United Nations ECE/CES/2016/24

Economic and Social Council

Distr.: General
16 February 2016

Original: English

Economic Commission for Europe

Conference of European Statisticians

Sixty-fourth plenary session
Paris, 27-29 April 2016
Item 6 of the provisional agenda
Geospatial information services based on official statistics

A common international conceptual framework for geospatial and statistical data acquisition, data management, and data use: goals and barriers

Note by the United States Census Bureau

Summary

Statistical agencies have many of the same goals and barriers in sharing geospatially enabled data and developing best practices. The challenges span data collection, management and use. Issues of data acquisition include proposed common standards and frameworks, basic units of collection, alignment of geographic entities, and temporal cycles of data collection and updates. Methodologies of geospatial integration, coordination of geospatially enabled data, considerations of national and international legislation and policy, and assurances of confidentiality and privacy are the critical concerns of data management. Data use for geospatial analysis and area definitions (e.g., urban/rural) aligns with national and international trends in geographical and statistical modelling (e.g., smart cities). In addition, the international geospatial community continues to search for common ground on the issues of data integration between nations. This paper will discuss the impact of these issues on developing a common international conceptual framework for geospatial and statistical data acquisition, management, and use.

The paper is presented to the Conference of European Statisticians seminar on "Geospatial information services based on official statistics" for discussion.

GE.16-02220(E)

I. Introduction

1. While sharing and using diverse data within an organization is a manageable endeavour, attempting to extend this goal in a global context is a formidable challenge. The design, production, and use of integrated core statistical and geospatial datasets follows a predictable sequence as data from a survey lifecycle are used in a national and international context.

2. In 2014, Lars Backer from Sweden presented the United Nations Global Geospatial Information Management (UN-GGIM) challenge to "produce systems of qualified information to serve as foundation for direct and indirect government action on all levels of public authority from local to global." Each government has a survey lifecycle to collect, tabulate, disseminate, and analyse information collected by their National Statistical Agencies.

3. These national datasets serve as the basis for an integrated national and international dataset (or resource?) that combines statistical and geospatial data. What follows is disseminated data used by all levels of public authority from local to global. The formation of research networks and references disseminate this knowledge base. The public is sometimes involved in partnerships to improve collection and dissemination activities. Decision makers ultimately use these integrated core datasets for global sustainable development goals. A natural consequence occurs through the management, accountability, monitoring of global datasets, standards and best practices.

4. While many nations are successful at data integration within their borders, when the data needs to be integrated with neighbouring nations for a global view challenges are realized or are introduced.

II. Issues and challenges

A. Defining a common or standard framework

5. The successful combination of statistical and geospatial data in a global context is made possible by a common or standard framework to support the collection, integration and use of statistical and geospatial data. Statistical data usually are referenced geographically, but are not integrated with the geospatial data.

6. Geospatial data form the geographic infrastructure are used by innumerable programs and exist in their own domain. In addition, each data type can have its own framework within an organization that needs to be integrated into a common or standard framework to facilitate data sharing. This is crucial for success in a global context. This data integration process will uncover differences in the data that will require action to smooth the data for use. Owners of the data will need to determine priorities and if necessary make framework modifications. Without solving the problem at its source, repeatable corrections could lead to resource inefficiencies.

7. To achieve a common framework cooperation and compromises are required. One approach is to follow existing standards (of the International Standards Organization (ISO), Open Geospatial Consortium, etc.) However, it is conceivable that long established terms and definitions of an organization will require modifications in a global context. If changes impact the core mission of an organization, then a parallel data management approach for the affected circumstances may be required. To effectively manage this, metadata serves an important part in this process.

8. A formalized set of descriptive properties which is shared by a community to include guidance on expected structures, definitions, repeatability, and conditionality of elements is made possible by robust metadata. Information on sources, accuracy, precision, spatial characteristics such as the datum, coordinate system, and projection and data types are examples of metadata content.

9. The commitment to building and managing metadata is foundational for the integration of statistical and geospatial data. In a shared global environment, consensus is needed on basic statistical and geospatial content. For agreement on basic content, the integrators must ask what drives the decision. This quandary is easier to address now with agreement on the Sustainability Development Goals (SDGs). The SDGs help to focus separate interests and challenges. What remains is prioritizing how to proceed. To assist in that regard, determination of any mandated content elements are a first consideration. While this prioritization process evolves, specific content that adds value in the context of a global network is taken into account. For example, the success in the United Nations General Assembly's interest in the importance of a common geodetic reference frame serves as a model for potential agreements on core components of data integration.

B. Data at different levels of geography

10. In the statistical domain, data are collected and disseminated at different geographic levels. There is no standard geographic unit that exists where a global view of a statistical theme is accepted and used, aside from the national level. For example, some data collected within an enumeration district (ED), which, in the absence of individual person-based registers or housing unit addresses, is the smallest geographic unit. This ED may be the smallest level in which the data are disseminated or the data could be aggregated to a higher level of geography for reporting purposes.

11. Determination of a standard level of reporting is paramount for data use in a global context. Generally, the level of geography in which the data are collected determines the smallest geographic unit in which to report the data, barring data interpolation. As it currently stands, there is no common framework for administrative units. For example, legal government areas are determined by the laws and ordinances defining those areas. Cities vary greatly in geographic area, characteristics, and reported statistical values. Statistical areas like enumeration districts and census blocks vary based on criteria, but those criteria are oftentimes in the control of the statistical agency. It is conceivable that a geographic unit could be developed that either allows for direct use of existing geography or aggregates to a common geographic area.

12. Currently, most countries have their own statistical geography that does not align with that of their neighbours'. This is sometimes complicated by different levels of geography that do not always nest within each other by design and purpose. This is an area for discussion, proposals, collaboration and consensus.

C. Lowest common geographic denominator for collection and dissemination

13. Addresses are becoming increasingly important across the globe. In a national and global context an address (basic location information) is the most elemental geospatial unit. Addresses have characteristics that describe their type (housing vs. business vs. governmental, etc.). Location information is represented in some form of a coordinate system such as by latitude/longitude. Identifiers are commonly used to differentiate one address from another. For example, the address for a multi-unit structure with apartments

on each floor may be represented by a basic street address (123 Arbor Drive) which may comprise 14 separate units within the same building (all housing, all businesses or a mix of each type). Each unit has its own address identification to distinguish it from another address. The unit ID does not necessarily require that geography be built into it. In fact, it is preferable not to embed geography in the ID to minimize the maintenance required when boundaries change.

14. Geographic linkage occurs in matching the ID to a series of geographic codes (geocodes) that apply to the location of the address. This scenario is an example of a proposal for discussion, collaboration, and decision on a way forward for a global conceptual framework, beginning at the most elemental level and building upwards. There are differences in addresses and addressing systems.

15. These topics are part of the agenda on current efforts of international address standardization from a use perspective in the statistical and geospatial arenas. This does not necessarily take into account address data management from a mail delivery perspective. This is an example where defined priorities aid in managing the scope of attention to assure forward progress on the framework. For areas without housing unit and/or business addresses, this approach can serve as a topic for planning future requirements for that nation.

16. In the context of statistical and geospatial data, addresses are not limited to structures like housing units and businesses. Statistics are collected and disseminated on many different themes ranging from environmental to cultural to natural sciences. The locations of water wells and fire hydrants, while serving similar basic functions in providing water, may serve different purposes such as a source for drinking water vs. emergency response in putting out a fire.

17. Each of these examples (housing unit, business establishment, water well and fire hydrant) can serve as a basic unit of data collection in the statistical and geospatial realm.

D. Frequency of data availability

18. The frequency of data collection and the vintage of disseminated data are crucial in appropriate data use. Temporal characteristics of the data are a basic form of information often included in metadata. Having data without reference to its time stamp is problematic. The adage “something is better than nothing” is oftentimes true even with dated information, but users need to know the temporal state of the data for their planned use.

19. A determination on the frequency of collecting data is part of the planning process. In the case of censuses (for different topics such as population, economic, environmental and agricultural), most nations either have a census every ten years or five years. The longer the period, the more dated the information becomes unless it is supplemented by other sources of information such as a periodic statistical survey.

20. Decisions on frequency are determined by factors such as a nation’s constitution, laws, practice, and need. Cost sometimes affects plans for the extent and scope of data. This is important in integrating statistical and geospatial data. For various types of geospatial data, it must be determined if the collection of the data needs to align with the reference date of statistics. One could argue that earlier dates of land surveys used to determine elevation and slope are not as important as having synchronized legal boundaries for which statistical data are disseminated for cities and towns.

E. Enumeration procedures

21. In addition to how data is collected, the rules of enumeration help users better understand the condition of the data. For example, assuming complete coverage of a nation, a door-to-door count results in a full enumeration of the country. As labour is one of the greatest costs of a census, with time, this approach may not be sustainable for most countries. Another option includes mailing questionnaires which requires an up-to-date address list and also requires some form of field follow-up for any non-respondents. Participation in many censuses is accomplished through self-enumeration; in other words, the assumption is that the respondent is answering the questions truthfully and correctly. There are usually efforts at testing the geographic coverage for a complete count of the population without evaluating the correctness of the response.

F. Data sources

22. With geospatial data, a relatively new technique of crowdsourcing is made possible through access to GIS technology where interested individuals contribute to a geospatial database by adding missing information or by correcting existing data. Crowdsourcing is serving either as a source of geospatial data where there is an absence of information, as happened in Haiti and other locations where disasters took place, or is being used as an alternative to geospatial infrastructures as in the case of OpenStreetMap.

23. This type of volunteered geographic information (VGI) is obviously helpful and useful where the crowdsourced data is better (more complete, more current and/or more precise). However, this raises the question of the value of crowdsourced information versus authoritative information. Crowdsourcing has occurred during the absence of authoritative data and in cases where authoritative data existed. In this latter case, it is important to evaluate the reasons why existing authoritative data are not made available, particularly as increasing numbers of Nations support the notion of Open Data.

24. The state of geography, populations, economies, environment and so forth continually change which raises the question about the frequency of geospatial and statistical data update. New housing and new roads leading to housing developments occurs at irregular rates and diverse locations. Questions that need to be answered include: What is the process by which updates are made to include the new or altered data content? How often is information collected about the affected population to these changes? Do updates occur automatically where government levels responsible to managing the change report those instances to higher level organizations so there is an accurate set of information for the Nation? Or do updates occur periodically based on a prescribed schedule where that approach suffices?

25. There are important questions about data availability in which minimally, information about the plans, decisions and state of the data are known to the data user community.

G. Data integration

26. With geospatial data, where two different sources are integrated, oftentimes there are opportunities for data errors and inconsistencies. For example, one source is overlaid with another because data are combined together or adjoined at the edges of the geographic extent of one source with another.

27. Examples of challenges that might be encountered include the different accuracies of the data sources where obvious lines and areas do not line up thereby causing gaps, slivers

of new areas, and overlaps where lines extend beyond their true location. These differences require corrections. In most cases, the corrections involve visual inspection and some interactive steps using GIS software to apply the modifications. This takes time and is usually costly in terms of labour.

28. In integrating statistical data sets together, similar challenges surface. For example, are the definitions of statistical activity comparable? Is the vintage of each data set acceptable relative to the expected outcome? Before getting deeper into this topic, the notion of what is data integration needs to be first addressed. When two datasets are overlaid or combined, the process may be a simple draping of the data where one sees the effect of different data added to an existing set of data.

29. The integration and interpretation are visual. Mash-ups are an example of this approach. In a different context of data integration, data are unified by merging them together where conditions are imposed on the process to affect a result that alters aspects of the original sources. The changes occur through software based on, for example, business rules, interpretation of legal values from the metadata, and project requirements.

30. This approach requires greater control as the data are altered, in effect, creating new data. Integrating statistical and geospatial data uses both approaches, depending on the project objectives. As these are different data types, it is likely that additional data will be required when integrating them together. For example, if statistics apply to a geographic area not currently in the geospatial data set, the appropriate geographic area boundaries and their geographic relationships are needed to make the integration complete.

H. Confidentiality and privacy

31. Statistical organizations are familiar with confidentiality requirements for assuring that the identity of a respondent is protected from identification. With the advent of increasing amounts and exposure of geospatial information through various tools including those on the internet, questions about privacy have emerged from local to global concerns. Different policies and laws exist to protect individuals and their personal circumstances and these vary among different administrative levels.

32. In the statistical arena, there are established and documented practices for assuring confidentiality of the results. With geospatial data, concerns over privacy is an evolving situation. Questions are raised in advance of policies and laws. The challenge is an evolving environment, with questions often raised in legal proceedings prior to the existence of adequate codes of behaviour concerning the privacy of geospatial information. A question as simple as “if it is observable, is it public or private?” does not have a unified interpretation for geospatial information.

33. While much more effort will be exerted on the question of geospatial data privacy, one approach may be to begin with the most restrictive policy and/or law and determine if it will suffice in terms of common use. Where a different interpretation is needed, negotiations are necessary to determine where flexibility is required. The most compelling arguments should be applied in these examples to avoid inadequate judgements. For example, the need to know an address varies based on a variety of uses.

34. A more open use is to deliver mail to a household or business where a more protected view is to guard against illegal activity and protect the inhabitants. The compelling argument for revealing an address is to respond to a public safety emergency like fire prevention, ambulance assistance, or police protection. This is an example of a process that is needed to arrive at a workable solution.

I. Coordination of data collection and use

35. Statistical and geospatial data have distinct reasons for their use. Nations oftentimes legislate for the collection and dissemination of statistical data based on national needs, adherence to legal requirements, and support of national programs.

36. The development of national policies requires availability and access to data. Use of statistical data contributes to the adherence to program requirements, determination of program impacts, and support for open data initiatives. Geospatial data has its origins in the mapping of nations that began with content displayed on paper maps and evolved to digital geospatial data that includes data themes such as transportation, hydrography, boundaries, parcels and cadastres, and land use/land cover to name a few.

37. Nations are in various stages of developing national spatial data infrastructures (NSDIs) that include various themes of geospatial data. Beyond official uses of data, there are expectations from public data users, including that the data be accurate and authoritative. Data user expectations are sometimes in conflict with the realities of organizations offering data with respect to project scope, resources, and timing. Increased uses of integrated geospatial databases to support collection activities and analysis by other government agencies, commercial companies, businesses, and private individuals add new requirements as user needs increase.

38. Data quality is always an expectation of users as well as data accessibility. Some factors influence and impact data use. Use restrictions such as constraints used in "hold-harmless" clauses impose limits that can range from any use of the data to restrictions that prevent the public's access to the data. Use of restrictions inhibits the objectives and benefits of an open data policy.

39. As in any endeavour, there is a project beginning in which there is a challenge of getting organized. Support is needed at the highest levels of an organization and projects of this scope require governance. Who is responsible for the project and how are decisions made in terms of investment, cost, resources and schedules? If authoritative data is the responsibility at a level of government below the nation, then there likely is benefit of partnerships to strengthen the data integration enterprise.

40. Sharing and exchanging data with government departments and agencies, commercial companies, and professional and academic organizations and institutions adds sources for consideration in integrating data. Use of existing data avoids duplication. New data collections occur because there is a justifiable business need such as a requirement for more current and/or more accurate data.

J. Technology

41. Geospatial and statistical data acquisition, data management, and data use offer practical case studies for exploitation of applicable existing technology and the development of new technology. Every organization needs an IT strategy to plan an approach to acquire, process, manage, integrate and create products for data users. Some examples of IT considerations include: looking to the cloud first as a way to leverage management of IT resources, particularly if they are not constant; crowdsourcing for information and ideas, particularly if resources are constrained for data acquisition; reducing enterprise redundancies through shared services which eliminates duplicate investments and uncoordinated uses; emphasizing standard-based, commercial off-the-shelf solutions over custom development (buy vs. build); and adopting new technology while maintaining secure systems and information.

42. Technology can sometimes change within the course of the project, particularly if it is a longer term project. A challenge is how to respond and plan for technology advancements where new developments come on line.

III. Conclusion

43. The use of statistical and geospatial information for analysis is a primary expectation of decision makers and businesses that use data to manage demographic and economic changes over time. The United Nations Global Geospatial Information Management Committee of Experts recognizes the value in statistical and geospatial information. They also realize the benefits and power of the results when the two data types are used in combination.

44. A joint working group has been formed to pursue the issues associated with integrating statistical and geospatial data. Various ideas are presented in this paper as considerations for Nations as they conduct their efforts within their country and region. Using this information in the form of a checklist as a way forward in successfully implementing a framework for statistical and geospatial data is an action that may prove helpful.