doi:10.5194/piahs-368-9-2015

GéoSAS: A modular and interoperable Open Source Spatial Data Infrastructure for research

R. BERA[1,2,3], H. SQUIVIDANT[1,2,3] G. LE HENAFF[2,3], P. PICHELIN[1,2,3], L. RUIZ[2,3], J. LAUNAY[1,2,3,4], J. VANHOUTEGHEM[1,2,3,4], P. AUROUSSEAU[1,2,3,4] & C. CUDENNEC[1,2,3]

1 Agrocampus Ouest, UMR 1069, Soil Agrohydrosystem Spatialisation, F-35000 Rennes, France
roderic.bera@agrocampus-ouest.fr
2 INRA, UMR1069, Soil Agrohydrosystem Spatialisation, F-35000 Rennes, France
3 Université Européenne de Bretagne, F-35000 Rennes, France
4 Conseil Scientifique de l'Environnement de Bretagne, F-35000, France

Abstract To-date, the commonest way to deal with geographical information and processes still appears to consume local resources, i.e. locally stored data processed on a local desktop or server. The maturity and subsequent growing use of OGC standards to exchange data on the World Wide Web, enhanced in Europe by the INSPIRE Directive, is bound to change the way people (and among them research scientists, especially in environmental sciences) make use of, and manage, spatial data. A clever use of OGC standards can help scientists to better store, share and use data, in particular for modelling. We propose a framework for online processing by making an intensive use of OGC standards. We illustrate it using the Spatial Data Infrastructure (SDI) GéoSAS which is the SDI set up for researchers' needs in our department. It is based on the existing open source, modular and interoperable Spatial Data Architecture geOrchestra.

Key words OGC standards; Web processing service; spatial data infrastructure; spatial data architecture; GéoSAS; geOrchestra; MNTSurf; hydrological modelling; watershed

INTRODUCTION AND CONTEXT

Proper storage, sharing, and dissemination of data, tools and knowledge, is central to most research departments. Additionally science is often partly or totally funded with public money, hence has a moral obligation to communicate science to the wider society. Peer-reviewed paper publication is an answer and the most valued means to communicate output to the community. However, many by-products of research are not properly conveyed through this channel, leaving aside the fact that the public for such literature is mostly limited to a fringe of specialists.

This paper addresses the case of environmental data, processes and knowledge, which form a substantial part of our department's output. These data, processes and knowledge are spatially referenced, making pertinent the deployment of a spatial data infrastructure (SDI). However, too many SDIs simply fail to fulfil their role (i.e. convey data and knowledge) as the services they offer are not used widely enough for a number of reasons (user interface – UI, publicity, interoperability, openness, etc.). This points to the need to fully integrate SDI implementation as part of a dissemination strategy. We therefore advocate more openness, seamless exchanges between scientific SDIs, and wish to share the experience and thoughts gathered from the inception of GéoSAS (2010), the SDI of our research department SAS – Soil Agrohydrosystem Spatialisation.

In the next section we present the needs of environmental (hydrology, soil science, agronomy) research in our department from a spatial perspective. We then introduce the general purpose spatial data architecture (SDA) that we based our SDI on and developed to further serve the specific needs of environmental research. We then describe how an open ecosystem of research spatial data and process infrastructure could be designed and bring some conclusions.

STORING AND SHARING GEOSPATIAL DATA AND PROCESSES IN RESEARCH

Geographical information systems (GIS) are long established in research institutes dealing with spatial set-ups and processes. Similarly, the World Wide Web has become a natural extension of a researcher's desktop. Despite this, sharing data, tools and knowledge, though not a new concept,

does not lack caveats when it comes to practice: use of common standards (where available), organisational framework, interaction with other platforms.

In the field of GIS science there is an opportunity, however, with a conjunction of: (1) the existence of common grounds to exchange geospatial data and processes, especially trough the Web, with the Open Geospatial Consortium's (OGC, 1994) standards; (2) the definition of rules and a legal framework at the European scale to open and share geospatial data produced by the public sector, with the Directive setting an infrastructure for spatial information in the European Union (EU) (INSPIRE, 2007); (3) A general move towards the opening and diffusion of publicly-funded data. Since its inception in the EU (PSI, 2003, revised in PSI, 2013), several non-European governments have themselves proposed implementations and portals to grant access to public sector data, starting with the USA (data.gov, 2009, building on the science.gov, 2002, experience). As a consequence, the Open Data (OD) movement now percolates widely.

The three items above act at different levels to ease the access to data and knowledge: (1) OGC standards are on the technical side and set common ground for exchange of geospatial data and processes; (2) INSPIRE provides standards of a more organisational and operational nature, and sets rules for organisational interoperability; (3) OD in turn can be viewed as a "philosophical" standard, and sets general guidelines towards un-refrained openness, which for science also comes as an international recommendation (OECD, 2007; Pilat and Fukasaku, 2007).

These guidelines are therefore becoming widely accepted for data, metadata, and even related web-services, but few implementations are satisfactory as: (1) though OGC standards are quite easy to use separately, it still takes time to fully understand the logics and the best ways to have them interact. (2) INSPIRE only applies to the European Union, though the underpinning principles would apply globally. Moreover it may be a painful process for many data providers to comply with the Directive. (3) When research institutes implementing a SDI already struggle with the techniques and laws, aiming to open data and science may come second or be seen as a source for extra technical/legal trouble. As a result, everything implemented tends to be *a minima*, e.g. merely complies with the compulsory requirements, leaving the rest unaddressed. Indeed, there is a cost to assimilate, a cost to implement, and an extra top-up cost as the time thus spent is not dedicated to pure science. But we posit there are also benefits to come resulting from the increased availability and accessibility of research data, processes and services (as OGC Web Services – OWS).

Services are not merely about providing data in a standardised way. The processes, models, algorithms and tools, are a central part of research too. On-line or cloud processing and modelling architectures are being explored to better share know-how and algorithms running on powerful distant servers, of which the Cybergis project (Wang, 2010; Wang *et al.* 2013) is an emblematic example. The approach has shown to be versatile and powerful, but only partly addresses the issues of interoperability, accessibility and openness. Indeed, processes, services and UI, are tightly coupled from a user's perspective, and though everyone can get access, one needs to log in, then comply with, and adapt to, the UI and working environment as is. We suggest users may sometimes seek to do things differently to what developers had in mind, and we think this should be encouraged or at least made possible. Therefore users should not be compelled to using a given portal to access processes and models. Instead, most effort must be put into services (variety of processes, models, tools) and sharing (through standards especially designed to foster interoperability). We therefore advocate a wider use of WPS in combination with other OGC standards. The WPS standard is now mature and several solutions exist to implement such services on the server side: PyWPS, ZooWPS, and GeoServer WPS, to name a few; as well as on the client side: FOSS GISs QGIS, GRASS, or GvSIG, all have their WPS client extension, whereas 52°North WPS is a standalone WPS client. Squividant *et al.* (2015) demonstrate the use of Web browsers as WPS clients.

However, having services that follow the same standards is not a sufficient condition for them to effectively interoperate. Organisational standards have their role and research institutes need to agree on common rules and behaviours. In the European Union INSPIRE sets the foundations for

such an agreement. INSPIRE schematically states that publicly-funded data is public-sector and must be accessible to the public. This involves publishing the data and means to easily find and explore datasets, on the Web. Thus, datasets must be described by strictly formatted metadata. Metadata must be at the public's disposal in online catalogues reachable from a number of well known Web portals.

Beyond the obligation to comply with technical guidance or organisational rules there must also be a will to extensively share (in a manner that is as versatile as possible) scientific outputs. This is a crucial point, as research scientists must wholeheartedly accept the handing of the results of their efforts to others with no guarantee of proper acknowledgement (it is expected that a wider diffusion increases the chances for improper use or citation: though scientists are accustomed to good practices this may not hold for the general public), nor a certitude that their output will be used the expected way. This conception of the scientific process is still being debated but gradually gains proponents as several works have shown the comparative advantages (Uhlir and Schröder, 2007, and Willinsky, 2005, amongst others) of unrestricted cooperation. It has come into an OECD (2007) recommendation to make the global scientific process more effective. Effectiveness must also encompass numerous alternative ways to process, display, explore and extract knowledge. A lot remains to be done in various domains: spatial representation, time-series exploration, analysis, modelling. A special attention should be drawn to UI for those (scientists, and whenever possible, the wider public) less conversant with technology or with a different background. These ideas are applied to environmental science with proof-of-concept implementations of and around the SDI GéoSAS. We are not considering the fields of linked data and the semantic web, though our approach must eventually converge with, and complement, them. In addition, our proposals only address the case of data, processes, models and services with spatial components, inputs, or outputs, which make up most of our production.

BUILDING ON GEORCHESTRA

geOrchestra is a free and open source (FOSS), modular and interoperable Spatial Data Architecture (SDA) primarily developed in 2009 by Camptocamp for the Council of Brittany to meet the requirements of INSPIRE (geOrchestra, 2009). geOrchestra features a security proxy and a single-sign-on authentication system, and consists of independent and interoperable modules, which can be selectively deployed: (1) a metadata catalogue: GeoNetwork; (2) a map and features server: GeoServer; (3) a map tile server: GeoWebCache; (4) a set of advanced web map viewers (swiewer, mapfishapp…); (5) a data extractor; and (6) several other admin-oriented modules (user and group manager, log parser) plus a data access list manager. A number of add-ons to these components are also available, contributed by the communities involved (e.g. GeoServer add-ons), or by the geOrchestra community itself. These components interact through OGC interfaces and standards, or REST APIs where no OGC standard applies. To further develop and keep geOrchestra thriving, a dynamic community of users and developers is dedicated to the project, and can rely on the communities around the specific modules. A number of geOrchestra-based SDIs are now run by local and national governments, research institutes and teams, and in academia.

SDA geOrchestra uses OGC standard web services to discover (CSW), view (WMS, WMTS), symbolize (SLD), query and download (WFS), edit (WFS-T), and even process (WPS) geographical features. It also provides powerful web applications making use of, and combining, these services. It operates perfectly with GIS desktop software such as QGIS, GvSIG, and ArcGIS.

geOrchestra proves to be suitable for scientific use as an effective way to publish and share outputs, our primary concern, and foster dissemination towards the wider society, our moral obligation. In addition the geOrchestra project revolves around principles (openness, interoperability, modularity) and a community that constitute assets for any structure implementing an SDI.

Openness The code is made public from the early stages of development to favour exchange and widen participation. This is important to have developers as well as the wider community adhere to the project, have their say, and contribute, hence foster on-going development, securing geOrchestra as a long-term solution.

Interoperability This mostly relies on compliance to standards, whether OGC or INSPIRE. OGC technical standards apply to services, whereas INSPIRE sets guidelines to properly publish spatial data, with associated metadata sharing a common structure. They can be exchanged ("harvested") between SDI's catalogues.

Modularity geOrchestra is deliberately kept as a set of weakly coupled modules exchanging through OGC standards/REST APIs; depending on needs, one or more components can be skipped or replaced with others fulfilling the same role, as long as they comply with the standards. geOrchestra stands on the shoulders of giants: it builds on projects like GeoServer or GeoNetwork, and direct contributions from the geOrchestra community to these projects are encouraged wherever relevant. Indeed it is a preferred option to build matching bricks which then tightly hold together with a drip of mortar, rather than put loads of mortar onto whacky building blocks.

Community Development and strategies are shared and decided upon publicly and exchanges are based on mutual trust. For scientists this is a change of ethos, a ceasefire in the publish-or-perish war. In any case the help and support we receive from this community (users and developers) is a real asset. Benefiting from the experience of others (whether from the geOrchestra community or from "components" communities, especially GeoServer and GeoNetwork) has boosted the build-up of our SDI a great deal.

GéoSAS Following early investigations (CSEB, 2008), the SAS research department became in 2010 the second organisation after GéoBretagne (2009) to run a geOrchestra SDI (the first one for scientific purposes), under the name GéoSAS. Rather than deploying the whole architecture at once we started making use of the different components gradually, starting with GeoServer, GeoWebCache and mapfishapp, then GeoNetwork. The central authentication service was deployed only to restrict access to data layers where privacy was at stake. We did not deploy a data extractor for the following reasons: (1) download is possible directly from the viewer, and (2) we consider that best practices should expunge the widespread habit of systematically downloading (and hence producing duplicates) data, as an SDI is meant to persistently give access to it.

A progressive SDI build-up was preferred to full deployment in-a-row as it enables the accumulation and assimilation of local competencies through a learning-by-doing process, assisted by the then nascent geOrchestra community. We therefore soon had opportunities to both contribute to geOrchestra and GeoServer.

In parallel we explored the potential of an adjunction of OGC Web processing services (WPS) to our SDI, and now have a WPS server as part of GéoSAS. Processes can be tuned and triggered from a bespoke add-on to the viewer with direct cartographic output. To-date the WPS-server solution retained is PyWPS, though others would also be convenient. A WPS server inserts into the overall geOrchestra architecture in the same way as other components: it is weakly coupled and can easily be removed or replaced as inputs/outputs (i/o) flow as OGC standards.

A watershed delineation WPS (WS-WPS) has been in production since 2011, and add-on WS-AO since 2012 (Squividant *et al.*, 2013). The OGC standards and the generic approach to develop add-ons for geOrchestra's main viewer make sharing with other SDIs easier: WS-AO was made available on GéoBretagne in 2014. Other hydrology-related WPSs (e.g. modelling of hydrographic networks) were released at the same time (2011, 2012), and a project stemming from this approach is under way to bring functionalities to officers in charge of water management with the support of the French national agency for water and aquatic environments (ONEMA).

Spatial monitoring and display of the physical and chemical characteristics of water is also a centre of interest. Data and model output displays that include alternative graphical views were explored as part of project Vidae (2009, 2013), and its sequels with the implementation of the OGC Sensor Observation Service (SOS) using the Sensor Observation Service Data Management System (ISTSOS), combined with Dygraphs (a dynamic JavaScript charting library). It was applied to environmental research observations on agro-hydrological systems AgrHyS. All these elements were integrated into GéoSAS.

Applications are also being diversified with the development of WPS, and add-ons for: (1) soil types database exploration add-on, project Sols de Bretagne (SdB, 2013); (2) a remote-sensing

WPS and add-on to detect and monitor invasive aquatic and amphibious vegetal species based on imagery and on-site knowledge.

All the developments we have made (proof-of-concept an production alike) are either directly publicly shared or included in FOSS, with possible contributions back from the community to help improve the piece of software. This is an important element that helped and helps turning the efforts of a very few people (our small team) into valuable outputs without much funding.

SPREADING THE WORD: OPEN, MODULAR, INTEROPERABLE

Open Sharing is not just about putting material (data, processes, or models) on-line, and must also encompass re-use and exchange. To this end, experience shows it is extremely important not to conceive an SDI as (or, merely as) a portal. Rather, every SDI's services must be: (1) autonomous with i/o using OGC standards where applicable; (2) implemented so as not to narrow the scope and potential for service use/re-use, as one can (or must!) never be certain of the way these are going to be used. When SDIs centre on standardised services instead of portals, different services from different SDIs seamlessly interoperate, and SDIs truly operate as an ecosystem. Conversely, thinking portal first bears a risk of achieving poorly interacting SDIs running independently as silos. The question: "do we want a silo or instead to enrich an ecosystem?" needs to be considered in-depth, as when overlooked (as is often the case), the outcome consistently turns out to be a silo. We posit that science at large has more to gain with an open ecosystem. In our experience a way to lower the risk of an SDI becoming a silo is to avoid brandishing banners: logos and acknowledgements must remain discrete, and put on data, process, and service metadata rather than on viewers and portals (metadata are here to credit adequately the producer, manager, and/or owner, of the data, service, viewer, … and metadata itself). Also, portals must not be exclusive. A portal is an access to data, metadata and processes (all being services) but it should never impede alternative accesses to favour the use/re-use of metadata, data or processes.

Open metadata In the INSPIRE system (and for SDIs in general) metadata hold a central place as they are the way of choice to search and discover data; metadata are to be included in searchable catalogues, and catalogues can in turn harvest and enrich each other's contents.

Open data Until now, only access to metadata is made compulsory by the INSPIRE Directive. However, the obligation will gradually extend to data. Furthermore, most data produced or acquired on public funds is subject to the PSI directive, which already claims a public access to data (and this, without delay, i.e. within 20 days from the date of production whenever possible).

Open processes We envision that whatever holds for data and metadata can soon extend to services and processes. As an example, WS-AO is available on GéoSAS, but also GéoBretagne, and soon on other SDIs. Up-to-now there is no framework to foster process re-use, but we view this as a natural move towards maximal interoperability.

To summarise, the term "Open" applies to the SDA and software (FOSS), but it must be extended to the way a SDI and the SDI ecosystem function.

Modular Building SDIs as autonomous components and services too appears as a best practice that fosters the emergence of a truly interoperable SDI ecosystem. As such a geOrchestra-SDI can work without having all its components deployed, depending on needs and local capacity. This allows partial or progressive deployment (as we did), and whenever needed relying on other, pre-existing SDI components of the ecosystem (e.g. not deploying a catalogue and using another SDI's catalogue instead).

Interoperable The SDIs and SDI components making up an ecosystem do not have to be geOrchestra. Interoperability for a cluster of elementary bricks is ensured by OGC standards, which also means that an SDI can use bricks from another interoperable SDI as a replacement. This also means that a brick may be substituted with another that is not part of geOrchestra's original architecture, as long as it complies with the standards.

CONCLUSION

In this paper we have expressed the needs (generic and specific) of a department in environmental science in terms of SDI. Such an SDI in itself must serve a double purpose: to publish and share data with the science community and to present output to the wider public. We stated that it is also convenient to publish and share processes and models. geOrchestra was then introduced and we explained why it is especially convenient and fulfils our needs. We also explored the various ways a geOrchestra-based SDI can be further tuned for specific scientific needs (online processing, charting, etc.) and showed the importance of openness, modularity and interoperability, for SDIs, individually and collectively. We clearly support the maximum of interoperability and openness, together with a modular architecture, as it induces best practices to foster advances in knowledge sharing and discovery, and ultimately advertise and advance science.

On the technical side, geOrchestra has proved to be a versatile and sizeable SDA, suitable for various scientific applications. The experience and all the thought put into the geOrchestra project by the whole geOrchestra community (GéoSAS included) also improved our comprehension and helped identify best practices from the methodological and organisational, if not theoretical and philosophical point-of-view. Though a supposedly distant and secondary goal while struggling to implement an SDI, one should constantly keep in mind how it shall be inserted into the whole SDI ecosystem, how it shall interact with it, and what it shall share. We believe the more it does (insert, interact, and share) the better for the SDI as well as for the ecosystem at large.

REFERENCES

CSEB (2008) CSEB portal for the dissemination of information on water quality (nitrogen, pesticides) in the catchments of Brittany. *Conseil Scientifique de l'Environnement de Bretagne, Aquascop*. http://tiny.cc/geosas-cseb

data.gov (2009) The home of the U.S. Government's open data. http://data.gov/

GéoBretagne (2009) http://geobretagne.fr/mapfishapp/?lang=en; http://cms.geobretagne.fr/

geOrchestra (2009) http://www.georchestra.org/

GéoSAS (2010) http://tiny.cc/geosas

INSPIRE (2007) Directive 2007/2/EC of the European Parliament and of the Council of 14/03/2007 establishing an Infrastructure for Spatial Information in the European Community. http://inspire.ec.europa.eu/

OECD (2007) OECD Principles and Guidelines for Access to Research Data from Public Funding. http://www.oecd.org/sti/sci-tech/38500813.pdf

OGC (1994) Open Geospatial Consortium. http://www.opengeospatial.org/

Pilat, D. and Fukasaku, Y. (2007) OECD Principles and guidelines for access to research data from public funding. *Data Science Journal* 6, Open Data Issue, 2007.

PSI (2003) Directive 2003/98/EC of the European Parliament and of the Council of 17/11/2003 on the re-use of public sector information. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:345:0090:0096:EN:PDF

PSI (2013). Directive 2013/37/EU of the European Parliament and of the Council of 26/06/2013 amending Directive 2003/98/EC on the re-use of public sector information. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:175:0001:0008:EN:PDF

Science.gov (2002) A gateway to government science information and research results. http://www.science.gov/

SdB (2013) Sols de Bretagne (Websol Bretagne). http://tiny.cc/geosas-sdb

Squividant H., Bera R. and Aurousseau, P. (2013) WPS Bassin Versant: un outil de modélisation hydrologique intégré à une infrastructure de données spatiales. *M@ppemonde*. 112, 4-2013. http://mappemonde.mgm.fr/num40/fig13/fig13402.html

Squividant H., *et al.* (2015) Online watershed boundary delineation: sharing models through Spatial Data Infrastructures. Proc. *3rd Remote Sensing and Hydrology Symposium* and *3rd International Conference of GIS/RS in Hydrology, Water Resources, and Environment*, Guangzhou, China, 24-27 August 2014. This issue. IAHS Publ. 368.

Uhlir, P., and Schröder P. (2007) Open data for open science. *Data Science Journal* 6, Open Data Issue, 2007.

Vidae (2009) VIsualisation de Données Agro-Environnementales. http://tiny.cc/geosas-vidae2009

Vidae (2013) http://tiny.cc/vidae2013

Wang, S. (2010) A CyberGIS framework for the synthesis of cyberinfrastructure, GIS, and spatial analysis. *Annals of the Association of American Geographers* 100(3), 535–557. doi:10.1080/00045601003791243

Wang, S., *et al.* (2013) CyberGIS software: a synthetic review and integration roadmap. *International Journal of Geographical Information Science* 27(11), 2122–2145. doi:10.1080/13658816.2013.776049

Willinsky, J. (2005). The unacknowledged convergence of open source, open access, and open science. *First Monday*, 10, 8-1 August 2005. http://ojphi.org/ojs/index.php/fm/article/view/1265/1185. doi:10.5210/fm.v10i8.1265