Extending the formal model of a spatial data infrastructure to include volunteered geographical information

Antony K Cooper*, Petr Rapant[†], Jan Hjelmager[‡], Dominique Laurent[§], Adam Iwaniak[#], Serena Coetzee[$], Harold Moellering[¶] and Ulrich Düren[‡]

*Logistics and Quantitative Methods, CSIR, PO Box 395, Pretoria, 0001, South Africa
[†]Institute of Geoinfomatics, VSB-Technical University of Ostrava, 17. listopadu 15, Ostrava-Poruba, Czech Republic
[‡]Kort & Matrikelstyrelsen, Rentemestervej 8, Copenhagen, DK-2400 NV, Denmark
[§]Institut Géographique National, 4 Rue Pasteur, 94165 Saint Mandé, France
[#]Katedra Geodezji i Fotogrametrii Akademii Rolniczej we Wrocławiu, ul. Grunwaldzka 53, 50-357 Wrocław Poland
[$]Department of Computer Science, University of Pretoria, Pretoria, 0002, South Africa
[¶]Department of Geography, Ohio State University, Columbus, OH, 43210, USA
[‡]Bezirksregierung Köln, GEObasis.NRW, Muffendorfer Strasse 19-21, 53177 Bonn, Germany

9 October 2010

Abstract

A spatial data infrastructure (SDI) is an evolving concept for facilitating, coordinating and monitoring the exchange and sharing of geospatial data and services. In earlier work, we developed a formal model for an SDI from the Enterprise, Information and Computational Viewpoints of the Reference Model for Open Distributed Processing (RM-ODP). We identified six stakeholders: Policy Maker, Producer, Provider, Broker, Value-added Reseller and End User. The Internet has spawned the development of virtual communities or virtual social networks, which share data with one another and with the public at large. This user-generated content is most obvious in web sites such as Wikipedia to which the general public, rather than domain experts, contribute information. Similarly, the term volunteered geographic information (VGI) is used for geospatial data contributed to datasets by the general public. Increasing costs of official mapping programmes coupled with the availability of high volumes of quality and up-to-date VGI, have led to the integration of VGI into some SDIs. Therefore it is necessary to rethink our formal model of an SDI to accommodate VGI. We started our rethinking process with the SDI stakeholders in an attempt to establish which changes are required to the stakeholders for including VGI in an SDI. The influence of VGI did not necessitate new stakeholders but rather the specialization of them, by defining a number of subtypes for each.

Theme: T6 SDI, Standards, Ontologies, Integration

Keywords: Spatial data infrastructures (SDI); Cartographic ontology and terminology; INSPIRE.

1. Background and Objectives

A *spatial data infrastructure (SDI)* is an evolving concept for facilitating, coordinating and monitoring the exchange and sharing of geospatial data and services, and the metadata about both. It encompasses stakeholders from different levels and disciplines. An SDI is more than just the technology of a geographical information system (GIS): it is a collection of technologies, policies and institutional arrangements and provides the basis for the discovery, evaluation and application of geospatial data and services (adapted from Hjelmager *et al* [2008] and Nebert [2004]). One SDI can be part of another SDI, either functionally (such as a national water SDI within a general national SDI) or hierarchically (such as the Europe-wide SDI, INSPIRE (Infrastructure for Spatial Information in the European Community), which is based on the national SDIs of Member States [European Parliament 2007]). The Commission on Geospatial Data Standards of the International Cartographic Association (ICA) has been using the Reference Model for Open Distributed Processing (RM-ODP) [ISO/IEC 10746-1:1998] and the Unified Modelling Language (UML) [ISO/IEC 19501:2005] to develop formal models of an SDI. We have described an SDI from the Enterprise and Information Viewpoints of RM ODP [Hjelmager *et al* 2005, Hjelmager *et al* 2008], and from the Computational Viewpoint [Cooper *et al* 2007, Cooper *et al* 2009].

The Internet has spawned the development of *virtual communities* or *virtual social networks*, which share data with one another, and with the public at large. This *user generated content* is most obvious in web sites such as Wikipedia [Wikimedia 2010], the free, online encyclopaedia in many languages, consisting of contributions mainly from the public at large, rather than from domain experts (though it does also include much content from encyclopaedias that are out of copyright and other expert sources). Similarly, virtual communities have also facilitated *folksonomies* or *collaborative tagging*, which are the classification and identification of content by the general public, rather than by domain experts [Cooper *et al* 2010]. Within *geographical information science*, user generated content is also known as *volunteered geographical information (VGI)* [Goodchild 2007] and is made available as base maps on public websites, such as Tracks4Africa [2010] and OpenStreetMap [2010], or as third party data overlaid on *virtual globes*, such as Google Earth [2010].

Traditionally, the data for an SDI have come from official or recognised professional producers of geospatial data, such as national mapping agencies. However, because of the costs of official mapping programmes and the volume of quality and up-to-date VGI becoming available, the custodians of SDIs are starting to admit VGI into their SDIs. This could be in the form of revision requests or notices submitted to an SDI through its web site by the public [Guélat 2009], or potentially even using large quantities of VGI. Emergency services need to react quickly to deal with emergencies such as fires, earthquakes, storms, floods or crime. Hence, they need to update their SDIs suddenly and rapidly: for responding to the earthquake in Haiti in January 2010, the relief agencies of the United Nations depended on VGI from OpenStreetMap, Ushahidi and others [Duvall 2010]. An obvious concern with VGI is how its quality compares with official information [Haklay 2010].

Conceptually, an SDI can exist without users, but VGI needs users, by definition! It is possible for an SDI to fail, such as by restricting the use of data (eg: for security reasons), ignoring the requirements of end users (as opposed to those of institutions), having a faulty business model (eg: without adequate funding sources), lack of resources (funding, skills, equipment, connectivity, data, metadata, services, etc), or lack of cooperation from key stakeholders. Using VGI in an SDI highlights the importance of the user as a stakeholder, particularly for improving the SDI. Hence, it is necessary to rethink the SDI concept from the VGI point of view to bring new responsibilities to the stakeholders (or develop a hierarchical structure of stakeholder's subtypes). An effective SDI should generate participatory VGI because it provides value to end users and hence stimulates them to contribute to the SDI.

2. Approach and Methods

Hjelmager *et al* [2008] identified six types of stakeholders that have roles in a spatial data infrastructure (SDI). An individual or organization can perform one or more of these roles. All of these stakeholders could deal with volunteered geographical information (VGI), for example as follows:

1) **Policy Maker**: A stakeholder who sets the policy pursued by an SDI and all its stakeholders, such as developing policies for VGI, such as soliciting for VGI, acceptance criteria, quality assurance (eg: verification against other, independent VGI), etc.
1) **Producer**: A stakeholder who produces SDI data or services, such as a lay person who generates VGI.
2) **Provider**: A stakeholder who provides data or services, produced by others or itself, to users through an SDI. Examples include an aggregator of VGI, such as Ushahidi, and the provider of the infrastructure for collecting VGI, such as OpenStreetMap.
3) **Broker**: A stakeholder who brings End Users and Providers together and assists in the negotiation of contracts between them. They are specialised publishers and can maintain metadata records on behalf of an owner of a product. Their functions include harvesting metadata from Producers and Providers, creating catalogues, and providing services based on these catalogues. An example for VGI is a community-based organisation that enables the members of its community to provide updates and corrections to the published information of their local authority, such as addresses.
4) **Value-added reseller (VAR)**: A stakeholder who adds some new feature to an existing product or group of products, and then makes it available as a new product. An example is searching for, evaluating and integrating VGI (possibly also with official information), to create a new data set or product. It is important to realize that a VAR does not necessarily sell its products, but could generate its income from other sources (eg: support services).
5) **End user**: A stakeholder who uses the SDI for its intended purpose. Many End Users cannot differentiate between VGI and official information, unless they are told explicitly, and hence would use VGI transparently. End Users tend to use VGI for “quick and dirty” purposes, such as navigation, because there are no issues of copyright or liability.

These stakeholders are the same for both traditional SDIs and SDIs that use VGI, but the importance of each stakeholder will vary, and one organization could be a combination of several stakeholders. An official SDI will generally have a rigid, well-defined framework, whereas an SDI dominated by VGI could be fluid and unconstrained. The strengths of VGI include openness, market-orientation and interaction between stakeholders, while the weaknesses of VGI include heterogeneous data (e.g. VGI coverage mainly where young and well-educated people live – creating a digital divide within countries), lack of metadata (some contributors are anonymous) and uncertainty over the reliability of the data in comparison to official data. We would suggest that SDIs are evolving from a rigid traditional framework (of which there might be few left now) towards a mixed VGI model.

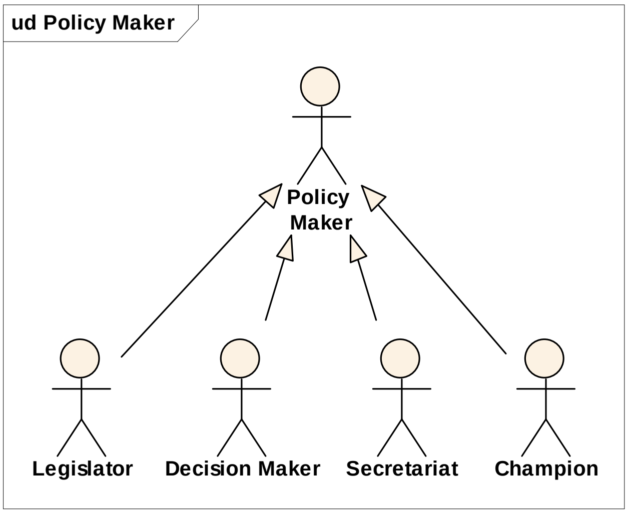
3. Subtypes of the stakeholders in an SDI

Our rethinking of SDI should start from the Enterprise Viewpoint again. In assessing the original model of the Enterprise Viewpoint of an SDI [Hjelmager *et al* 2008], we identified several functions or roles that we could not place immediately with any of these six stakeholders. The problem is that these stakeholders had not been expanded upon, and we realised that it would be appropriate to develop subtypes of these six. In the same way that one person or organization can perform the role of several stakeholders; these subtypes can overlap with one another. We have illustrated these subtypes with examples, some drawn from the European SDI, INSPIRE [European Parliament 2007], as it is well known and comprehensive. INSPIRE is being established to support European environmental policies, and policies or activities which may have an impact on the environment. It is based on the SDIs of the 27 Member States and addresses 34 spatial data themes.

These subtypes are not given in any particular order and it is not certain that there should be any ranking of the subtypes, though clearly some can have a greater impact on an SDI than others.

Subtypes of the **Policy Maker** are:

- **Legislator** – an "external" authority (not obviously perceived as being part of the SDI, but in practice, a key stakeholder) that determines the framework within which the SDI has to exist, but the Legislator does not necessarily understand anything about the SDI. For INSPIRE, this would be the European Parliament.
- **Decision Maker** – a participant in the SDI who makes policies (including initiating the SDI) and who understands geospatial data and the applications, constraints, etc. The Decision Maker is often a committee of representatives of stakeholder communities. For INSPIRE, this would be the INSPIRE Committee (IC).
- **Secretariat** – the 'glue' of the SDI keeping it all together. The Secretariat is often a department in government with the mandate and budget to support the SDI, and that can contract out services. Especially for an SDI of VGI, the Secretariat can start informally and then crystallize once funding is available to pay for participation (as happened with OpenStreetMap, for example, which only received core funding in its second year of operations [OpenStreetMap 2010]). For INSPIRE at the European level, this would be the Joint Research Centre (JRC), as the overall technical coordinator, and Eurostat, as the overall implementation coordinator. Specific roles of the Secretariat include:
    - Supporting and monitoring the implementation of policies, etc.
    - Facilitating communication between stakeholders, particularly to provide feedback (eg: quality or popularity of a data set, viability of a data product specification, responses to draft policies).
    - Building the actual SDI (generally through contractors).
    - Ensuring the smooth running of processes.
    - Classification of stakeholders.
- **Champion** – promotes the SDI, such as encouraging citizens to contribute VGI. The Champion does not necessarily have a mandate, but could be motivated by the need to promote social justice, by environmental awareness, or by commercial interest. The Champion could be the initiator of the SDI.

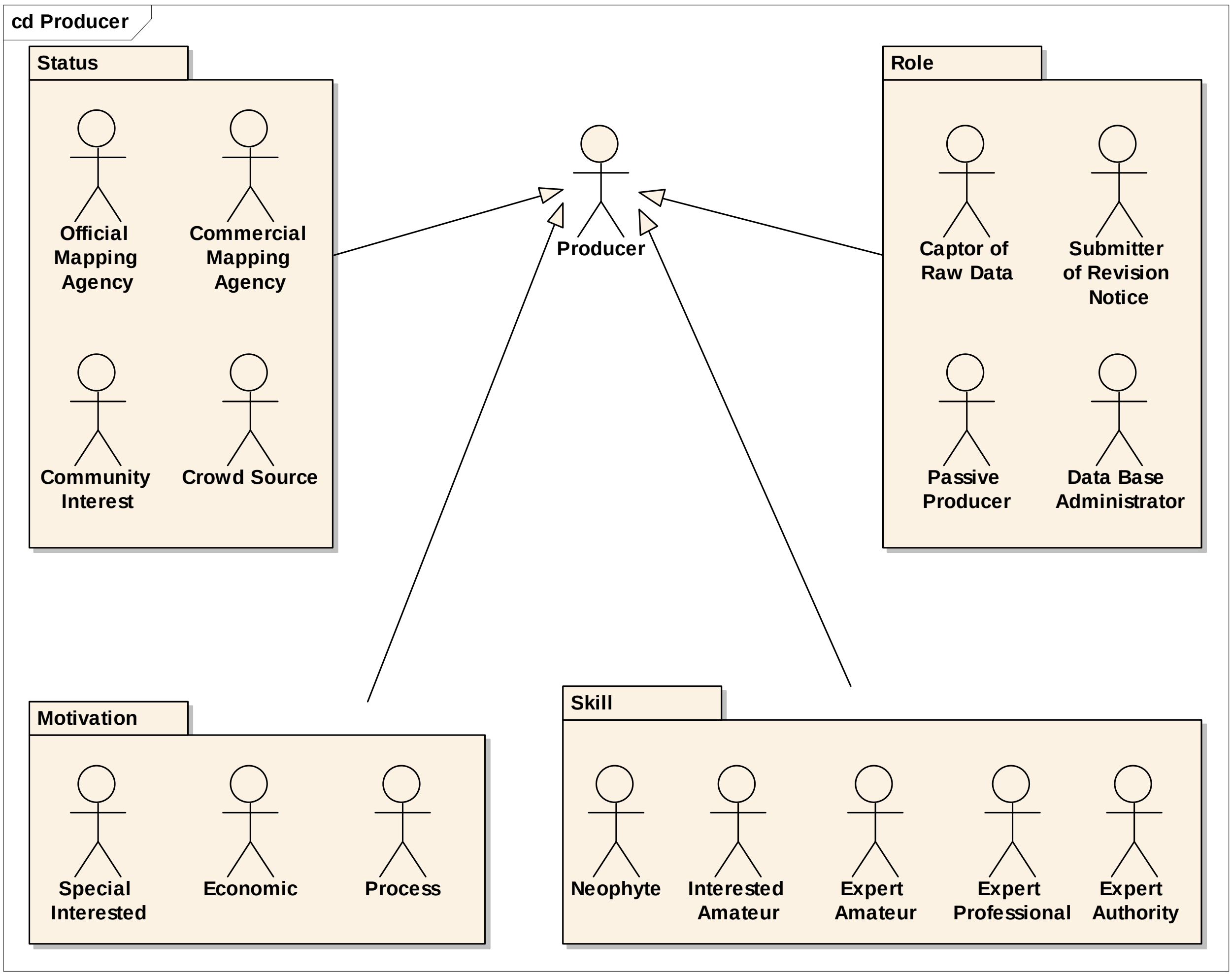


We can classify subtypes of the **Producer** by their status, motivation, roles and skills. Clearly, these subtypes overlap one another:

- **Status**

  - **Official Mapping Agency** – an organization with the budget, resources, expertise and mandate to perform mass data production across the whole of the area of interest, normally to a consistent specification across the whole area. These include topographical, cadastral, hydrographic, meteorological, geological, hydrological, social statistical, environmental and other mapping agencies. These are at all levels of government (local, provincial, national, regional and global).
  - **Commercial Mapping Agency** – a for-profit organisation producing data and products for their identified markets.
  - **Community Interest** – produce general base data or specialized data with broad or narrow coverage, especially as VGI. Exhibits the "long tail", with many contributors of small data sets and few contributors of most of the data. There will be many more End Users than Producers.
  - **Crowd Source** – issue an open call for data to anyone (the crowd), often according to a specification and often with a reward (not necessarily financial). This includes citizen science projects.
- **Motivation**
  - **Special Interest** – produce data for their local area and/or for a narrow interest, such as to protect the environment, empower a community (e.g. asset-based community development) or counteract bias in official sources of data.
  - **Economic** – produce data for economic or financial reasons, such as for direct financial reward (eg: as an employee, on contract or to sell), promoting awareness of a business (locations, products, services, special offers and opening hours), and End Users unwilling to pay for institutional data.
  - **Process** – produce data because of particular interest in the data capture processes per se, such as training for students (as a way to motivate them), or the mapping parties that combine data capture with social events.
- **Role**
  - **Captor of Raw Data** – produce data such as locations measured by GPS or drawn from background images, categorization and description of features, photos and images.
  - **Submitter of Revision Notice** – submit a notice to revise or correct data in an SDI, performed most often by citizens to improve the data of their immediate environment. An example is swisstopo [Guélat 2009]. This would comprise many contributors of very small data sets.
  - **Passive Producer** – produce data through their mobile devices being tracked by a service provider, such as cellular telephones or in-car navigation devices, to monitor traffic flows, assess telecommunication network congestion, or for other purposes. Clearly, this raises ethical issues concerning informed consent, uninformed consent, surreptitious tracking and privacy.
  - **Data Base Administrator** – ensure that the data base specifications are respected (eg: by providing rules to integrate data in the data base and by checking these rules are respected, by ensuring consistency checks, etc).
- **Skill**: Coleman *et al* [2009] categorise the skill levels of users that are producers (which they identify with the neologism, *produsers*), as (in their ordering):
  - **Neophyte** – no formal background in a subject, but with the interest, time and willingness to offer opinions or data.
  - **Interested Amateur** – "discovered" an interest in a subject and begun reading background literature, consulting colleagues and experts, experimenting with applications and gaining experience in appreciating the subject.
  - **Expert Amateur** – may know a great deal about a subject and practice it with passion on occasion, but does not rely on it for a living.

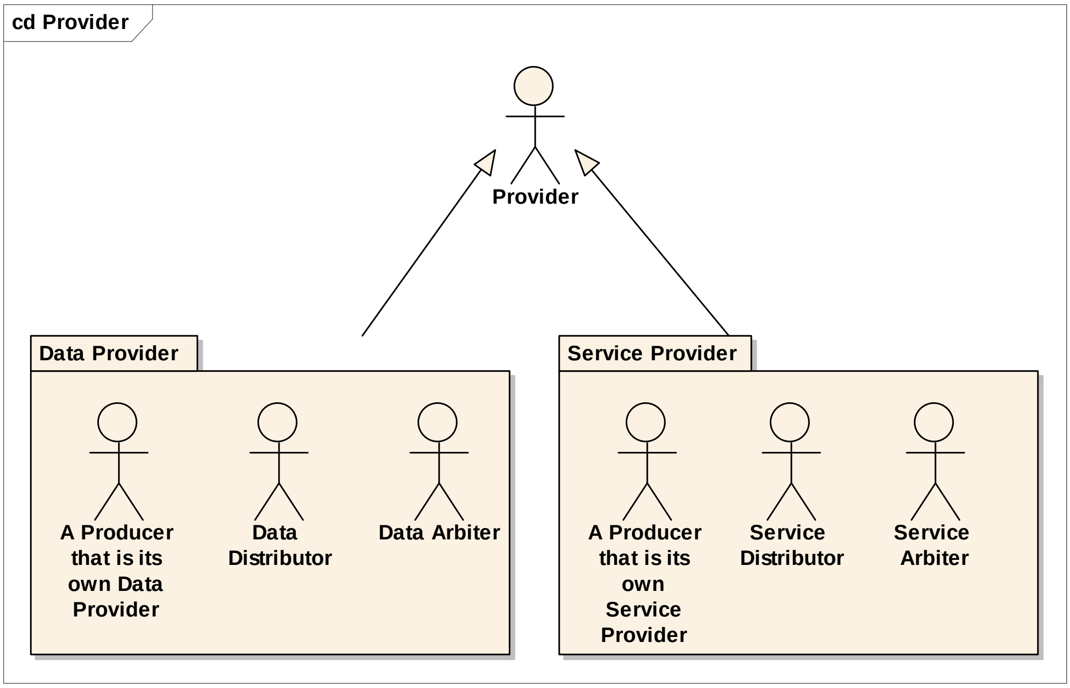
- **Expert Professional** – studied and practices the subject, relying on that knowledge for a living, and may be sued if their products, opinions and/or recommendations are proven inadequate, incorrect or libelous.
- **Expert Authority** – widely studied and long practiced a subject and now recognized to possess an established record of providing high-quality products and services and/or well-informed opinions – and stands to lose that reputation and perhaps their livelihood if that credibility is lost, even temporarily.



We can classify subtypes of the **Provider** by:

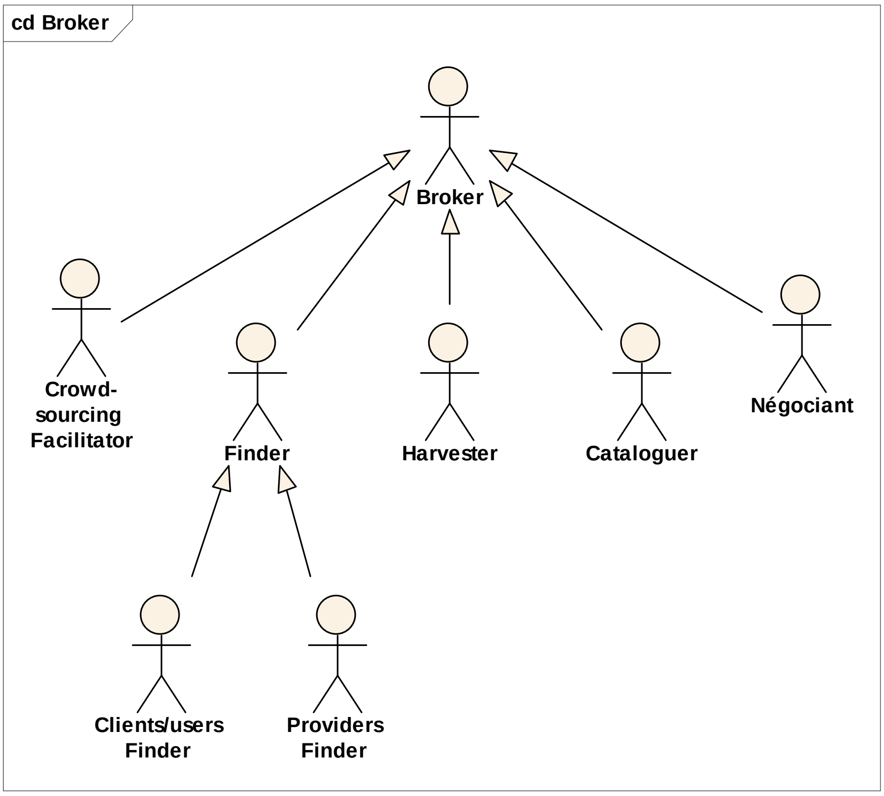
- **Data Provider**
  - **A Producer that is its own Data Provider** – this is the classical model used by a national mapping agency.
  - **Data Distributor** – holds the catalogues and data of Producers, to take the administrative burden away from the Producers in dealing with users. The Distributor does not assess the data they are redistributing; they are merely an agent for the Producer. This would include dissemination through a web site or on CD-ROM, etc.
  - **Data Arbiter** – selects data sets from Producers according their published criteria (ie: performing quality assurance and even certification), but does not add value in any other way.
- **Service Provider**
  - **A Producer that is its own Service Provider** – this is the typical model used by a location-based service (LBS) provider (eg: find a service or facility available where I am now).

- **Service Distributor** – makes services available through their web site or runs the services internally for clients. The cloud computing model is typical.
- **Service Arbiter** – selects services from Producers according their published criteria (ie: performing quality assurance and even certification) and provides them through their web site, but does not add value in any other way.



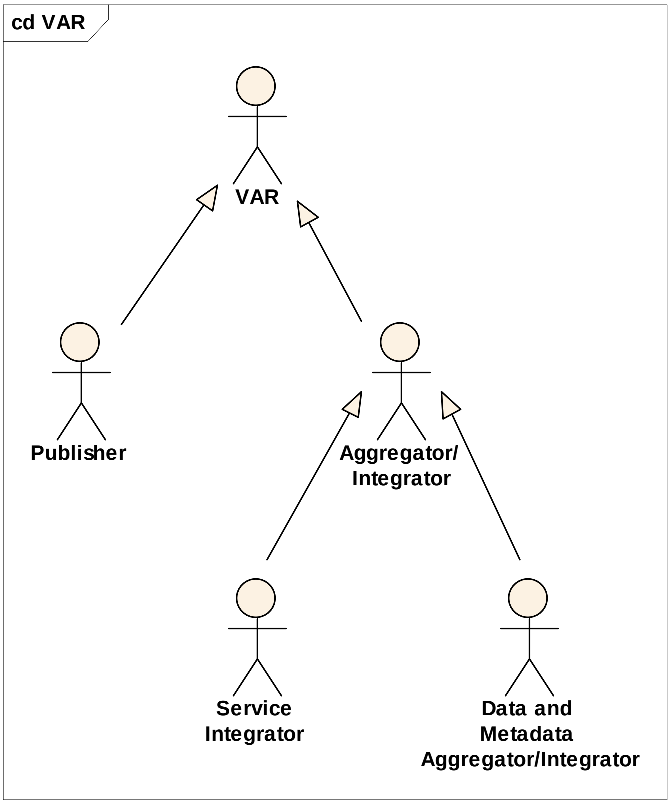
Subtypes of the **Broker** are:

- **Crowd-sourcing Facilitator** – such as Amazon Mechanical Turk, which allows businesses to access an on-demand, scalable work force by advertising small "human intelligence tasks" to be completed [Amazon 2010].
- **Finder**:
  - **Clients/users Finder** – promotes and sells a portfolio of data and services from Producers, Providers and VARs, to End Users.
  - **Providers Finder** – sources data or services for an SDI. In South Africa, for example, the State Information Technology Agency (SITA) has a mandate to procure services for government departments, providing tender evaluation and management, etc.
- **Harvester** – harvest metadata on data and services and integrates them.
- **Cataloguer** – build and maintain a catalogue.
- **Négociant** – a stakeholder who brings End Users and Providers together and assists in the negotiation of contracts between them. They are specialised publishers and can maintain metadata records on behalf of an owner of a product. Their functions include harvesting metadata from Producers and Providers, creating catalogues and providing services based on these catalogues. A VGI example is a community-based organisation that enables the members of its community to provide updates and corrections to the published information of their local authority.



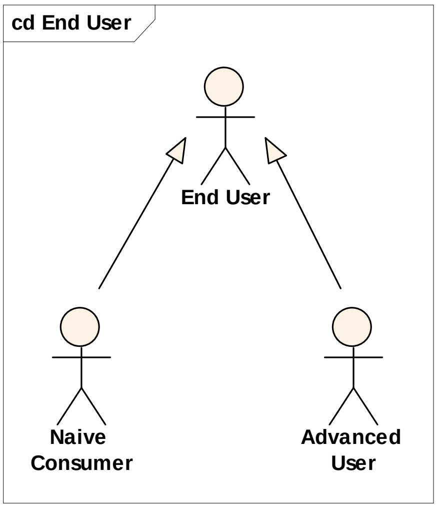
Subtypes of the **Value-Added Reseller**:

- **Publisher –** takes data from various sources, and integrates and edits them to produce a new product, such as an atlas or a location-based service (LBS). A Publisher could add some of their own data.
- **Aggregator/Integrator**
  - **Service Integrator –** chains services together. Would often reside in the cloud.
  - **Data and Metadata Aggregator/Integrator –** selects, edits, enhances and combines data into a new offering:
    - Conflation of data sets (selecting the “best” versions of features and attributes from across several data sets)
    - Aggregation of metadata (more complex to do for VGI because of the multitude of Producers and the patchwork nature of their contributions)
    - Integration of different data sets and their metadata



Subtypes of **End User**:

- **Naive Consumer** – uses whatever is available with limited ability to determine the quality of the data or services.
- **Advanced User** – has expert domain and/or geospatial expertise and hence can make informed decisions about the data and services to use and can provide informed, technical criticism of the data and services. They often use a GIS or other advanced software.



4. Conclusion and Future Plans

We have found that the initial model we developed for an SDI [Hjelmager *et al* 2008] is robust enough to include the contribution from VGI. However, the influence of VGI has lead us to specialize the roles of the six stakeholders and to improve our model. The most significant impact of this has been on the Producer, which is unsurprising as contributors of VGI are Producers. The impact of VGI is possibly less on the other stakeholders, as for them, VGI would tend to be mixed in with official data.

Questions to consider concerning the significance of using VGI contributions in an official SDI:

- What is the continuum of different types of VGI, say from contributed randomly (eg: someone adding geospatial data to their blog) through to crowd-sourced (typically with a predetermined specification and standards)?
    - Where does an open data repository fit in?
- VGI can need metadata at the feature or attribute level, because it is likely to be contributed piecemeal by many people (it is essential to understand this from the standards development perspective).
- To what extent can GPS vendors be encouraged to provide standard metadata automatically as part of any capture of VGI? This would help to improve the acceptability of VGI.
- SDI has an administrative focus and VGI has a business or social responsibility focus.
    - How can these different foci be merged?
    - What about overlapping responsibilities and gaps between responsibilities?
    - What about liability or responsibility?
- Does the nature of VGI make it difficult for a VAR to use VGI?
- Is anyone actually aggregating metadata at this stage? It might be easier using ontologies and technologies such as RDF.
- An SDI consisting primarily of VGI develops organically, not necessarily with a mandate, but driven by a perceived need.
- The contribution of VGI to official SDI is a promising idea to be further investigated.

5. Acknowledgements

We would like to acknowledge the following for their funding that made the work on this paper possible:

- Council for Scientific and Industrial Research (CSIR), South Africa.
- South Africa/Poland Agreement on Cooperation in Science and Technology Joint Research Grant for the project, *Volunteered Geographical Information (VGI) for Spatial Data Infrastructures (SDIs) and Geoportals*.
- Institute of Geoinfomatics, VSB-Technical University of Ostrava, Czech Republic.
- Kort & Matrikelstyrelsen, Denmark.
- Institut Géographique National (IGN), France.
- Katedra Geodezji i Fotogrametrii Akademii Rolniczej we Wrocławiu, Poland.
- Department of Computer Science, University of Pretoria, South Africa.
- Department of Geography, Ohio State University, United States of America.

6. References

Amazon, 2010, Amazon Mechanical Turk: Artificial Artificial Intelligence. Home page. Accessed 4 October 2010 at: http://www.mturk.com/

Coleman DJ, Georgiadou Y & Labonte J, 2009, “Volunteered geographic information: The nature and motivation of produsers. International Journal of Spatial Data Infrastructures Research, Special Issue on GSDI-11, 4. URL http://ijsdir.jrc.ec.europa.eu/index.php/ijsdir/article/view/140/198.

Cooper AK, Moellering H, Delgado T, Düren U, Hjelmager J, Huet M, Rapant P, Rajabifard A, Laurent D, Iwaniak A, Abad P, & Martynenko A, 5 August 2007, "An initial model for the computation viewpoint of a spatial data infrastructure", 23rd International Cartographic Conference, Moscow, Russia.

Cooper AK, Moellering H, Hjelmager J, Rapant P, Laurent D, Abad P and Danko D, 2009, "Detailed Services in a Spatial Data Infrastructure from the Computation Viewpoint", Proceedings of the 24th International Cartographic Conference, Santiago, Chile.

Duvall L, 2010, "Crisis-mapping Puts Fletcher Students at the Forefront of a Revolution in Humanitarian Aid", Fletcher Features, the Fletcher School, Tufts University. Accessed 10 October 2010 at: http://fletcher.tufts.edu/news/2010/02/features/ushahidi.shtml

European Parliament, 2007, "Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)", OJ L 108/1. Accessed 9 March 2009, from: http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:108:0001:0014:EN:PDF.

Goodchild MF, 2007, "Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0", International Journal of Spatial Data Infrastructures Research, Vol 2, pp 24-32.

Google. 2010, Google Earth: Explore, Search, and Discover. Home page. Accessed 4 October 2010 at: http://earth.google.com/

Guélat J.-C, 2009, "Integration of user generated content into national databases – Revision workflow at swisstopo", 1st EuroSDR Workshop on Crowd Sourcing for Updating National Databases, Wabern, Switzerland.

Haklay Mordechai, 2010, "How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets", Environment and Planning B: Planning and Design, vol 37, no 4, pp 682-703.

Hjelmager J, Delgado T, Moellering H, Cooper AK, Danko D, Huet M, Aalders HJGL & Martynenko A, 2005, "Developing a Modeling for the Spatial Data Infrastructure", Proceedings of the 22nd International Cartographic Conference, A Coruña, Spain.

Hjelmager J, Moellering H, Delgado T, Cooper AK, Rajabifard A, Rapant P, Danko D, Huet M, Laurent D, Aalders HJGL, Iwaniak A, Abad P, Düren U & Martynenko A, (2008), "An initial Formal Model for Spatial Data Infrastructures", International Journal of Geographical Information Science, Vol 22, No 11, pp 1295 — 1309.

ISO/IEC 10746-1:1998, Information Technology --- Open Distributed Processing --- Reference Model: Overview, International Organization for Standardization (ISO), Geneva, Switzerland.

ISO/IEC 19501:2005, Information Technology --- Open Distributed Processing --- Unified Modeling Language (UML) Version 1.4.2, International Organization for Standardization (ISO), Geneva, Switzerland.

Nebert, Douglas D, 2004, "Developing spatial data infrastructures: The SDI Cookbook". Accessed 23 August 2009 from: http://www.gsdi.org/docs2004/Cookbook/cookbookV2.0.pdf.

OpenStreetMap, 2010, “OpenStreetMap: The Free Wiki World Map”. Home page. Accessed 4 October 2010 at: http://www.openstreetmap.org/

Tracks4Africa, 2010, “Tracks4Africa: Mapping Africa, one day at a time”. Home page. Accessed 4 October 2010 at: http://www.tracks4africa.co.za/

Wikimedia, 2010, “Wikipedia”. Home page. Accessed 4 October 2010 at: http://en.wikipedia.org/