats: N/A**	N/A		
5d.	**Title descriptor: N/A**	N/A		
6a.	**Alternate title:**	N/A		
6b.	**Alternate title UK protective marking:**	N/A		
6c.	**Alternate title national caveats: N/A**	**N/A**		
6d.	**Alternate title descriptor: N/A**	**N/A**		
7.	**Authors:**	[redacted]		
8.	**Name and address of publisher:**	[redacted]	9.	**Name and address of sponsor:** [redacted]
10.	**Sponsor contract:**	DSTLX-1000095527		
11.	**Dstl project number:**	TIN 3.165		
12.	**Work package number:**	N/A		
13.	**Other report numbers:**	N/A		
14a.	**Contract start date:**	**21/01/2015**	14b.	**Contract end date:** **31/03/2015**
15a.	**IPR:**	705 FULL		
15b.	**Patents:**	No		
15c.	**Application number:**	N/A		

16a. Abstract:



16b.	**Abstract UK protective marking:**	OFFICIAL
16c.	**Abstract national caveats:**	N/A
16d.	**Abstract descriptor:**	N/A

17. Keywords:

Big Data, Natural Language Processing, Geo-coding, Intelligence, Text, Hadoop, Gazeteer, Geo-mapping

18. Report announcement and availability

	Announce to?	Available to?	
18a.	☒	☒	UK MOD has unlimited distribution rights
18b.	☐	☐	UK MOD has no rights of distribution
18c.	☐	☐	Can be distributed to UK MOD and its agencies
18d.	☐	☐	Can be distributed to all UK government departments
18e.	☐	☐	Can be distributed to all UK defence contractors
18f.	☐	☐	Can be distributed to all foreign government departments

18g. Additional announcement:

18h. Additional availability:

18i. Release authority role: