Data Harmonization Principles and Development Approaches as Applied to INSPIRE SDIs

Dean HINTZ
Safe Software · 7445 132nd Street · Surrey · BC V3W 1JB · CANADA
E-Mail: dean.hintz@safe.com

The core driver for INSPIRE is the need for spatial data harmonization to better support decision making in areas such as environment and resource management, sustainable development and disaster response. Integration across the diversity of themes needed implies challenges that require careful planning to mitigate. Fortunately INSPIRE provides a framework to guide this process (EU-INSPIRE 2010). Below are some insights into the harmonization principles and best practices that are worth considering as part of the INSPIRE implementation process, supported by SDI experiences from both a European and a global perspective.

1 How Do SDI Data Harmonization Principles Apply to the INSPIRE Context?

Key to harmonization for INSPIRE is the use of a common data model in the context of an open standard, service oriented environment. This implies the need for tools that support schema mapping from internal to INSPIRE data models. It also implies support for open standards and web services to allow systems to readily interact with minimal reconfiguration. Systems which load data into INSPIRE structures need to preserve metadata, semantics and rich geometric structures and ensure overall quality and compliance with INSPIRE standards. Solutions need to be model driven and scalable to support the level of maintainability and performance production environments require.

Typically, five stages are involved in harmonization processes: evaluation, assembly, transformation, validation and publication. These steps are also sometimes referred to as Spatial ETL – Extract Transform and Load. First, it is essential to fully evaluate the existing spatial information context. Source, target schemas and actual data should be closely examined before design begins.

Data assembly involves extraction of data from required sources, often with some combination of queries and translation. Format translation needs to take into account the diversity of data sources implied in the wide range of INSPIRE themes. These may include a combination of CAD, GIS, vector, raster, database, text, XML, web, 3D, sensor and non-spatial source data. Given the rich INSPIRE data models, often data required for a specific INSPIRE theme comes from multiple sources, requiring many joins, whether relational or spatial. Lithuania Geographic Information Infrastructure (LGII) is Lithuania's SDI built by a consortium led by HNIT-Baltic, and assisted by con terra. It harmonizes data between government agencies, business, education, research institutions and NGOs, using 38 spatial ETL schema transformation models along with other con terra and ESRI tools. This supports automated data conversion from a wide variety of CAD and GIS formats, coordinate system reprojection and schema mapping to a common data model based on INSPIRE (WAGNER 2009).

Koch, A., Kutzner, T. & Eder, T. (Hrsg.) (2012): Geoinformationssysteme.

Core to the harmonization workflow is the transformation process which reshapes source schema and geometry to match the required destination structure. Disparate data sources imply different data models which must be mapped to a common destination model (MAGUIRE et al. 2008). One of the most labour intensive processes to configure is schema mapping, which includes processes such as feature type, attribute, and code list mapping, new attribute creation, and conditional value mappings. In addition, often some type of geometry transformation is required, whether coordinate systems reprojection (ED50 to ETRF89), or type conversion (CAD lines to GIS polygons; non-spatial text coordinates to point geometry), generalization or interpolation. Datasiel built a system to support Nature SDIplus harmonization for Regione Liguria (PARODI 2011), Italy. Their transformation model extracts data from an Oracle database, performs the required joins, and then uses schema mapping models to transform the data structure to the INSPIRE Protected Sites schema. The system generates INSPIRE compliant GML for publication via WFS (PARODI 2011).

Once data is assembled and transformed, a validation process is essential to ensure quality. This includes validation against standards such as the INSPIRE schemas, and also general validation processes to ensure data integrity, which may help detect upstream problems in the extraction and transformation process. Checks for unique ids, geometric integrity, null values, domain codes, realistic data ranges, data gaps, tolerances, and bounds are often needed. A project by State Office for Nature Environment and Consumer Protection North Rhine-Westfalia and con terra provides a good example (HINTERLANG 2011). This operational system includes a validation process which ensures the data uploaded meets specific data model and quality requirements. Then the con terra INSPIRE Solution Pack for FME deployed on FMEServer is used to transform and load into a compliant staging geodatabase, which in turn serves as the foundation for INSPIRE web services by ArcServer.

After data is transformed and loaded into the common INSPIRE data model, the task of publishing services is the main challenge. Central to the spirit of INSPIRE is accessibility – how will the users get at and interact with the data? While OGC services are mandated, augmenting these with support for more ubiquitous clients and de facto industry standards and API's is advisable. A good example of integration between a vendor system, open standards and open source software is a system developed by Spatialworld for the National Land Survey of Finland (NLSF) (TANI 2011). NLSF preferred an open source solution where possible, so the geoportal was augmented with deegree's WPS (Web Processing Service) in order to provide transformation services. The Open Layers client is configured to generate WPS requests. These are received by the deegree WPS and passed via API to FME Server. Based on the request type, FME Server then runs the appropriate transformation model and provides the resulting GML or raster data stream to the deegree WPS for streaming back to the client. This way NSLF can transparently host an open WPS transformation service supported by any data source or spatial ETL transform that FME supports.

Finally, production systems require some assurance of performance, security and access control. There is usually the need for role based security as different groups require different levels of access. Level of service requirements need to be met. Service process should be readily scalable, easily replicated as new instances, and if possible deployable via cloud to maximize flexibility. Failover configurations are needed to avoid any single point of failure (MAGUIRE et al. 2008). Logging and reporting is useful to review service history and diagnose performance problems.

2 What Are some Key Implementation Risks, and Best Practices to Mitigate Them?

First, from a management perspective, it is essential to identify early on the resources required (MAGUIRE 2008). For INSPIRE projects, IT, GIS, domain and INSPIRE experts are all needed. Domain experts are often not IT experts. One solution is to break down the harmonization problem and let data experts use the tools they are familiar with, such as spreadsheets, to describe their data, rather than forcing them to learn new modeling languages or interfaces.

This approach helped manage the complexity of an ongoing project Metria is building for the Swedish EPA (GIM 2010). Metria developed spatial ETL tools (using FME) to extract and join the required datasets together, before transforming them into the INSPIRE Protected Sites data model in a staging database. Data from Europe's Natura 2000, regional Helsinki Commission, and the Swedish EPA database (NVR) is mapped to the INSPIRE schema using attribute and code list mapping tables stored in spreadsheets external to the ETL models. Domain experts can share and modify schema mappings without having to understand the rest of the transformation model. The staging database then serves as the basis for INSPIRE Protected Sites OGC services as part of the Swedish national SDI (GIM 2010), (MAGUIRE 2008).

FilterAttribute	Filter Value	SourceAttribute	Field_Inspire	Destination AttributeValue	DestinationAttribute
				Full	INSPIRE_APPLICATIONSCHEMA
				SE	INSPIRE_NAMESPACE
		DID	INSPIRE_LOCALID		
		OBJECTNAME	INSPIRE_SITENAME		
		DECISIONDATE	INSPIRE_DATE		
				creation	INSPIRE_DATETYPE
IUNC_CATEGORY	0			Empty value	INSPIRE_DESIGNATION
IUNC_CATEGORY	Ia			strictNatureReserve	INSPIRE_DESIGNATION
IUNC_CATEGORY	Ib			wildernessArea	INSPIRE_DESIGNATION
IUNC_CATEGORY	II			nationalPark	INSPIRE_DESIGNATION
IUNC_CATEGORY	III			naturalMonument	INSPIRE_DESIGNATION
IUNC_CATEGORY	null			Empty value	INSPIRE_DESIGNATION
PROTECTIONTYPE	1			natureConservation	INSPIRE_PROTECTIONCLASSIFICATION
PROTECTIONTYPE	2			natureConservation	INSPIRE_PROTECTIONCLASSIFICATION

Fig. 1: Partial schema mapping table from Swedish NVR to INSPIRE Protected Sites

Overdesign is a potential hazard given the complexity of INSPIRE. Large projects can get bogged down, run out of resources, or lose sponsorship. The answer is to start simple, start small, and with what is readily available. Staged, iterative design and development limits risk and involves starting with a small proof of concept and working through to prototype, before going to production (KUHL 2002), (en.wikipedia.org/wiki/Project_management), (FOOTE & CRUM 2009). Buy-in can be improved by soliciting input from key stakeholder groups and making sure they receive value added results early (en.wikipedia.org/wiki/Software_prototyping) (MAGUIRE 2008).

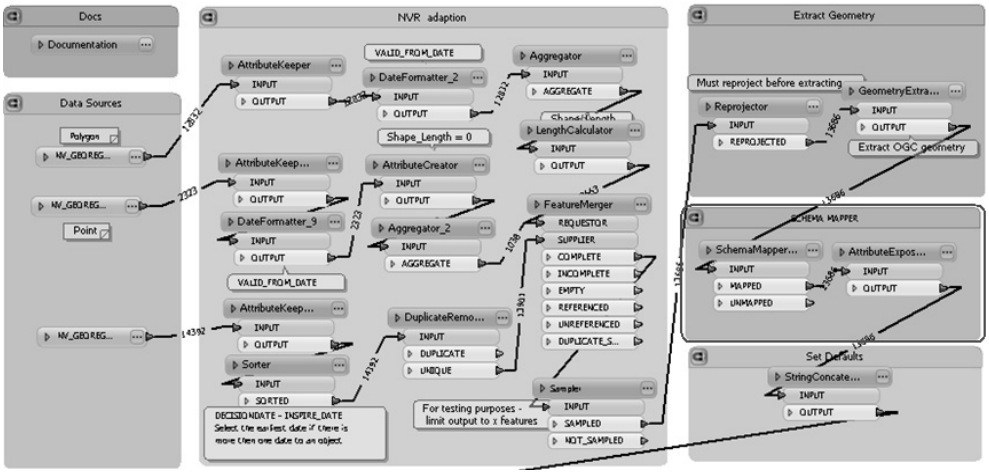

Fig. 2: Data harmonization model showing schema mapping from Swedish NVR to INSPIRE Protected Sites

Another way to minimize risk is to test with real data and environments, soliciting user feedback early and often (MAGUIRE 2008). Too often systems are designed in a vacuum, and look great only until exposed to the harsh demands of an operational environment. Also data, standards, requirements, and environments all can change. Successful systems are designed to be adaptive (en.wikipedia.org/wiki/Software_prototyping), (FOOTE & CRUM 2009). For example, migration methods should support updates as well as loads. It's advisable to develop modular systems that communicate via standardized APIs and keep schema mapping separate from code. Document workflows to avoid a system that is not extendable because a key architect leaves. Automate where viable to minimize manual effort.

3 How Does INSPIRE's Approach Compare with other NSDIs around the World?

Like INSPIRE, many SDIs tend to include extensive use of standards such as ISO and OGC, and had an early emphasis on metadata, discovery, and catalogue services. In contrast to INSPIRE, many world regions outside of Europe are focusing their efforts more on national and state level SDIs (www.gsdi.org/SDILinks), (MAGUIRE 2008). Canada has the CGDI (Canadian Geospatial Data Infrastructure) hosted by GeoConnections (www.geoconnections.org/) and the US has the NSDI developed by the FGDC (www.fgdc.gov/). In these SDIs, data models are not as strictly prescribed as in INSPIRE. While capabilities and capacity is continually growing, the primary focus has been on enabling data sharing by setting up standards for metadata and developing catalogue services. The US system allows for downloading of a wide array of datasets through sites like data.gov. Authoring agencies of Canadian datasets are often at the provincial level. Direct access to some national scale datasets is offered, though access to higher precision provincial and regional data usually involves linkages from the CGDI catalogue service to the regional data management authority. Most of these systems are not as prescriptive as INSPIRE. INSPIRE defines the

precise data structure including every table and field required. North American SDIs tend to provide an infrastructure for discovery, and leave it to the individual authoring agencies to define standards for their data (USGS defines standards for base mapping).

For many, INSPIRE is seen as taking a pioneering role in the establishment of an ambitious transnational SDI which defines common data model and content standards across all member states, even to the level of the precise application schemas. To date there are not many other examples of mature pan-national SDI's. Currently under development are the Permanent Committee on Geographic Information for Asia and the Pacific (PCGIAP), the Permanent Committee for the Americas (PC IDEA), the UNGI Working group (MASSER 2005), and the African Clearinghouse for Spatial Data. The Global SDI (GSDI) (www.gsdi.org/SDILinks) is more of a global association that promotes cooperation and communication between various SDI efforts by producing guidelines and publications (NEBERT 2004), and hosting conferences and working groups (MAGUIRE 2008). On the other hand the number of national scale SDI efforts are clearly increasing globally. As Masser notes, in 1996 there were 11 SDI projects, then 50 in 2000 growing to 120 in 2003 (MASSER 2005).

The leadership embodied by INSPIRE's transnational efforts entails both advantages and risks. The advantages are implicit: once data is compliant it is easily shared across borders and disciplines promoting greater productivity. The risk is the level of investment required up front to transform the data into a compliant state before the benefits are seen. While some may be sceptical about how much investment this will take, many are watching closely to see what lessons will be learned through the INSPIRE development process, to build on INSPIRE successes and perhaps avoid some of the potential pitfalls. At a recent GIS conference in the UAE (GISWORX) there was significant interest in INSPIRE. ADSIC (Abu Dhabi Systems and Information Centre) is in the midst of developing the AD SDI (Abu Dhabi SDI) which is a comprehensive spatial data sharing initiative aimed at linking 49 different government agencies, albeit over an area limited to the UAE (AD-SDI 2011).

Another emerging trend is crowd sourcing, or the ability of citizens to contribute data directly into SDIs (MAGUIRE 2008). In Bulgaria, a recent e-government project was implemented with the participation of 23 local municipalities and utilities that allows users on mobile devices to flag physical infrastructure problems such as pipe leaks, pot holes, fallen trees or sagging power lines (AKTIKEN 2012). Public administrators are alerted to problems much more rapidly than if limited to conventional inspections only. As support is implemented for INSPIRE themes related to sensors and events, care will be needed to ensure that such listening services are implemented in a way that maximizes ease of public interaction and responsiveness.

4 What's Next for INSPIRE?

As INSPIRE efforts at the national agency level mature there will emerge more demand for lower level agency adoption by regions, cities and utilities. This will necessitate greater integration with existing vendor systems (WILLIAMSON 2007). While national agencies may be able to afford complete INSPIRE centric systems, local agencies and businesses need to work within their existing architecture and make minimal investments to integrate with

SDIs. Thus systems capable of integration between open standards, de facto standards, proprietary systems and open source software will be needed (MAGUIRE 2008).

Implementation demands will inevitably encourage the convergence of approaches and refinement of INSPIRE guidelines. Performance will be a problem in some cases such as on-the-fly translation or downloads for large datasets. To date most testing has taken place using discovery and view services and small downloads with few clients, most of which do not require much band-width. To meet volume demands, production level systems will require publication optimization methods such as staging databases and caching. Common practices will emerge that demonstrate efficient approaches to implementing INSPIRE standards and services in the context of production systems.

Key to demonstrating the value of INSPIRE is making its data more widely accessible. So far, many efforts have focused on how to make data compliant for collection by central EU authorities. Not enough focus has yet been placed on distribution of INSPIRE data for daily applications (WILLIAMSON 2007). Vendors will need to improve the ability of their systems to consume INSPIRE data and services directly (MAGUIRE 2008). Data distribution services should provide data via both OGC services and CAD and GIS file formats, national data models such as AAA NAS GML, and common browser accessible streams such as PDF, PNG or KML. In an age of Google Maps, geo-enabled smart phones and social media, consumers of geo-information often know nothing about GIS and expect services that are seamlessly integrated with the environments they are familiar with.

To some extent, the degree to which INSPIRE is implemented across the EU will depend on the effectiveness of monitoring compliance, quality and performance. Given current tight budgetary conditions and existing institutional barriers, there appears to be a disparity emerging between those who have sought to plan ahead for compliance, and those who have taken more of a wait and see approach. Compliance monitoring tools will help identify where services are not in place or up to standard. For example, automated web service queries can be used to record level of service statistics.

However, the push for compliance should be tempered by practical considerations of each nation's context. The sophistication of each agency's solution will necessarily depend on their available resources and the relative demands of their audience. There will be increasing pressure to look for technical solutions that allow implementers to do more with less. Staged development, early value demonstration, and partnership with local agencies and the private sector will go some way towards building the support needed to mitigate and offset these challenges (MAGUIRE 2008), (WILLIAMSON 2007). Combining INSPIRE development with that required for other mandated SDI and e-government projects will help share the burden of this investment among partnering agencies.

Thus, these data harmonization approaches of evaluation, assembly, schema transformation, validation and publication provide a path for confronting the daunting challenges INSPIRE poses. Careful consideration of SDI best practices, both near and far, can help mitigate implementation risks. Solutions are needed that bridge gaps between the complexities of INSPIRE standards and the daily requirements of end users working with legacy systems. Only then will the vision of community wide spatial data sharing for better decision support based on the common INSPIRE model be realized.

References

AD-SDI (2011), Abu Dhabi Spatial Data Infrastructure sdi.abudhabi.ae/Sites/SDI/Navigation/EN/root.html.

AKTIKEN.BG (2012), System for Citizen Reporting of Irregularities. Bulgarian Municipalities and ESRI Bulgaria.

EU INSPIRE (2011), inspire.jrc.ec.europa.eu/index.cfm/pageid/48.

FEDERAL GEOGRAPHIC DATA COMMITTEE (2011), www.fgdc.gov/.

FOOTE, K. E. & CRUM, S. L. (2009), Project Planning Life Cycle, The Geographer's Craft Project, Dept of Geography, The University of Colorado at Boulder. www.colorado.edu/geography/gcraft/notes/lifecycle/lifecycl_f.html.

GEOCONNECTIONS CANADA (2011), www.geoconnections.org/.

GIM INTERNATIONAL (2010), "INSPIRE Prototype", Geomares Publishing, Netherlands, www.gim-international.com/news/id4719-INSPIRE_Prototype.html; www.geodata.se/en.

GLOBAL SPATIAL DATA INFRASTRUCTURE (2011), www.gsdi.org/SDILinks.

HINTERLANG, D. (2011), State Office for Nature, Environment and Consumer Protection North Rhine-Westphalia. conterra.de/de/service/download/cs/Finals/case_study_data_harmonization_EN.pdf.

KUHL, J. J. (2002), Project Lifecycle Models: How They Differ and When to Use Them www.business-esolutions.com/islm.htm.
"Project Management": en.wikipedia.org/wiki/Project_management.
"Software Prototyping": en.wikipedia.org/wiki/Software_prototyping.

MAGUIRE, B., CAMPBELL, M., RIMKUVIEN, J., SVERNDAN, L. & SANKALAS, V. (2008), Structure of Geographic information infrastructure. National Land Service, Vilnius, Lithuania, and Malaspina University-College, Nanaimo, Canada, 2008. www.geoportal.lt/download/gii_mokymai/GII_08_mokomoji_medziaga/En/Paskaitu_konspektai/GII-08_training_material.pdf.

MASSER, I. (2005), GIS Worlds: Creating Spatial Data Infrastructures. Redlands, California, ESRI Press.

NEBERT, D. D. Editor (2004), Developing Spatial Data Infrastructures: The SDI Cookbook, Version 2.0. GSDI, www.gsdi.org/docs2004/Cookbook/cookbookV2.0.pdf.

PARODI, S. (2011), INSPIRE data harmonization. Nature SDI plus in Regione Liguria, Italy. www.sysgroup.it/sysgroup/files/u1/datasiel.pdf.

SPATIAL DATA INFRASTRUCTURE (2012), Wikipedia. en.wikipedia.org/wiki/Spatial_data_infrastructure.

TANI, L. (2011), Spatial Data Analysis Through Web Services, GI Norden Conference. www.lounaispaikka.fi/paikkaoppi/presentations/GI_Norden_Tani.pdf & National Land Survey of Finland geoportal: http://www.paikkatietoikkuna.fi/web/en.

WAGNER, M. J. (2009), From Silos to Open Data Fields: Lithuania's INSPIRED SDI. GeoInformatics, 12.

WILLIAMSON, I., RAJABIFARD, A. & BINNS, A. The Role of Spatial Data Infrastructures in Establishing an Enabling Platform for Decision Making in Australia, Centre for SDIs and Land Administration, Geomatics, University of Melbourne, Victoria, Australia www.gsdidocs.org/gsdiconf/GSDI-9/papers/TS40.4paper.pdf.