Paper attached to:
Interoperability of statistical and geospatial data and metadata standards

Connecting Geographic and Statistical Information Standards

The case for, and proposed early steps toward, more explicitly and consistently connecting Geographic and Statistical Information Standards

I. Context

The third meeting of the UN-GGIM on 24-26 July 2013 adopted decision 3/107 on *Linking of geospatial information to statistics and other data*.

This decision:

- acknowledged the critical importance of integrating geospatial information with statistics and socioeconomic data and the development of a geospatial-statistical framework.

- supported holding an international conference on the topic, and urged the expert group to look into issues such as institutional arrangements and standards that would facilitate improved data integration.

The following paper considers how the "information standards" commonly used within the statistical community and within the geospatial community could be applied in a way that would make it simpler for statisticians, geographers and end users to represent and analyse data from a statistical perspective as well as a geospatial perspective.

As outlined in Sections III to VII of this paper, geospatial and statistical information standards enable combining, analysing and presenting data that relates to various "real world" domains in order to generate valuable insights. The standards from both communities achieve this by associating foundational concepts associated with their particular perspective on the world.

- Statistical information standards focus on the *statistical unit* that data is collected from or about (e.g. Persons, Buildings/Dwellings and Businesses) and on the *populations* that statistics, derived from the statistical units that represent that population, are produced for (e.g. Persons in a region, houses in a city, businesses in an industry). Any information associated with these *statistical units*, including location, is stored as a characteristic of each unit. The statistical aggregates or estimates produced for *populations* are subset using variables, with location being only one of a larger set of variables that can be cross classified, with the metadata for the location stored within multidimensional datasets.
- Geographic information standards use geographic features as a foundation (e.g. Land Parcels, Address Points, Rivers and Roads) and these are represented by various forms of geometry (i.e. areas, points, lines/arcs). Information about observations related to these features is then stored in direct reference to the feature and/or the geometry that represents it.

Each community associates information linked to their relevant foundational concepts by applying similar approaches of “entities”, “relationships” and “attributes” (and, ultimately, data related to these), which are specific to the “real world” domain under consideration. More detail on these approaches is included in Appendix A - High level modelling of real world phenomena.

Both communities are describing the same “real world”, albeit from different perspectives. It should be possible to relate the two frames of reference at a high level by considering – and connecting – how each community models and describes real world phenomena.

Working

- **from** a high level, real world based, understanding of the connections between the geospatial and statistical perspectives,
- **towards** a concrete understanding of how the same set of data could be described, analysed and presented from both a statistical and geospatial perspective,

is considered more promising than:

- trying to identify someone who has deep technical expertise related to both sets of standards who can then advise everyone else how to connect them, or

- having technical specialists in geospatial and statistical standards talk to each other and hope that the experts are able to reach an understanding of how the standards fit together and communicate it to everyone else, despite:
    - complexity in both sets of technical standards
    - differences in terminology
    - differences in geospatial and statistical perspectives
    - differences in technical implementation.

The latter approaches have been explored in a number of initiatives in the past. No examples of a resultant resounding breakthrough in terms of being able to design and communicate a generalised approach to interoperability between statistical and geographical perspectives is known to the authors.

II. Structure

The next section of the paper starts by briefly considering what is meant by “geographic information”. It notes that for many practical applications, geographic information – which conforms to the relevant geographic information standards – is a means to an end. The goal of many users is to “geo-enable” data which is otherwise not primarily “geographic” in nature. This allows the users to then undertake new forms of analysis and presentation, leading to new insights

- within their domain of interest, and
- across multiple domains whose data would otherwise be difficult to connect.

The paper then considers what is meant by “statistical information”. It notes that “statistical information” is broader than, but includes, “statistics” as:

> *aggregated and representative information characterising a collective phenomenon in a considered population.*

The paper considers the role of statistical frameworks in modelling “real world” subject matter domains for statistical purposes.

The example of the “Health Domain” is then considered in regard to:

- the way operational information is standardised and exchanged within the domain (eg using the HL7 family of standards)
- application of statistical perspectives and standards (eg SDMX)
- application of geospatial perspectives and standards (eg GML)
- needs to exchange information with other domains (eg Emergency Management).

The conceptual bases for statistical perspectives and geospatial perspectives on domains such as Health are then reviewed briefly. There is an initial focus on the Generic Statistical Information Model (GSIM) and the General Feature Model (GFM), and their application to modelling entities, relationships and attributes within a domain.

The paper concludes by proposing further work on reaching a common understanding of how GSIM and GFM relate to each other (in concept and in practice) when modelling information within domains of interest.

Having agreed the broad “fit” between the statistical and geographic perspectives on domains it is likely to become possible to recommend updates to GSIM and its implementation (eg through standards such as SDMX and DDI) that make it easier to connect statistical information with geographic information.

The “top down” approach to establishing the “fit”

- between statistical and geographic perspectives; and, therefore,
- between the applied information standards such as SDMX, DDI and GML which support these perspectives.

It is also expected to benefit information managers in domains such as health informatics for whom specialised standards such as SDMX and GML appear complex, foreign, daunting and completely unrelated to their day to day experiences of information standards used in their domain (eg HL7).

III. What is meant by Geographic Information?

ISO 19101 (Geographic information—Reference Model) defines geographic information as:

> *information concerning phenomena implicitly or explicitly associated with a location relative to the Earth.*

Most phenomena can be associated with one or more locations relative to the earth. For example, a person will typically be associated with:

- a location where they were born
- a location where they live currently (a residential address)
- a place where they are located at the present moment
- a long history of locations where they have been previously and “journeys” they have taken from one location to another.

A person, as a real world phenomenon, therefore has geographic information associated with them.

They also have a lot of information associated with them that is not particularly geographic in nature, for example:

- sex and age (demographic information)
- annual personal income, an occupation, a level of education, assets/wealth they own (socio-economic information)
- relationships with other people (eg spouse, children, parents, colleagues, friends)
- medical information.

Similar examples exist for economic and environmental information. More detail on geospatial application schema is included in Appendix B.

Geographic information standards are capable of defining, representing and exchanging information that is not specifically geographic in nature. In fact, relating “non-geographic” data to geographic information can provide new means to combine, analyse and present “non-geographic” information from various sources (eg connecting information from disparate sources about the natural environment, the built environment, demography and socio-economic conditions).

IV. Geographic Information vs Geo-enabled Information

Example: Role & relevance of geographic information standards in health informatics

The fact that geographic information standards are capable of defining, representing and exchanging information which is not specifically geographic in nature does not mean these standards provide the ideal means of defining, representing and exchanging all data for all purposes.

For example, defining and managing all information related to the health domain, including its operations and outcomes, entails managing information related to:

- healthcare establishments (eg hospitals and clinics), their organisational structure, medical capabilities (assets/ resources, services offered), staffing, patients, financial operations
- medical practitioners (qualifications, specialisations)
- patients (health conditions, medications, medical outcomes)
- etc.

It can be noted that, because they are people, medical practitioners and patients are also associated with a range of demographic and socioeconomic information which is not specifically “medical” in nature.

Geospatial information standards are not designed to define, manage and exchange the totality of "health information" to the level of detail and consistency required by the health community and interoperable ICT systems used in the health domain.

In fact, just as ISO/TC 211 (Geographic information/Geomatics) oversees a structured set of standards for digital geographic information (the 19100 series of ISO standards), ISO/TC215 (Health informatics) oversees standardisation in the field of health information (to support, for example, interoperable Electronic Health Records).

The HL7 (Health Level 7) suite of standards[1] perform a similar data definition and interchange role for the health community as Geography Markup Language (GML) and related standards perform for the geospatial community.

While geospatial information standards are not designed for defining, managing and exchanging the totality of "health information", there are many cases where a geospatial perspective on "health information" is invaluable, including for relating "health information" to other relevant information (eg Environmental information).

One example would be epidemiology, when studying the distribution of medical conditions over space and time. The geospatial component might be based on a variety of data elements, such as:

- the location of the medical establishment where the condition was diagnosed/managed
- the residential address of the patient.

This highlights that it is not always easy to predetermine exactly which aspect of "health information" will be useful for a particular "geospatial analysis" purpose. It provides maximum opportunity for flexible, cost effective, future analysis if all attributes of the health information which **can** be associated with relevant geographic information **are** associated with that information.

It is also worth noting that epidemiology draws regularly on demographic and socio-economic information about populations that are subject to medical conditions, and the outcomes that population experiences. In other words, epidemiologists also draw on (and produce) statistical information (as defined in Section V of this paper).

A second example is Emergency Management. When responding to an emergency, knowing the capacity of medical establishments to aid victims, the proximity of these establishments to the emergency and the transportation corridors for reaching these establishments is critical.

A second example is Emergency Management. When responding to an emergency, knowing the capacity of medical establishments to aid victims, the proximity of these establishments to the emergency and the transportation corridors for reaching these establishments is critical.

The Emergency Management community has its own primary information standard – NIEM (National Information Exchange Model)[2]. NIEM is distinct from HL7 or the 19100 series of Geographic Information Standards.

[1] http://www.hl7.org/implement/standards/index.cfm?ref=nav

[2] https://niem.gov/communities/em/Pages/about-em.aspx

Concept of geo-enabling data from any, and every, domain

As the example of "health information" illustrates, in many cases the aim is not to use geographic information standards to define and manage all the information related to a particular domain (eg health, education, trade, agriculture, the economy). The aim, instead, is to "geospatially enable" (commonly shortened to "geo-enable") the data related to a particular domain to facilitate geospatial analysis.

Geo-enabling[3] can be considered as:

- the application of location or geospatial information as part of business processes or using 'location intelligence' to augment non-spatial information systems
- the act of deriving and utilising geography within non-spatial information.

The US Federal Geographic Data Committee (FGDC) states[4]:

- Geo-enabling is to take loosely geo-referenced information… and automatically join it up with the representation of that geography… to support visual and GIS analysis against other data.
- Geo-enabling only has value when some added benefit comes from it - to do something *faster*, *more accurately*, or something that was *not previously possible.*

Geo-enabling ensures relevant elements of a National Spatial Data Infrastructure (NSDI) can be applied in an efficient, flexible and useful manner to data from a particular domain to support:

- new, geospatially based, analysis and presentation of data from the domain, and
- combining data from that domain with data from other domains and then analysing and presenting the combined data using geospatial information, techniques and tools.

Both of these purposes allow additional insights to be obtained from the information which is defined and managed within the domain.

All National Spatial Data Infrastructures, and their applications, are underpinned by Geospatial information standards,

- as standardised through ISO/TC 211 (the 19100 series of ISO standards), and
- as developed and applied through OGC (Open Geospatial Consortium) and other standardisation and implementation communities.

[3] http://en.wikipedia.org/wiki/Geo-enable
[4] http://www.fgdc.gov/geospatial-lob/documents/geb-factsheet.pdf

V. What is meant by Statistical Information?

Legislation with European Union[5] defines "statistical information" as:

> *Aggregated and individual data, indicators and related metadata.*

The definition of "statistical information" is broader than simply "statistics"[6].

The legal definition of "statistics" within the European Union[7] is:

> *Quantitative and qualitative, aggregated and representative information characterising a collective phenomenon in a considered population.*

The OECD definition of "statistics" likewise refers to populations ("aggregates of individuals").

The definition of "statistical information" is broader because it encompasses the data from which statistics are produced, as well as the statistics themselves.

The data from which statistics are produced often relates to individuals rather than populations.

The individual which a data record describes is typically termed a (Statistical) "Unit". The Generic Statistical Information Model (GSIM) states that a *Unit represents a real world object (or "entity") of interest when producing statistics.*

The data may, for example,

- be collected directly from a Unit (eg person or business) for statistical purposes
- sourced from information about the Unit which was created for another purpose, eg
  - Registrations of businesses for taxation and other regulatory purposes
  - Registrations of individual births, building approvals, motor vehicles, trade transactions, etc.

The fact statistics themselves relate to populations, however, has prominent implications for how statistical information is described and managed.

Statisticians seldom, if ever, have "perfect" information in regard to all the units within a population of interest. (These statistical methods are also applied in environmental studies, which generally use geospatial data models, and the resulting measurement issues are often present in geospatial datasets.)

Where data is collected directly, it is often only cost effective to "sample" a number of units from within the population of interest rather than obtain data for all units. Designing samples, and producing estimates for a population based on a sample, are key aspects of statistical methodology.

[5] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:269:0001:0006:EN:PDF

[6] As Eurostat's Concepts and Definitions Database documents, the broadness of this definition is consistent with the definitions of "Statistical Information" used by the United Nations and by the SDMX initiative (introduced in Section V).

[7] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:087:0164:0173:EN:PDF

Even where data is obtained from an existing source, producing estimates for the population of interests requires understanding, and addressing, limitations in the data which is available.

For example:

- When a business registers for regulatory purposes, accurately identifying their primary business activity (based on the Australian and New Zealand Standard Industrial Classification - ANZSIC) may not be their highest priority. While data is recorded for the individual units, statisticians cannot simply assume everything recorded is correct. Sources of error need to be understood.

- If the statistics are intended to relate to the population of businesses operating at the current time, it needs to be taken into account that:
  - some registered businesses will have stopped operating
  - some operating businesses will not have up to date details on the register.

Different statistical estimates (for different sub-populations and for different variables) within a single dataset often have different levels of precision/confidence (eg different relative standard errors) associated with them.

Understanding the “statistical quality” of estimates is essential when seeking to draw sound conclusions from, and make informed decisions based on, statistics.

While “statistical quality” is not a completely separate concept to “data quality” more generally, expression of “statistical quality” is detailed and standardised within the statistical community. Generic data quality frameworks tend not to be designed to express “statistical quality” with the level of structured detail that expert producers and users of statistics expect.

VI. Statistical perspectives on domains

As noted in Section III, geospatial perspectives often support additional insights on domains (eg Health) which are not primarily “geospatial” in nature.

This applies at least equally to statistical perspectives.

The UN Economic Commission for Europe states that:

> *The purpose of official statistics is to produce and disseminate authoritative results designed to reliably reflect economically and socially relevant phenomena of a complex and dynamic reality in a given country*[8]*.*

In other words, statistics are intended to provide reliable insight (a perspective) on “real world” phenomena in a complex reality.

[8] http://www.unece.org/fileadmin/DAM/stats/documents/applyprinciples.e.pdf

The international Statistical Data and Metadata eXchange (SDMX) initiative formally recognises a set of Statistical Subject-Matter Domains[9]. The "subject-matter" domains span demographic, social, economic and environmental considerations. It also recognises that statistics drawn from multiple domains are relevant to matters such as sustainable development (which takes into account social, economic and environmental dimensions).

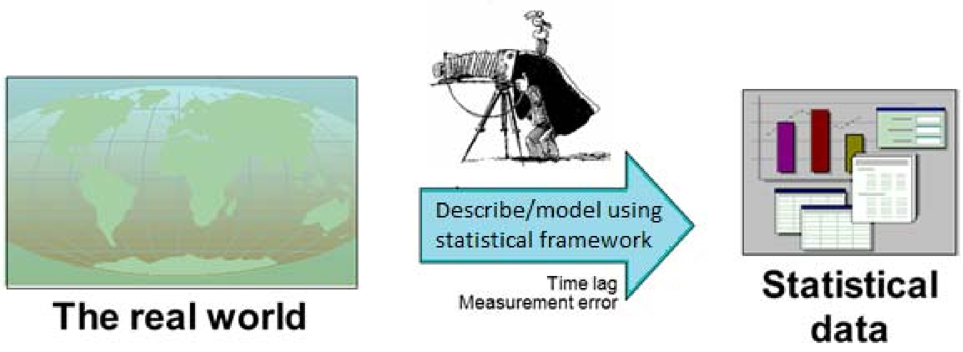
Within a statistical subject matter domain, a common starting point is defining and agreeing a statistical framework.

A statistical framework represents an agreed way of thinking about an area of interest. It promotes consistency and comparability across national data collections and between countries.

A framework delineates the important concepts associated with a domain and organises these into a logical structure. It helps identify, and standardise, important types of units, populations and variables (measures) within the domain.

Statistical frameworks provide a means for relating a domain of interest in the real world to a consistent statistical "portrait" of that domain.

The concept is illustrated by Statistics Netherlands in the following manner.



The most prominent statistical framework is the System of National Accounts (SNA)[10].

> *The comprehensive statistical framework for economic statistics that provides a consistent and flexible set of macroeconomic accounts for policymaking, analysis and research.*

SNA is fundamental to ensuring economic statistics are reported on a transparent and comparable basis by the countries of the world.

The data available within counties often does not align with SNA definitions. For example, the definitions used nationally for accounting and regulatory compliance purposes may not align with SNA definitions. Countries are required to report the methods used to estimate statistics based on the SNA framework when starting from national data sources which are incomplete or misaligned.

[9] http://sdmx.org/wp-content/uploads/2009/01/03_sdmx_cog_annex_3_smd_2009.pdf

[10] http://unstats.un.org/unsd/pubs/gesgrid.asp?id=419

For domains such as Education[11], statistical frameworks help to identify:

- how the (education) system operates in regard to resources, activities and outputs and outcomes.
- what statistics are of most value in informing government and citizens in regard to operations and outcomes.
- what administrative data used within the domain could contribute to the production of statistics.

Defining, exchanging and analysing statistics on a meaningful basis requires information about relevant aspects of the "domain model" associated with the statistics. In the case of GSIM and SDMX, *Data Structure Definitions* are used to associate *Datasets* (containing codes and numbers) with relevant constructs used in modelling the domain, such as *Variables* and *Statistical Classifications.*

VII. Considering the example of the health domain

In Section III it was noted that the Health Domain has its own information standards (eg HL7) which support the exchange of information (eg for administrative and management purposes) between ICT systems in the domain.

It was also noted that a geospatial perspective on information from the health domain often "adds value". It was noted, also, that selected information from the health domain is useful for emergency management purposes.

Statistical perspectives on demands, resources and outcomes associated with the health domain are likewise valuable for informing:

- policy making,
- the strategies of health and allied industries, and
- public understanding and debate.

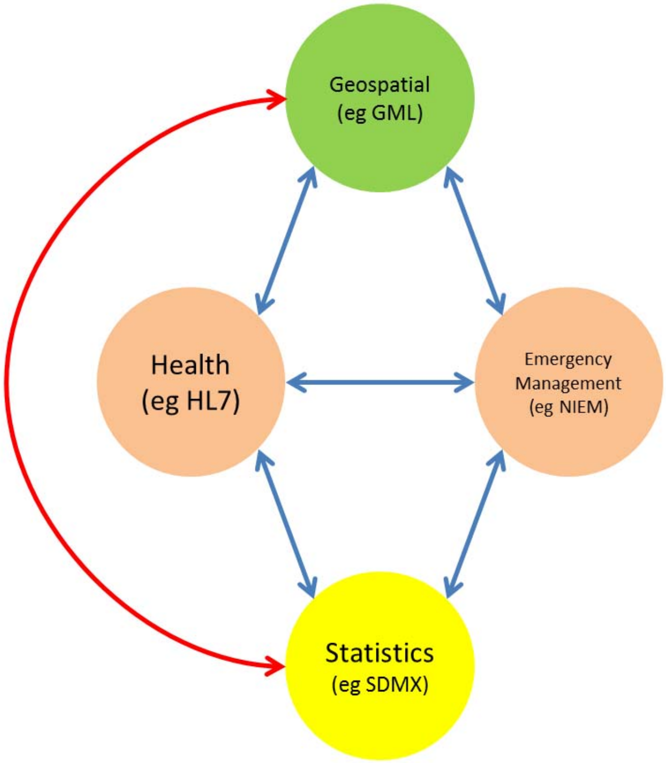
A comparable statistical framework across different health systems (eg within one country and internationally) supports comparison - including assessing and sharing what appear to represent "best practice" benchmarks for health systems and their operations.

SDMX-HD (Health Domain) was developed by the World Health Organisation (WHO) and partners to serve the need of the "Monitoring and Evaluation" community[12]. It recognises that "individual data" will likely be carried within the health domain using domain standards such as HL7. Aggregate data for statistical purposes would then be generated based on that "individual data". The aggregate data would be represented and exchanged using SDMX-HD.

[11] http://www.ausstats.abs.gov.au/ausstats/free.nsf/0/223D40E6FAF99324CA256CBC0078E114/$File/42130_2003.pdf

[12] http://www.sdmx-hd.org/

This leads to the following web of connections between information domains and disciplines:



In regard to the blue arrows in the diagram above (from left to right and top to bottom)

- There is a "Health Domain Working Group" under OGC[13].
    - The group is considering SDMX-HD, in addition to HL7.
- There is the "Monitoring and Evaluation" community within the health domain that developed SDMX-HD.
- There is a "Health Community" within NIEM to support health information exchange[14].
- There is an Emergency & Disaster Management Working Group under OGC[15].
- Statistical agencies work with the emergency management sector to make demographic information available.

The red arrow represents the addition of UN-GGIM and UNSC seeking to have the geospatial and statistical community "connect" more consistently in regard to information standards. From a domain perspective, key benefits of this approach will be:

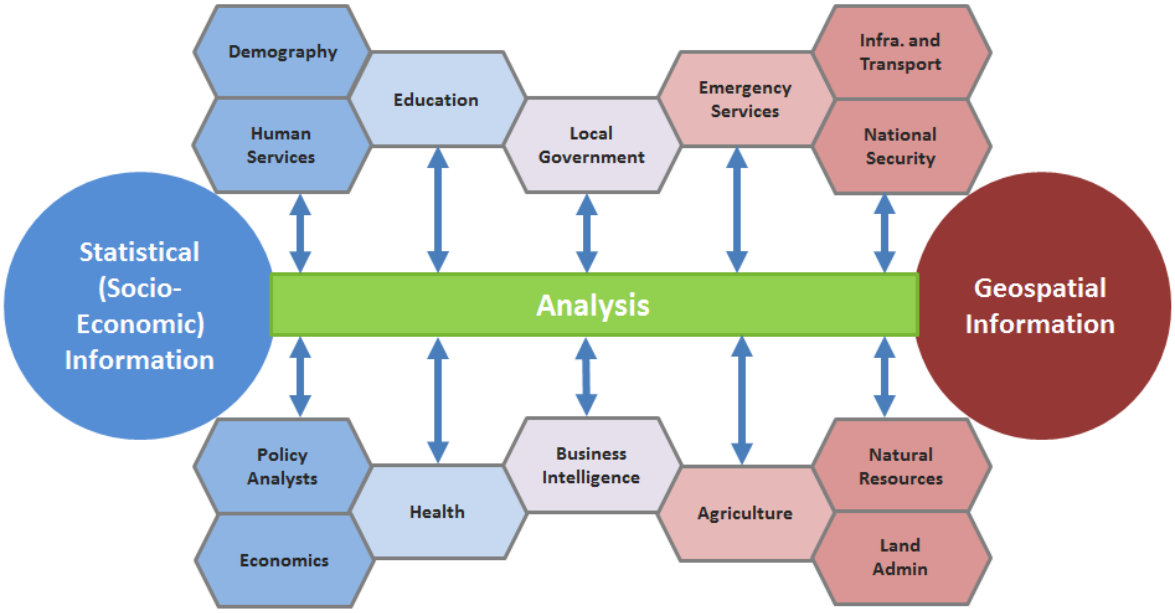
- Data will be more easily geo-enabled at the source.
- The geo-enabled data produced will be more seamless used in geospatial analysis.
- The dataset will be more easily integrated into statistical analysis.
- Most statistical analysis completed will be able to be visualised spatially.

[13] http://www.opengeospatial.org/projects/groups/healthdwg
[14] https://www.niem.gov/communities/emc/health/Pages/about-health.aspx
[15] http://www.opengeospatial.org/domain/eranddm#edmwg

The diagram below illustrates that geographic and statistical information standards represent "information perspectives" rather than "information domains". It also shows that they share common interests in supporting analysis in a range of information domains.

While only two information domains (health and emergency management) are shown in the diagram on the previous page, there are many more domains served by statistical and geospatial perspectives.

Some domains (eg macro-economics) tend to have more of a focus on statistical perspectives, while other domains (eg geosciences) tend to have more of a focus on geospatial perspectives. Many domains (eg information related to industry sectors such as agriculture or education), however, have a strong interest in both statistical and geospatial perspectives.



VIII. Basis for statistical and geospatial "information perspectives"

As the previous sections illustrate, both geospatial and statistical information standards enable combining, analysing and presenting data from various domains in order to generate valuable insights.

Both achieve this by associating foundational concepts associated with their particular perspective on the world. These are:

- statistical units, populations and variables in the case of statistical information standards, and
- features, geometry in the case of geographic information standards.

Both apply similar approaches of "entities", "relationships" and "attributes" (and, ultimately, data related to these) which are specific to the "real world" domain under consideration. More detail on these approaches is included in Appendix A - High level modelling of real world phenomena.

In the case of statistics, at a conceptual level, domains can be described using "information objects" from the concepts grouping within the Generic Statistical Information Model. This includes identification of *UnitTypes* and *Populations*, *Concepts*, *Variables* (including representation and allowable values), *StatisticalClassifications, CategorySets* and *CodeLists* associated with the domain.

As mentioned in Section V, *DataStructureDefinitions* are then used to associate the conceptual model of the domain with datasets that contain statistical information. SDMX is ideal for physically encoding and exchanging datasets containing statistical aggregates. The DDI (Data Documentation Initiative) standard is more typically used for the purpose of encoding and exchanging datasets related to individual *Units*.

At a conceptual level, the following geospatial information standards define the General Feature Model (GFM).

- ISO 19101—Reference Model, and
- ISO 19109—Rules for Application Schema.

The GFM provides a generic means of modelling different types of features *(FeatureTypes)* (ie "entities"), their relationships and attributes. GFM is very high level and flexible.

FeatureTypes as applied in the GFM do not need to be geospatial in nature, and are different to the concept of a feature type used in some GIS applications, which define different geometry types. That said, *FeatureTypes* as defined in GFM that are not geospatial would have one or more geometries associated with them through *Properties,* such as:

- *Attributes* (e.g. a residential address *and the Mesh Block that address falls within* for a "Person" *FeatureType*), or
- *Associations* with other *FeatureTypes* that either are geospatial in nature or have geospatial *Attributes (e.g. a "Persons" ownership of a "Property" is an association, with the address and Mesh Block being attributes of that "Property").*

For readers who focus more on technical implementation than on GFM as a conceptual model, implementation standards reflect, and build upon, GFM's conceptual description of the relationship between *FeatureTypes* and geometries. For example, as described in Section 2.5 of the OGC Reference Model[16], at a technical encoding level, it is possible to define a GML Application Schema[17] which associates standard features supported within GML with domain specific constructs; for example, feature types, attributes and relationships which are specific to a domain (such as Health) rather than those supported generically in GML.

IX. Proposed next steps

It is rare to find individuals who have a detailed and applied understanding of implementation standards associated with both statistical information (eg SDMX and DDI), and geographic information (eg GML).

Standards at these levels of detailed encoding are relatively complex and technical.

[16] http://rap.opengeospatial.org/orm.php
[17] http://www.ogcnetwork.net/gmlprofiles

In addition, while various technical approaches can be taken (for example) to associate information encoded using SDMX with information encoded using GML, it becomes relatively difficult to "work backwards" from the technical approach to confirm the approach taken is conceptually sound and appropriate from a statistical and a geospatial perspective.

It appears more feasible and appropriate to start by reaching a common understanding of

- how GSIM models statistical subject-matter domains and their data from a conceptual perspective, compared with,
- how GFM models "application" domains and their data.

Preliminary analysis suggests that relating GSIM *Units* to GFM *Features* tends to be relatively (but not always) straightforward.

Relating GSIM *Populations* to GFM *Features* and their *Properties* appears more problematic. For example, presently *Dimensional Data* related to the *Population* of persons resident in a particular Local Government Area (LGA) tends to be translated to a string of attributes associated with that LGA. If the statistical data were, for example:

- Counts of persons, by:
  - Occupation (10 Categories, including total), by
  - Age (10 Categories, including total), by
  - Sex (3 Categories, including total ("Persons)).

This would typically be translated to a "flat" list of 300 (10 x 10 x 3) attributes associated with the LGA (as a Feature).

It then becomes very hard to compare statistical "sub-populations" (eg patterns for males vs females across occupations) if *Dimensional Data* related to the *Population* becomes "flattened" in this way.

Even if further analysis shows there are no viable options other than this "flattening", it may be possible to provide recommendations on how a "flattened" geospatial view of a *Population* can be linked back to a statistical "dimensional" view of the same *Population*.

Having agreed the broad "fit" between the statistical and geographic perspectives on domains, it is likely to become possible to recommend updates to GSIM and its implementation (eg through standards such as SDMX and DDI) that make it easier to connect statistical information with geographic information. For example:

- guidelines for relating *UnitTypes* to *FeatureTypes,* and to *Geometries* associated with *FeatureTypes*
- guidelines for describing and presenting information about populations and subpopulations from a geospatial perspective
- guidelines for associating relevant *Variables* (eg addresses, locations coded to statistical geographies[18]) with geographic information that supports a geospatial perspective

[18] For example, the Australian Statistical Geography Standard http://www.abs.gov.au/websitedbs/d3310114.nsf/4a256353001af3ed4b2562bb00121564/c453c497aadde71cca2576d300026a38/$FILE/ASGS%202011%20Structure%20and%20Summary.pdf

- guidelines for documenting “statistical quality” considerations when presenting statistics on a geospatial basis.

A practical benefit from this approach is that it should make it easier to provide information managers in various domains (eg Health) with advice that helps them both “geo-enable” their data and produce statistics for monitoring and evaluation purposes. It should also facilitate simpler, more efficient and more effective integration of information from both statistical and geospatial perspectives for analysis and, ultimately, informed decision making.

Currently, both geospatial information standards and statistical information standards appear complex, foreign and daunting to information managers who are used to defining and using data on a domain specific administrative and operational basis. Being able to provide “top down” guidance that avoids geospatial and statistical perspectives requiring completely separate, and disconnected, learning and modelling should prove invaluable.

It would be possible to use a domain such as Health as a test case for exploring the connecting of statistical and geospatial perspectives on information, based on GSIM and GFM.

Appendix A – High level modelling of real world phenomena

i. Geographic – Features and Properties

The starting point for modelling geographic information is a *Feature*.

A *Feature* is a digital representation of a real world entity or an abstraction of the real world. It has a spatial domain, a temporal domain, or a spatial/temporal domain as one of its attributes. Examples of features include almost anything that can be placed in time and space, including buildings, cities, trees, ecosystems, vehicles, routes, oil spills, and so on.

The GFM provides a generic means of modelling different types of *Features*. The GFM is very high level and flexible.

It is possible, for example, to define a *FeatureType* “Café/Restaurant”. Each individual café/restaurant (eg “Jumping Java Coffee Shop”) would be a *Feature* of type “Café/Restaurant”

A *FeatureType* has *Properties*.

At least one of the *Properties* would be expected to be spatial in nature. For example, the “Café/Restaurant” *FeatureType* might have two spatial *Properties* associated with it - A Location (eg based on street address) and a Floor Plan.

Properties associated with a *FeatureType* may include relationships with other *FeatureTypes*. For example, a “Café/Restaurant” might be associated with one or more “Proprietors”. (“Proprietor” as a *FeatureType* may, in turn, have spatial *Properties* such as Residential Street Address and Business Street Address.)

“Café/Restaurant” may also have *Properties* associated with it that are not spatial in nature, such as opening hours, a star rating and a phone number. These types of *Properties* relate to *Attributes* rather than relationships (*Associations*) with other *FeatureTypes*. For example:

- a phone number might be represented as a simple string (starting with “+” and a country code) that meets appropriate validation rules,
- a star rating might be represented with a code from “0” to “5”,
  - alternatively, it might be represented with a decimal data type, with a value between 0 and 5, if ratings from different reviewers have been averaged.

The GFM allows the data type of *Attributes* to be specified and constrained appropriately.

ii. Statistical – Units and Variables

In very general terms, the concept of *Unit* in GSIM is the closest analogue to *Feature* in GFM.

In GSIM a (Statistical) *Unit* represents a real world object (or “entity”) of interest when producing statistics.

A *Unit* is associated with a *UnitType,* similarly to the way a *Feature* is associated with a *FeatureType*.

In the 2006-07 Cafes, Restaurants and Catering Services Survey in Australia, for example, the statistical units – in most cases – were of *UnitType* “ABN units”[19], which:

- had registered for an Australian Business Number (ABN) and appeared on the Australian Business Register (ABR)
- were within scope of survey
  - the predominant business activity associated with the ABN, as recorded on the ABR, was categorised as
    - Class 4511 Cafes and Restaurants
    - Class 4513 Catering Services.

In approximate terms, “ABN units” represent “businesses”.

If “Jumping Java Coffee Shop” had operated in Australia in 2006-07 then it might have been in scope, as a *Unit*, for this survey.

It can be noted in Table 3 of the statistics produced[20] that “businesses” as the main *UnitType* for this survey do not equate to Locations. There were estimated to be 13 987 cafes and restaurants operating from 15111 locations. In other words, some businesses operated from more than one location.

While it would be possible to define *FeatureType* on the same basis (a *Feature* could be associated with multiple operating locations), there may be an understandable tendency in similar cases to associate the *FeatureType with* “business location” rather than the legal definition of a business.

As is the case with *FeatureTypes, UnitTypes* can have relationships with other *UnitTypes*. Examples include the:

- Australian units model[21]
  - The model addresses, for example, how an “enterprise group” relates to the individual “businesses” that belong to the enterprise group
- As the main *UnitTypes* of interest within Australia’s Population Census[22], “Persons” are associated with “Families” and “Families” are associated with “Dwellings”.

Relationships between *UnitTypes* are somewhat analogous to *Association* type *Properties* within GFM. At the present time relationships between *UnitTypes* are not modelled explicitly in GSIM, but they are reflected in *Record Relationships. RecordRelationships* describe how to connect data related to one *Unit* (eg a person) to data for another *Unit* of a different *UnitType* (eg data for the dwelling in which that person lives).

[19] http://abs.gov.au/AUSSTATS/abs@.nsf/Lookup/8655.0Explanatory%20Notes12006-07?OpenDocument

[20] http://www.ausstats.abs.gov.au/Ausstats/subscriber.nsf/0/208E6A8D6857022ACA25743500119553/$File/86550_2006-07.pdf

[21] http://www.abs.gov.au/AUSSTATS/abs@.nsf/Latestproducts/CE8C1B19FB7E6C79CA257B9500133D5F?opendocument

[22] http://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/2901.0Main%20Features252011?opendocument&tabname=Summary&prodno=2901.0&issue=2011&num=&view=

Other "*Properties*" of statistical *UnitTypes* are defined as *Variables*. These are somewhat analogous to *Attribute* type *Properties* in GFM.

Each *Variable* defines a "characteristic" associated with a *UnitType* which is to be measured. The *Value Domain* associated with a *Represented Variable* establishes the valid values the attribute being measured can assume for any *Unit* of that *UnitType*.

In the case of the 2006-07 Cafes, Restaurants and Catering Services Survey, a wide range of *Variables* were collected from the Units (businesses) that were in scope and sampled for the survey. These *Variables* included:

- number of premises in metropolitan and non-metropolitan locations and total available seating (at 30 June 2007)
- number of people working for the business, by sex and employment type and by occupation (at 30 June 2007)
- income and expenses by type over the past financial year.

iii. Statistics – Populations and Dimensions

A key statistical concept that has not yet been mentioned is *Population*. A *Population* comprises the complete set of *Units* that share the attribute values which define the population.

For example, the target *Population* for the Australian Labour Force statistics is resident, civilian (persons) aged 15 years and older. This is also known as the "civilian population".

Subpopulations of interest can be defined. For example "Unemployed Young People" could be defined as members of the civilian population who have a Labour Force Status of "Unemployed" (persons without work who are seeking work and available to work) and an age between 15 and 24.

Subpopulations can then be compared with other populations and additional attributes analysed. An example might be examining the prevalence of youth unemployment in different regions based on the place of residence attribute of unemployed young persons.

Most statistical outputs present data related to *Populations* rather than data for individual *Units* within the *Population*. Data supplied by, or in regard to, individual *Units* (eg individual persons or businesses) is typically confidential – and unable to be disseminated in its own right.

In addition, it is common to collect data for a sample of *Units* in a *Population* and then estimate measures for the Population as a whole. Statisticians apply expertise to sample design and estimation methods – including assessing the quality of estimates.

This is illustrated in the outputs from the 2006-07 Cafes, Restaurants and Catering Services Survey.

- The explanatory notes include detailed consideration of sampling and non-sampling error.
- Individual estimates are annotated where they have a "relative standard error" of 10% or more.

Within GSIM, datasets which contain data "about" individual units are known as *UnitDatasets*. Datasets which contain data "about" *Populations* and subpopulations are known as *DimensionalDatasets*. *DimensionalDatasets* are commonly considered "data cubes"[23].

An example of a *DimensionalDataset* might be:

- Employed persons by Sex, Occupation, Industry and State/Territory.

The overall *Population* is "Employed persons". Measures are provided, however, for many subpopulations – e.g.

- Male, Managers, Employed in the Mining Industry, in the Northern Territory.

Multiple measures may be provided for each sub-population – e.g.

- the estimated count of employed people within that sub-population, and
- the average hours worked by employed persons within that sub-population.

Statistical data is often collected and presented as "time series". For example, the *Dimensional Dataset* might contain quarterly observations for a span of 25 years (ie 100 quarters). The time series approach allows, in this cases, changing patterns in employment to be analysed over time.

The total number of sub-populations described within this *DimensionalDataset* could be considered to be:

- 3 (Male and Female and Total) x 20 (19 Industry Divisions and Total) x 9 (8 Major Group Occupations and Total) x 9 (8 States/Territories and the Total for Australia) = 4860

There are two measures recorded for each sub population, observed for 100 quarters, giving 972,000 cells in total.

iv. Challenges when representing Populations and Dimensions spatially

As described below, *Properties* of *Populations* (including sub-populations) typically pose more challenges to describe geospatially than *Properties* of individual *Units*.

Challenge 1: Populations are not features

Firstly*, Populations* tend to be more awkward than *Units* to define as *Features* in their own right. Statistical *Populations* (and subpopulations within them) often, instead, become represented as *Attributes* of a *Feature* which is spatial in nature (usually a polygon).

From a geospatial perspective, *Dimensions* that relate to spatial attributes of *UnitTypes* in the *Population* tend to receive "primary" attention.

For example, in the case of the *DimensionalDataset* introduced in the previous section, this would be the *Dimension* containing the State/Territory where each subpopulation lives. (*DimensionalDatasets* often include *Dimensions* related to smaller scale geographic entities – such as statistical areas

[23] http://www2.cs.uregina.ca/~dbd/cs831/notes/dcubes/dcubes.html

defined on various bases[24]. *DimensionalDatasets* may also contain *Dimensions* related to larger scale geographic entities such as "Country of Origin".)

From a statistical perspective, however, the State/Territory dimension is not intrinsically "primary" compared with the sex, occupation or industry (or time) dimensions. The relative importance of the dimensions tends to vary according to the particular analysis the statistician wishes to undertake. (For example, in Time Series Analysis the time dimension will be "primary".)

From a geospatial perspective, the *DimensionalDataset* might be interpreted as describing 540 (3x20x9) subpopulations within each State/Territory. For each subpopulation we have two measures:

- the estimated count of persons in the subpopulation, and
- the average hours worked by those people.

This gives us 1080 "facts" about each State/Territory – each of which could be represented as a different attribute of that particular State/Territory (*Feature*). (The number can potentially be multiplied by 100 because we have the 1080 "facts" for 100 points in time.)

Managing so many "attributes" – based on subpopulations which were structured based on independent dimensions (rather than one fixed hierarchy) – tends to be challenging in a geospatial context. Typically, in practice, only a small number of the potential subpopulations will be represented geospatially at any one time. The choice of which subpopulations to represent and analyse on a spatial basis is typically driven by user needs.

As (sub)*Populations* don't readily translate directly to *Features*, bulk conversion of *DimensionalDatasets* to geospatial datasets should be approached with care.

Nevertheless, explicitly identifying – using appropriate metadata - which *Dimensions* relate to spatial attributes of *UnitTypes* would appear to be a valuable practical extension for GSIM.

Challenge 2: Measures as estimates, not observations

The concept of sampling and estimation in statistics was discussed briefly in the context of the 2006-07 Cafes, Restaurants and Catering Services Survey.

As was shown by that example, estimates for different sub-populations (and different measures for that sub-population) vary markedly in terms of precision (eg as measured by Relative Standard Error). A single tolerance (whether expressed in absolute terms or as a percentage) cannot be provided for all the data in a dataset (or all data related to a particular measure represented in a dataset).

Understanding, and accounting, for error/imprecision is a critical element of statistical analysis.

Geospatial information standards provide support for documenting data quality considerations – particularly (although not exclusively) in regard to geographic data.

[24] For example, Statistical Areas (or even Mesh Blocks) defined within the Australian Statistical Geography Standard. http://www.abs.gov.au/websitedbs/d3310114.nsf/4a256353001af3ed4b2562bb00121564/c453c497aadde71cca2576d300026a38/$FILE/ASGS%202011%20Structure%20and%20Summary.pdf

They are not designed, however, to represent as much information about “statistical quality” as statisticians typically wish to provide – and use – in regard to statistical estimates.

At the same time, it is anticipated that many references to geographic *Features* and *Properties* in statistical datasets probably fail to provide as much information about “geographic data quality” as geospatial experts would wish to have available.

This indicates that improved mutual understanding of what “quality” means to statisticians and geospatial experts, and how both perspectives can be supported in practice, is a desirable early step.

Appendix B – More on geospatial application schema

Geospatial standards (eg ISO 19109 - "Rules for Application Schema", ISO 19136 "Geography Markup Language (GML)) recognise that datasets in these domains are likely to contain both "geographic data" and "other data".

Taking an example from the Health domain, a dataset might contain "geographic" data related to hospital emergency departments – e.g.

- address/location
- "catchment area" from which patients are expected to present at that department rather than another hospital

and also "non-geographic" data – e.g.

- contact information
- number of beds
- types of medical care which are available
- average wait times for cases at each level of urgency (National Triage Scale).

From a geospatial perspective, "Emergency Department" would be a *Feature Type*, and each item of information recorded would be a *Property*.

Each "geographic" *Property* would be directly associated with *Geometry.* These properties for each Emergency Department (*Feature*) could be described in terms of detailed geographic information (data and metadata), ideally in a standard form.

The "non-geographic" properties would not be directly associated with a *Geometry*.

It is possible to define an "application schema" which is specific to a particular domain of interest and which describes the feature types whose data the community is interested in. In the health domain example, features of interest might include:

- hospitals
- specialised departments within the hospital (eg Emergency Departments)
- medical staff
- patients
- medical services that are available from departments
- individual assets owned by the hospital.

Geographic properties would be associated with these features (eg patients would have home addresses) as well as properties that are not intrinsically geographical in nature (eg the association between a patient and the medical conditions for which they have been treated).

A number of communities have formally defined and implemented application schema so they can define and exchange data on a consistent, "spatially enabled" basis. These include Geoscience, Meteorology and Aviation.