International Journal of Digital Earth

ISSN: 1753-8947 (Print) 1753-8955 (Online) Journal homepage: http://www.tandfonline.com/loi/tjde20

Wāhi, a discrete global grid gazetteer built using linked open data

Benjamin Adams

To cite this article: Benjamin Adams (2016): Wāhi, a discrete global grid gazetteer built using linked open data, International Journal of Digital Earth, DOI: 10.1080/17538947.2016.1229819

To link to this article: http://dx.doi.org/10.1080/17538947.2016.1229819

INTERNATIONAL JOURNAL OF DIGITAL EARTH, 2016
http://dx.doi.org/10.1080/17538947.2016.1229819

Wāhi, a discrete global grid gazetteer built using linked open data

Benjamin Adams

Department of Computer Science, Centre for eResearch, The University of Auckland, Auckland, New Zealand

ABSTRACT

Discrete global grid systems have become an important component of Digital Earth systems. However, previously there has not existed an easy way to map between named places (toponyms) and the cells of a discrete global grid system. The lack of such a tool has limited the opportunities to synthesize social place-based data with the more standard Earth and environmental science data currently being analyzed in Digital Earth applications. This paper introduces Wāhi, the first gazetteer to map entities from the GeoNames database to multiple discrete global grid systems. A gazetteer service is presented that exposes the grid system and the associated gazetteer data as Linked Data. A set of use cases for the discrete global grid gazetteer is discussed.

ARTICLE HISTORY
Received 15 June 2016
Accepted 24 August 2016

KEYWORDS
Discrete global grid; gazetteer; linked data; place; toponym; Digital Earth

1. Introduction

One of the mainstays of the Digital Earth vision has been the importance of using discrete global grids as a framework to represent spatial information (Goodchild 2000; Goodchild et al. 2012). A geodesic discrete global grid is a partitioning of the surface of the globe into a grid. An indexed hierarchy of grids at multiple interrelated levels of resolution is called a discrete global grid system. Discrete global grids are preferred over other projected vector and raster representations of spatial information for many geo-statistical operations, and they provide a common spatial framework to combine data from multiple, heterogeneous sources and thus support data interoperability (Sahr, White, and Kimerling 2003; Nguyen, Cressie, and Braverman 2012; Tong et al. 2013). Progress has been made to develop a standard for discrete global grid systems, and open-source and commercial software systems to work with discrete global grids are advancing (Vince and Zheng 2009; Amiri, Samavati, and Peterson 2015; Sahr 2015; Gibb, Purss, and Samavati 2016). In this paper, we introduce Wāhi (http://dg3.cer.auckland.ac.nz), the first open discrete global grid gazetteer that maps toponyms (named places) to cells on a discrete global grid. The name Wāhi comes from the Māori noun for *place* and the verb meaning *to partition, subdivide*.

To date, most of the prominent research applications of discrete global grids relate to landcover, remote sensing, and other geoscience data sets (Kerr et al. 2001; Schipper et al. 2008; Strassburg et al. 2010; Dulvy et al. 2014). Typically, these applications do not refer to regions of the Earth that are socially defined, such as countries, administrative regions, cities, neighbourhoods, national parks, and marine reserve areas. When we compare this to Gore's original vision of Digital Earth, with its references to 'Yellowstone National Park' and the historical 'trip' through Paris, it is clear that use of discrete global grid systems in the context of the Digital Earth vision is more limited than it could be (Gore 1998). The idea of Digital Earth as an interactive exploratory tool for education, social understanding, and shared decision-making requires the ability to bring traditional

CONTACT Benjamin Adams b.adams@auckland.ac.nz Department of Computer Science, Centre for eResearch, The University of Auckland, Private Bag 92019, Auckland 1142, New Zealand

geosciences data together with other kinds of social data (Craglia, Ostermann, and Spinsanti 2012; Janowicz and Hitzler 2012). Incorporating social data into discrete global grids provides a mechanism to bring these different kinds of data into the same spatial framework, a prerequisite to more comprehensive semantic integration. Furthermore, it is at the interface of human and environmental/ecological systems that many of the most difficult problems of the twenty-first century will need to be investigated, and Digital Earth systems should therefore be able to support complex socio-environmental analysis of data synthesized from disparate social and geoscientific sources (Berkes, Folke, and Colding 2000; Turner et al. 2012).

A key defining feature of spatial data that represents socially demarcated features, also known as a fiat features (Smith and Varzi 2000), is the use of toponyms or place names to refer to the features. Although it is difficult to measure, there is a massive amount of unstructured geographic information that is generated every day, which is only described in terms of place names. This includes references to places found in scientific articles, data abstracts, online media, social media, and other forms of human communication (Sui and Goodchild 2011). Despite this wealth of geographic information, there is no straightforward mechanism for folding this information into a Digital Earth based on a discrete global grid system. In order to bring place-based geographic information together with a spatial representation, gazetteers are utilized to associate the unstructured data with the spatial footprint of the referenced place.

A gazetteer is a dictionary of place names combined with information about their spatial footprints, feature types, and other associated properties (Hill 2000). Gazetteers vary widely in terms of how they represent the spatial footprint (points, polygons), though most are vector-based. Over the years numerous organizations have developed digital gazetteers, including the Alexandria Digital Library developed at the University of California, Santa Barbara (Smith and Frew 1995), the GEOnet Names Server developed by the National Geospatial Intelligence Agency (National Geospatial Intelligence Agency 2016), and the Getty Thesaurus of Geographic Names (Getty Research Institute 2016). Crowdsourced gazetteers have also been built from authoritative data augmented with additional user-generated content that is often more up-to-date than the content that comes from the authoritative sources (Goodchild 2007). The most prominent of these crowdsourced gazetteers are the GeoNames database (GeoNames 2016), which currently has over 11 million place names, albeit with simple point-based representations, and OpenStreetMap (Haklay and Weber 2008). In addition, to these general-purpose gazetteers, a number of special-purpose digital gazetteers have been developed focusing on regions (e.g. US Census TIGER gazetteer files) or historical eras (Southall, Mostern, and Berman 2011).

The fuzzy and sometimes contested nature of the spatial representation of many places (especially non-canonically defined regions) means that multiple spatial representations of a place are sometimes necessary (Montello et al. 2003; Adams 2015). Recently, meta-gazetteers have been developed that acknowledge this heterogeneous nature of place-representation and which are designed to establish links between different spatial representations stored in different gazetteers (Cope and Kelso 2016). In some cases these relations are defined using the principles of Linked Data (Stadler et al. 2012).

In the context of Digital Earth, discrete global grid systems were originally contrasted with place name-based representations of locations on the Earth's surface (Goodchild 2000). Perhaps as a result of this stance, today a gap exists between the efforts to build ever more sophisticated digital gazetteers and the development of discrete global grid systems. The subject of this paper is the development of Wāhi, an open *discrete global grid gazetteer*, designed to fill this gap and to allow users to easily map between named places and the cells on a discrete global grid. Currently, the gazetteer is built using source material from the GeoNames gazetteer plus additional areal data from other available sources detailed below. One of the primary goals in creating this discrete global grid gazetteer is to provide a low barrier-to-entry for researchers who are already working with place-based data and want to use discrete global grids. Toward this end, the gazetteer data are provided through a web service that can also export geographic information encoded using the practices of Linked Data (Bizer, Heath, and

Berners-Lee 2009). The service uses RESTful (representational state transfer) design, a lightweight software architecture protocol for building services on the web based on standard hypertext transfer protocol (HTTP) verbs, such as GET and POST.

In the following section, we describe the discrete global grid systems used to build the gazetteer. In Section 3, the methodology to create the gazetteer is detailed, and in Section 4, the gazetteer service, which exposes the information in the gazetteer as Linked Data, is described. The paper closes out with some use cases for the gazetteer in Section 5, followed by the conclusion and next steps in Section 6.

2. Discrete global grids

There are a number of different frameworks for decomposing the surface of the Earth into a discrete grid starting with a base polyhedon. The procedures for generating these grids are defined based on a set of parameters, including the choice of base polyhedron, its orientation, the partitioning scheme used, and transformation function to project the cells to the globe (Sahr, White, and Kimerling 2003). The grids are constructed by picking a base polyhedron (cube, octahedron, dodecahedron, or icosahedron), and the faces of the polyhedron are oriented to align with specific locations on the Earth. Once this is done, each face of the polyhedron is hierarchically partitioned into cells, with the most common shape of the cells being triangles, diamonds, and hexagons. Each cell can be hierarchically decomposed into smaller cells, where the factor of increase in the number of cells from one level in the hierarchy to the next is called aperture. Therefore, an aperture four triangular grid, is composed of triangular cells which are divided into four smaller triangles and so on. Finally, a projection must be chosen to map points in the cells to points on the globe, where the projection can privilege equal area or shape. HEALPix is alternative framework originally developed for astrophysics research, which decomposes the sphere into an iso-latitudinal equal area diamond grid (Gorski et al. 2005).

In addition to the discrete global grid systems based on polyhedral solids already described, a number of other grid-based spatial index methods have been in use for many years, including the Geohash standard (http://geohash.org/site/tips.html#format), which is popular in spatial information retrieval applications, and the c-squares spatial index developed by Australia's Commonwealth Scientific and Industrial Research Organisation (Rees 2003). The key disadvantage to both c-squares and geohashing, is that the cells are based on a longitude–latitude grid and therefore vary greatly in terms of the area covered depending on their location on the globe, making them unsuitable for use with many spatial statistical operations. We made a design decision to focus on building Wāhi based on equal area (or near equal area, e.g. Fuller Dymaxion) grid systems but do not rule out expanding the service to include other non-equal area grid systems in the future, given demand by the community.

Comprehensive surveys have been performed that discuss the relative merits different configurations for generating discrete global grids, so we will not do a full review in this section and refer the readers to several surveys on the topic: (Goodchild and Yang 1992; Kimerling et al. 1999; Clarke 2002; Sahr, White, and Kimerling 2003; Mahdavi-Amiri, Alderson, and Samavati 2015). To generate the grids used in the discrete global grid gazetteer, we utilized the DGGRID software, version 6.2 (Sahr 2015), which creates grids based on an icosahedron solid, and which allows for the use of two common transformation projections: (1) Icosahedral Snyder Equal Area (ISEA) (Snyder 1992), and (2) icosahedral Dymaxion projection (Fuller 1975; Gray 1995; Crider 2008). The ISEA projection is the most commonly used projection for research that utilizes discrete global grids, because the cells are equal area, making subsequent statistical analysis more sound.

The current version of the discrete global grid gazetteer has spatial footprints on ISEA aperture 3 hexagonal (ISEA3H), aperture 4 hexagonal (ISEA4H), and aperture 4 triangular (ISEA4T) grids. The ISEA3H level 4 and ISEA4T level 3 grids are shown in Figure 1.

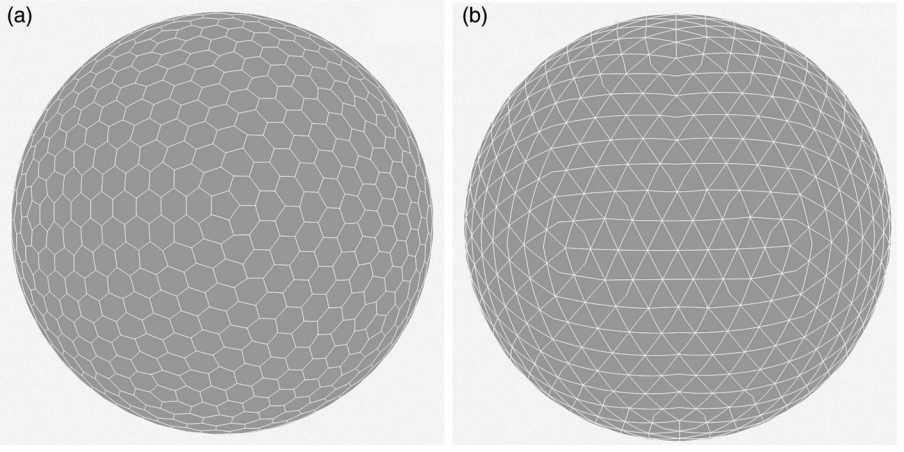

Figure 1. Two sample icosahedral equal area grids shown in orthographic projection. (a) ISEA aperture 3 hexagon level 4. (b) ISEA aperture 4 triangle level 3.

2.1. Representation of the cells

Using DGGRID, discrete global grids for the entire Earth were generated for ISEA aperture 3 hexagon grid levels 0–12, ISEA aperture 3 hexagon grid levels 0–10, and ISEA aperture 4 triangle grid levels 0–10. The default orientation with point 0 of the icosahedron set to latitude 58.2825 and longitude 11.25 was used, because it minimizes the number of icosahedron vertices on land. Densification was applied to the polygon representations of each cell to add additional points along the geodesic edges, which allow for the cells to be projected with less shape distortion. Table 1 shows statistics on the cells generated by DGGRID for each of these levels (see Sahr 2015 for additional statistics).

The grids were imported into a PostgreSQL PostGIS database using the WGS84 reference ellipsoid. Each layer is represented as table in the database, with a row for each cell, comprising an identifier and the geometry of the cell encoded using the PostGIS Geometry type. The geometry column is spatially indexed using PostGIS's generalized search tree structure. In addition, to the cell geometry, the centroid point geography for each cell is stored in a separate table. A diagram of the database schema for a discrete global grid is shown in Figure 2.

2.2. Relation graphs

Key relationships between cells are encoded in the database using a sparse graphical representation. Adjacency relationships between cells on the same level are precomputed and stored for each cell in a new table using the PostgreSQL intarray module. For the hexagonal grids this means, in most cases, six adjacent cell identifiers are stored for each cell, and for triangular grids three adjacent cell identifiers are stored. Hierarchical (parent–child) relationships between cells at different levels are stored in another table. These relations are established based on spatial overlap. Thus, hexagonal grid cells can have multiple parent cells because they overlap more than one hexagonal cell at a coarser resolution. We chose to directly store these relationships, rather than use an indexing encoding scheme in the cell identifiers, because the same method could be applied to both triangular and hexagonal grids. Furthermore, for the granularity of grids we were using, the storage requirements to record this information as a graph posed no problem, and showed no practical disadvantages when performing spatial queries. As the gazetteer progresses to include higher granularity grids, it may be necessary to employ a more sophisticated indexing scheme such as hexagonal connectivity maps (Mahdavi-Amiri, Harrison, and Samavati 2014).

Table 1. Discrete global grid statistics for the grids currently represented in the gazetteer.

Projection	Level	# cells	Cell area (km^2)	Densification
ISEA3H	1	32	17,002,187.39	12
ISEA3H	2	92	5,667,395.80	12
ISEA3H	3	272	1,889,131.93	12
ISEA3H	4	812	629,710.64	12
ISEA3H	5	2432	209,903.55	10
ISEA3H	6	7292	69,967.85	10
ISEA3H	7	21,872	23,322.62	8
ISEA3H	8	65,612	7774.21	8
ISEA3H	9	196,832	2591.40	7
ISEA3H	10	590,492	863.80	6
ISEA3H	11	1,771,472	287.93	4
ISEA3H	12	5,314,412	95.98	3
ISEA4H	1	42	12,751,640.54	12
ISEA4H	2	162	3,187,910.14	12
ISEA4H	3	642	796,977.53	12
ISEA4H	4	2562	199,244.38	12
ISEA4H	5	10,242	49,811.10	10
ISEA4H	6	40,962	12,452.77	10
ISEA4H	7	163,842	3113.19	8
ISEA4H	8	655,362	778.30	8
ISEA4H	9	2,621,442	194.57	4
ISEA4H	10	10,485,762	48.64	3
ISEA4T	1	80	6,375,820.27	12
ISEA4T	2	320	1,593,955.07	12
ISEA4T	3	1280	398,488.77	12
ISEA4T	4	5120	99,622.19	12
ISEA4T	5	20,480	24,905.55	10
ISEA4T	6	81,920	6226.39	10
ISEA4T	7	327,680	1556.60	8
ISEA4T	8	1,310,720	389.15	8
ISEA4T	9	5,242,880	97.29	4
ISEA4T	10	20,971,520	24.32	3

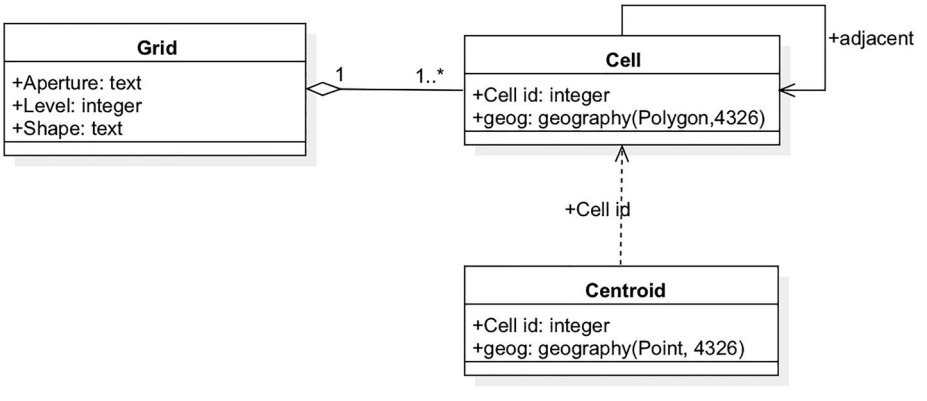

Figure 2. Database schema for a discrete global grid.

3. Discrete global grid gazetteer

The GeoNames gazetteer, with over 11 million place entries, served as the starting point for creating the discrete global gazetteer. The coverage of GeoNames primarily includes geographic features at the meso-scale, such as cities, other populated settlements, and administrative regions. However, it does include a selection of points of interest, such as museums and places of historical interest. Unlike OpenStreetMap, it does not represent nodes and edges in street networks and individual

buildings, other than the aforementioned points of interest. Each entry in the GeoNames database has a primary place name, a set of alternate names, feature type and feature code based on a shallow taxonomy of 645 feature types, and a spatial footprint in the form of a latitude and longitude. In addition, it includes parent–child relationships between entities. The assessment of the quality of the source GeoNames data remains an open research problem, but a main concern includes the shortening of decimal coordinate values at the sub minute level in decimal degrees (Ahlers 2013; Janowicz et al. 2016). Thus, while GeoNames is perhaps a good starting point for grids at a relatively coarser resolution, including the levels currently represented in the gazetteer, better source data will be required for future versions of Wāhi that include finer-grained grids.

GeoNames serves as a hub in the web of Linked Data, which means that we can maintain a connection between the discrete global grid gazetteer entries and GeoNames through the use of unique resource identifiers (URIs) (Bizer et al. 2008). However, because GeoNames has very simple point-based spatial representation of geographic features, it is necessary to supplement many entries with better spatial footprint information. Vector-based geometry of concepts in place-based language are often more accurately represented as polygons (or polylines), especially in cases when one wants to spatially reason over the knowledge to obtain relevant results (Martins, López-Pellicer, and Ahlers 2015). In the case of an equal area grid-based representation for places this is doubly important, because the notion of accurate representation is closely tied to accurate area.

In order that the discrete global grid gazetteer be open, we limited our data sources to ones that have no restrictions on re-use. Polygon data for all countries and their administrative regions at level 1 (e.g. US States) were obtained from Natural Earth (http://www.naturalearthdata.com/), for the United States TIGER data were collected for all counties and places data that is available. In addition, polygonal regional data for were imported from Quattroshapes (http://quattroshapes.com/), a gazetteer developed by FourSquare, and additional polygon data were obtained from national data sets from around the world. We aligned these various polygon datasets with the GeoNames identifiers for 222,331 entries.

Spatial footprints for all of the places in the gazetteer were generated for each level of the three discrete global grid systems. Two kinds of spatial footprints were created. The first kind of footprint is defined by the grid cells that spatially intersect with the source geometry – polygons when available, otherwise points – and will always cover the source polygon. The second kind of footprint is defined by the spatial intersection of the centroids of grid cells with the geographic features with polygon representation, unless no centroid intersects. In that case, the single cell with the closest centroid is matched to the feature. This second kind of footprint will always be smaller or equal in size to the first kind and is not guaranteed to cover the entire source polygon. Figure 3 shows some sample footprints generated for the country of New Zealand.

Augmenting places from GeoNames that do not fit well into an existing administrative hierarchy with source data that is polygon-based is major challenge going forward. The work already performed to conflate gazetteer entries with other polygon representations consisted of a great deal of hands-on data matching and cleaning. Custom scripts were employed that used the geographic hierarchy to match administrative units and cities. For example, TIGER files for populated places in the United States are organized by state, which allowed us to disambiguate between similarly named places in different states. While these methods work reasonably well for regions at the country and administrative levels, which are currently included in the gazetteer, it is unlikely that such hands-on methods will work for larger, more fine-grained databases of places, especially when matching polygons to point-of-interest features and others at the scale of street-level information. In this case, we foresee needing to employ more sophisticated automated methods, e.g. using machine learning (Hastings 2008; Martins 2011; Gelernter et al. 2013). In addition, historical information presents unique challenges requiring specialized context rules (such as based on temporal information) to disambiguate between entities (Smith and Crane 2001). Efforts are underway, however, to improve the quality of vector-based gazetteers for historical data, and we will be able to use the principles of linked data to connect the Wāhi gazetteer to these ongoing projects (Isaksen et al. 2014).

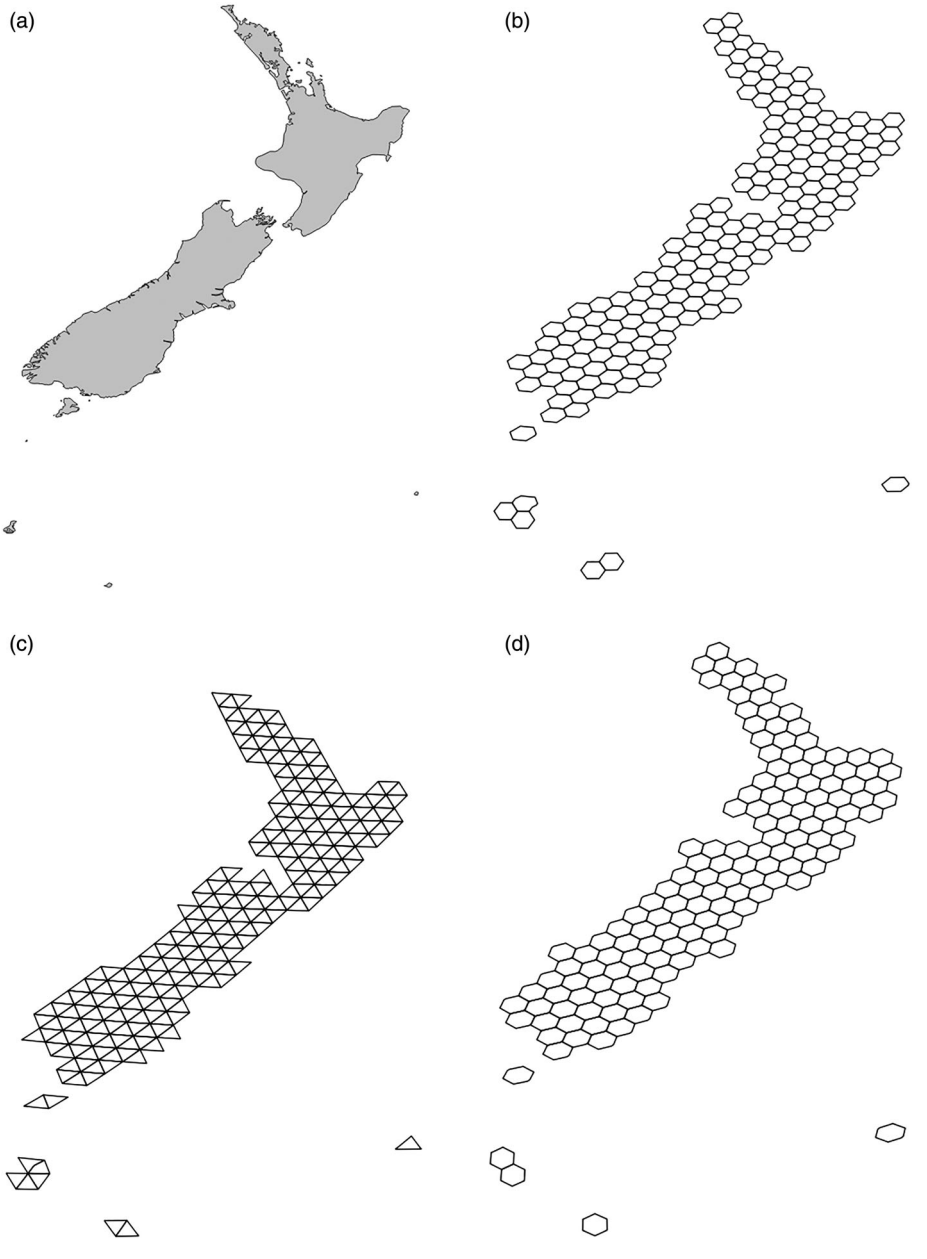

Figure 3. Source polygon of New Zealand (country) from the Natural Earth dataset and the resulting grid-based spatial footprints generated for ISEA3H9, ISEA4H7, and ISEA4T7. (a) Source polygon for the country of New Zealand from Natural Earth. (b) ISEA aperture 3 hexagon level 9. (c) ISEA aperture 4 triangle level 7. (d) ISEA aperture 4 hexagon level 7.

At present, the storage requirements for the server are on the order of 130 GB, so well within the reach of a small server. However, the storage requirements will grow as higher and higher resolution grids are created and may require alternative methods of serving the data, e.g. by using the same approach as map tile servers, which only generate and cache the data for a tile once it has been requested. Currently, the largest feature represented in the gazetteer in terms of cell count is the region for Russia in the ISEA aperture 4 hexagon level 10 grid, which has 353,699 cells and is

over 300 MB in size when represented in GeoJSON-LD format. However, the vast majority of the other features in the gazetteer are substantially smaller in download size.

4. Linking the gazetteer to the web of data

In this section, we describe a process to publish the discrete global gazetteer database as linked open data. We first describe the procedure to generate linked data serializations of the grid cell geometry data, followed by a description of a linked data service for gazetteer entries. In order to support integration of the gazetteer data with other geospatial data sets, we also present a service that finds the spatial intersection between ad hoc regions and grid cells for a given DGGS level. Finally, we discuss issues related to the governance of the service and plans for future growth.

4.1. GeoJSON-LD encoding

The geometry of the grid cells can be exported as Linked Data from the gazetteer using GeoJSON-LD encoding, which is a JSON serialization for describing geographic features and linking the data to other Linked Data sources (http://geojson.org/vocab). GeoJSON-LD combines JSON-LD, a W3C recommendation for serializing Linked Data using JSON, with GeoJSON, a commonly used data interchange format for representing geographic entities in web content (Butler et al. 2008; Sporny et al. 2014). Because GeoJSON-LD data can also be read as regular GeoJSON, the geometry of the cells can be used outside of Linked Data applications as well.

Each grid cell has an URI that allows it to be referenced by other linked data sets. By linking the gazetteer entries to other data sources on the web, beginning with the GeoNames database, the discrete global grid gazetteer can be combined with related web services. For example, a user can utilize the GeoNames web service API (http://www.geonames.org/export/web-services.html) to search for GeoNames identifiers based on name or locational proximity, and additional place attributes (such as feature type) that are stored in the database can be mapped to the grid cells.

The URIs for the grid cells take the form of http://dg3.cer.auckland.ac.nz/{*projection*}/{*orientation*}/{*level*}/cell/{*idx*}, where (1) *projection* is one of the `isea3h`, `isea4h`, or `isea4t`; (2) *orientation* is a point represented by longitude and latitude concatenated with an underscore character (e.g. `11.25_58.2825`); (3) *level* is a numeric representation of the hierarchical level in the grid system (e.g. `7`); and (4) *idx* is the unique index for the cell in the level of the grid system. For example, http://dg3.cer.auckland.ac.nz/isea3h/11.25_58.2825/7/cell/1001 is the URI for cell #1001 for the ISEA aperture 3 hexagon grid at level 7 with point 0 of the icosahedron with the default orientation for DGGRID. The initial version of the gazetteer is based on the icosahedral equal area projection, which is reflected in the `isea{aperture}{shape}` form in the URI. As the gazetteer is expanded in the future to include additional discrete global grid systems based on different projections and polyhedral solids (e.g. cube or octahedron), we will use a similar URI scheme, where the `isea` part of the URI will be modified to reflect the different projections. Figure 4 shows the (abidged) GeoJSON-LD encoding of the cell that is returned when the URI is looked up using the gazetteer service. The centroid of a cell is referenced in a similar manner with the form http://dg3.cer.auckland.ac.nz/{*proj.*}/{*orientation*}/{*level*}/point/{*idx*}.

4.2. Gazetteer service

In addition to individual cells, each gazetteer entry can be exported as a GeoJSON FeatureCollection type composed of an array of polygon features corresponding to the discrete grid cells. Figure 5 shows a sample export of a New Zealand entry in the gazetteer. The link to GeoNames is established using the `sameAs` relation to match the geographic feature to an entry in the GeoNames database.

```
{
  "@context": {
    "dg3": "http://dg3.cer.auckland.ac.nz/",
    "adjacent": {
      "@id": "dg3:vocab#adjacent",
      "@container": "@set"
    },
    "centroid": "dg3:vocab#centroid",
    "projection": "dg3:vocab#projection",
    "geojson": "http://ld.geojson.org/vocab#",
    "Feature": "geojson:Feature",
    "Polygon": "geojson:Polygon",
    "geometry": "geojson:geometry",
    "properties": "geojson:properties",
    "coordinates": "geojson:coordinates",
    "id": "@id"
  },
  "type": "Feature",
  "id": "dg3:isea3h/11.25_58.28252559/7/cell/1001",
  "crs": {
    "type": "name",
    "properties": { "name": "urn:ogc:def:crs:EPSG::4326" }
  },
  "geometry": {
    "type": "Polygon",
    "coordinates": [[[-146.66894,33.77059], [-146.74807,33.69806],
                     ... , [-146.66894,33.77059]]]
  },
  "properties":  {
    "id": 1001,
    "adjacent": [
      "dg3:isea3h/11.25_58.2825/7/cell/947",
      "dg3:isea3h/11.25_58.2825/7/cell/974",
      "dg3:isea3h/11.25_58.2825/7/cell/975",
      "dg3:isea3h/11.25_58.2825/7/cell/1028",
      "dg3:isea3h/11.25_58.2825/7/cell/1029",
      "dg3:isea3h/11.25_58.2825/7/cell/1056"
    ],
    "centroid": "dg3:isea3h/11.25_58.2825/7/point/1001",
    "projection": "isea3h"
  }
}
```


Figure 4. Example GeoJSON-LD export of a grid cell geometry.

This link, in turn, can be used to connect to a wealth of other Linked Data using GeoNames connections to other hubs in the Linked Data web, such as the DBpedia database (Auer et al. 2007).

4.3. Matching ad hoc regions

All of the URIs defined for the grid cells and gazetteer entries are resolvable, which means an HTTP request to the URI will return the GeoJSON result. In addition, a web service has been created that allows for the matching of ad hoc regions to grid cells using a POST request. The URL of the service has the form http://dg3.cer.auckland.ac.nz/`{projection}/{orientation}/{level}/geojson`.

Figure 6 shows an example request to return the geometry of cells that intersect with a 1 by 1 degree bounding box. Note, the intersection operation is calculated on the spheroid, so the geometry would appear as a 'curved' box on a Mercator projection.

4.4. Governance of the service

The current version of the Wāhi gazetteer represents only the first version with the goal that the gazetteer will grow with additional source material and to reflect changing technology standards.

```
{
  "@context": {
    "dg3": "http://dg3.cer.auckland.ac.nz/",
    "geojson": "http://ld.geojson.org/vocab#",
    "Feature": "geojson:Feature",
    "FeatureCollection": "geojson:FeatureCollection",
    "Polygon": "geojson:Polygon",
    "geometry": "geojson:geometry",
    "properties": "geojson:properties",
    "coordinates": "geojson:coordinates",
    "features": {
      "@container": "@set",
      "@id": "geojson:features"
    },
    "id": "@id",
    "Place": "http://schema.org/Place",
    "toponym": "http://schema.org/name",
    "geometry": "geojson:vocab#geometry",
    "sameAs": {
      "@id": "http://www.w3.org/2002/07/owl#sameAs",
      "@container": "@set"
    }
  },
  "toponym": "New Zealand",
  "sameAs": [ "http://sws.geonames.org/2186224/" ],
  "crs": {
    "type": "name",
    "properties": { "name":"urn:ogc:def:crs:EPSG::4326" }
  },
  "type": "FeatureCollection",
  "@type": ["FeatureCollection" , "Place"],
  "features": [
    {
      "type": "Feature",
      "@type": ["Feature", "Place"],
      "id": "dg3:isea3h/11.25_58.2825/6/cell/6458",
      "geometry": {
        "type": "Polygon",
        "coordinates": [[[167.64907,-51.00816],[167.80841,-51.08452],
                         ... , [167.64907,-51.00816]]]
      }
    },
    {
      ...
    }
  ]
}
```


Figure 5. Example GeoJSON-LD export of the New Zealand country entry in the gazetteer.

```
curl -X POST -d '{"type": "Polygon", "coordinates": \
  [[[30, 10],[31.0,10.0],[31.0,11.0],[31.0,10.0],[30,10]]]}' \
  http://dg3.cer.auckland.ac.nz/isea3h/11.25_58.2825/7/geojson \
  --header "Content-Type:application/json"
```

Figure 6. Example POST web request for the grid cell geometries intersecting a one degree bounding box.

In the latter case, the GeoJSON-LD specification used to represent the information in the gazetteer as linked data is in flux and should it change, then we will update the gazetteer service to reflect that. The intention going forth is to maintain historical versions of the gazetteer that are referenced via a versioning system. The URIs as described in Section 4.1 will always refer to the latest version and as older versions are retired we will publish new URIs representing the older version of the data.

5. Use cases

A discrete global grid gazetteer has several use cases, and in this section we highlight a few of the active areas of research where place-based data are being combined with grid-based information. Spatial decomposition of areal social science data, such as census population data, onto grid-based systems to increase the spatial resolution of the data is a well-studied area and has been applied to build databases such as the Landscan USA (Flowerdew and Green 1993; Bhaduri et al. 2007). The global grid gazetteer and associated services provide a spatial framework for performing this kind of analysis on any kind of place-based data.

Access to a discrete global grid gazetteer also enables the ability to map georeferenced textual documents to Digital Earth systems. Gazetteers are an essential component to a number of geographic information retrieval applications that rely on georeferencing textual documents (Santos, Anastácio, and Martins 2015). An isocahedral Fuller Dymaxion triangular projection precursor to the Wāhi gazetteer database was demonstrated to be valuable in the building of the Frankenplace thematic map-based exploratory search system (Adams, McKenzie, and Gahegan 2015). By mapping place names to a multi-resolution grid, it enabled the ability to geographically search through the contents of a collection of millions of textual documents in real time. Using machine learning methods, natural language text can also be mapped to a probability distribution over the cells in a discrete global grid, enabling predictive analytics to determine the likelihood of where a text is written about (Adams and Janowicz 2012; Wing and Baldridge 2014; Melo and Martins 2016).

As with other kinds social science data, analyzing text based on place names can be viewed as a spatial decomposition problem. Because textual sources can reference places at multiple scales (everything from continents down to points of interest), associating all place names to a hierarchical grid system provides a common framework for performing analysis. Beyond these specific examples, making place-based data available to be spatially analyzed on Digital Earth systems with other sensor-based observations, such as remote sensing data, will support analysis in the context of a wide variety of emergency response, public health, and socio-environmental applications.

6. Conclusion and future work

In this paper we described the development of Wāhi, a discrete global grid gazetteer designed to enable the incorporation of place-based analysis into Digital Earth systems. The gazetteer is built using source data from GeoNames and other open spatial data sets from around the world. A complementary web service was also implemented, which exposes the discrete global grid and Wāhi gazetteer data as Linked Data, so it can be connected to many other data sets about places that are currently available on the web. The discrete global grid gazetteer is designed to support multiple use cases, including modeling demographic data on a grid, mapping and spatially indexing unstructured textual data, and integrating social and environmental data.

The current state of the Wāhi gazetteer represents the first version, and we will continue to improve its coverage and link in other data sources. Although, we have included many administrative regions based on areal representations, there remains a great deal of work to find better-than-point source data for a number of features, especially non-political natural features, such as islands, rivers, and lakes. We will also broaden the scope of the gazetteer to include other grid systems, such as the icosahedral Fuller Dymaxion projection, additional aperture 3 and 4 combinations for ISEA grids, and higher granularity grids. The latter will be of particular importance as more sources of points of interest and finer scale features such as buildings and roads are added. This will allow us to import many more datasets including OpenStreetMap. Depending on need expressed by the community of users, it might also be necessary to create additional grids with different orientations. Several indexing schemes have been developed for different kinds of grids, and future work will include creating mappings between the data in the gazetteer and the various indexing schemes

that have been developed (Sahr, White, and Kimerling 2003; Tong et al. 2013; Mahdavi-Amiri, Harrison, and Samavati 2014).

Acknowledgments

We would like to thank three anonymous reviewers for their thoughtful suggestions, which helped to improve the manuscript. We would also like to thank the Centre for eResearch at the University of Auckland for providing a server to host the gazetteer service.

Disclosure statement

No potential conflict of interest was reported by the author.

References

Adams, B. 2015. "Finding Similar Places using the Observation-to-Generalization Place Model." *Journal of Geographical Systems* 17 (2): 137–156.

Adams, B., and K. Janowicz. 2012. "On the Geo-Indicativeness of Non-Georeferenced Text." Sixth international AAAI conference on web and social media (ICWSM 2012), 375–378, Dublin, Ireland.

Adams, B., G. McKenzie, and M. Gahegan. 2015. "Frankenplace: Interactive Thematic Mapping for ad hoc Exploratory Search." In Proceedings of the 24th international conference on World Wide Web, edited by A. Gangemi, S. Leonardi, A. Panconesi, K. Gummadi, and C. Zhai, 12–22. Florence: International World Wide Web Conferences Steering Committee.

Ahlers, D. 2013. "Assessment of the Accuracy of GeoNames Gazetteer Data." In Proceedings of the 7th workshop on geographic information retrieval (GIR '13), Orlando, FL, edited by R. Purves and C. Jones, 74–81. New York: ACM.

Amiri, A. M., F. Samavati, and P. Peterson. 2015. "Categorization and Conversions for Indexing Methods of Discrete Global Grid Systems." *ISPRS International Journal of Geo-Information* 4 (1): 320–336.

Auer, S., C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. 2007. "DBpedia: A Nucleus for a Web of Open Data." The Semantic Web: 6th international semantic web conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11–15, 2007. Proceedings, edited by K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K-I. Lee, L. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, and Ph. Cudré-Mauroux, 722–735. Berlin: Springer. http://dx.doi.org/10.1007/978-3-540-76298-0_52.

Berkes, F., C. Folke, and J. Colding. 2000. *Linking Social and Ecological Systems: Management Practices and Social Mechanisms for Building Resilience*. Cambridge: Cambridge University Press.

Bhaduri, B., E. Bright, P. Coleman, and M. L. Urban. 2007. "LandScan USA: A High-resolution Geospatial and Temporal Modeling Approach for Population Distribution and Dynamics." *GeoJournal* 69 (1): 103–117.

Bizer, C., T. Heath, and T. Berners-Lee. 2009. "Linked Data-the Story So Far." In *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, edited by A. P. Sheth, 205–227. Hershey, PA: Information Science.

Bizer, C., T. Heath, K. Idehen, and T. Berners-Lee. 2008. "Linked Data on the Web (LDOW2008)." Proceedings of the 17th international conference on World Wide Web, edited by J. Huai, R. Chen, W.-Y. Ma, A. Tomkins, and X. Zhang, 1265–1266. Beijing: ACM.

Butler, H., M. Daly, A. Doyle, S. Gillies, T. Schaub, and C. Schmidt. 2008. "The GeoJSON Format Specification." Accessed June 7, 2016. http://geojson.org/geojson-spec.html

Clarke, K. C. 2002. "Criteria and Measures for the Comparison of Global Geocoding Systems." *Discrete Global Grids: A Web Book*. Accessed June 3, 2016. http://www.geog.ucsb.edu/~kclarke/Papers/GlobalGrids.html

Cope, A., and N. Kelso. 2016. "Who's On First." Accessed June 2, 2016. https://mapzen.com/blog/who-s-on-first/

Craglia, M., F. Ostermann, and L. Spinsanti. 2012. "Digital Earth from Vision to Practice: Making Sense of Citizen-generated Content." *International Journal of Digital Earth* 5 (5): 398–416.

Crider, J. E. 2008. "Exact Equations for Fuller's Map Projection and Inverse." *Cartographica: The International Journal for Geographic Information and Geovisualization* 43 (1): 67–72.

Dulvy, N. K., S. L. Fowler, J. A. Musick, R. D. Cavanagh, P. M. Kyne, L. R. Harrison, J. K. Carlson, et al. 2014. "Extinction Risk and Conservation of the World's Sharks and Rays." *Elife* 3: e00590.

Flowerdew, R., and M. Green. 1993. "Developments in Areal Interpolation Methods and GIS." In *Geographic Information Systems, Spatial Modelling and Policy Evaluation*, edited by M. M. Fischer, and P. Nijkamp, 73–84. Berlin: Springer.

Fuller, R. B. 1975. *Synergetics*. New York: MacMillan.

Gelernter, J., G. Ganesh, H. Krishnakumar, and W. Zhang. 2013. "Automatic Gazetteer Enrichment with User-geocoded Data." Proceedings of the second ACM SIGSPATIAL International Workshop on Crowdsourced and

Volunteered Geographic Information (GEOCROWD '13), Orlando, FL, edited by D. Pfoser and A. Voisard, 87–94. New York: ACM.
GeoNames. 2016. "GeoNames." Accessed June 2, 2016. http://www.geonames.org/
The Getty Research Institute. 2016. "Getty Thesaurus of Geographic Names Online." Accessed June 2, 2016. http://www.getty.edu/research/tools/vocabularies/tgn/
Gibb, R., M. Purss, and F. Samavati. 2016. "Discrete Global Grid Systems SWG." Accessed May 31, 2016, http://www.opengeospatial.org/projects/groups/dggsswg
Goodchild, M. F. 2000. "Discrete Global Grids for Digital Earth." Accessed June 2, 2016. http://www.ncgia.ucsb.edu/globalgrids/papers/goodchild.pdf
Goodchild, M. F. 2007. "Citizens as Sensors: The World of Volunteered Geography." *GeoJournal* 69 (4): 211–221.
Goodchild, M. F., H. Guo, A. Annoni, L. Bian, K. de Bie, F. Campbell, M. Craglia, et al. 2012. "Next-generation Digital Earth." *Proceedings of the National Academy of Sciences* 109 (28): 11088–11094.
Goodchild, M. F., and Y. Shiren. 1992. "A Hierarchical Spatial Data Structure for Global Geographic Information Systems." *CVGIP: Graphical Models and Image Processing* 54 (1): 31–44.
Gore, A. 1998. "The Digital Earth: Understanding Our Planet in the 21st Century." *Australian Surveyor* 43 (2): 89–91.
Gorski, K. M., E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelmann. 2005. "HEALPix: A Framework for High-resolution Discretization and Fast Analysis of Data Distributed on the Sphere." *The Astrophysical Journal* 622 (2): 759–771.
Gray, R. W. 1995. "Exact Transformation Equations for Fuller's World Map." *Cartographica: The International Journal for Geographic Information and Geovisualization* 32 (3): 17–25.
Haklay, M., and P. Weber. 2008. "Openstreetmap: User-generated Street Maps." *IEEE Pervasive Computing* 7 (4): 12–18.
Hastings, J. T. 2008. "Automated Conflation of Digital Gazetteer Data." *International Journal of Geographical Information Science* 22 (10): 1109–1127.
Hill, L. L. 2000. "Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints." In *Research and Advanced Technology for Digital Libraries*, edited by J. Borbinha, and T. Baker, 280–290. Berlin: Springer.
Isaksen, L., R. Simon, E. T. E. Barker, and P. de Soto Cañamares. 2014. "Pelagios and the Emerging Graph of Ancient World Data." In Proceedings of the 2014 ACM conference on Web Science (WebSci '14), Bloomington, IN, USA, edited by F. Menczer, J. Hendler, W. Dutton, M. Strohmaier, E. T. Meyer, and C. Cattuto, 197–201. New York: ACM.
Janowicz, K., and P. Hitzler. 2012. "The Digital Earth as Knowledge Engine." *Semantic Web* 3 (3): 213–221.
Janowicz, K., Y. Hu, G. McKenzie, S. Gao, B. Regalia, G. Mai, R. Zhu, B. Adams, and K. Taylor. 2016. "Moon Landing or Safari? A Study of Systematic Errors and their Causes in Geographic Linked Data." Geographic Information Science 9th International Conference, GIScience 2016, Montreal, QC, Canada, September 27–30, 2016, Proceedings, edited by J. Miller, D. O'Sullivan, and N. Wiegand, 275–292. Berlin: Springer.
Kerr, Y. H., P. Waldteufel, J.-P. Wigneron, J. Martinuzzi, J. Font, and M. Berger. 2001. "Soil Moisture Retrieval from Space: The Soil Moisture and Ocean Salinity (SMOS) Mission." *IEEE Transactions on Geoscience and Remote Sensing* 39 (8): 1729–1735.
Kimerling, J. A., K. Sahr, D. White, and L. Song. 1999. "Comparing Geometrical Properties of Global Grids." *Cartography and Geographic Information Science* 26 (4): 271–288.
Mahdavi-Amiri, A., T. Alderson, and F. Samavati. 2015. "A Survey of Digital Earth." *Computers & Graphics* 53 (Part B): 95–117.
Mahdavi-Amiri, A., E. Harrison, and F. Samavati. 2014. "Hexagonal Connectivity Maps for Digital Earth." *International Journal of Digital Earth* 8 (9): 750–769.
Martins, Bruno. 2011. "A Supervised Machine Learning Approach for Duplicate Detection over Gazetteer Records." GeoSpatial Semantics: 4th International Conference, GeoS 2011, Brest, France, May 12–13, 2011, Proceedings, edited by C. Claramunt, S. Levashkin, and M. Bertolotto, 34–51. Berlin: Springer.
Martins, B., F. J. López-Pellicer, and D. Ahlers. 2015. "Expanding the Utility of Geospatial Knowledge Bases by Linking Concepts to WikiText and to Polygonal Boundaries." Proceedings of the 9th Workshop on Geographic Information Retrieval (GIR '15), Paris, France, Vol. 1, 1–1:2. New York: ACM. http://doi.acm.org/10.1145/2837689.2837693
Melo, Fernando, and Bruno Martins. 2016. "Automated Geocoding of Textual Documents: A Survey of Current Approaches." *Transactions in GIS* n/a–n/a.
Montello, D. R., M. F. Goodchild, J. Gottsegen, and P. Fohl. 2003. "Where's Downtown?: Behavioral Methods for Determining Referents of Vague Spatial Queries." *Spatial Cognition & Computation* 3 (2–3): 185–204.
National Geospatial Intelligence. 2016. "NGA GEOnet Names Server (GNS)." Accessed June 2, 2016. http://geonames.nga.mil/gns/html/
Nguyen, H., N. Cressie, and A. Braverman. 2012. "Spatial Statistical Data Fusion for Remote Sensing Applications." *Journal of the American Statistical Association* 107 (499): 1004–1018.
Rees, T. 2003. "C-squares, A New Spatial Indexing System and its Applicability to the Description of Oceanographic Datasets." *Oceanography* 16 (1): 11–19.

Sahr, K. 2015. "DGGRID Version 6.2b: User Documentation for Discrete Global Grid Software." Accessed May 31, 2016. http://webpages.sou.edu/~sahrk/sqspc/pubs/dggridManualV62.pdf

Sahr, K., D. White, and A. J. Kimerling. 2003. "Geodesic Discrete Global Grid Systems." *Cartography and Geographic Information Science* 30 (2): 121–134.

Santos, J., I. Anastácio, and B. Martins. 2015. "Using Machine Learning Methods for Disambiguating Place References in Textual Documents." *GeoJournal* 80 (3): 375–392.

Schipper, J., J. S. Chanson, F. Chiozza, N. A. Cox, M. Hoffmann, V. Katariya, J. Lamoreux, et al. 2008. "The Status of the World's Land and Marine Mammals: Diversity, Threat, and Knowledge." *Science* 322 (5899): 225–230.

Smith, D. A., and G. Crane. 2001. "Disambiguating Geographic Names in a Historical Digital Library." Research and Advanced Technology for Digital Libraries: 5th European Conference, ECDL 2001 Darmstadt, Germany, September 4–9, 2001, Proceedings, edited by Panos Constantopoulos and Ingeborg T. Sølvberg, 127–136. Berlin: Springer.

Smith, T. R., and J. Frew. 1995. "Alexandria Digital Library." *Communications of the ACM* 38 (4): 61–62.

Smith, B., and A. C. Varzi. 2000. "Fiat and Bona Fide Boundaries." *Philosophy and Phenomenological Research* 60: 2401–420.

Snyder, J. P. 1992. "An Equal-area Map Projection for Polyhedral Globes." *Cartographica: The International Journal for Geographic Information and Geovisualization* 29 (1): 10–21.

Southall, H., R. Mostern, and M. L. Berman. 2011. "On Historical Gazetteers." *International Journal of Humanities and Arts Computing* 5 (2): 127–145.

Sporny, M., D. Longley, G. Kellogg, M. Lanthaler, and N. Lindström. 2014. "JSON-LD 1.0: A JSON-based Serialization for Linked Data." Accessed June 7, 2016. https://www.w3.org/TR/json-ld/

Stadler, C., J. Lehmann, K. Höffner, and S. Auer. 2012. "Linkedgeodata: A Core for a Web of Spatial Open Data." *Semantic Web* 3 (4): 333–354.

Strassburg, B. B., A. Kelly, A. Balmford, R. G. Davies, H. K. Gibbs, A. Lovett, L. Miles, et al. 2010. "Global Congruence of Carbon Storage and Biodiversity in Terrestrial Ecosystems." *Conservation Letters* 3 (2): 98–105.

Sui, D., and M. Goodchild. 2011. "The Convergence of GIS and Social Media: Challenges for GIScience." *International Journal of Geographical Information Science* 25 (11): 1737–1748.

Tong, X., J. Ben, Y. Wang, Y. Zhang, and T. Pei. 2013. "Efficient Encoding and Spatial Operation Scheme for Aperture 4 Hexagonal Discrete Global Grid System." *International Journal of Geographical Information Science* 27 (5): 898–921.

Turner, W. R., K. Brandon, T. M. Brooks, C. Gascon, H. K. Gibbs, K. S. Lawrence, R. A. Mittermeier, and E. R. Selig. 2012. "Global Biodiversity Conservation and the Alleviation of Poverty." *BioScience* 62 (1): 85–92.

Vince, A., and X. Zheng. 2009. "Arithmetic and Fourier Transform for the PYXIS Multi-resolution Digital Earth Model." *International Journal of Digital Earth* 2 (1): 59–79.

Wing, Benjamin, and Jason Baldridge. 2014. "Hierarchical Discriminative Classification for Text-Based Geolocation." In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, edited by A. Moschitti, B. Pang, and W. Daelemans, 336–348. Stroudsburg, PA: Association for Computational Linguistics.