5-1-2017

HGC-Description

Michael Polczynski
Georgetown University

Mark Polczynski
Marquette University, mark.polczynski@marquette.edu

The OpenOttoman Historical Gazetteer of the Crimea

Mark Polczynski[1] ~ Michael Polczynski[2]

July 24, 2017

Introduction

Every day, Ottomanists around the world generate valuable materials as components of their scholarly research, but these materials may not be readily available to scholars with similar interests. The purpose of the OpenOttoman[3] is to stimulate generation of such materials and provide open and convenient access to these materials. The OpenOttoman Portal[4] (OOP) is one vehicle for accessing these materials. Materials made available through the OOP include databases, with a prime example being gazetteers of Ottoman world places. Here, we describe the Historical Gazetteer of the Crimea (HGC), a prototype OOP gazetteer that includes populated places, districts, and provinces of Crimea in existence at the end of the Crimean Khanate. While providing a useful gazetteer, a primary purpose of the HGC is to serve as a test bed and use case for how other Ottoman world gazetteers could be incorporated into the OOP.

This article provides an overview of the HGC, starting with a review of the information sources used to create the gazetteer followed by descriptions of the HGC databases and database format. The means of accessing the HGC databases are then presented, three examples of HGC database visualization are provided, and directions for future work are outlined. To aid in providing a test bed and use case for future OOP gazetteers, Appendix A summarizes the workflow used to create the HGC databases, Appendix B shows how KML format versions of the databases were created, and Appendix C provides additional detail on how the GeoNames database[5] was used to create the HGC databases.

Data Sources

The primary source of data for the HGC databases is Jankowski's dictionary of pre-Russian places in Crimea[6]. This source contains over 1300 unique dictionary entries for populated places plus a map showing the locations of 137 populated places. Of the populated places that have dictionary entries and that are shown on Jankowski's map, 99 have been associated with existing places contained in the GeoNames database. These are the populated place entries in the HGC, with latitude and longitude for these entries being taken from the GeoNames database.

Jankowski's map also shows boundaries for six provinces and 43 districts. Locations for the province and district entries in the HGC are region centroids approximated using boundaries shown on Jankowski's map.

Of the 16 rivers shown on Jankowski's map, 14 have existing equivalents. The HGC contains entries for these 14 rivers plus 61 additional rivers included to aid in locating populated places and district and province boundaries. Gazetteer locations for these rivers are river mouth latitude and longitude taken from the GeoNames database entries for these rivers.

Terrain elevation data[7] was incorporated into a base map of the region covered by Jankowski's map. This allows terrain features such as valleys and ridge lines to be used when locating district and province boundaries.

Toponyms for all populated places, districts, and rivers included in the HGC utilize the spelling and character set used in Jankowski's work. Where Jankowski's map toponyms differ in spelling from Jankowski's dictionary spellings, dictionary spellings are used. Rivers not shown on Jankowski's map use modern English names and are enclosed in parentheses () to indicate this. All toponym spellings use Unicode-8 characters. Table 1 shows the special characters used in Jankowski's work and the Unicode hex codes for each (needed to search databases for toponyms with special characters).

---

[1] mark.polczynski@marquette.edu.

[2] mjp225@georgetown.edu.

[3] Amy Singer, OpenOttoman: A Collaborative Platform for Digital Scholarship, https://networks.hnet.org/node/11419/discussions/166360/openottoman-collaborative-platform-digital-scholarship, 2017.

[4] http:// www.openottoman.org.

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

[6] Henryk Jankowski, Historical-Etymological Dictionary of Pre-Russian Habitation Names of the Crimea, Brill, Leiden, 2006.

[7] Shuttle Radar Topography Mission (SRTM) data, available at https://lta.cr.usgs.gov/SRTM1Arc.

Character	Unicode hex code	Character	Unicode hex code
Ç	00C7	ç	00E7
Ö	00D6	ö	00F6
Ü	00DC	ü	00FC
Č	010C	č	010D
Ğ	011E	ğ	011F
Ŋ	014A	ŋ	014B
Š	0160	š	0161
Ɣ	0194	ɣ	0263
Χ	03A7	χ	03C7

Table 1: Special characters used in Jankowski's work.

Databases and Format

The HGC includes three separate sets of databases as summarized in Table 2. The first set consists of text files, the second has KML files that can be displayed using applications such as Google Earth[8], and the third contains shape files and associated files for use in GIS systems such as ArcGIS[9] and QGIS[10]. Each set has separate databases for populated places, provinces, districts, and rivers. The table also shows the names of the database .zip or .kmz folders that can be accessed as described in the following section.

	Text files	KML files	Shape files
Populated places	HGCK-Populated-Places	HGCK-Populated-Places	HGCK-Populated-Places
Districts	HGCK-Districts	HGCK-Districts	HGCK-Districts
Provinces	HGCK-Provinces	HGCK-Provinces	HGCK-Provinces
Rivers	HGCK-Rivers	HGCK-Rivers	HGCK-Rivers
Folder	HGCK-TXT.zip	HGCK-KML.kmz	HGCK-SHP.zip

Table 2: Gazetteer databases.

The HGC fields used for the populated places text database are shown in Table 3. Note that the text database format shown in the table provides fields for a single-point feature longitude and latitude. This adequately accommodates locations for populated places, district and province centroids, and river mouths, but the KML and shape files cited in Table 2 include line features for rivers and polygon features for districts and provinces, thereby allowing river courses and district and province boundaries to be displayed using applications such as ArcGIS, QGIS, and Google Earth. For the KML and shape file databases, rivers were traced in QGIS using the Google Maps physical map[11]. District and province boundaries were approximated using Jankowski's map plus terrain features such as rivers, ridge lines, and valleys where boundaries on Jankowski' map were difficult to trace.

Referring to Table 3, note the manner in which a unique ID is assigned to each place in the database. Since each Jankowski place is associated with a place in the GeoNames database, and since each place in the GeoNames database has a unique ID given as a URL, the GeoNames URL for the place serves as a unique identifier for the HGC place.

The fields for district, province, and river databases are the same as the fields for populated places except that the *Page* field is left blank, since this field contains the page number in Jankowski's dictionary where a populated place is described, but districts, provinces, and rivers do not appear as entries in Jankowski's dictionary. Districts, provinces, and rivers are assigned URL IDs in the same way that they are assigned for populated places. Thus, in Table 3 the *Part of* field for the populated place *Qarağy* is the GeoNames URL for *Taryan*, the district in which this place is located.

The *Code* and *Class* fields use the nomenclature employed in the GeoNames database. Per the example, the *P* code applies generally to populated places, and the *PPL* class (sub-code) applies more specifically to existing

[8] https://www.google.com/earth
[9] http://www.arcgis.com
[10] http://www.qgis.org
[11] https://maps.google.com

populated places vs., for example, the class *PPLH* which is used for historical populated places that no longer exists. Since all populated places currently in the HGC have been associated with existing places, all populated places have the class *PPL*. Note, however, that potential future HGC populated place entries where associations between Jankowski places and GeoNames places are not possible will have the class *PPLH*. For rivers, the code is *H* for hydrographic feature and the class is *STRM* for stream. The code used for districts and provinces is *A* for administrative region. The class is *ADM**x**H* for historical administrative divisions, where *x* represents the administrative level. The HGC uses the codes *ADM4H* for districts (*kadylyk*) and *ADM3H* for provinces (*kaimakamlyk*). Extending this approach, the codes *ADM2H* = Crimean Khanate and *ADM1H* = Ottoman Empire can be used in other gazetteers of the Ottoman world. The GeoNames code ADMH5 is also available for any administrative levels below district.

Field	Description	Example
ID	Unique identifier as GeoNames URL	http://www.geonames.org/461727
LONGITUDE	Longitude of place	32.53333
LATITUDE	Latitude of place	45.38333
TOPONYM	Name of place	Qarağy
BEGIN_DATE	Start year for which this entry applies	1450
END_DATE	End year for which this entry applies	1800
CODE	GeoNames code for feature type	P
CLASS	GeoNames class for feature type	PPLH
PART_OF	ID of district where populated place is located	http://www.geonames.org/11494617
CONTRIBUTOR	Unique identifier for this entry's contributor as URL	http://www.technologyforge.net/MHP
SOURCE	Unique identifier for Jankowski's publication as URL	http://www.brill.com/historical-etymological-dictionary-pre-russian-habitation-names-crimea
PAGE	Page in source where populated place is cited	106
CONFIDENCE	Level of confidence in this entry's information	1

Table 3: Gazetteer text file fields for populated places.

In Table 3, the *Confidence* field reflects the confidence level of the data captured in a particular database entry. Because all populated places in the HGC can be associated with existing places in the GeoNames database, they are assigned a confidence level of 1. Rivers are also assigned a confidence level of 1, since they have been traced directly from Google Maps. Confidence levels for HGC district and province entries have been given a value of 2, since the locations of the borders for these areas have been approximated from Jankowski's map.

Accessing and Using HGC Databases

As mentioned, OpenOttoman is focused on providing open and convenient access to various types of materials. To this end, the HGC databases and support materials are being made available through the HGC website[12] under a Creative Commons Attribution 3.0 license[13]. Users may copy and redistribute the material in any medium or format, and may remix, transform, and build upon the material for any purpose. Per the license, users must cite the HGC website as the source of a HGC database. HGC and related materials are also available through the OpenOttoman data repository[14].

HGC Visualization and Analysis

While a text-based gazetteer can provide value in its own right, this value can be leveraged significantly through the application of visualization and analytical tools. Three examples will now be described.

Per Table 2, the HGC provides access to the folder *HGC-SHP.zip* which contains shape files and associated files for populated places, districts, provinces, and rivers, all of which can be opened in tools such as ArcGIS and QGIS, thereby providing users with a host of mapmaking and analysis tools. In addition, the HGC website provides access to the folder *HGC-QGIS-Quick-Start.zip*. Unzipping this folder and opening the project named *HGC-QGIS-Project.qgs* in QGIS yields the result shown in Figure 1.

[12] http://www.technologyforge.net/HGC

[13] https://creativecommons.org/licenses/by/3.0/

[14] http://epublications.marquette.edu/ottoman

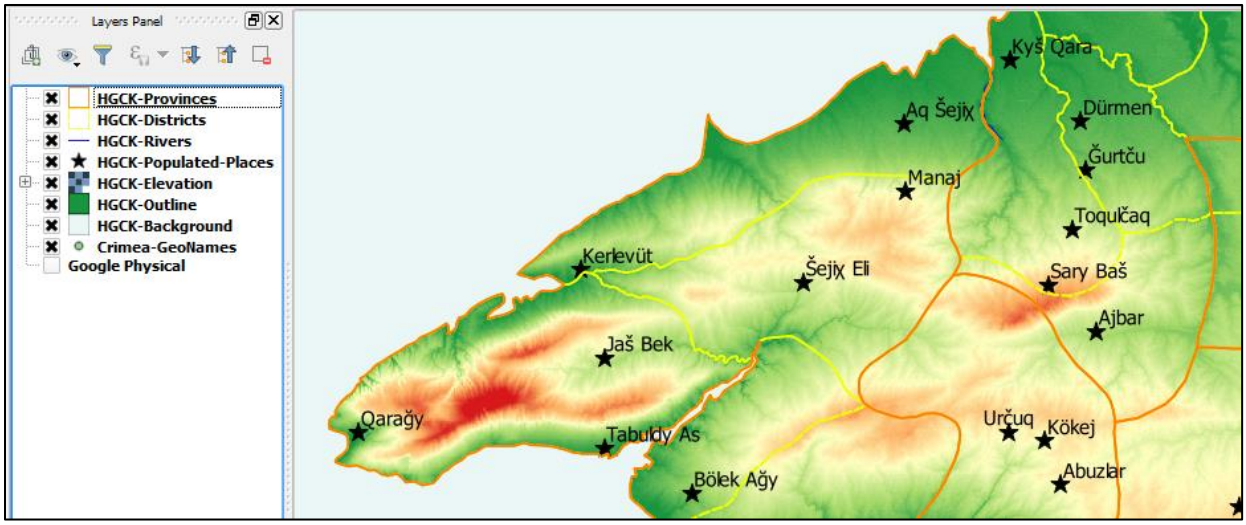

Figure 1: *HGC-QGIS-Project.qgs* opened in QGIS.

Figure 2 shows the *HGC-KML.kmz* file cited in Table 2 displayed in GoogleEarth, with a balloon opened for the populated place named *Qarağy*. This folder contains layers for populated places, districts, provinces, and rivers, all of which have all been turned on in the figure.

Figure 2: *HGC-KML.kmz* displayed in GoogleEarth

In addition to providing a means of specifying unique IDs for database entries, a place's GeoNames URL identifier can be opened in a web browser, thereby automatically displaying a GeoNames map with the populated place at its center. Figure 3 shows such a map for the example of Table 3, where *Olenevka* is the GeoNames place associated with Jankowski place named *Qarağy*. Note here that *Qarağy* is shown as an alternate name for Olenevka. See Appendix C for how to create alternate names for GeoNames gazetteer entries.

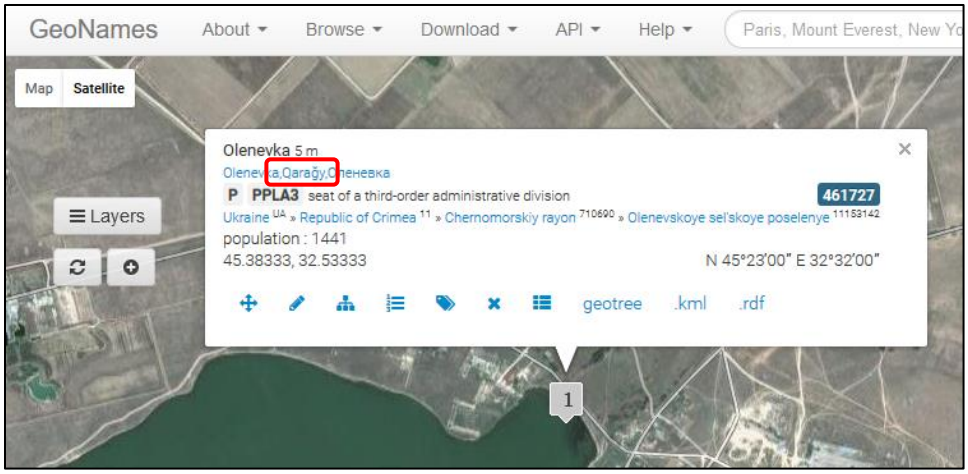

Figure 3: Opening the URL ID for *Qarağy* in a web browser.

Future Work

The HGC currently contains only places found on Jankowski's map that are included in his dictionary and that are also included in the GeoNames database. A next step in the development of the HGC is addition of the places shown on Jankowski's map that cannot be positively associated with places found in the current GeoNames database. Beyond this, additional sources will be used to add places included in in Jankowski's dictionary but not shown on his map.

Regarding the *Confidence* field of Table 3, populated places currently included in the gazetteer have high confidence due primarily to strong similarities between Jankowski toponyms and GeoNames toponyms. As the HGC is expanded to include Jankowski places that do not have GeoNames equivalents, using locations of surrounding existing places and topographical features such as rivers will allow approximation of the locations of these places, but the confidence level of such gazetteer entries will, of course, be lower. While assigning appropriate confidence levels to future database entries can be straightforward within the limited context of the HGC, it is highly-desirable that a clear, concise, consistent, and generally applicable means of quantifying confidence be developed so that it can applied beyond just the HGC. Ultimately, assignment of a particular confidence level should have the same meaning for all gazetteers.

Another issue that must be dealt with in future versions of the HGC is the necessity to accommodate time-variant many-to-many relationships between place names and locations. The use of unique URL IDs and mapping tools such as GoogleEarth allow differentiation and disambiguation for different places with the same name, but the HGC must ultimately accommodate time-variant multiple names for the same place.

The issue of time-variant many-to-many relationships among database entries leads to the long-term objective of formulating HGC databases as linked data[15] and incorporating HGC data into the sematic web[16]. The use of URLs for place identifiers, contributors, and information sources as illustrated in Table 3 anticipates this objective. An example of how linked data concepts could be incorporated into future versions of the HGC is shown in Table 4.

SUBJECT	PREDICATE	OBJECT	BEGIN_DATE	END_DATE	CONTRIBUTOR	SOURCE	PAGE	CONF
http://www.geonames.org/461727	TOPONYM	Qarağy	1450	1800	http://www.technologyforge.net/MHP	http://www.brill.com/historical-etymological-dictionary-pre-russian-habitation-names-crimea	106	1
http://www.geonames.org/461727	TOPONYM	Olenevka	1950	2017	http://www.technologyforge.net/MHP	http://www.geonames.org	461727	1
http://www.geonames.org/461727	LOCATION	32.53333, 45.38333	1450	2017	http://www.technologyforge.net/MHP	http://www.geonames.org	461727	1

Table 4: Example of possible linked data format for HGC.

As shown in the table, HGC data could be broken down into subject-predicate-object triples, where the subject field contains an entry's unique ID URL. This example shows how different toponyms as applied over different time spans can be applied to Jankowski's place named *Qarağy*.

[15] https://www.w3.org/wiki/LinkedData
[16] https://www.w3.org/wiki/SemanticWeb

Appendix A – HGC Work Flow

In its role as test bed and use case for future OOP gazetteers, significant effort was placed on establishing an effective workflow for creating gazetteer databases. Here, the workflow for creating a populated places database is outlined. It is assumed that the reader has some familiarity with GIS systems such as ArcGIS or QGIS. The various layers created per the following can be seen in the QGIS project shown in Figure 1.

1. Obtain external data
    a. Download digital elevation model (DEM) for Crimea[6]
    b. Download GeoNames database for Ukraine[5]

2. Create QGIS project named *HGC-QGIS-Project.qgs*

3. Create base maps
    a. Load DEM and save as *HGC-Elevation*
    b. Load GeoNames database and save as *Crimea-GeoNames*
    c. Load *Google Physical* map layer
    d. Create *HGC-Rivers* line layer with attributes shown in Table 3
    e. Using *Google Physical* map layer and *HGC-Elevation* layer, trace rivers and save in *HGC-Rivers*
    f. After all rivers are traced, search GeoNames database for rivers and use GeoNames URLs as river IDs
    g. If a river cannot be found in GeoNames, create new GeoNames entry for the river per Appendix C

4. Add populated places
    a. Create *HGC-Populated-Places* point layer with attributes shown in Table 3
    b. Visually search *Crimea-GeoNames* layer for places associated with places on Jankowski's map
    c. When a match is found, create point feature for the place in *HGC-Populated-Places* layer
    d. When all possible Jankowski places are found
        i. Open GeoNames website
        ii. Copy GeoNames toponym from *Crimea-GeoNames* and paste it into GeoNames search box
        iii. Do search, copy GeoNames URL from search results, and paste it into the place ID of the associated *HGC-Populated-Places* entry
    e. If the GeoNames database does not include the Jankowski name as an alternate name, add the Jankowski name to the GeoNames database per Appendix C

5. Create additional databases
    a. In QGIS, save *HGC-Populated-Places* shape file as KML file (see Appendix B)
    b. In QGIS, save HGC-Populated-Places shape file as text file

Appendix B – Generating KML Files

As described in Appendix A, HGC databases originate as shape files, with QGIS having an option to save files in KML format[17]. For the KML files provided as part of the HGC, the QGIS *Save as - Datasource Options* of Figure B1 were chosen. When opening a file in Google Earth, these options cause a place's *TOPONYM* and *ID* to appear in a balloon per Figure 2. Clicking the ID URL opens up a GeoNames map like that shown in Figure 3.

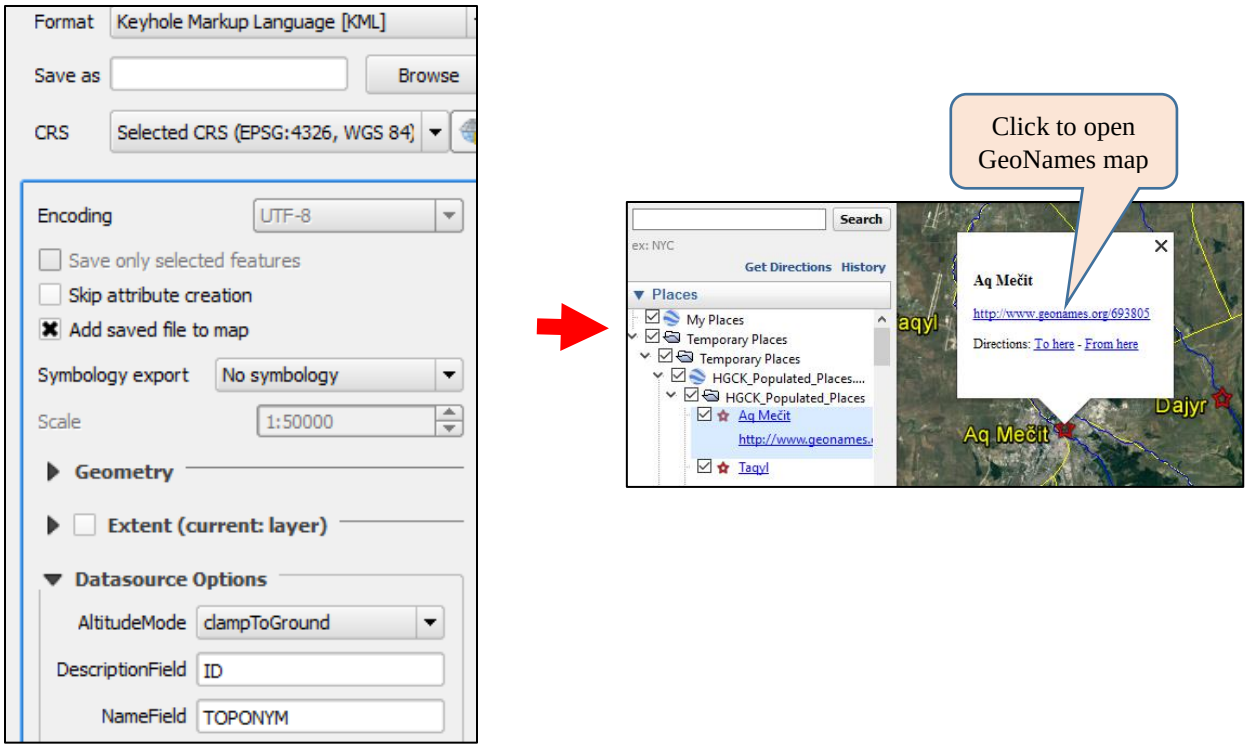

Figure B1: QGIS options chosen when saving a shape file as a KML file.

The more informative balloon description shown in Figure B2 can be obtained by using the *KML2KML.xlsm* Microsoft Excel Visual Basic application generated as part of the HGC and available for download via the HGC website. Note that in addition to providing a more detailed description of a place, this application generates a time slider which causes places to appear and disappear on the map according to their database BEGIN_DATE and END_DATE values.

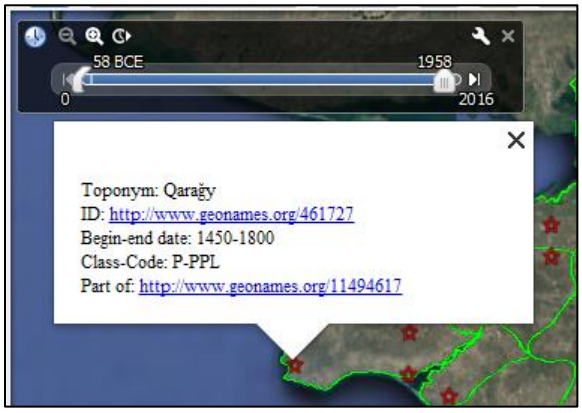

Figure B2: Balloon description and time slider available using the KML2KML.xlsm application.

[17] Note that KML files can also be generated in QGIS using the MMQGIS plug-in available at the Quantum GIS Plugin Repository (plugins.qgis.org).

Appendix C – Using the GeoNames Database

When using the GeoNames database to locate places on Jankowski's map and to create unique IDs for gazetteer entries, two situations can arise: 1) the GeoNames toponym does not match Jankowski's toponym; 2) there may be no place in the GeoNames database that corresponds to a place to be included in the HGC database.

For the first condition, consider the example where Jankowski's dictionary cites *Qarayurt* as corresponding to the GeoNames place *Dolinka*[18]. Figure C1 shows a search of the GeoNames database on *Dolinka*, and Figure C2 shows that *Qarayurt* is not listed as an alternate name for this place.

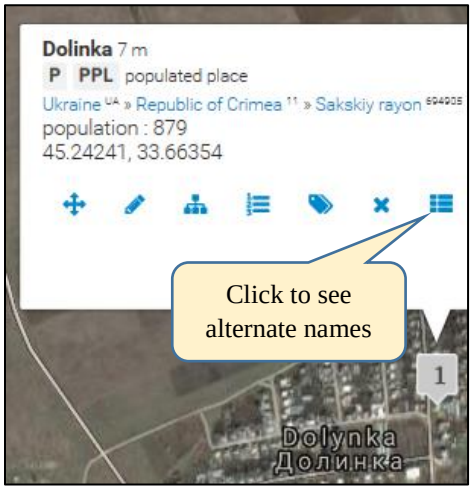

Figure C1: GeoNames search on *Dolinka*. Click as shown here to see alternate names for *Dolinka*.

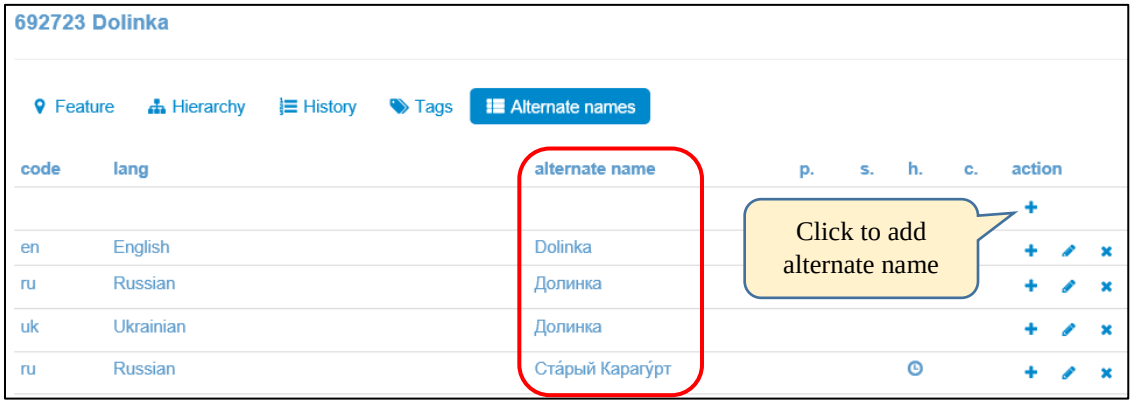

Figure C2: Alternate names for *Dolinka*. Click as shown to add an alternate name for *Dolinka*.

Figure C3 shows the alternate name information to be added to the GeoNames database, where *crh* indicates the toponym language is Crimean Turkish and *h* designates this to be a historical name.

[18] Henryk Jankowski, Historical-Etymological Dictionary of Pre-Russian Habitation Names of the Crimea, Brill, Leiden, 2006, p 827.

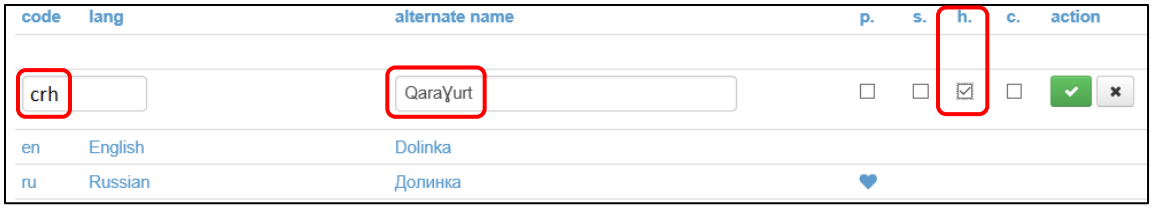

Figure C3: Adding *Qaraɣurt* as an alternate historical name for *Dolinka*.

Rivers provide convenient landmarks when locating populated places and administrative boundaries, but a number of Crimea rivers were not included in the GeoNames database when work on the HGC started. This represents the second condition cited at the beginning of this appendix.

Figure C4 provides the first step in adding new places to the GeoNames database. Open a web browser to the GeoNames map at http://www.geonames.org/v3 and then navigate to the location where a new place is to be added. Click on the add location icon, right-click on the map at the location of the new place, and choose *Insert new name here*. This opens up the box shown in Figure C5, where you provide a name, class, and code for the new place.

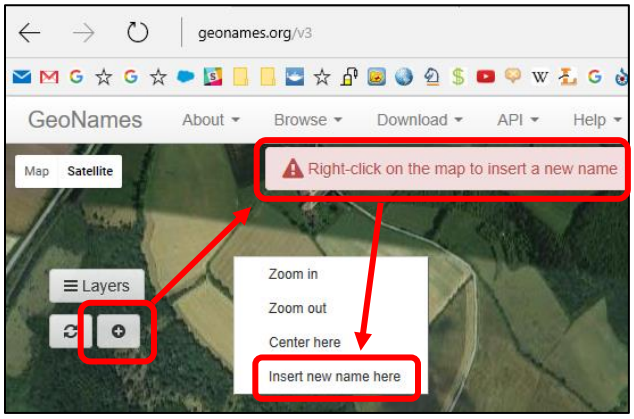

Figure C4: Adding a place to the GeoNames database.

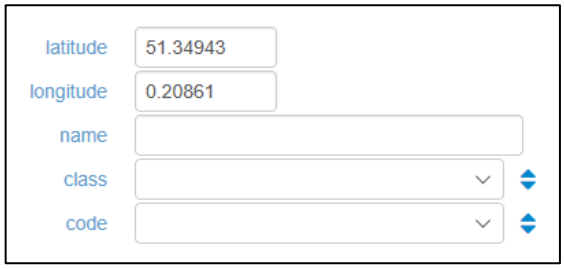

Figure C5: Naming the new GeoNames place.

Clicking the code and class dropdown box in Figure C5 reveals available values, with the specific values used for HGC database entries having been described in the main body of this discussion.