Appendix 3 Usage case for the design of an Asian Regional Economic Integration Observatory

We are in the midst of a data science revolution (Mayer-Schönberger and Cukier, 2013) that is transforming what we might observe and infer about the world around us; deepening the level of insight that we can build in specific domains of interest; expanding the number and reach of connections that we can establish between people, places and things; and growing our collective ability to ask questions and solve problems. In particular, two catalysts have emerged that have shifted how we now think about data and how we might use it. The first is the widespread availability of massive silos of big data; the second is the expansion of the resources now available to query, interpret and add value to those data. Together, these developments have created efficiencies and economies of scale that have never before been presented, while also lowering many traditional barriers to data resources that have long persisted (Frankel and Reid, 2008; The Economist, 2010).

Nevertheless, at the same time, a growing 'data deluge' (Baraniuk, 2011; Bell et al. 2009) is emerging, and presenting new problems. Chief among these are issues of data privacy and ethics amid data abundance (Dobson and Fisher, 2003; Jacobs, 2009; Mayer-Schönberger and Cukier, 2013; Tene and Polonetsky, 2012). New concerns have arisen around ownership and control over the shadows that our big data cast (Clarke, 1994) and this is particularly salient when data are associated with places and times, as is usually the case with geographic information (Bilton, 2011; Dobson, 2009; Goodchild, 2011; Lessig, 2000; Monmonier, 2002; Seely Brown and Duguid, 2000; Smith et al., 2005). While data has grown more abundant, and the machinery to 'feed on' data has grown more ravenous, so too has the intractability, intricacy and complexity of those data and the connections between them (Ouellette, 2013; West, 2013). Long the domain of official data-collection agencies, much big data is now being generated within the commercial sector, with concerns that the knowledge

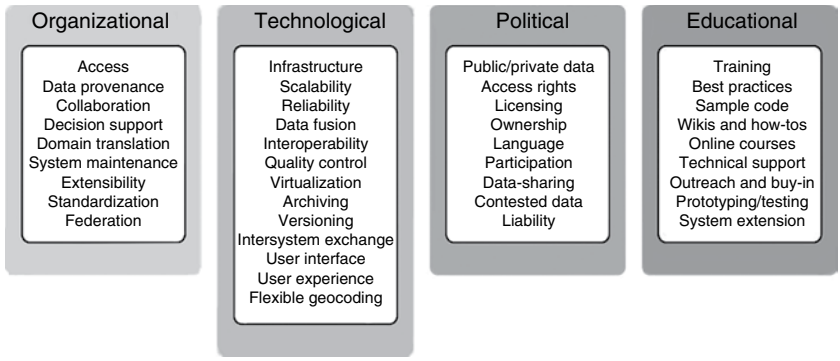
and computer resources to leverage the benefits of those data might be cloistered behind the companies that have the scale to search and serve it (Picciano, 2014). Similarly, many worry that traditional forms of training and education might be outmoded in the face of entirely new ways of doing business (Anderson, 2008; Nature Publishing Group, 2008; Trelles et al., 2011).

While these concerns must be considered when planning any data platform, there remain significant opportunities for the Asian Development Bank (ADB) and its partner communities to build a landmark knowledge platform that, with careful planning and robust design, could leverage many of the benefits of the emerging data science revolution, while mitigating potential complications. The development of a knowledge platform design with an open architecture, atop diverse community data resources, with flexibility to adapt to future innovations and scale for growth, and providing varied paths to entry and use could establish ADB as a leader in next-generation architectures and platforms for regional cooperation and integration.

Achieving this practically and usefully will require some work. It necessitates innovative – and innovatively applied – *dataware*. By dataware, we refer to the tools necessary for collecting, collating, managing and communicating data, as well as techniques for adding value to data, via models, dashboards, statistical analyses, economic metrics, impact assessments, visualization and so on. Developing this dataware and leveraging it in service of an agile and scalable knowledge platform requires negotiation of technical and operational challenges that could potentially span diverse geographies, organizations, data sources, data types, systems and interaction schemes, as well as languages, cultures, topics, ontologies and domains of expertise. Moreover, navigating these aspects of the system design requires solutions that are in equal parts organizational, technological, political and educational (Figure A3.1).

The existing use-case and development-case scenarios for geographic information systems (GIS)-based and data-based regional observatories and exploration platforms fall into the following taxonomy; however, there is increasing interoperability across these taxonomies, with the result that systems with massive reach, scope and potential utility are now feasible.

Development agencies Several development agencies have instituted systems that are map-based and GIS-based (by GIS-based, we mean systems that are or may be map-based at the level of their interface, but offer the additional ability to download, repurpose and/or analyze the data using geographic-based queries on the database side of the system). They are producing and publishing – often on public-facing websites – much of



Source: Author.

Figure A3.1 Initial considerations and opportunities for designing and positioning knowledge platform

their development data (investments, targets, priorities, impacts and indicators) for public communication. In several instances, these data may be downloaded in a variety of formats for further analysis.

Universities Academic research and development efforts are also underway to produce GIS-based development data portals, either to support (1) open GIS, citizen mapping, or volunteered geographic information, or (2) to *use* the portals in service of substantive investigation into human geography, economics, international affairs, government and politics, disaster relief and so on. Many of these groups have partnered with the other use cases to collaborate (The College of William and Mary, Brigham Young University, and The University of Texas at Austin with United States Agency for International Development (USAID) and Environmental Systems Research Institute (ESRI) on the *AidData* portal, for example).

OPENAID PARTNERSHIP

Citizen advocacy There is a relatively long tradition of map-based portal development for public participation purposes (Elwood, 2010). Recently, this has evolved into dedicated citizen mapping initiatives (Hodson, 2013), based largely around the idea of first crowdsourcing the *task* of data collection (Bonney et al., 2009; Cohn, 2008; Goodchild, 2007; Hand, 2010) and portal development (Coast, 2011; Haklay, 2010; Haklay et al., 2008;

Haklay and Weber, 2008), and second, using the products of that work in an open fashion to foster citizen engagement in development decisions and impacts (Eagle, 2009; Hagen, 2011; Nelson, 2011).

Software companies Providing support platforms (databases, data models, map layers, data access schemes, base data, visualization, virtualization, servers) in support of these activities. These companies are behind (or partnering with) the other use-case scenarios, for example, ESRI, CartoDB, MapBox. In other cases, they have built value-added platforms atop these primary providers (for example, Development Gateway using ESRI services).

Government agencies As part of open government initiatives, many local, regional and national governments have begun to produce data portals, organized around GIS. Development groups within governments have also begun to produce these for their investments and priorities (for example, the United Kingdom's Department for International Development's *Development Aid Tracker*). Recently, similar initiatives have appeared at national and international level. Chief among these is the European Union's portals to their existing datasets, which have evolved from GIS data *infrastructures* developed for client-server interactions (for example, the Infrastructure for Spatial Information in the European Commission), to fully fledged publically oriented portals and platforms for widespread access and reusability (see the latest version of Eurostat, for example, or plans for the second generation of the European Cluster Observatory (http://www.tci-network.org/news/776)). At global level, the United Nations has also developed several exploratories. Some are focused on specific UN missions (global health, for example), but they have recently published a wide-reaching initiative, *UN Global Pulse*, designed as an umbrella portal to a wide-reaching set of data, with many paths to entry to those data across media and datasets, with the aim of scaling to massive datasets across interest domains (United Nations, 2012). Their recent work in docking streaming social media data to Global Pulse has attracted significant attention for its innovation and the potential range of applications to which it can be applied (Lohr, 2013).

Universities There are also a number of university and academic groups that have built, or are building cluster-type observatories as part of their research exploration. In the United States, the recent (October, 2014) *Cluster Mapping* initiative from the Harvard Business School is one such high-profile effort (http://clustermapping.us/).

DATA

Data Sources

While the knowledge platform will provide the scaffolding for information-sharing, use and contextualization, it must be sourced in *raw data* before the information can be produced. These data are likely to come from a variety of sources, some of which may currently be known and for which design goals can be established, but also for other data that are yet to be known, such that the platform should support extensibility in a way that future-proofs its development, use and media.

Legacy data products It is most likely that a large volume of previously developed, previously produced and previously collated datasets are 'floating around' already. These could, if corralled and reconciled into a cohesive system, provide much of the foundation for the knowledge platform. Indeed, the fact that the Asian Development Bank is likely to have these data on hand, or within reach, could establish significant headway in building the platform. These data could be in digital format. Data in paper format can be digitized and, in some cases, geo-referenced, if in map form. At a minimum, data that are in the form of documents or reports can be digitized to portable format and *indexed* in their native format to the GIS side of the platform, simply by assigning them a location and storing them in the spatial database.

Official sources Several datasets will be likely to be sourced from official data collection efforts. These would probably include sources from governmental agencies, such as censuses of population and economic activity; tax and property records; city and regional planning surveys; road network files; building locations; land use, political, and legal boundaries and so on. Other official sources could come from non-governmental agencies (NGOs), and this is a significant point as many NGOs are beginning to get involved in data collection and certification of spatial data records, with specific efforts to foster public participation and transparency in the production of data (Crooks and Wise, 2013; Hagen, 2011; Nelson, 2011). International agencies (or quasi-international, or international-facing agencies) are also a significant source of official data. This is particularly true for remotely sensed data, such as land use and land cover, or environmental characteristics of the Earth's surface. Several other official datasets could come from companies and commercial entities, and this is significant for many aspects of economic development.

Unofficial sources In many cases, unofficial data sources may be a significant product for the platform (or at least could be usefully allied to other datasets on the platform for the purposes of analysis). In particular, *case studies* could provide a significant source of insight. These could come from a variety of sources: students working on projects, local groups with site-specific insight, records of local agricultural or industrial companies and so on. Negotiating access to these data could be challenging, but the platform could be developed to encourage these groups to *volunteer* their data to the platform, in the interest of the public good, perhaps, but more likely as a mechanism to *add value* to those data, by connecting them to the other data sources provided by the platform, and by connecting them to the tools for analysis that the platform might afford.

Machine-sensed data Data from machines, particularly orbiting imaging platforms, can be critical in providing (1) a steady-stream of data that is up to date; (2) an unbiased (or at least objective) view of conditions on the ground; and (3) a 'big picture' view of Earth surface features and human activity. Of course, these assertions must be tempered with the reality that the choice of machine, the view, the time of data capture, and the lens through which things are seen or data are collected may themselves be subject to bias. Moreover, machine-sensed data must often be processed (as imagery, classified variables such as land-use and land cover, interpreted features) and this also introduces the notion of bias and uncertainty. However, there have been significant advances in using remote sensing, in particular, to gather data about a variety of human (Sutton et al., 2001), built (Elvidge et al., 1999; Elvidge et al., 2007; Herold et al., 2002), infrastructural (Elvidge et al., 2011), physical (Akgun et al., 2012; Hodgson and Bresnahan, 2004; Moore et al., 1993; Townsend and Walsh, 1998), environmental (Curran, 1989; Gao, 1996; Schmidt and Karnieli, 2000; Voogt and Oke, 2003), development (Elvidge et al., 2009; Foody, 2003; Xiuwan, 2002), and economic attributes (Field et al., 1995; Jensen and Cowen, 1999; Moran et al., 1997; Sutton et al., 2007). In many instances these data are proprietary to the commercial companies that collected them, but many others are free for use in the public commons. Moreover, there are a variety of national agencies that have collected these data, over long periods, in their own national interest. Similarly, there are cooperative agreements across many agencies to collect such data in the global public good. Sensed data could also come from *sensor grids* that are positioned on or below the Earth's surface, or in and around water bodies. Generally, such systems are proprietary in nature, but in some instances they are available for public use (Wright and Bartlett, 2000; Wright and Goodchild, 1997). Data are also increasingly available from instrumented

built and transport infrastructure (smart roads, tool systems, logistics pipelines, environmental sensor web, and so on) and these can similarly be integrated into GIS-based platforms (or streamed dynamically to them) where available (McCullough, 2004).

Social media The rapidly growing volumes of data that are now being produced (passively or actively) by social media platforms and technologies are of great relevance to the development of the platform. Such data represent the most voluminous and rapidly generated sources of data for many social factors and transactional attributes that we have ever encountered. Increasingly, these can be reconciled to common data platforms via GIS, which provides spatial structure across such data (Elwood, 2010; Goodchild, 2007). (However, there is growing realization that these data are often highly biased, and that the quality of their geographical identifiers can be problematic (De Longueville et al., 2010; Elwood, 2008; Flanagin and Metzger, 2008; Haklay, 2010; Haklay et al., 2010).) Nevertheless, a huge variety of data products can be assembled from such data, spanning from human demography (Frias-Martinez et al., 2012b; Frias-Martinez et al., 2010) and activity (Frias-Martinez and Virseda, 2013; Liu et al., 2010), to economics (Frias-Martinez and Virseda, 2012; Frias-Martinez et al., 2012c), culture (Croitoru et al., 2012), politics (Ratti et al., 2010; Sobolevsky et al., 2013), development (Frias-Martinez et al., 2010), sociality (Croitoru et al., 2013; Stefanidis et al., 2013; Vieira et al., 2010), movement and migration (Girardin et al., 2008; Rubio et al., 2010), and land use (Frias-Martinez et al., 2012a). The inclusion of social media data to the *UN Global Pulse* platform, for example, has generated significant coverage for and interest in the system (Lohr, 2013).

Stages in the Data Development Pipeline

At face value, the knowledge platform will be a tool for communicating (visualizing) data to a variety of users. However, within that role, the platform should support a seamless transition from data to information to knowledge to understanding. At each transition point in the chain, different components of the platform should support the transition, and should do this is in different ways.

For spatial data, the primary mechanism for transforming it into information is to place it in its geographical context, by georeferencing it relative to universal spaces (geometry, cartography, topology, networks, time geography, systems diagrams and so on), or to domain-relevant spaces (human geography, urban geography, economic geography, political geography, historical geography, social geography, physical geography, transport

geography, biogeography and so on). However, there are, of course, many other domains that can add value to raw data, well beyond geography, and these can be made geographical by mapping them (counting them in particular places and times, looking for clusters or their absence, performing buffering operations, examining heterogeneity and homogeneity, assessing adjacency and boundary effects, exploring space–time distributions, and so on). This can be quite concerning for the design of information systems, as a potentially massive array of data sources might need to be considered, and multiples of that array may need to be treated to accommodate transformations between them. The geographical sciences have long grappled with this issue (Kwan, 2002, 2012; Kwan and Schwanen, 2009), and to some extent geographic information science has emerged as the most universal solution, to a large degree because of the ability of geography to structure otherwise unstructured data (Blumberg and Atre, 2003; Leavitt, 2010; Mansuri and Sarawagi, 2006; Rao, 2003; Sester, 2000). As data and information systems (considered generally) have grown into large and even massive silos, spatial data handling has emerged as a special data scheme for coping (Gieryn, 2000; Goodchild et al., 2000).

It is therefore prudent (perhaps crucial) that a deliberate (and extensible) georeferencing plan and actionable scheme be developed to handle and grow spatial data from the platform's first principles (see later section on Georeferencing).

Data Types and Data Models

A successful platform should support a wide variety of data types of both a spatial and non-spatial nature. There are potentially a wide variety of data types to be accommodated, but a minimal taxonomy of types would include those listed below. (Note that most mapping services, such as Google Maps and Bing Maps, realistically allow users only to manipulate geometry and attribute data at the interface and application programming interface level of the services that they offer.)

Location data At its simplest level, the platform should provide two-way interaction with location data, that is, it should allow for the querying, display and manipulation of data by location, and it should allow for the data to be uploaded to the system and registered to the system via location. This latter point is significant: the data should have unique location identifiers, where possible, of resolution, accuracy and precision (Goodchild and Gopal, 1989) appropriate to the source and to its use. Moreover, they should be extensible enough to accommodate the expression of location in as wide a variety of contexts and formats as possible, so that a broad

range of geographies can be employed in adding value to the data. This can be difficult when the data are not naturally or natively spatial in nature (for example, when they are produced non-spatially, and rendered spatial after the fact), and here the roles of metadata and geocoding become significant (as we discuss shortly). Furthermore, the nature of location data is currently shifting dramatically, as data streamed from location-aware technologies and services become part of the spatial data ecosystem, and as such data become quasi-ubiquitous for many usage scenarios (Borriello et al., 2005). Thus, the platform needs to be able to rapidly ingest and reconcile data (Hazas et al., 2004; Muthukrishnan et al., 2005).

Relationship data A central component of supporting robust analysis on the platform, and using the data that it provides, as well as docking the platform with related model-based or statistics-based analyses will be to handle relationship data carefully. Here, we refer to the spatial connections (distance (Sui, 2004; Tobler, 1970), adjacency (Anselin, 2003), flow (Tobler, 1987), barriers (United Nations Development Programme, 2009), within and without (Karanja, 2010; Thurstain-Goodwin and Unwin, 2000), isolation (Anselin, 1995), connectivity (Liben-Nowell et al., 2005; Welch and Mishra, 2013) and so on) between data points. Many of these can be treated with modern georelational models (Dueker, 1985), and can be optimized for large datasets using clustering on spatial data access schemes (using quadtrees (Samet, 1984), for example). Ideally, these should also work with standard relationship schemes for other information systems, including object-oriented hierarchy (Gamma et al., 1995), entity-relationship models (Peckham et al., 1995), topology (Ellul and Haklay, 2006), and newly emerging formats such as the Resource Description Framework (Miller, 1998). A resource description framework can operate on metadata (which we discuss shortly) and is therefore a candidate for suprarelationships. Hypergraphs (Gunopulos et al., 1997) can provide similar functionality for network data.

Network data Network data constitute something of a special case of relationship data for the platform because of their significance in ascribing variables and structure to linkages in the system, between entities, and across space and time. Dedicated spatial network data types are possible in most GIS, although they are generally limited to geometry and topology and therefore constrained in the range of operations that they afford in spatial analysis and spatial data access. New forms of spatial network data model, such as SANET (Okabe et al., 2006a) are beginning to be used, and are beginning to be folded into spatial analysis routines (including spatial statistics) (Shiode and Shiode, 2010), but they are academic in nature.

Similarly, graph-based network data types can be employed (and spatial networking could be considered as one property of the graph). Graph structures are straightforward to implement in database systems, but connections to GIS thus far have been rather experimental (Butts, 2009).

Attributes Attributes of the data types already discussed can be represented in the platform via GIS in a straightforward manner, particularly if a georelational data model is employed: the 'geography' can be held in one file, the attributes in a database, and the glue to fashion spatial data exchange can be developed between them. This facilitates the separation and specialization of all three, without necessarily sacrificing interoperability. Georelational techniques also allow users to create their own databases, and then connect them to the platform in a unified (and structured) fashion by performing spatial joins (Patel and DeWitt, 1996), or similar operations (that is, matching and then merging database records via their attribute data type – text, image, values, documents, video media and so on – to the location data type via common indices). These joining and merging operations are now scalable to huge databases and have been optimized for efficient operation over diverse data input streams (Jacox and Samet, 2007).

Place names Place names (*toponyms*) are a special form of attribute data in GIS (Vögele et al., 2003). They are both attribute data and location data, but they are often not unique, and their meaning is often significant across many (sometimes conflicting) axes of consideration. This is further complicated by language and differences in the expression of place names in different dialects or vernacular (Berg and Vuolteenaho, 2009). That place names are sometimes contested or have diverging cultural or historical meaning (Rose-Redwood et al., 2010) is a significant consideration when developing a platform that crosses cultural and political spaces. These issues are of long-standing concern in geography and in GIS and are not well-reconciled. Recent developments in ontologies (database classes for ascribing meaning to data items held within them, as style sheets or equivalents, for example) provide one possible path for reconciling the diverse treatment of place names within a structured spatial data platform: particular names can be invoked when toponyms are well defined or allied to a particular language class or location container (Agarwal, 2004).

Objects Object data types can be reconciled to the spatial database using standard object-oriented schemes, with the advantages of polymorphism, hierarchy and encapsulation that they afford (Gamma et al., 1995; Microsoft Corporation and Digital Equipment Corporation, 1995).

These can also be registered to other data types, in the GIS, if the objects are indexed with functional location types (using schemes such the Component Object Model (Ungerer and Goodchild, 2002), for example). However, objects often require special treatment in GIS, particularly when they have boundaries that are expressed in the GIS: these boundaries can be indeterminate (Burrough and Frank, 1996; Cohn and Gotts, 1996; Schneider, 1996), fuzzy (Schneider, 1999), and contested (Paasi, 1998). They are also subject to change (Galton, 2004) and can move (Gidófalvi and Pedersen, 2009; Wolfson et al., 1998). Issues of defining what, exactly, an object is and what its bounds might be are also often subjective and will probably depend on the context in which the object is placed, and the use for which it is considered (Guesgen and Albrecht, 2000; Guo et al., 2008; Jacquez et al., 2000).

Surfaces Objects and fields have something of a long-standing conflict in GIS (Couclelis, 1992). While objects (market areas, sovereign boundaries, property bounds) are often distinct in their identities (ownership, land use, address) at particular scales of space and time, fields (temperature, travel time, soil moisture) are continuous and subject to sampling, scale and observation in many ways. Fields are thus difficult to represent in object-based GIS platforms, such as the geometry-focused GIS that are predominately used. This makes it difficult to represent surfaces, with related difficulties for supporting conditions as they appear on the ground, and for supporting many data types that one may wish to reconcile to GIS (particularly data from remotely sensed platforms (Fisher, 1997)). In these cases, it is necessary to use spatial analysis and spatial statistics to sample fields as geometries that can be stored and manipulated in a GIS (Anselin, 1995; Anselin et al., 2006; Clark and Evans, 1954; Cressie, 1991; Getis and Ord, 1992; Moran, 1950), or geostatistics that can interpolate and probably estimate surfaces (Fotheringham et al., 2004; Oliver and Webster, 1990; Shepard, 1968). Once derived, surfaces can be stored using mesh data types or raster data types (Goodchild, 1992), and a suite of operators can be employed to process them in those formats (Lu et al., 2008; Mennis, 2010; Yu et al., 2003).

Three-dimensional A decision to incorporate three-dimensional (3D) data types, such as 3D geometries (which can be reconciled in GIS using common data models such as FBX (Filmbox proprietary file format, .fbx), for example, and handled using open source scene graphs (Sun et al., 2014)) or Triangulated Irregular Networks (TINs) (Peucker et al., 1978) is significant. Including these details would render the platform incredibly valuable as a tool, as the incorporation of 3D facilitates a much richer

set of representations of data, and it allows for a wider range of analyses, both on three dimensions of spatial data (aspect analyses, terrain generation, least-cost traversal paths and so on) (Fowler and Little, 1979; Nagy, 1994), but also on multiple dimensions of *any* data (Chen and Guevara, 1987). However, few systems are available to support this. One approach to circumnavigate the issue is to develop extrusion of 2D features into 'two and a half dimensional' data that illustrate variable value. This approach, for example, is common on virtual globe platforms such as *Worldwind* and *Virtual Earth* (Butler, 2006).

Graphs Much of the data to be reconciled in and generated by the platform may usefully be represented by graph data types (Amin and Hakimi, 1973; Dijkstra, 1959; Watts, 2003), that is, as vertices (points, nodes, locations, entities) and edges (links, paths, connections, roads, rails, corridors). In many cases, graphs can be represented natively in a GIS, if we consider them as geographic objects. On the database side, most GIS can handle massive graphs and data structures for big data over graphs (Gupta et al., 2014; Quamar et al., 2014). However, performing graph analyses over them in non-geographical ways (using social network analyses, for example) can be difficult in these cases (Faust et al., 2000; Liben-Nowell et al., 2005; Singleton and Longley, 2009; Ter Wal and Boschma, 2009; Waaserman and Faust, 1994). It should be noted that graph data structures are often significant for data that will be shared, accessed, hosted and reconciled on the Web (including cloud resources that store data in several locations and must treat reconciliation across these databases and locations concurrently or reasonably in synchrony) (Broder et al., 2000).

Time GIS have long grappled with how to represent time, and particularly how to build data models for time that can 'play well' with data models for space (Miller and Wu, 2000). Several schemes for achieving this exist in the field of time geography (Peuquet, 2002; Timmermans et al., 2002). These include transforming time to a third dimension, and docking it with planar geographies to produce space–time paths, space–time prisms, space–time aquariums, and so on. The benefit of this approach is that it opens up time to a rather full range of GIS and database operators and facilitates accessibility (Miller, 1999), sufficiency (Brimicombe and Li, 2006; Miller, 2005) and event-based queries (Chen and Kwan, 2012), such as 'where do these two things intersect in space and time, and for how long?'; 'given this much space and time, how far can this object span?'; 'what is the potential roll-out range for this particular diffusion event?' and so on. Dedicated data-access (Rey and Janikas, 2006) and visualization systems have also been

developed to handle time in this way, such as space–time cubes (Kraak, 2008; Kristensson et al., 2009), and these have been polished for a wide variety of substantive applications (Huisman et al., 2009). Recently, there has been considerable effort to build space–time GIS schemes to handle big data and streaming data (Shaw et al., 2008). This has led to the development of a next generation of space–time data models that should be considered for the platform, including space–time cluster models (Diggle et al., 1995; Wayant et al., 2012; Yamada et al., 2009), space–time network models (Shiode and Shiode 2009), trajectory models (Buchin et al., 2008; Demšar and Virrantaus, 2010), space–time shape models (Gorelick et al., 2007), space–time interest points (Laptev, 2005), and dedicated space–time event and action models (Gatalsky et al., 2004). In many of these cases, the models have been developed to analyze and structure space–time data in dynamic feeds (Wang et al., 2011).

Change It is likely to be critical that the platform treat change in a variety of fashions and dedicated data models to handle change can be introduced. Several schemes have been employed in database theory (Cho and Garcia-Molina, 2000, 2003), and dedicated methods have been developed for change detection and change reconciliation in GIS (Ahlqvist, 2008; Fisher et al., 2006; Goldsberry and Battersby, 2009; Hornsby and Egenhofer, 2000; Lambin, 1996; Moreno et al., 2008; Yi et al., 2014), although many of these are experimental. The current standard for handling change in most GIS platforms is via attributes, metadata, updates and animation (Koussoulakou and Kraak, 1992), which in turn are based on the space–time data models discussed above.

Mark-up The inclusion of a dedicated mark-up scheme (or a set of interoperable schemes) is critical for developing a unified platform, across several axes of consideration: (1) in enabling communication between datasets (particularly those with different knowledge domains); (2) in allowing for data parity between different systems (within the platform or federated to the platform); (3) enabling Web functionality for the platform; and (4) semantically enabling the platform. Various domain-specific mark-up schemes and languages are available, for example, for transport (Cambridge Systematics Inc. et al., 2006), urban environments (Kolbe, 2009), for planning (Hopkins et al., 2003), public policy (Schill et al., 2007), and traffic (Gu et al., 2004). Similarly, mark-up schemes are available for data collection methods and representation, independent of domain, for example, for sensors (Botts et al., 2008), computer animated design and drafting (CAD) (Döllner and Hagedorn, 2007), and virtual reality (Wu et al., 2010) in GIS settings. These are potentially commensurate with

similar developments in other information systems and other domains, for example, with event mark-up for logistics (Mendling and Nüttgens, 2006), decision-making (Tang and Meersman, 2009), banking (Barnes and Corbitt, 2003), and government services (Kavadias and Tambouris 2003). Given the diversity of these mark-up schemes, recent efforts have focused on developing interoperable and extensible mark-ups (that is, the mark-up encodes its own semantics as part of its scheme), with some progress in doing so in a dedicated fashion for GIS (Badard and Richard, 2001). Much of this is targeted toward moving GIS and their functionality to the Web and cloud (Shanzhen et al., 2001). An emerging standard – Geographic Markup Language (GML) – is coalescing around these efforts, and is gaining support from international standards agencies (Kolbe, 2009; Lake, 2005; Peng and Zhang, 2004). Moreover, dedicated data operators are beginning to be developed specifically for and with GML (Boucelma and Colonna, 2004).

Metadata Metadata, that is, data about data is critical for managing a seamless platform across a variety of users, uses, media and datasets (Tsou, 2002). This includes general metadata, such as access rights, ownership, dates, edit history and provenance, source, and so on (Edwards et al., 2011). However, specific treatment of geographic metadata is critical. Indeed, there is growing agreement that robust geographic metadata are essential to providing a public commons for both spatial data and spatial data infrastructure (Onsrud et al., 2004). Key here are issues of transparency (the metadata should allow users to fully understand the limitations and opportunities that the data provide) and interoperability (the metadata should allow users to transfer data between systems, or could allow systems to do this automatically) (Lacasta et al., 2003). This includes cartographic metadata: projection used, units used, place name style sheets, data models, mark-up languages, timing, change, directionality and so on. It may also extend to the data collection scheme (whether this is from a sensor Web with particulars of its engineering and precision, or from survey instruments and the sampling and biases involved) (Nogueras-Iso et al., 2005). These may also be domain-specific (Petras et al., 2006). Recent developments are focused on automated extraction of metadata from spatial data (which is of relevance to the legacy data sources and international sourcing of data for the platform that we discussed earlier) (Manso et al., 2004), on developing national and international metadata standards (Zarazaga-Soria et al., 2003), and on building knowledge domains (again, automatically) from metadata (Ahlqvist et al., 2000; Albertoni et al., 2005; Schuurman and Leszczynski, 2006).

Ontology One of the end stages in the pipeline of use for a successful platform should be the generation of knowledge (Gantner et al., 2013). Ontologies codify this, for data, for users, for uses and for systems into formal 'way of knowing' about things. These include classifications, algorithms, semantics, ground truth, typologies, hierarchies, heuristics, best practices and so on. There has been a long-standing interest in ontologies for GIS (Frank, 1997; Schatzki, 1991; Winter, 2001) and the topic has recently advanced with some rapid progression, thanks in large part to the move of spatial data and GIS to the Web, where those data and systems have come into contact with other ontologies. This is a huge topic for consideration, but a successful knowledge platform should treat metrics of success and ways of knowing in some formal, structured fashion (Bateman and Farrar, 2004). Moreover, it should be extensible enough to support the development of new ontologies. Several operational ontological schemes are available to achieve this, with the Web Ontology Language (OWL) (Bechhofer, 2009) and Protocol and Resource Description Framework Query Language (SPARQL) (Pérez et al., 2009) being among the most widely used. Schemes for developing these ontologies *with* metadata (and bundling the two) are also being developed (Schuurman and Leszczynski, 2006).

Social media Social media, as a potentially valuable data source, have already been discussed above. However, it is worth mentioning that specific data types *for* social media could and should be considered as part of the platform. This would include types for docking social media data to the system (and GIS is one possible universally structuring container for those types, particularly for social media data that have been produced using location-aware technologies). However, social media should also be considered as an output for the platform, that is, as one of many potential media and interaction schemes that the platform could consider.

Databases

Once the data types and data model have been settled upon (which is no small undertaking, as the discussion above probably conveys), these need to be implemented and instantiated as physical models as databases. Here, the discussion grows further. Database methods for GIS or hybridized information systems that dock or 'talk' with GIS are well developed (ArcGIS offers many formats, as do open source GIS, and big data systems usually have spatial database structures – Oracle Spatial is an example (Kothuri et al., 2007)). However, if data and databases span several systems and organizations, this can become a thorny issue far

beyond the spatial nature of the physical database used. That said, most physical spatial databases should be able to function at the 'enterprise' level (Qi et al., 2003), and many more can scale over big data and multisite schemes via conventional high-performance computing (Behzad et al., 2011; Wang, 2010; Wang et al., 2013; Wang and Armstrong, 2003), Web services (Zhang et al., 2007; Zhang and Tsou, 2009), and virtualization schemes (Bhat et al., 2011; Degen and Qin'ou, 2012; Jinnan and Sheng, 2010; Shekhar et al., 2012).

GEOREFERENCING

Georeferencing (Cramer and Stallmann, 2002; Hill, 2009) is essential to the development of a usable and scalable platform, as we discussed throughout the previous section. In essence, it provides the foundation and scaffolding for (1) the data that make up the system's resources; (2) the interfaces that are capable between users and the data; (3) the operators and queries that are possible on the data and within the databases; and (4) the interoperability of the platform with other related and dependent systems. Because of the diverse nature of the data that are likely to be included in the platform, and the diversity of uses and interpretations that the platform should support atop those data, the georeferencing scheme should be both well grounded and flexible.

Base Maps

There are many diverse pathways for achieving grounding and flexibility. The most common would be to establish a series of base maps (Frank, 1992) as ground truth, and a series of projections or transformations from that base to flexible further forms (Griffin, 1980). These may be sourced in common geometry (Buttenfield, 1991), and then specialized to be domain-specific and several possible basemaps may need to coexist in the platform to accommodate this, for example, demography (Bhaduri et al., 2002), digital elevation (Adkins, 2002), physical features (Dikau, 1992), roads (Khan et al., 2010), land parcels (Bishop et al., 2000), utility networks (Knecht et al., 2001), address files (Drummond, 1995) and so on. Coexistence can be negotiated by several further schemes, such as layering (MacDougall, 1975) and map algebra (Mennis, 2010; Takeyama and Couclelis, 1997; Tomlin, 1990). While the notion of 'layer-caking' basemaps is rather well developed across a diverse set of GIS suites, schemes for transforming between basemaps are less mature and often require commercial solutions (Griffin, 1980). Many cities, regions, states

and nations have settled upon basemaps that need to be reconciled when geography and data crosses their boundaries, and so this issue of transformation between them is a significant component of interoperability for the platform (Mennis, 2010; Takeyama and Couclelis, 1997; Tomlin, 1990). Similarly, the rate of refresh and update of the basemaps and ground truth need to be considered. Many organizations may update their maps on a regular cycle, while others may take decades. Recent developments in geosocial media and remote sensing have been targeted at addressing the problem of updating basemaps for this reason (Arai and Shikada, 2001).

Accuracy and Uncertainty

Issues of accuracy and uncertainty almost always need to be addressed when basemaps are developed, or when they are reconciled. Again, this is a topic of long-standing concern in the geographic information sciences (Ahlqvist, 2004; Ahlqvist et al., 2000; Guo et al., 2008; Hunter and Goodchild, 1993; Jones et al., 2008; Liu et al., 2009b; Prager, 2007; Spielman et al., 2014; Voudouris, 2010; Wieczorek et al., 2004). Recent developments have focused on map-matching as a technique for (automatically) performing this (Drummond, 1995; Greenfeld, 2002; McKenzie et al., 2013; Power et al., 2000; Pyo et al., 2001; Quddus et al., 2007; Sobolevsky et al., 2013; Yin and Wolfson, 2004). Other techniques are focused on crowdsourcing the problem (Elwood et al., 2013; Fritz et al., 2009; Gao et al., 2011).

Geocoding

Geocoding is a special case for georeferencing. It involves the conversion of place name data (or sometimes other address-based attribute data) into location data types (points, polygons, lines, objects). For some countries, address systems have been developed to perform this on a quasi-automatic basis (the United States' zone improvement plan (ZIP) code +4 system, or the United Kingdom's Ordnance Survey's postal code system are examples). However, in many places in the world, such systems are not in place and many replicated geocodes or variable vernaculars must be negotiated. This becomes even more problematic when such data present in multiple languages and alphabets. Moreover, geocoders developed to machine-learn resolution schemes are often proprietary. As in other cases, this can also be semantically specific and domain specific (Larsson, 2014). Recently, schemes have been developed to produce universal geocoding (geonames, for example), although this is still in relative infancy (Goldberg, 2011).

VISUALIZATION

Visualization is a critical component of most information systems, and of GIS, in particular, as it serves as the *interface* to the data, as the main *interactive modality* for interacting with the system, and as a central *communication medium* for the system (Card et al., 1983). Most users of the system are unlikely to interact directly with the underlying data and in many cases the visual interface and the user experience (UX) (Garrett, 2010) that it provides *is* the system.

Cartography

The mainstay of both the visual interface to the observatory system and the interaction scheme for making extensible use of its functionality as a planning support system or decision support system will be cartography. Details of what could, should or ought to be included in the cartographic design of the system are perhaps voluminous in their axes of consideration. At a minimum, and given the immediately known needs of the observatory, they should include: (1) boundaries (Pundt and Brinkkötter-Runde, 2000); (2) networks and relationships (Okabe et al., 1992; Okabe et al., 2006b); (3) surfaces and/or fields (whether as dynamically generated surfaces sourced from a strong GIS, or sampled surfaces in raster or image form) (MacEachren and Davidson, 1987); (4) attribute display (Leitner and Buttenfield, 2000; Volta and Egenhofer, 1993); and (5) layering by data, feature class and particularly by theme (population, trade, economy, finance, environment, transport, sociology and so on) (Foody, 1999). As with most conventional Web-based cartography, the system should also accommodate (6) linking and brushing between datasets, data types, and view windows directly through the interface (Cook et al., 1997).

Visualization for Change and Process

Time Much of the data to be displayed and exchanged via the observatory may have historical components, future components, or may be tied or allied to particular processes and policies with change attributes. It is therefore critical that the visualization design accommodate this. However, as noted above, most GIS are not well equipped to handle temporal components of data beyond their attribute cases, and even less well equipped to treat *spatiotemporal* data (Andrienko et al., 2000; MacEachren, 1992). Strategies for tackling this at the data model scheme are discussed above. There are, also, several strategies for visualizing change on Web-based GIS, including animation schemes (Harrower, 2003; Ogao and Kraak,

2002), using dithering and change vectors (Acevedo and Masuoka, 1997; Ehlschlaeger et al., 1997), transition probabilities (Logsdon et al., 1996); rhythms and motifs in timing (Edsall et al., 2000); interactive timeline scrubbing and data entry schemes (Shepherd, 1995), space–time transformation of GIS geometries (Ahmed and Miller, 2007), and space–time paths for trajectory data (Aigner et al., 2007; Chen et al., 2011; Kwan and Lee, 2004).

Scaling Given the multiscale focus of ADB's interests (world, nation, region, city, town, locality, neighborhood), it is also critical that the visualization scheme be sensitive to, responsive to, and flexible relative to scale. This can be accommodated at a simple level using zooming and zoom-dependent data abstraction (that is, features only relevant at a particular scale appear only at that zoom level), which is easily accomplished using conventional tiling schemes (Liu et al., 2007), *Leaflet* being the most commonly used (Crickard III, 2014; Derrough, 2013). Tiling of this nature, while (and sometimes because, particularly when datasets are complicated and large in volume) visually oriented, can have significant impacts on load balancing on the computational side of the system (Fox and Pierce, 2009), and so the choice of fetching schemes and caching (Kang et al., 2001; Talbot and Talbot, 2013), topology between tiles and patches (Li et al., 2009), *and* rendering options (Liu et al., 2013; Sorokine, 2007; Zhang and You, 2010) for the tiler need to be carefully considered (Lee et al., 2002; Li et al., 2009).

Processes Because of ADB's initiatives on cross-border factors, it is critical that the observatory develop visual schemes for handling flow, diffusion and movement. Traditionally, flow has been accommodated with cartographic techniques for representing line–link relationships, for generalizing lines, for adding detail and enhancing lines, for expanding and shrinking boundary polygons and so on (see Tobler, 1987; 2005 for an overview). More recent work (Andrienko and Andrienko, 2012) is focused on visualizing dynamic processes implied in flows, spillovers, trade, traffic, diffusion and movement either using animation or through creative revising of traditional line–link static relationship representations. These include ringmaps (Battersby et al., 2011; Zhao et al., 2008), velocity and diffusion fields (Blaise and Dudek, 2013), trajectories (Demšar and Virrantaus, 2010), sequencing and events (Vrotsou et al., 2009), routing (Liu et al., 2011), and dynamic bounding (Murray et al., 2012).

COMPUTING

Several computing considerations also present with unique, or at least special, relevance for GIS-based observatories.

Application Programming Interfaces Much of this is already discussed above in the section relating to visualization schemes for development. However, there should be careful consideration of application programming interface (API) standards, particularly for the computing components of the observatory. In particular, should the observatory be based on existing commercial and off-the-shelf software, or on mashups via widely available Web mapping APIs from major search companies (Lee, 2009; Miller, 2006), the peculiarities of those APIs will need to be considered in the *systematic* design of the observatory, as all other users and uses will have to negotiate them. Another option may be to base the observatory on free and open-source APIs. Several such APIs are available for mapping (Ames et al., 2007; Chow, 2008), or can be adapted from remote-sensing data handling. However, increasingly there are robust API suites available as free and open-source for either GIS specifically (Warmerdam, 2008), or based around existing spatial database APIs. Increasingly, such APIs are being developed for 'big data' and 'big data access' computing (Anselin et al., 2006). In particular, much of the activity in this area is focused on (1) service-oriented architectures (Coetzee and Bishop, 1998; Kim and Kim, 2002; Paul and Ghosh, 2006; Sha and Xie, 2010) (and Web services especially (Anselin et al., 2006; Sayar et al., 2006)), and (2) high-performance computing on distributed networks.

Virtualization, mirroring and distribution Much of the functionality of the system becomes critically dependent on computing when it is published and used in real time (Zhang and Li, 2005). This presents several computational challenges that are sometimes intertwined with the system software but at other times a function of the base computing and networking on which the system functions. In particular, a dedicated strategy must be considered when implementing the observatory, to consider virtualization, mirroring and distribution. Virtualization refers to the need to provide a duplicate experience for each access and each user of the system, regardless of the load that the system is enduring. For this reason, the system may be served from multiple sites. This can be complex when dealing with GIS-based systems; however, as data exchanges are often large in size and rapid in transactional update, data may be hosted in different physical locations and databases, and the system is likely to be under continual update with requirements for reconciling those dynamics. Mirroring refers to the need

to host (or serve) data from multiple sites, so that many users can access them, or because particular data owners may need or prefer to have them located in particular physical locations or configurations. Distribution refers to both the distributed nature of users (and their media for access, particularly if they are accessing the system via *mobile devices*), data, but also of the data processing required by such systems. Recent developments in this area have seen much of the commercial and off-the-shelf architecture migrating to commercial distributed computing and virtualization services (Blower, 2010). Much of ArcGIS functionality is now available within Amazon Web Services and its Elastic Compute (EC2) resources (Shao et al., 2011), for example, and among academic GIS there is a movement to replicate this functionality in free and open-source form via cyberinfrastructure (Wang, 2010; Wang et al., 2013), with some tie-in to commercial resources (Microsoft's Azure platform, for example (Behzad et al., 2011)). This is not as easy as copying systems to cloud computing platforms, however, as it often requires specific treatment of spatial data organization (Papadopoulos and Katsaros, 2011) and access atop those resources (Cary et al., 2010; Wang et al., 2009). Increasingly, there is a recognition that fundamental operators for GIS and related spatial data query may also need to be treated specially in cloud contexts (Agarwal et al., 2012), and there is a need to rethink spatial analysis (particularly on big data and distributed big data) in a cloud environment (Rezgui et al., 2013). Other developments have seen the separation of geoprocessing functionality (basic operators, spatial analyses (Kerry and Hawick, 1998), database processing (Frye and McKenney, 2015) (via MapReduce and Hadoop for GIS, in particular (Aji et al., 2013; Dittrich and Quiané-Ruiz, 2012; Liu et al., 2009a; Wang and Wang, 2010; Weng and Liu, 2013)), and update functions (Müller et al., 2013) and so on) to high-performance computing schemes (Stojanovic and Stojanovic, 2013). This, in turn, then creates the necessity for high-performance networking considerations that can keep pace with the data exchange requirements from a distributed system. Networking becomes critical, in particular, when the system needs to offload geoprocessing (Wolf and Howe, 2009) while also handling asynchronous update (Rodrigues and Rodrigues, 2009), by human users as well as Web services (Yang et al., 2010) and sensor networks that might stream data to the system (Gadea et al., 2010).