United Nations Group of Experts on Geographical Names

UNGEGN World Geographical Names Database

DRAFT Position Paper on the development of a global unique identifier for cities

By the UNGEGN Working Group on Geographical Names Data Management

Paper Structure

Timeline and Actions

#	Task/Action	Responsibility	Status
1	Discussion of ideas for a unique identifier for cities concept for the WGN DB with colleagues from DE (BKG) and NL (Dutch Kadaster)	Convenor WG GNDM	Done
2	DRAFT Position Paper on the development of a unique identifier for cities, version 0.1	Convenor WG GNDM	Done
3	Discussions about the WGN DB and the DRAFT Position Paper with UN SD / DESA and UNGEGN Bureau members, version 0.3	UN SD	Done
4	Exchange of the DRAFT Position Paper, version 0.4, with the WG GNDM	Convenor WG GNDM	Done
5	Update of the DRAFT Position Paper, version 0.5	Convenor WG GNDM	Done
6	Discussions about the WGN DB and the updated DRAFT Position Paper with UN SD / DESA and UNGEGN Expanded Bureau members, version 0.6-0.9	UN SD	Done
7	Circulate version 0.9 to • UNGEGN WG GNDM --- for information • UNGEGN Expanded Bureau --- for information and discussion at its next meeting in February 2024 • Expert Group on the Integration of Statistical and Geospatial Information (EG ISGI)	Convenor WG GNDM	January 2024 (most probably)
8	Endorsement of the Position Paper by the bodies mentioned under #7	Convenor WG GNDM	Q1/2024

Done	In progress	Delayed

The WG GNDM seeks the review and comment of relevant groups on the concepts outlined within this paper. Please contact Mr Pier-Giorgio Zaccheddu (Pier-Giorgio.Zaccheddu@bkg.bund.de) and Mr Mark Iliffe (mark.iliffe@un.org) with your comments.

1. Introduction

The revised United Nations Group of Experts on Geographical Names (UNGEGN) World Geographical Names Database (WGN DB) is a reliable and extendable technical platform that promotes and demonstrates how geographical names connect people to geographic locations. It also highlights why standardised geographical names preserve and protect the culture, heritage, and languages of nations. Further the WGN DB enhances the capability of the Group of Experts to implement its strategic plan and programme of work for the period 2021-2029, including managing authoritative short and full country names of 193 United Nations Member states and two non-Member States observers, their capitals, and major cities (with population over 100,000), and other types of geographic names in a multilingual, multiscriptual and geo-referenced format. National Names Authorities provide authoritative city endonyms and sound files, which are being added to assist users with pronunciation. The WGN DB contains endonyms, and exonyms of geographical features, detailing over 2,700 country names, some 6,100 names for 3,362 cities, with more than 970 audio files.

The UNGEGN WGN DB was initiated in July 2004 at the Group of Experts' twentieth-second session, guided by the recommendation that asked the UNGEGN Secretariat to take the lead in developing an authoritative database on country and major city names[1]. The WGN DB was to be made accessible from UNGEGN's website and the information would help to respond to toponymic questions received by the UNGEGN Secretariat and would provide a vehicle for countries to have their city names displayed in standardised form within a worldwide framework. The creation of the WGN DB was subsequently supported with resolution IX/6, adopted at the ninth Conference on the Standardization of Geographical Names, in August 2007.

Further, at its 2021 session, the Group of Experts adopted decision 2/2021/2[2], which "appreciated the work done by the secretariat in maintaining the website of the Group of Experts and reviving the online WGN DB, noted that the completion of those modifications and access to the database as soon as possible would be of great benefit". In the intersessional period, up to and beyond the 2023 session of the Group of Experts, the Working Group on Geographical Names Data Management (WG GNDM), supported by convenors of other functional Groups of the Group of Experts and the UNGEGN Secretariat, have continued working on the WGN DB's data model, with one outcome of this work discussed in this paper; while there are several unique identifier systems for countries (including M49 and ISO 3166 standards), there is no unique identifier system for cities. Thus, by examining good practices of data management, principles of Linked Data, and other relevant concepts through the lens of geographical names, this paper discusses factors around a unique identifier for cities as a means of responding to Recommendation 2[3] and Decision 3/2023/17[4] of the Group of Experts and recommends entrusting this decision to the Secretariat, based on available resources and technical capacity of the WGN DB.

2. The WGN DB and the Group of Expert's Mandate from the 2023 Session

In 2023, the Group of Experts was invited to consider and provide further guidance regarding the development of a unique identifier for cities and to express its views on how Linked Data and the semantic web can be integrated into the WGN DB. During this work, it was discovered that cities were not covered by existing UN geographical/area code standards, such as the case for countries, cf. the M49 Standard. Given that cities (generally with populations over 100,000) are one of the two primary data sources managed in the WGN DB, this needs to be addressed. Subsequently, the development of a "unique identifier for cities" was taken on by the WG GNDM, including this as part of Action Item 1-i-5 of its workplan ("Promote and support the establishment of an authoritative international

[1] see GEGN/22 Report of the United Nations Group of Experts on Geographical Names on the work of its twenty-second session
[2] see E/CONF.98/136 Ninth United Nations Conference on the Standardization of Geographical Names
[3] see E/2023/84 Report of United Nations Group of Experts on Geographical Names on its 2023 Session
[4] Specifically items (i) welcomed the concept of a unique and standardized identifier for cities and other geographical features within the Database; and, (ii) requested UNSD, within available resources, to enhance the functionality of, and data in, the Database

database/gazetteer of geographical names to be disseminated to the public through UNGEGN")[5]. The unique identifier concept for cities needs further consideration and elaboration, including developing a set of conditions and prerequisites to technically realise the concept, supported by a unique identifier concept for a database structure which is as flexible (and extendable) as possible.

Looking forward, the WG GNDM aims to promote the WGN DB to relevant working groups of the Group of Experts and other subsidiary bodies of the Economic and Social Council, including the Statistical Commission and the UN Committee of Experts on Global Geospatial Information Management (UN-GGIM), following the development of a consistent and reliable unique identifier for UUID for cities and present its progress to the Group of Experts at its 2025 session, and then other relevant entities and groups deemed appropriate.

3. Background and prerequisites of the World Geographical Names Database

3.1 General purpose of the WGN DB

The primary objective of the WGN DB suite of applications is to serve as a tool for managing geographical name data for cities and countries with the aim of providing UNGEGN experts, governments, and the public at large, access to an authoritative global repository of standardised geographical names of countries, capitals and other cities (primarily those with populations over 100,000). Given the database structure, pending the availability of resources and the needs of users, it will be possible in the future to extend the scope of WGN DB to include the names of indigenous places and geographical features of specific national importance. However, this would require the ownership of relevant bodies to define appropriate data workflows and other cartographic concerns.

3.2 Users of the WGN DB

The main users of the WGN DB are geographers, toponymists, cartographers, planners, researchers, official and public institutions, Non-Governmental Organisations (NGOs), and the general public wishing to use the names of countries and major cities. In the short term, the integration of geographical name data from UN System sources will focus on the Demographic Yearbook (which contains toponymic data provided by Member States for the purpose of dissemination of demographic statistics), as well as the UNTERM[6] database. In time, the is the aim to integrate with other authoritative UN sources such as the Second Administrative Level Boundary (SALB) initiative, the UN Gazetteer, and the UNLOCODE database, but this will be dependent on resourcing.

3.3 The journey of the WGN DB

A roadmap for the implementation of the WGN DB was presented on 4 May at the 2023 UNGEGN session. The Roadmap[7] consists of three parts:

1) **WGN DB launch** (completed at the 2023 Session of UNGEGN, following Recommendation 2 and Decision 3/2023/17 of the 2023 Session see E/2023/84)
2) **Data integration**: Ensure full geographic completeness of all countries and cities in scope (i.e. currently of over 100,000 population). This will be achieved through integrating, following quality review and assurance, toponymic data provided by (1) Member States, and (2) other authoritative sources from across the UN System.
3) **Data governance, feature improvement and future development**. This paper articulates some of the data governance needs (i.e. a UUID), but the majority of feature improvement and development will be undertaken by the Secretariat within available resources.

[5] See report GEGN.2/2023/73/CRP.73
6 The UN Terminology database: https://unterm.un.org/unterm2/en/
[7] The WGN DB management responsibility for all three parts relies on UNSD – with the assistance of UNGEGN and contributions of data by the Member States

Moreover, it is hoped that further engagement with Member States will help identify national focal points to establish and/or enhance coordination and data update mechanisms. A point of note is that while the WGN DB system is extendable, current resources prevent the addition of system features.

4. Universal and unique identifier management considerations

4.1 Towards implementing globally unique identifiers for the WGN DB

Identifiers serve as a fundamental infrastructure component, facilitating the linkage of data across various domains and communities for collaborative, cross-disciplinary work. These identifiers must adhere to several key principles. Firstly, they should be **unique**, ensuring that no two objects share the same identifier and that each object possesses only one identifier. They also need to be **universal**, meaning every object should have an identifier. These **Universal Unique Identifiers (UUID)** should be **immutable**, retaining their original value throughout the object's lifecycle. These UUIDs are crucial to the database management. Once issued, an identifier should never be re-assigned to another object, even if the original object is decommissioned.

Unlike descriptive metadata, identifiers should be **meaning-independent**, relegating any interpretive information to separate, user-friendly textual descriptions. The governance of identifiers should be well-documented for interoperability across systems. Issuance should be centralised through an authoritative entity to maintain consistent assignment and management. Lastly, to maximise utility and integration, these identifiers should be readily available to all relevant systems, including third-party applications. There are various options to ensure the existence of globally unique identifiers, primarily through Uniform Resource Identifiers, Universal Unique Identifiers or a combination of both.

Using Uniform Resource Identifiers "URIs" as global resource identifiers
In the realm of web-based data resources, URIs (Uniform Resource Identifiers) play a critical role in providing globally unique identifiers. Structurally, a URI comprises four parts: [scheme]://[host]/[path]/[local identifier].

When building a robust system of URIs, there are various options for choosing a namespace (the "scheme]://[host]/" part), including:

a) Creating a namespace centrally controlled by the UNGEGN Secretariat (e.g., UN Statistics Division, https://unstats.un.org/...)
b) Establishing a new subdomain exclusive to the dataset (e.g., https://gwn.un.org/...)
c) Collaborate with a highly recognised partner organisation that already serves as a custodian for authoritative metadata (e.g., UN Dag Hammarskjold Library, https://metadata.un.org/...).

While each option had its merits, the first one option offers a way of centrally managing the namespace, without reliance on external parties, helping ensure consistency and reliability that is crucial for cross-disciplinary work and collaboration.

Below is one URI format using the UUID value as expressed below:
https://gwn.un.org/GN/NP/e41c7aacd5dd49eea604ff8c4f88a988

Using Universal Unique Identifiers "UUIDs"
UUIDs are a registry-style string of letters and numbers totalling e.g. 36 characters. The 36-character, or 128-bit integer, style of these identifiers means there is almost no chance they will ever be duplicated. *"It would take somewhere around a quintillion UUID values to have a 50% change of*

collision (or duplicate values), which ultimately means giving a table row a UUID value will allow it to remain unique for its entire lifespan."[8]

There are several services to establish UUIDs (see format below). These values can be formatted a few different ways depending on the software used and the presence/absence of uppercase letters, curly braces, and hyphens. Below are four different formats of the same UUID value:

- {E41C7AAC-D5DD-49EE-A604-FF8C4F88A988}
- {e41c7aac-d5dd-49ee-a604-ff8c4f88a988}
- E41C7AAC-D5DD-49EE-A604-FF8C4F88A988
- e41c7aacd5dd49eea604ff8c4f88a988

4.2 Assumptions for the globally unique identifiers (either URIs or UUIDs) of the WGN DB

The development and implementation of a globally unique identifier concept requires knowledge of the WGN DB model/structure and its content (see Annex 1) on the WGN DB Data model. Flexibility for the future to incorporate new forms of data of new geographies (for example indigenous names) is also a consideration. If the feature type 'NamedPlace' is the main object/feature that the identifier can refer to, each 'NamedPlace' would simply get one globally unique identifier and each data type 'GeographicalName' (together with its attributes) would be linked to it without additional information. If each data type 'GeographicalName' is associated would have a global unique identifier, and 'GeographicalName' would have separate attributes, these attributes could be linked separately to additional information. These latter considerations[9] would add complexity, but are worth noting.

5. Summary and Point of Decision

This paper has discussed concepts for the creation of a unique identifier for cities and identifies URIs and/or UUIDs as sustainable and suitable mechanisms for their creation. This is a necessary step that must be taken for the WGN DB to deliver on its mandates including Recommendation 2 and Decision 3/2023/17 of the Group of Experts.

Now, the WG GNDM requests the Secretariat, with existing resources and technical capacity to pilot a unique identifier based on the URI or UUID approach, in alignment with Recommendation 2 and Decision 3/2023/17.

[8] https://www.cloudpointgeo.com/blog/2022/1/17/is-your-data-unique

[9] An example is the attribute 'language'. If 'language' is a simple attribute ("string to the name") to each 'GeographicalName' without any linkage to further information on language a unique identifier for the 'NamedPlace' is sufficient. If 'language' shall provide additional information or be linked to additional information on the 'language', it gets more complex. Then, each 'GeographicalName' shall get a unique identifier.

6. Annex 1: The (new) WGN DB model

6.1 WGN Architecture

The WGN DB architecture is shown in Figure 1 *below.*

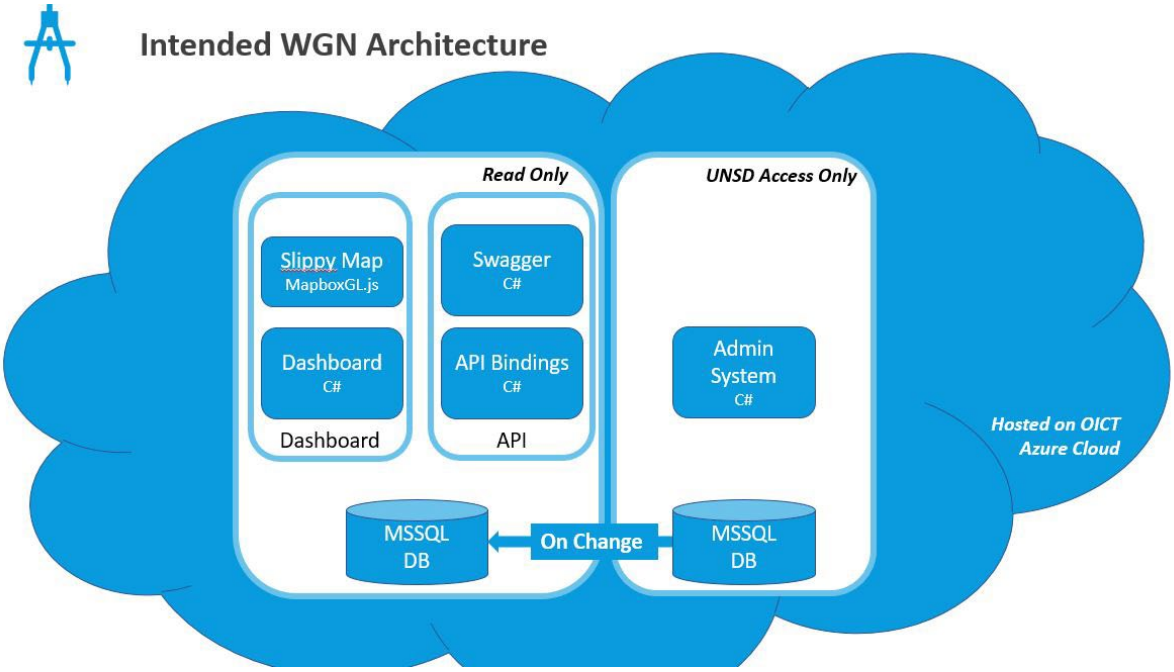

Figure 1: WGN DB Architecture

6.2 Core Data Schema

The core of the geographical names' application schema (UML class diagram) extends the core schema from the European INSPIRE specification (see Figure 2).

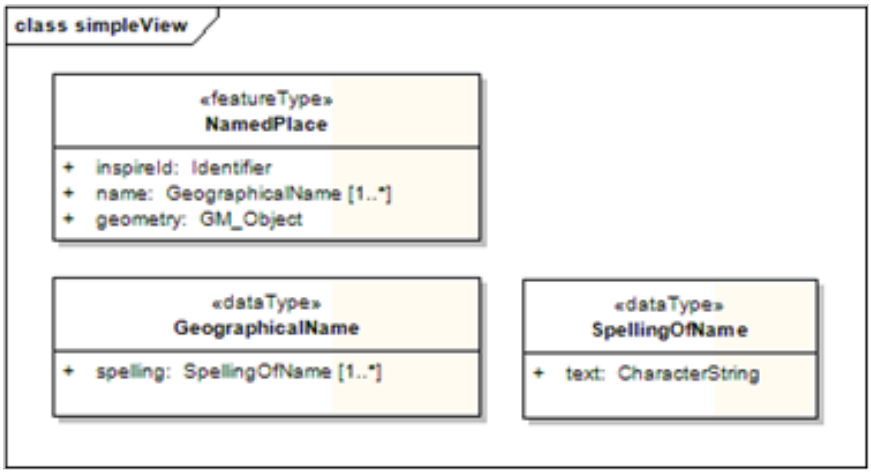

Figure 2: Simple view on the European GN data schema

The only feature type of the schema is 'NamedPlace' (main object / feature), representing any real-world entity referred to by one or several proper nouns.

- Each 'NamedPlace' is associated with one or several geographical names, i.e., proper nouns applied to the spatial object, modelled with the data type 'GeographicalName'. The different geographical names of one given spatial object may be, for example, the names in different languages or in different forms (e.g., complete and short forms of country and administrative unit names).

- Each 'GeographicalName' may have one or several spellings, i.e., proper ways of writing it, in one or several scripts like the Latin/Roman, Greek and Cyrillic scripts, modelled with the data type SpellingOfName.

For example:

- The city of Athens may be modelled in the schema as one 'NamedPlace'.
- The endonym "Athína" (Greek language) and exonym "Athens" (English language) are two different 'GeographicalName' of this unique 'NamedPlace'.
- "Αθήνα" (Greek script) and its standard romanisation "Athína" (Latin script) are two different 'SpellingOfName' of the same 'GeographicalName' "Athína".

Complete and precise definitions of the types and attributes must be determined for the WGN DB content. This is relevant for the UUID concept development.

6.3 Content (attribute) requirements[10]

As a repository for standardised names, including country names in all six UN languages, the WGN DB must provide the correct representation of all scripts, both on the map and in the database. A preliminary selection of potential mandatory and optional attributes is summarised in Table 1. This Table 1 highlights mandatory information and potential "future" considerations. It is intended that these mandatory fields will be confirmed as part of this review process, with future considerations allowing participatory and inclusive discussion on the ongoing direction of the WGN DB.

Country name	**Mandatory** • ISO 3-letter country and language codes • M49 Country and Geographic regions code • Short and formal country names in all six UN languages (Arabic, Chinese, English, French, Russian, and Spanish) • Short and formal endonym • For endonyms, the romanised form for non-Roman script names and romanisation system used • Name(s) of official/national language(s) • Country names geographic coordinates • Source • Date of last update
	Future Considerations • UNGEGN geographical / linguistic division (a country may be in more than one division) • UN Economic Division / Region • Audio files of country name • Historical names (changes over time) and time period
Names of cities	**Mandatory** • **UNGEGN UUID / URI City code** • Endonym(s) – as supplied by the country • Source of endonym • Name form(s) as used in all six UN languages (where these differ). • Geographic coordinates

[10] Based on UNGEGN Geo names requirements specification of 01.03.2021

	• Source of geographic coordinates • For endonyms, the romanised form for non-Roman script names and romanisation system used • Language identification • Date of last update • Whether it is a Capital City (or not)
	Mandatory • Country • Names of authority/authorities • Date founded/established
National Names Authority	**Future Considerations** • Country • Names of authority/authorities • Date founded/established • Legislation • Website • Source • Date of update former names of the authority • Toponymic guidelines • Existence of geographic names database and online access • Geographic/linguistic division(s)

Table 1: Example WGN DB data schema

Future Decision Point: Determine the content (attribute) information required for the WGN DB within an Online Consultation in 2024. Confirm (1) **mandatory as 'appropriate'** and (2) **agree on a prioritisation (and support with resources) of 'future considerations'**

Acknowledgements

The Convenor of the UNGEGN Working Group on Geographical Names Data Management (WG GNDM), Mr Pier-Giorgio Zaccheddu, Federal Agency for Cartography and Geodesy, GERMANY, expresses its gratitude and thanks to Ms. Alexandra Rowland, Dutch Kadaster, THE KINGDOM OF THE NETHERLANDS, as well as Ms. Regine Elling, Mr. Tim Trautmann and Mr. Jörgen Spradau, Federal Agency for Cartography and Geodesy, GERMANY, for their input and ideas to draft the first version 0.1 of the position paper (UUID concept).

Members of the UNGEGN Working Group on Geographical Names Data Management (WG GNDM) and UNGEGN Bureau, as well as staff members from UNSD and OICT, contributed to the position paper, which was finally agreed to in early 2024.