Historical Gazetteer of the Crimea

Mark Polczynski ~ Orcid-0001-5427-7601 Michael Polczynski ~ Orcid-003-2098-9461

Introduction

Henryk Jankowski's historical-etymological dictionary of pre-Russian habitation names in Crimea[1] contains toponyms of approximately 1,300 unique habitations that existed in this region prior to Russian annexation in 1783. The Historical Gazetteer of Crimea (HGC) is an open-access georeferenced gazetteer of historical places derived from Jankowski's work. The HGC includes 135 habitations, 43 districts, 6 provinces, and 17 rivers shown on Jankowski's map accompanying his book. Here we provide an overview of the HGC, starting with a review of the information sources used to create the gazetteer, followed by descriptions of the HGC datasets. The HGC dataset format is then presented, along with directions on how to access the dataset. Next, implications of associating gazetteer places with the GeoNames database are discusses. This document concludes with suggestions for future work.

The authors encourage students, researchers, and the public to use HGC materials in their investigations of this historically and currently critical region of the world. To this end, all HGC materials are being made available through a Creative Commons CC0 1.0 Universal Public Domain Dedication license. For geographers and historians in general, the HGC also provides an example of an approach to creating an open-access georeferenced gazetteer of historical places.

Data Sources

Jankowski's dictionary of pre-Russian places in Crimea contains over 1,300 unique dictionary entries for populated places, plus a map showing the locations of 135 habitations. Of the habitations shown on Jankowski's map, 104 have been positively associated with existing places contained in the GeoNames database (www.geonames.org), with the HGC gazetteer's longitude and latitude location of these places being taken from GeoNames. The location of the remaining 33 map habitations were approximated using surrounding landmarks.

Jankowski's map shows boundaries for six provinces (kaymakamlyk) and 43 districts (kadylyk). Polygons outlining boundaries of these places traced from Jankowski's map are included in the HGC. Locations for the province and district entries in the HGC gazetteer are the polygon centroids. Jankowski's map shows 17 rivers that have also been associated with GeoNames entries and are included in the HGC as points, with the location of these point features being at the mouth of the river.

The HGC gazetteer is augmented with two base maps. The courses of the 17 rivers shown on Jankowski's map plus 56 additional Crimea rivers are traced from OpenStreetMap (https://www.openstreetmap.org). Terrain elevation data (https://www.earthdata.nasa.gov/sensors/srtm) for Crimea allows terrain features such as valleys and ridge lines to be used when approximating habitation, district and province boundaries.

Dataverses and Datasets

As will be described later in this document, all HGC materials are available through a Harvard Dataverse Repository (https://dataverse.harvard.edu/dataverse/harvard). The Harvard repository is structured as *dataverses* and *datasets*. A *dataverse* is a container for *datasets*. A dataverse functions like a folder or collection that can hold multiple datasets, as well as other dataverses (sub-dataverses). Each dataverse can be customized with its own branding, metadata, and access settings. Dataverses are often used to organize and share related datasets. A *dataset* is an actual collection of data within a dataverse. It includes the data files, metadata, documentation, and any related materials. Datasets are the fundamental units of data storage in the Harvard Dataverse Repository.

The structure of the HGC dataverse is shown in Table 1. Under the top-level HGC dataverse are two dataverses – one for HGC gazetteer materials and one for Crimea base maps. Under the gazetteer dataverse are dataverses for the four elements of the gazetteer: habitations, districts, provinces, and rivers. Under each of these dataverses are three datasets: a .zip folder containing files needed to open the shape file version of the dataset in applications such as ArcGIS and QGIS, an OpenDocument spreadsheet version of the dataset, and a .kmz folder that can be opened in applications such as GoogleEarth. Figures 1, 2, and 3 show examples of each of these datasets for the three habitation datasets. The two datasets in the base maps dataverse are shown in Figures 4 and 5.

---

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

Dataset Format

All of the datasets in the Gazetter dataverse conform to the format shown in Table 2. Per the table, each entry has a unique sequential *ENTRY* value starting at 1 and going through 201 (135 habitations + 43 districts + 6 provinces + 17 rivers = 201 entries). The *ID* field is different from the *ENTRY* field in that each *place* (vs. each *entry*) has a unique *ID* value. This accommodates situations where a particular place may be associated with multiple entries as would occur if the place has multiple spellings or appears on multiple maps. Including both *ENTRY* and *ID* fields is one way to handle the many-to-many situations that often occur in digital gazetteers[2].

The *TOPONYM* field contains place names as shown on Jankowski's map and utilizes 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. All toponym spellings use Unicode-8 characters. Table 3 shows the special characters used in Jankowski's work and the Unicode hex codes for each (needed to search datasets for toponyms with special characters).

START_YEAR contains the approximate start of the Crimean Khanate, and *END_YEAR* is the year of Russian annexation. As noted, *LONGITUDE* and *LATITUDE* entries are those of the associated GeoNames dataset entry or locations approximated using surrounding landmarks.

The *NAME CONFIDENCE*, *DATE CONFIDENCE*, and *LOCATION CONFIDENCE* fields reflect the confidence level associated with *TOPONYM*, *START_YEAR*, *END_YEAR*, *LONGITUDE*, and *LATITUDE* entries, respectively[3].

CLASS and *CODE* values conform to the GeNames definitions (https://www.geonames.org/export/codes.html)[4]. *GEONAMES* contains a link to the GeoNames dataset entry for the associated place, with *(none)* being entered when no such association could be established. *PART_OF* contains the GeoNames link of the entity in which the place is found. So, for example, the *PART_OF* value for a habitation is the GeoNames link of the HGC district in which the habitation is located.

SOURCE is the Worldcat entry for Jankowski's book, and *LOCATION* is the page in the book where the dataset entry can be found. *SOURCE* is *(none)* for districts, provinces, and rivers, since these are not contained in Jankowski's book, only on his map. Finally, *CONTRIBUTOR* contains a link to the person providing the associated dataset entry.

Accessing and Using HGC Datasets

As noted, the HGC datasets and support materials are being made available through a Harvard Dataverse repository (https://dataverse.harvard.edu/dataverse/HGC). A video describing how to use the repository is available at: https://youtu.be/Vf0dhAAhAV0.

All HGC materials are licensed with a Creative Commons 0 Public Domain Dedication Waiver (https://creativecommons.org/publicdomain/zero/1.0/), which means that all materials are released into the public domain to the greatest extent possible under the law, allowing for free use, reuse, and distribution without any restrictions or need for attribution.

Association With GeoNames Database

Association of HGC elements with the GeoNames database has several advantages. First, use of the GeoNames class and code nomenclature aligns HGC feature types with a rich and well-conceived ontology of categories, relationships, and properties of spatial entities. Further, use of GeoNames feature longitude and latitude reduces the potential for location ambiguity by using an existing location designation. Beyond these advantages, GeoNames is editable by users, thus allowing Jankowski's toponyms to have been added to places positively associated with

---

[2] For a discussion of how to handle many-to-many relationships in gazetteers, see Polczynski, M., & Polczynski, M. (2021). Lessons learned from using historical maps to create a digital gazetteer of historical places. International Journal of Cartography, 8(3), 326–342.

[3] For a full description of confidence levels as used in the HGC gazetteer, see Polczynski, M., & Polczynski, M. (2019). Beauplan's Ukraine: open access georeferenced databases for studies of early modern history of Central and Eastern Europe. Miscellania Geographica, 23(3), 185-193.

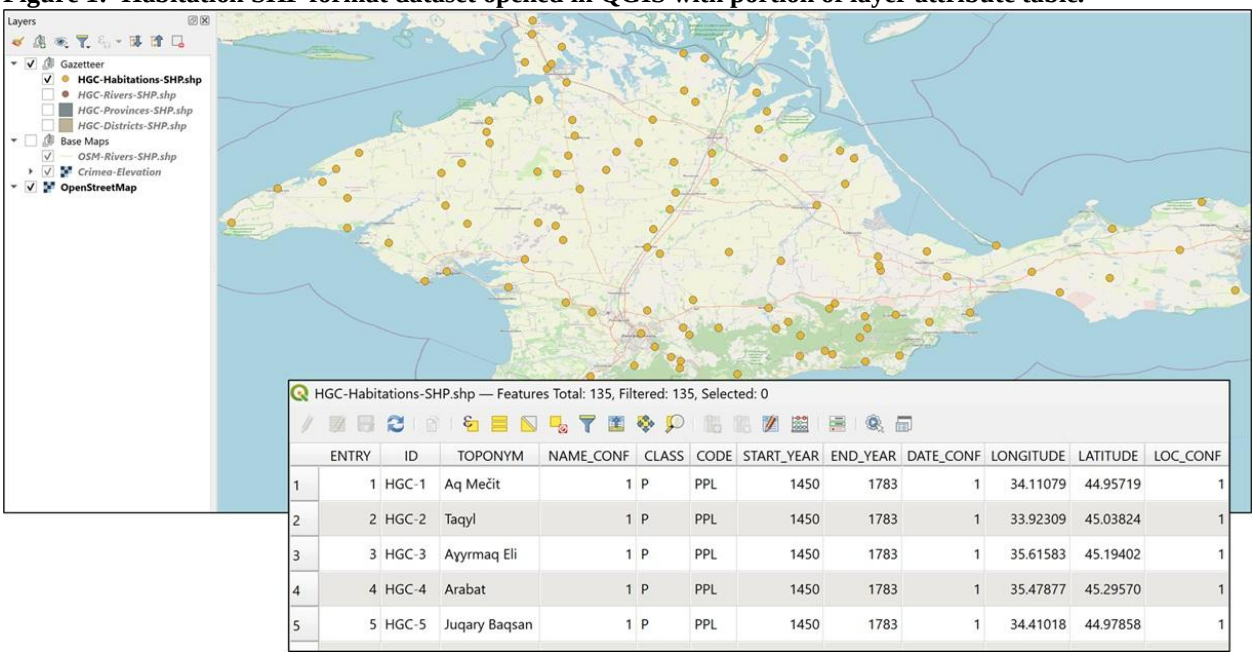
[4] The code used for districts and provinces is *A* for administrative region, and the class is *ADMxH* for historical administrative divisions, where *x* represents the administrative level. The HGC uses the codes *ADM4H* for districts (*kadylyk*) and *ADM3H* for provinces (*kaimakamlyk*). Extending this approach, the codes *ADM2H* and ADM1H can be used for the Crimean Khanate and the Ottoman Empire, respectively. The GeoNames code ADMH5 is also available for any administrative levels below the HGC district level, should such information become available.

existing places in the GeoNames database as part of HGC Gazetteer development. When adding toponyms, GeoNames allows a date range to be specified. This means that Jankowski's toponyms added to GeoNames with appropriate dating information appear as alternate names in GeoNames searches using modern names of places. Finally, this ability to accommodate multiple alternate names for places eliminates the need to include alternate names for entries in the HGC gazetteer.

Future Work

Two future work directions present themselves. Obviously, the over 1,300 habitations included in Jankowski's book but not shown on his map could be incorporated into the HGC gazetteer. Beyond that, the gazetteer format illustrated in Table 2 could be improved and generalized to conform to current and anticipated future gazetteer best practices, thereby making the HGC gazetteer interoperable with other gazetteers. For example, it is highly desirable that a clear, concise, consistent, and generally applicable means of quantifying confidence levels for gazetteer toponym, location, and date field entries be developed so that ultimately assignment of a particular confidence level would have the same meaning for all gazetteers.

Figure 1: Habitation SHP format dataset opened in QGIS with portion of layer attribute table.



	ENTRY	ID	TOPONYM	NAME_CONF	CLASS	CODE	START_YEAR	END_YEAR	DATE_CONF	LONGITUDE	LATITUDE	LOC_CONF
1	1	HGC-1	Aq Mečit	1	P	PPL	1450	1783	1	34.11079	44.95719	1
2	2	HGC-2	Taqyl	1	P	PPL	1450	1783	1	33.92309	45.03824	1
3	3	HGC-3	Ayyrmaq Eli	1	P	PPL	1450	1783	1	35.61583	45.19402	1
4	4	HGC-4	Arabat	1	P	PPL	1450	1783	1	35.47877	45.29570	1
5	5	HGC-5	Juqary Baqsan	1	P	PPL	1450	1783	1	34.41018	44.97858	1

Figure 2: Portion of habitation ODS format dataset opened in Excel.

	A	B	C	D	E	F	G	H	I	J	K	L
1	ENTRY	ID	TOPONYM	NAME_CONF	CLASS	CODE	START_YEAR	END_YEAR	DATE_CONF	LONGITUDE	LATITUDE	LOC_CONF
2	1	HGC-1	Aq Mečit	1	P	PPL	1450	1783	1	34.11079	44.95719	1
3	2	HGC-2	Taqyl	1	P	PPL	1450	1783	1	33.92309	45.03824	1
4	3	HGC-3	Ayyrmaq Eli	1	P	PPL	1450	1783	1	35.61583	45.19402	1
5	4	HGC-4	Arabat	1	P	PPL	1450	1783	1	35.47877	45.2957	1
6	5	HGC-5	Juqary Baqsan	1	P	PPL	1450	1783	1	34.41018	44.97858	1

Figure 3: Habitation KMZ format dataset opened in GoogleEarth with example feature attributes.

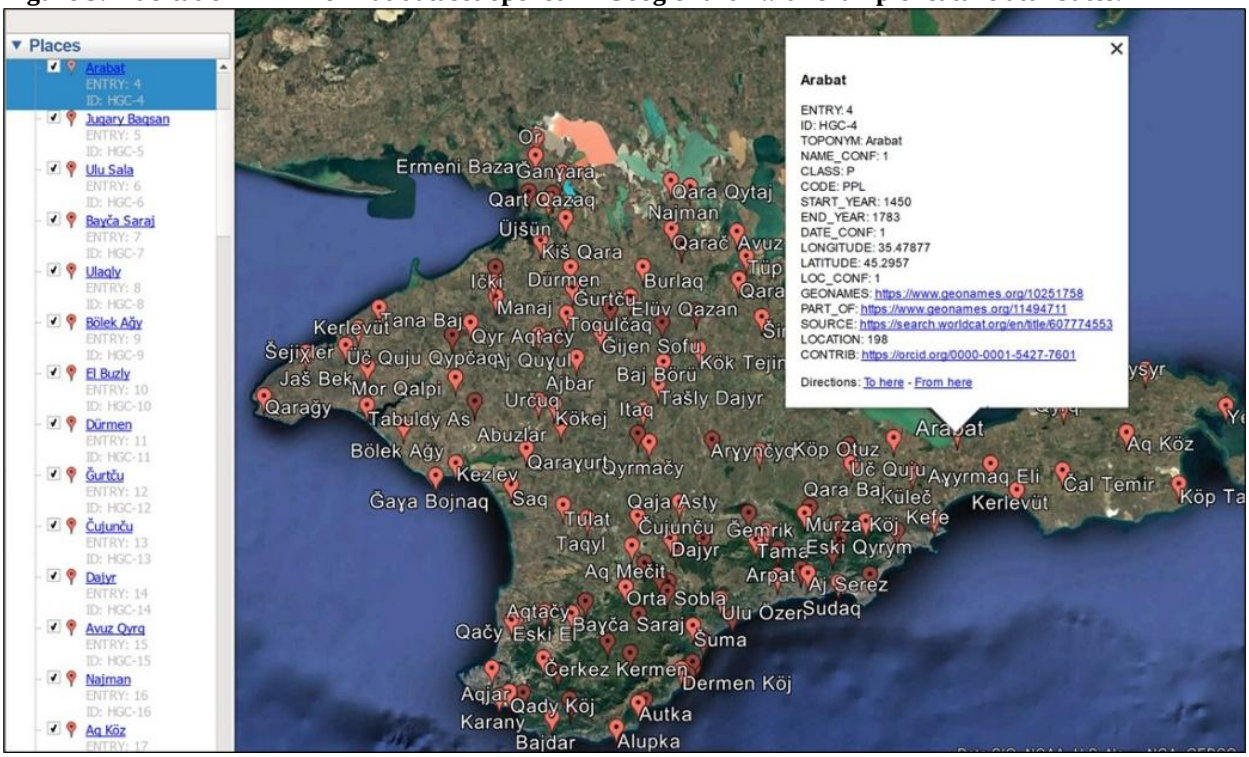


Figure 4: OSM Rivers SHP format dataset and Crimea Elevation tif format datasets in opened in QGIS.

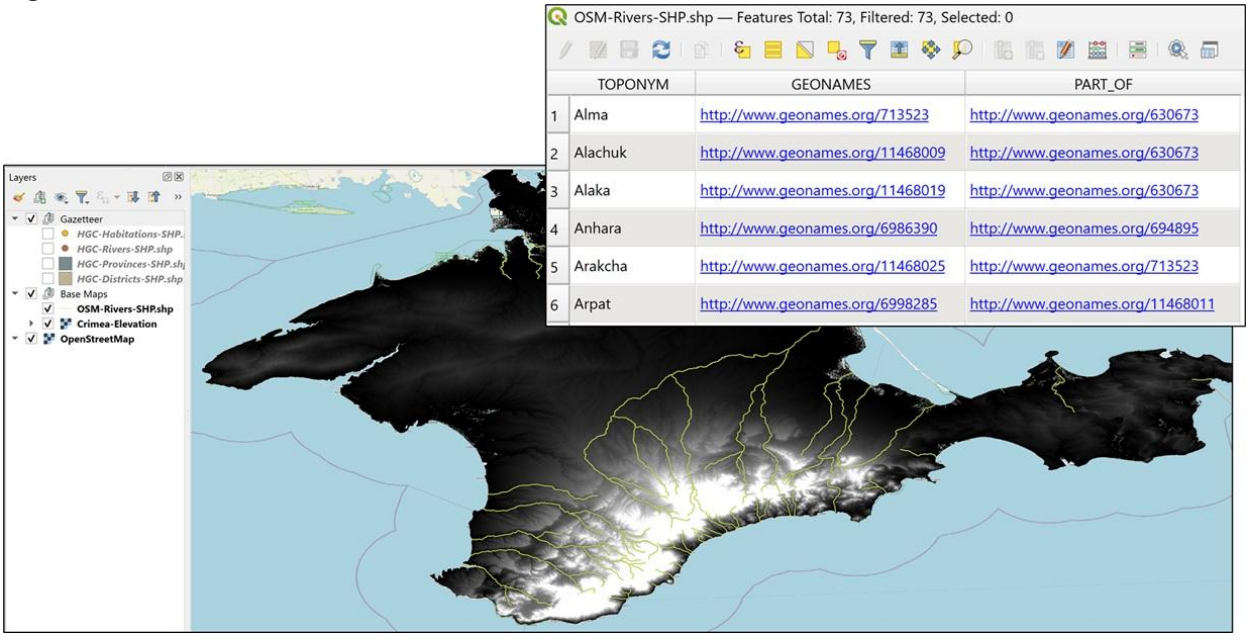


Figure 5: OSM Rivers KMZ format dataset opened in GoogleEarth with example feature attributes.

Table 1: Dataverse Structure - Red = dataverse: Green = dataset

HGC – An Historical Gazetteer of Crimea: https://dataverse.harvard.edu/dataverse/HGC

- Gazetteer - https://dataverse.harvard.edu/dataverse/HGC-Gazetteer
  - Habitations - https://dataverse.harvard.edu/dataverse/HGC-Gazetteer-Habitations
    - HGC-Habitations-SHP.zip
    - HGC-Habitations-ODS.ods
    - HGC-Habitations-KMZ.kmz
  - Districts - https://dataverse.harvard.edu/dataverse/HGC-Gazetteer-Districts
    - HGC-Districts-SHP.zip
    - HGC-Districts-ODS.ods
    - HGC-Districts-KMZ.kmz
  - Provinces - https://dataverse.harvard.edu/dataverse/HGC-Gazetteer-Provinces
    - HGC-Provinces-SHP.zip
    - HGC-Provinces-ODS.ods
    - HGC-Provinces-KMZ.kmz
  - Rivers - https://dataverse.harvard.edu/dataverse/HGC-Gazetteer-Rivers
    - HGC-Rivers-SHP.zip
    - HGC-Rivers-ODS.ods
    - HGC-Rivers-KMZ.kmz
- Base Maps - https://dataverse.harvard.edu/dataverse/HGC-Base-Maps
  - Crimea-Elevation.tif
  - OSM-Rivers-SHP.zip.zip
  - OSM-Rivers-KMZ.kmz

Table 2: HGC gazetteer dataset fields

Field Name	Example
ENTRY	1
ID	HGC-1
TOPONYM	Aq Mečit
NAME CONFIDENCE	1
CLASS	P
CODE	PPL
START YEAR	1450
END YEAR	1783
DATE CONFIDENCE	1
LONGITUDE	34.11079
LATITUDE	44.95719
LOCATION CONFIDENCE	1
GEONAMES	https://www.geonames.org/693805
PART OF	https://www.geonames.org/11494769
SOURCE	https://search.worldcat.org/en/title/607774553
LOCATION	184
CONTRIBUTOR	https://orcid.org/0000-0001-5427-7601

Table 3: Special characters used in Jankowski's work.

Character	Ç	ç	Ö	ö	Ü	ü	Č	č	Ğ	ğ	Ŋ	ŋ	Š	š	Ɣ	ɣ	Χ	χ
Unicode	00C7	00E7	00D6	00F6	00DC	00FC	010C	010D	011E	011F	014A	014B	0160	0161	0194	0263	0387	03C7