Emerging data challenges for next-generation spatial data infrastructure

Benjamin Adams   Mark Gahegan

Centre for eResearch
The University of Auckland, New Zealand
{b.adams, m.gahegan}@auckland.ac.nz

Abstract

The landscape of spatial data infrastructures (SDIs) is changing. In addition to traditional authoritative and reliably sourced geospatial data, SDIs increasingly need to incorporate data from non-traditional sources, such as local sensor networks and crowd-sourced message databases. These new data come with variable, loosely defined, and sometimes unknown provenance, semantics, and content. The next generation of SDIs will need the capability to integrate and federate geospatial data that are highly heterogeneous. These data comprise a vast observation space: they could be represented in many forms, will have been generated by a variety of producers using different processes and will have originally been intended for purposes that may differ markedly from their later use. There are several discriminative dimensions along which we can describe the properties of the data found in SDIs, such as the data structure, the spatial framework (e.g., field, image, or object-based), the semantics of the attributes, the author or producer, the licensing, etc. These dimensions define a universe of model possibilities for data in an SDI, known as a *model space*. A core research challenge remains to recognise and resolve - to the degree possible - a comprehensive set of *model dimensions* that will enable us to characterise the many possible models by which geospatial data can be represented. A second challenge is to describe the transformations within and between models, and the ways in which these transformations change aspects of the underlying model. Despite recent movement toward semantically described services for SDIs, the scope and range of descriptive dimensions for geospatial data are underspecified. In this paper we present a diverse set of important dimensions that point to a series of challenges for data integration and then describe how both traditional and emergent datasets can be characterised within these dimensions, and point to some interesting differences.

1 Introduction: The evolving role of the SDI

National and regional spatial data infrastructures (SDI) were originally conceived to be centralised geospatial data repositories containing data that came largely from authoritative sources (Masser, 1999; Groot and McLaughlin, 2000; Jacoby et al., 2002). The advent of local sensor webs, web 2.0 and so-called volunteered geographic information (VGI) - i.e.,

In: S. Winter and C. Rizos (Eds.): Research@Locate'14, Canberra, Australia, 07-09 April 2014, published at http://ceur-ws.org

voluminous geospatial data that are made available from a multitude of sources of varying quality and that are often uncontrolled in terms of their: creation process, representation and content- has changed the landscape (Goodchild, 2007; Budhathoki et al., 2008). While not produced by authoritative agencies, these data can represent better coverage of specific geospatial phenomena or be more timely due to their distributed and unconstrained methods of generation (Coleman et al. (2009)). Thus in many cases they are, in fact, of higher value, e.g., for time critical tasks in emergency response. However, their domain content (or the tasks to which they can be usefully applied) may not be known in advance and may require post-processing to extract. What then is the role of the spatial data infrastructure in such an environment, given that many of the data that analysts and policy makers will find useful may come from such widely varying and incompatible sources? And how can its users understand the utility and reliability of the data products derived from mashing up such heterogeneous datasets? It is a challenging problem because in order to effectively match data to a specific application need we must consider several aspects of the data at once, including not only the spatial framework of the data but also the provenance, semantics, context of authorship, access rights, etc. (what we might call the *pragmatics* of the data to differentiate it from the geospatial semantics of the data) (Pike and Gahegan, 2007; Gahegan et al., 2009). Recent work in merging VGI with SDI has advocated for better semantic representation, using formal languages from the semantic web, and while this is a good step we argue that a more holistic approach is necessary (Janowicz et al., 2010). This is not a new research problem, but to date the relevant research in GIScience has focused on piecemeal solutions to specific strands of the problem, tackled in isolation that do not work together in the orchestrated way that would be necessary to build a more advanced SDI.

In the following section we will introduce our vision for a next-generation SDI. In section 3 we present a diverse set of important dimensions for next-generation SDI. We follow with example for how data transformation can be represented in terms of those dimensions and show examples for both authoritative and other less formal data. Finally, we conclude with a summary of why we think this is an important time to reconsider the role of SDI as a vital component in a connected approach to the science process: i.e., linked science or eScience (Hey and Trefethen, 2005; Mäs et al., 2011).

2 Next generation spatial data infrastructure

In a typical GIS problem-solving workflow, we typically encounter distinct steps such as the following:

1. Locate, gain access to and – to some extent – understand the limitations of each dataset we intend to use. Currently, SDI and specifically their data catalog and search tools can sometimes help here.

2. Transform the datasets we will use into a consistent form (model), for example by re-projecting, converting from raster to vector or harmonising the semantics. The decisions we make here can have profound implications for the quality of the data.

3. Combine the datasets via an analytical workflow of some kind.

4. Assess the accuracy and reliability of the result and (possibly) publish it back into the SDI.

The geospatial datasets that we might wish to combine could be highly heterogeneous. They will be represented in many forms, will have been generated by a variety of producers using different processes and may have originally been intended for purposes that are different from their present use. Each of these ideas, and others, form the dimensions along which we can describe the properties of a dataset found in an SDI, such as the data structure, the spatial framework (e.g., field, image, or object-based), the semantics of the attributes, the author or producer, etc. These dimensions define a universe of model possibilities, or *model space*, for datasets that the SDI interacts with. In order to capitalise on the value of such a wide variety of data, we need to detail the many ways in which we might integrate or transform data that reside at different points in model space. A core research challenge, therefore, is to recognise and resolve - to the degree possible - a comprehensive set of characteristic model dimensions[1] that enable us to characterise transformations within and between data models. These dimensions provide us with a conceptual framework to understand the ways that data are transformed and are made fit-for-purpose.

A data source, such as the Landsat 7 sensor or a crime logging system has the potential to create a series of datasets, so a source can be represented in this model space similarly to an individual dataset. But rather than being represented as one point in model space, a data source may be represented as ranges along certain dimensions, describing the potential values that a specific dataset may inherit. For example, each Landsat 7 dataset will have a unique timestamp and a spatial footprint drawn from a set of possibilities defined by the orbital characteristics. But the spatial framework will always be an image and the data will always be packaged into a raster data structure.[2]

[1] The term ‘dimension’ is used loosely here in a cognitive sense and does not imply an ordering of values, as in the mathematical sense of the word.

[2] Interestingly, the error characteristics of the datasets change over time as the sensor picks up damage, so this too is a range rather than a point.

Moving a dataset from one point in model space to another point will incur a series of costs related to: the work done, changes in accuracy or resolution, changes in semantics, etc. We routinely DO move data in model space but we typically do not account for all the changes that ensue. We aim to address this shortcoming by representing the model space and describing (as richly as we can) what happens to datasets that are transformed from one point in this space to another. We can assign a cost function for each dimension (i.e., a distance metric) that allows us to account for the cost of transforming data from one model to another. A data transformation is represented as a function that takes one or more datasets and their associated models and returns a tuple consisting of a new dataset and its location within the model space. Finally, each model in the space has its own sets of behaviours, translators (to other models), constraints, and supported data structures.

An important difference between the traditional SDI and next-generation SDI is that because of the heterogeneity of producers, unlike the traditional model of a centralised repository, the next-generation SDI will be distributed and federated. It will thus need to incorporate data from disparate sources that are not controlled from within the SDI, in line with the paradigm of linked data (Bizer et al., 2009; Schade et al., 2010). Perhaps more importantly, the idea of SDI as simply an ingester of data will change. Geospatial data sources such as sensor networks and social media feeds are increasingly real-time and configurable. For example, a sensor network may be able to sample some phenomenon every day, hour or minute and may be able to report the value in a variety of different units (e.g. Celsius or Fahrenheit). An SDI may also need to *negotiate* with its data sources on behalf of the user. This means that the tasks that the SDI performs are not just search and integration (i.e. pulling) but also communication and requests for re-configuration and new information (i.e. pushing). From the perspective of describing the characteristics of data, therefore, the purpose is not only for publishing, sharing, and integration of data from static sources but also for communicating "what the user wants" back to sources, so they can better meet the need. Developing this capability will become essential because we can easily imagine that a universal data harmoniser might require a combinatorial explosion of pairwise translators. It might be much easier just to ask the source again for the data to match the user's needs![3]

Figure 1 provides a schematic view of one way such a next-generation SDI might be architected. At the heart of the proposed infrastructure are three layers of functionality shown in shades of green. An outer **Federation and Analysis Layer**, in which all supported datasets are descriptively rich, interoperable, and can be readily combined in analysis. This federation layer serves as the shell in which all data known to the infrastructure can be discovered, queried, analysed or shared. A **Mediation Services Layer** comprising of software services to transform geospatial data sources with the supported conceptual models in the Kernel. Specific mediation services to harmonise a given data source are shown as jigsaw pieces in the figure. The services in this layer are created and maintained by geospatial knowledge engineers, and are used by domain experts to create the required mediation services. A **Knowledge Representation Kernel**, used to describe and create rich descriptions of geospatial knowledge and the conceptual models of geospatial information that underpin the various exchange formats and analysis methods, along with their description semantics. This layer also includes the fit-for-purpose reasoning, which uses rich descriptions to create a 'recommender system' for geospatial data selection.

3 Dimensions of model space

The volume, variety and velocity of geospatial data create an extremely large model space that must be navigated smartly to facilitate meaningful discovery and analysis activities (Cavoukian and Jonas, 2012). Any fruitful paths taken through the space in order to transform the data to make it commensurate will be dictated by the application contexts. From a theoretical point of view then, the core research problems require us to identify the useful morphisms[4] that map from one part of the model space domain to another. The difficult research challenges arise because geospatial data is often highly contextual (interpreted), and there are many models in use, often with missing or implied dimensions such as semantics, accuracy and authority.

The creation of harmonised information from heterogeneous datasets presents us with several research challenges. Table 1 summarises some of these challenges. Most are current research themes that are usually studied separately in the GIScience literature that will need to be: (i) extended where needed, then (ii) integrated together. For the purposes of this paper, we have chosen six different domains along which we can define characteristic dimensions for geospatial data: 1) spatio-temporal frameworks, 2) semantics, 3) access and licensing, 4) provenance, 5) authority, and 6) quality. These domains are not entirely separable, as values in one may have a bearing on others in many cases, but it serves as a useful hierarchical organisation for the myriad dimensions that can describe geospatial data. The first three challenges

[3] Of course it makes sense to avoid the need for pairwise translators by favouring a small number of models with well-understood paths between them, as we currently see in most GISystems.

[4] A morphism is a structure-preserving mapping between two abstract conceptual or mathematical structures, such as is used in set theory and various description algebras.

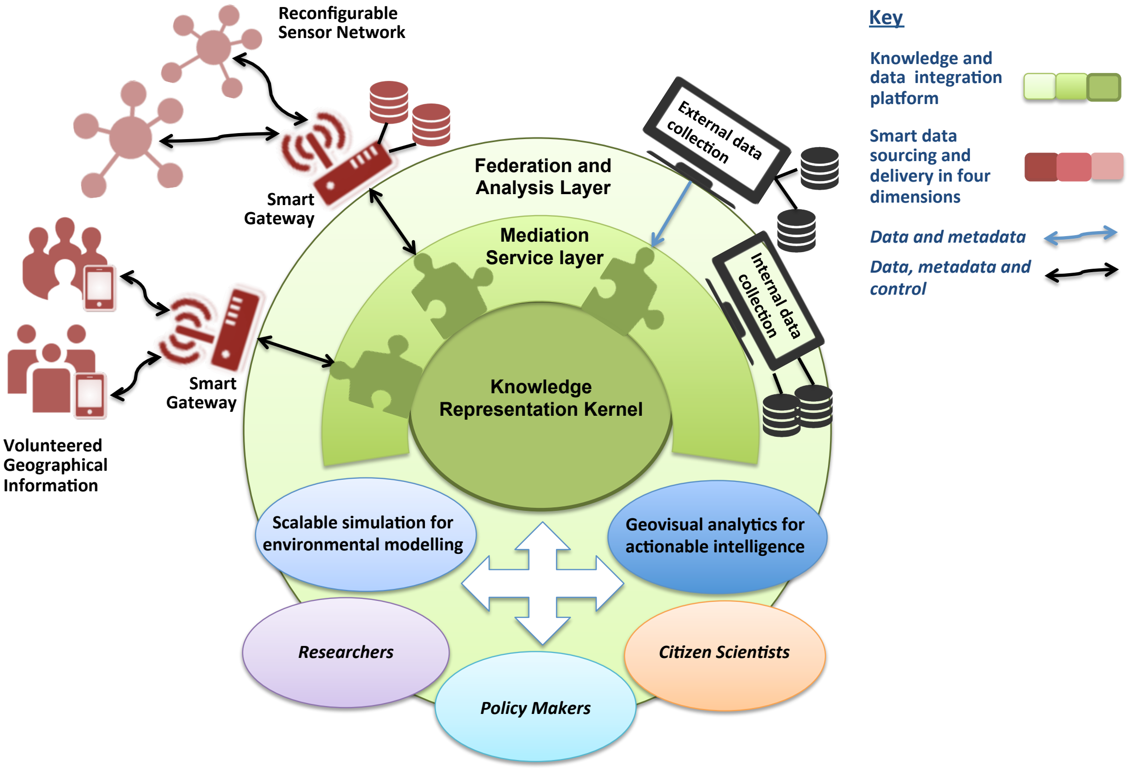

Reconfigurable Sensor Network
Smart Gateway
Smart Gateway
Volunteered Geographical Information
Federation and Analysis Layer
Mediation Service layer
Knowledge Representation Kernel
External data collection
Internal data collection
Scalable simulation for environmental modelling
Geovisual analytics for actionable intelligence
Researchers
Policy Makers
Citizen Scientists
Key
Knowledge and data integration platform
Smart data sourcing and delivery in four dimensions
Data and metadata
Data, metadata and control

Figure 1: An holistic view of how a next-generation SDI might be architected to harmonise heterogeneous data for multiple purposes.

address data harmonisation, the latter three expand to issues of governance and lead to a holistic measurement of *fitness-for-purpose* (Georgiadou et al., 2006; Devillers et al., 2007). Of course, many other domains could be chosen, these are not an exhaustive set, and new domains may emerge in the future (for example to support epistemology to go with ontology).

We consider the dimensions described here are important to most contemporary SDIs. The actual computational solutions for measuring along these dimensions will to some degree depend on the needs of the system being developed. Thus, we are not proposing a universal framework that will work for all data and every SDI, but rather a series of “best-practices” combined with a highlighting of what we see as the most pertinent research challenges to advance SDI. We fully anticipate that the structure of these dimensions will become more nuanced as we develop computable frameworks that can reason over all of them in concert.

3.1 Spatio-temporal frameworks

Spatio-temporal frameworks have been extensively studied in GIScience over that last couple of decades. These spatiotemporal models (which we term here spatial frameworks after Worboys and Duckham (2004)) concern the how the structure of space and time are represented in the data. While the format of the data (i.e., the syntactic description of the data) will often imply a specific spatial framework, that is not always the case – the same format can be used for data described by different models and vice versa. Some of the challenges for data harmonisation include the following aspects of the spatial and temporal framework in which the data reside. The tessellation of the space can be continuous, a regular grid, or an irregular grid. Worboys and Duckham (2004) describe the common bifurcation between field-based and object-based representations, though alternate models are proposed, including image-based, which has characteristics of both, as well as non-spatially explicit models, which are becoming more prevalent in research on place-based representation (Gahegan, 1996; Winter and Truelove, 2013). Other important spatiotemporal dimensions concern the projection, scale, and resolution of the data. Finally, the representation of time can be continuous or discrete.

Dimensions of *model space*	**Challenges for data harmonisation**	**Possible solutions**
Spatio-temporal frameworks (Worboys and Duckham, 2004; Gahegan, 1996)	• Tessellation of the space (continuous, regular grid, irregular grid), Field-, image-, object-based, or non-spatially explicit • Projection, scale, and resolution • Time	Devise a formal, holistic conceptual framework to encompass the variety of models along with the required tools to move data between these models.
Semantics of attributes (Egenhofer, 2002; Bishr and Kuhn, 2007; Brodaric and Gahegan, 2007; Gahegan et al., 2009; Adams and Janowicz, 2011; Janowicz et al., 2013)	• Measurement scale (ratio, interval, ordinal, categorical, unstructured, tuple) • Implied or missing semantics • Ontology alignment	Develop ontology creation and alignment tools, using both formal (top down) and informal (bottom-up, via use-cases) approaches.
Access and licensing (Onsrud et al., 1994; Miller et al., 2008; Cavoukian and Jonas, 2012; Hosking and Gahegan, 2013)	• Access rights to the data, according to purposes • Rights to update or propagate changes	Research suitable security models for use in a distributed setting.
Provenance (Clarke and Clark, 1995; Bose and Frew, 2005; Simmhan et al., 2005; Ludäscher et al., 2006; Belhajjame et al., 2013)	• Author and source • Workflow used to generate data	Extend current provenance research to explicitly represent key aspects of geospatial information processing.
Authority (Gahegan and Pike, 2006; Flanagin and Metzger, 2008; Coleman et al., 2009; Bishr and Kuhn, 2013)	• Authoritativeness of the source (top down) • Trustworthiness of the contributing individual or organisation (bottom up)	Develop models for digital governance that can encompass both *imposed* and *earned* authority
Quality (Chapman, 2005; Pike and Gahegan, 2007)	• Confidence in intended semantics • Confidence in the process used to generate data • Propagation of uncertainty through the analysis workflow	Extend and integrate spatial data accuracy methods to work within this context.

Table 1: Summary of some of the complex dimensions that comprise the model space for geospatial data and some of the related harmonisation challenges.

3.2 Semantics of attributes

Representing the meaning of attributes, i.e., the non-spatial data associated with features represented in a geographic dataset, presents a significant challenge to the next-generation of SDI. When present these semantics are often communicated informally, e.g., as table column labels, which poses a problem for building an SDI designed to do automated data transformation and harmonisation. Movement toward more formal representation of attribute semantics using the languages of the semantic web has been advancing but much VGI data are described in less structured ways (e.g., folksonomic tags) (Egenhofer, 2002; Bishr and Kuhn, 2007; Janowicz et al., 2013). Thus, while the semantics of the attributes are ideally well-understood by the creators of data, they are often missing when data are communicated. It is also the case that meanings of geographic concepts vary not only across but within communities (Brodaric and Gahegan, 2007). Toward the goal of calculating the "fitness-for-purpose" of a dataset, it will not always be the case that an estimation of attribute meaning can be made through formal reasoning and ontology alignment but rather will rely on other fuzzier rules and patterns (derived either through data mining and machine learning or via use-cases) to suggest better or worse fitness-for-purpose (Gahegan et al., 2009; Adams and Janowicz, 2011).

3.3 Access and licensing

With a more distributed and federated structure and subsequently less control over many data sources, the data processed through an advanced SDI will likely be governed by multiple and at times conflicting access and licensing restrictions. Despite the academic community's interest in open licensing for linked data, much geospatial data that we want to make accessible through SDI will be restricted in terms of use (Miller et al., 2008). This includes derived products from analyses of crowdsourced data collected through commercial applications such as Twitter[5]. We will need to develop ways of characterising the access and licensing models for secondary datasets that are derived from multiple, external sources and published through the SDI (Hosking and Gahegan, 2013). The SDI must be able to negotiate between a user profile model that describes access rights and the access model for the data. Another important aspect of access management is the need to build privacy-preserving mechanisms into the data by design as much as possible (Onsrud et al., 1994; Cavoukian and Jonas, 2012).

3.4 Provenance

In addition to modelling the state changes in the data model, a next generation SDI will also maintain descriptions of the provenance semantics of the data. These aspects of provenance include information about the author and source as well as the workflow used to generate the data. Knowing the path that a dataset has taken through the entire model space would provide very useful insight into its likely accuracy and utility for a specific task. Research on provenance representation in eScience and scientific workflow systems will be extended to represent key aspects of geospatial information processing operations and how they relate to the model space (Clarke and Clark, 1995; Bose and Frew, 2005; Simmhan et al., 2005; Ludäscher et al., 2006; Belhajjame et al., 2013).

3.5 Authority

Authority refers to a characterisation of the data producer in terms of its status within a Community of Practice (Gahegan and Pike, 2006; Coleman et al., 2009). Formal integration of authority models into SDI are increasingly needed now as we move away from the architecture of a centralised repository. We can describe authority of the source top down, or rather represent it bottom up in terms of trustworthiness of the contributing individual or organisation. Community and trust in VGI is often an emergent phenomenon where contributors gain trust through community interaction and past behaviour (Flanagin and Metzger, 2008). In absence of direct feedback on trust, proxies such as the spatial location of the contributor can be useful indicators (Bishr and Kuhn, 2013).

3.6 Quality

An important set of dimensions for describing data include measures of data quality that are independent of the task (Chapman, 2005). We distinguish measures along these dimensions of quality from evaluations of *fitness-for-purpose*, which we see as an outcome of all the dimensions described in this paper and based on a situational context (Pike and Gahegan, 2007). Examples of quantitative representations of quality are confusion (or error) matrices in a land-cover layers or surveying errors. Other measures of quality are more qualitative, e.g. confidence in the intended semantics of the attributes or confidence in the process that was used to generate the data.

[5] `https://twitter.com/`

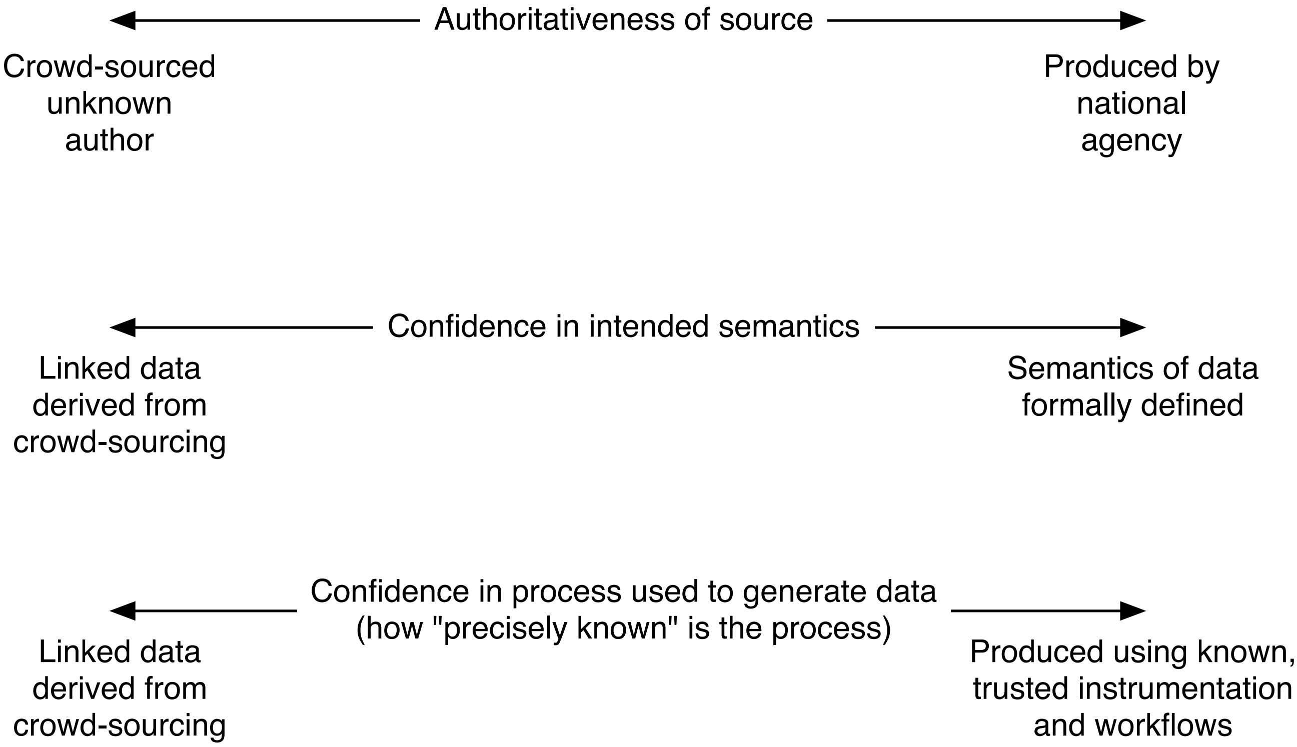

Authoritativeness of source
Crowd-sourced unknown author
Produced by national agency
Confidence in intended semantics
Linked data derived from crowd-sourcing
Semantics of data formally defined
Confidence in process used to generate data (how "precisely known" is the process)
Linked data derived from crowd-sourcing
Produced using known, trusted instrumentation and workflows

Figure 2: Some dimensions of authority and quality contributing to fitness-for-purpose.

The following are some examples of the many different factors that can influence the positioning of a dataset in model space along Quality and Authority dimensions:

1. Field guide describing how data are collected
2. Recorded workflow
3. Stamp of authority from a certifying organisation, which implies a rigorous process
4. Tiers of compliance
5. Peer approval ranking (trust)

Figure 2 illustrates three sample quality and authority dimensions of the model space, along which a dataset can be described. The first dimension is a measure the authoritativeness of the data source, which is defined within the context of a Community Of Practice. The second dimension characterises the semantics of the data; i.e., what they are intended to mean by the original source. The third dimension characterises how precisely known is the process by which the data came to be in its present form prior to being incorporated into the SDI. As with others, these dimensions might be correlated in certain contexts.

4 User Interaction

We envision that the primary function of a spatial data infrastructure will be to enable users to transform, combine, and fit geospatial data to the task at hand – i.e. to provide data that are fit-for-purpose, where the purpose is defined by the application context (Frank et al., 2004). In terms of the model space, this entails the following. First, we must identify where in the model space the user wants to be. Then the system needs to do one of the following: 1) transform data that exist in different models to the one the user wants (Figure 3a), 2) if the user is flexible in terms of the model, identify a set of candidate data from similar models (Figure 3b), or 3) communicate to a data source the need for *new* data that will match the needs of the user. In addition to a rich description of the data in model space, an essential component of an advanced SDI is a user profile that represents preferences in relation to the model space. User templates based on common categories of users and which learn based on previous behaviour of similar users will add efficiency to this functionality. To quote

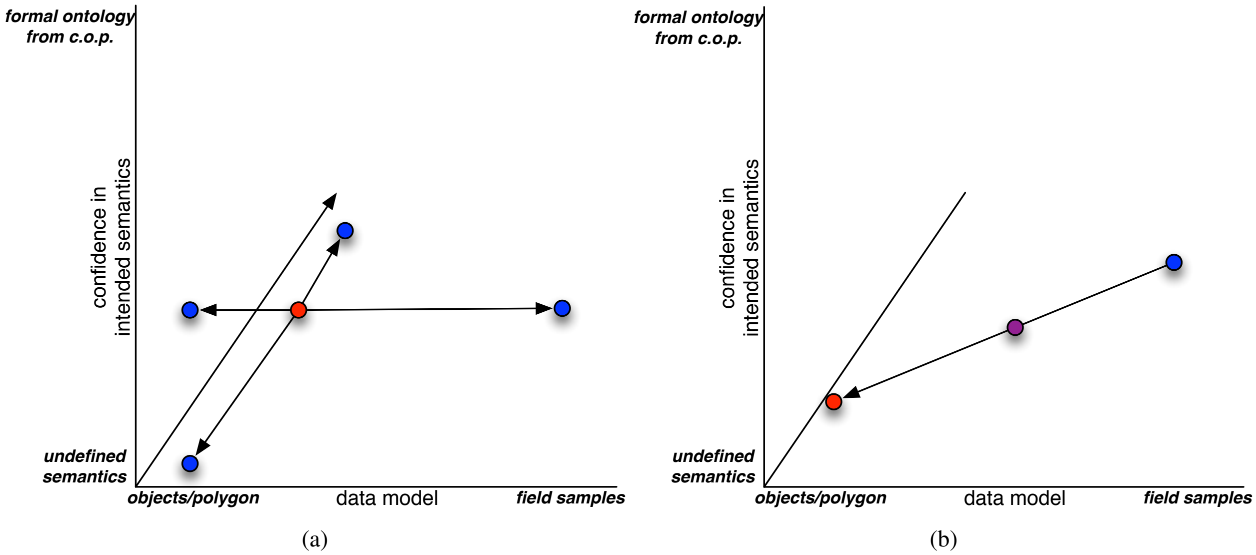


Figure 3: This figure is a schematised representation of example transformations in the model space. The red dot indicates the model desired by the user and the blue dots represent models of data that are available. (a) In the left picture, the system provides a set of data objects (shown in blue) that have similar models but which vary with respect to the dimensions along which they are different. (b) In the right picture, the data object is transformed to fit the desired model through a series of state changes.

the *Hendler hypothesis*[6] we anticipate that "a little semantics will go a long way" to help constrain the context for finding, transforming, and harmonising data. For example, knowing something of the scientific domain in which a user works, such as geodesy or criminology, can help determine where in the model space is best for providing useful data. Likewise, if a user is known to specialise on kinds of features found in specific localities (e.g., a vulcanologist) the SDI can favour data that match the appropriate scale and projection for those locations. The user profile might also provide restrictions on what data are available, based on, e.g., the access model of the data.

Because we describe the model space abstractly in terms of conceptual dimensions, it is a research challenge to develop methods to communicate these dimensions so that a user understands where in the model space they are and where they want to be. We propose that an exemplar-based system that provides visual cues to the user by attaching recognisable icons to common points in the model space will reduce the cognitive load somewhat. Figure 4 shows an example of icons that might be used to reference models that vary along dimensions for the spatio-temporal framework and the measurement scale of the attribute domain. These visual exemplars provide a frame of reference for the user.

4.1 Transforming data

Every GIS operation that transforms data can be more formally represented as movement along a state transition diagram. The state changes shown in Figure 5a are an example of the kinds of transformations that are possible between data represented with different spatial frameworks and attribute measurement semantics. State changes will incur information loss (or possibly gain) depending on the kind of operation done, which will contribute to the measure of quality (accuracy). The process of data transformation will be recorded as a workflow to enable users to reconstruct the steps used in analyses. Every state change is modelled as a binary property change in the data model, e.g., transforming from a continuous value to a categorical value, though richer descriptions may eventually be possible. The dashed connections between the discrete space / field models and continuous space / object models have a step in between, such as discrete space / object model. Rarely, the data is stored in these intermediate representations, it is more typical to use them only as a temporary step. For example, a raster to vector operation that converts integer-valued ordinals will first identify objects in the image space based on connected components with the same value (DS-O-Ord). Then, the space is made continuous by converting the edges of the regions to vector form (CS-O-Ord).

Figure 5b shows an example of how the transformation of sensor measurements of rainfall to polygonal regions can be modelled using this state diagram. The dataset begins as a set of rainfall measurements collected by a sensor network in the environment. The data are field-based measurements that are continuous-valued (e.g., in mm). In step 2, an operation

[6] `http://www.cs.rpi.edu/~hendler/LittleSemanticsWeb.html`

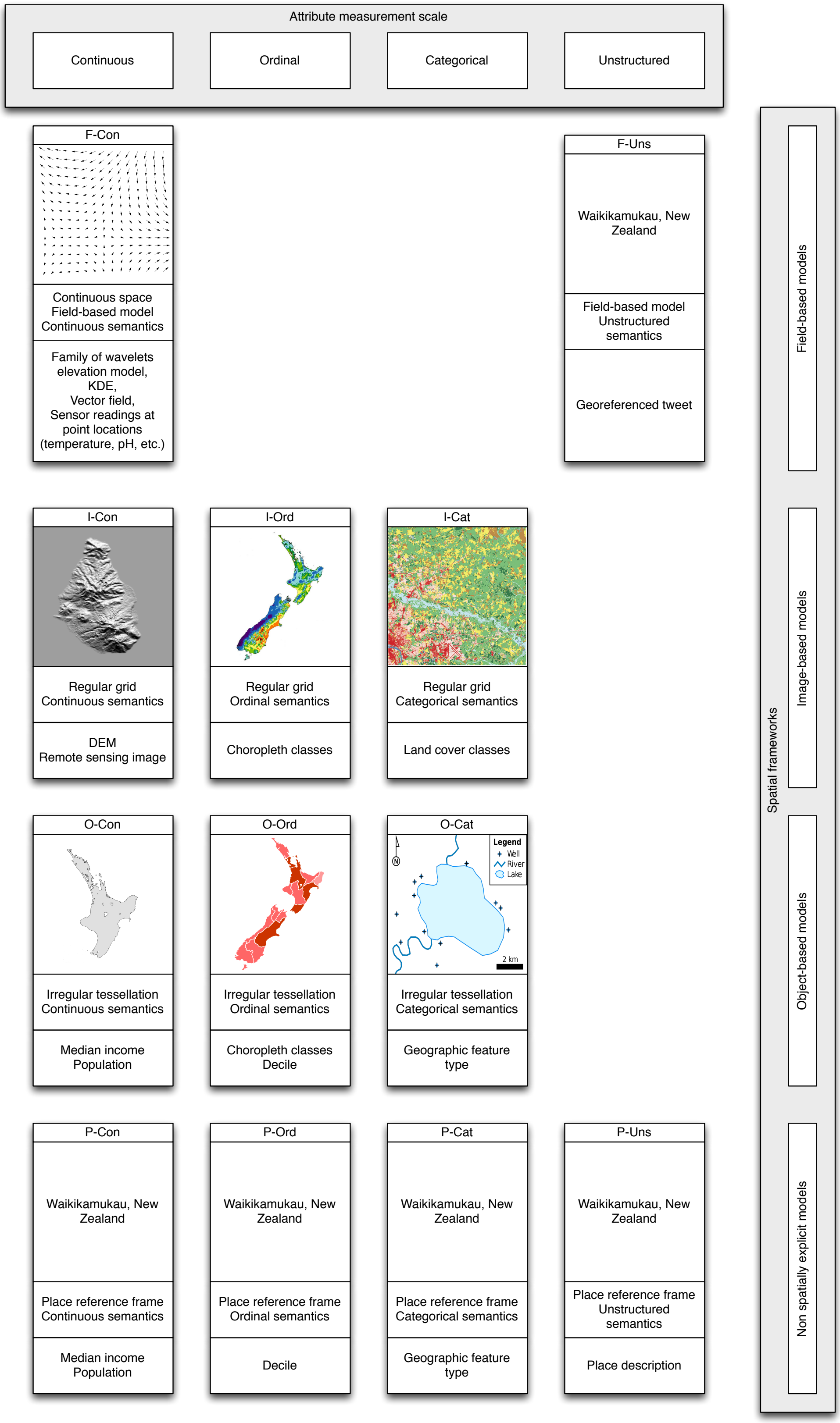
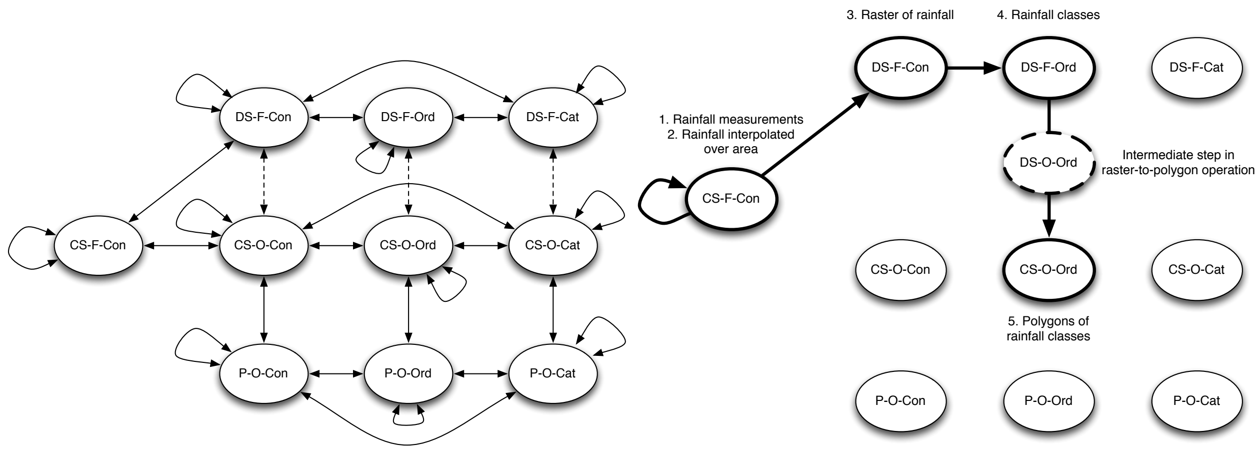


Figure 4: Example of conceptual data model icons that can be used by an SDI to communicate common points in the model space to data consumers. Visual exemplars provide a frame of reference for the consumer.



(a) (b)

Figure 5: Data transformation as a movement in a state transition diagram: (a) Data model state diagram showing data transformation on the edges; (b) Example of state transformation for rainfall data.

interpolates values for locations not being observed, and the result is converted to an image-space representation in step 3. The continuous values in the raster pixels are binned into ordinal classes in step 4, and a raster-to-polygon operation transforms the field-based view into discrete objects (represented by polygons) in step 5.

In order to combine data from different sources meaningfully in an analysis, it is often necessary that they be transformed to share the same spatio-temporal framework. Note of course that just because datasets share the same spatial framework they may still be incommensurable due to, e.g., different categorisation schemes, and so harmonisation will require transformation along other dimensions, such as semantics.

5 Summary and conclusions

If a researcher is able to clearly state their needs in terms of the dimensions of the model space, then the system can more clearly provide recommendations of data to use (or not use). The next-generation SDI we are proposing is moving away from the model of a centralised data warehouse toward a federated data network model, where the SDI acts as a mediator that harmonises data generated from heterogeneous sources and provides data to a user. New data sources, such as sensor networks and crowd sourcing, that provide streaming and high-resolution data in real-time will become important parts of the SDI. The data from these sources will be emergent and dynamic. At the edge of the system architecture these data will need to be translated to the model space adopted by the SDI. Despite the challenges of harmonising data from so many different kinds of sources, the dynamic nature of these sources provides opportunity to significantly expand the "control" functionality of the SDI to push data requests out to the sensors, thereby allowing a user to interact directly with the data providers. By specifying a data-need (in terms of location in model space) at the user side, an advanced SDI should be able to push a request for data that fits the specified parameters out to the data sources, which will reconfigure and acquire the needed data.

Formalising this control mechanism is an important research challenge and will require the development of "smart gateways" with the appropriate communication protocols that 1) broadcast the range of models available to the SDI, 2) reconfigure components as needed, 3) wrap the physical data output with a description vis à vis the model space, and 4) handle data requests from the SDI. For example, for a sensor network requests might involve reconfiguring to sample at higher temporal resolution. We can imagine these smart gateways working for human sensor networks as well. For example, spatially-referenced social media data has been shown to be a valuable data source for health-related information (Paul and Dredze, 2011). With a formal representation that maps Twitter hash tags to health-related concepts, a well-designed smart gateway could – based on the statistics of tags in the Twitter feed – communicate to an SDI that it is able of producing flu-related datasets on request, at certain spatial and temporal resolutions. A related challenge exists for the traditional SDI data catalog: how are the range of possibilities for such a dataset to be represented to the user? More generally, a number of usability challenges exist, including communicating the dimensions of the model space to consumers of data and facilitating interaction between consumers and producers.

Perhaps the biggest challenge to the current status quo concerns how we represent geospatial features and sets of features (datasets). Most data exchange formats, whether proprietary or open, do not support the inclusion of the kind of rich semantics and pragmatics described here. However, all that is required is that geospatial data carries with it a persistent link back to the rich descriptions formulated and maintained by the advanced SDI, such as a SPARQL endpoint. There is some movement towards richer descriptions in open exchange formats, The Open Geospatial Consortium (OGC) initiatives already support the notion of legends, and the beginnings of workflow and domain semantics to go with its feature and map exchange standards[7].

References

Adams, B. and K. Janowicz (2011). Constructing geo-ontologies by reification of observation data. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 309–318. ACM.

Belhajjame, K., J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, and J. Zhao (2013). PROV-O: The PROV ontology. Technical report.

Bishr, M. and W. Kuhn (2007). Geospatial information bottom-up: A matter of trust and semantics. In *The European information society*, pp. 365–387. Springer.

Bishr, M. and W. Kuhn (2013). Trust and reputation models for quality assessment of human sensor observations. In *Spatial Information Theory*, pp. 53–73. Springer.

Bizer, C., T. Heath, and T. Berners-Lee (2009). Linked data-the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS) 5*(3), 1–22.

Bose, R. and J. Frew (2005). Lineage retrieval for scientific data processing: a survey. *ACM Computing Surveys (CSUR) 37*(1), 1–28.

Brodaric, B. and M. Gahegan (2007). Experiments to examine the situated nature of geoscientific concepts. *Spatial Cognition and Computation 7*(1), 61–95.

Budhathoki, N. R., B. C. Bruce, and Z. Nedovic-Budic (2008). Reconceptualizing the role of the user of spatial data infrastructure. *GeoJournal 72*(3-4), 149–160.

Cavoukian, A. and J. Jonas (2012). Privacy by design in the age of big data. *Office of the Information and Privacy Commissioner*.

Chapman, A. D. (2005). *Principles of data quality*. GBIF.

Clarke, D. G. and D. M. Clark (1995). Lineage. *Elements of spatial data quality*, 13–30.

Coleman, D. J., Y. Georgiadou, and J. Labonte (2009). Volunteered geographic information: the nature and motivation of produsers. *International Journal of Spatial Data Infrastructures Research 4*(1), 332–358.

Devillers, R., Y. Bédard, R. Jeansoulin, and B. Moulin (2007). Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data. *International Journal of Geographical Information Science 21*(3), 261–282.

Egenhofer, M. J. (2002). Toward the semantic geospatial web. In *Proceedings of the 10th ACM international symposium on Advances in geographic information systems*, pp. 1–4. ACM.

Flanagin, A. J. and M. J. Metzger (2008). The credibility of volunteered geographic information. *GeoJournal 72*(3-4), 137–148.

Frank, A. U., E. Grum, and B. Vasseur (2004). Procedure to select the best dataset for a task. In M. J. Egenhofer, C. Freksa, and H. J. Miller (Eds.), *GIScience*, Volume 3234 of *Lecture Notes in Computer Science*, pp. 81–93. Springer.

Gahegan, M. (1996). Specifying the transformations within and between geographic data models. *Transactions in GIS 1*(2), 137–152.

Gahegan, M., J. Luo, S. D. Weaver, W. Pike, and T. Banchuen (2009). Connecting GEON: Making sense of the myriad resources, researchers and concepts that comprise a geoscience cyberinfrastructure. *Computers & Geosciences 35*(4), 836–854.

[7] `http://www.opengeospatial.org/standards/wmc`

Gahegan, M. and W. Pike (2006). A situated knowledge representation of geographical information. *Transactions in GIS 10*(5), 727–749.

Georgiadou, Y., O. Rodriguez-Pabón, and K. T. Lance (2006). Spatial data infrastructure (SDI) and e-governance: A quest for appropriate evaluation approaches. *URISA-WASHINGTON DC- 18*(2), 43.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal 69*(4), 211–221.

Groot, R. and J. D. McLaughlin (2000). *Geospatial data infrastructure: concepts, cases, and good practice*. Oxford university press Oxford.

Hey, T. and A. E. Trefethen (2005). Cyberinfrastructure for e-Science. *Science 308*(5723), 817–821.

Hosking, R. and M. Gahegan (2013). The effects of licensing on open data: Computing a measure of health for our scholarly record. In *The Semantic Web–ISWC 2013*, pp. 432–439. Springer.

Jacoby, S., J. Smith, L. Ting, and I. Williamson (2002). Developing a common spatial data infrastructure between state and local government–an australian case study. *International Journal of Geographical Information Science 16*(4), 305–322.

Janowicz, K., S. Schade, A. Bröring, C. Keßler, P. Maué, and C. Stasch (2010). Semantic enablement for spatial data infrastructures. *Transactions in GIS 14*(2), 111–129.

Janowicz, K., S. Scheider, and B. Adams (2013). A geo-semantics flyby. In S. Rudolph, G. Gottlob, I. Horrocks, and F. van Harmelen (Eds.), *Reasoning Web. Semantic Technologies for Intelligent Data Access*, pp. 230–250. Springer.

Ludäscher, B., I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. Jones, E. A. Lee, J. Tao, and Y. Zhao (2006). Scientific workflow management and the kepler system. *Concurrency and Computation: Practice and Experience 18*(10), 1039–1065.

Mäs, S., M. Müller, C. Henzen, and L. Bernard (2011). Linking the outcomes of scientific research: Requirements from the perspective of geosciences. In T. Kauppinen, L. C. Pouchard, and C. Keler (Eds.), *LISC*, Volume 783 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Masser, I. (1999). All shapes and sizes: the first generation of national spatial data infrastructures. *International Journal of Geographical Information Science 13*(1), 67–84.

Miller, P., R. Styles, and T. Heath (2008). Open data commons, a license for open data. In *LDOW*.

Onsrud, H. J., J. P. Johnson, and X. Lopez (1994). Protecting personal privacy in using geographic information systems. *Photogrammetric Engineering and Remote Sensing 60*(9), 1083–1095.

Paul, M. J. and M. Dredze (2011). You are what you tweet: Analyzing twitter for public health. In L. A. Adamic, R. A. Baeza-Yates, and S. Counts (Eds.), *ICWSM*, pp. 265–272. The AAAI Press.

Pike, W. and M. Gahegan (2007). Beyond ontologies: Toward situated representations of scientific knowledge. *International Journal of Human-Computer Studies 65*(7), 674–688.

Schade, S., C. Granell, and L. Diaz (2010). Augmenting SDI with linked data. In K. Janowicz, T. Pehle, G. Hart, and P. Maué (Eds.), *Proceedings of the Linked Spatiotemporal Data Workshop (LSTD 2010 GIScience 2010 Zürich)*, pp. 34–45.

Simmhan, Y. L., B. Plale, and D. Gannon (2005). A survey of data provenance in e-science. *ACM Sigmod Record 34*(3), 31–36.

Winter, S. and M. Truelove (2013). Talking about place where it matters. In *Cognitive and Linguistic Aspects of Geographic Space*, pp. 121–139. Springer.

Worboys, M. and M. Duckham (2004). *GIS: a computing perspective*. CRC press.