An open source strategy for public online algorithms and data services: the French water information system experience.

Alexandre Liccardi*[i], Laurent Coudercy[i], Samuel Dembski[i], Anthony Mauclerc[ii]

[i] AFB (French Biodiversity Agency), [ii] BRGM (French Geological Survey)

**alexandre.liccardi@afbiodiversite.fr*

Abstract

For almost two decades, the French water information system (F-WIS) provides references and datasets to a large audience of scientists, professional experts, decision makers and stakeholders. The first water open data repository was released in 2013. The transparency policy expands with the publication of many APIs and data services aimed at citizens and specialists. The wide scope of the water fields (hydrology, biology, chemistry, economics...) is concerned.

This paper focuses on three major projects: Hub'eau (a bigdata repository), the WILD programming library with the French WFD-reporting desk (an online data validator in line with EC obligations), SEEE (an online ecological assessment tool). These three projects are open source, available online and embody the will of their leaders (the French Agency for Biodiversity and the INSIDE cluster) to let a wider public invest the field of environmental data and to stimulate relevant stakeholders participation.

Keywords – Ecological assessment tools; Data collection and processing; Big data and large datasets repositories; Open source and open data public policies; Data access and online algorithms

1 Key tenets of a partnership-based strategy

Gathering environmental data from relevant public services and surveys requires a strong legal background and the means of its application (Toots, 2017). As far as water, the French water information system (F-WIS) emerged as a consortium first, then settled as an effective institutional framework.

1.1 The information framework: a top-down legacy?

Environmental data access for public information inherits a significant legal corpus at the intersection of two legislations. The first one is related to the production of monitoring data, a process at the basis of the implementation of an environmental policy, and which includes several thematic directives like the Water framework directive (WFD), the Nitrates directive, the drinking water directive, the Fish directive, the Marine strategy framework directive (…). The second legislation aims at improving the transparency of public policies and establishes a minimal regulatory framework that secures the sharing of data between public organisms and the citizens.

Both European Directives and national transpositions have been set up to help for a better efficiency in understanding the way decisions are made, and for the reuse of the knowledge gathered by local authorities[1].

At national and regional level, successive Laws and global initiatives allowed administrators and stakeholders to settle a real information framework (Attard et al., 2015). The French Water Law released in 1992 set the principles for concerted and participative river basins management plans, and the 2006 French Water Law transposed the WFD, reaffirming the needs in public data ownership. In lockstep with these top-down guidelines, the gathering of data and the building of information systems need every data producer involvement and a strong expression of needs of all kind of users. This bottom-up work took nearly 20 years to emerge as a comprehensive collaborative system, and to successfully rally interested communities (state officers, local agents, private engineers and scientists, recently small-business IT offices and civilian society) from the particularly complex field of environmental governance and sciences (Armitage & al, 2008) .

1.2 Building a momentum

Since 1992, national bodies and directions worked for a better data access: early technical efforts have been made thanks to a dedicated service (SANDRE[2]), providing reference and master data, and standard protocols for maximal technical access from collection to broadcasting, data master management, administration and monitoring. This first technical achievement provided a framework and a pragmatic starting point for any further reflection, and

[1] Official references can be found at PSI directive: 2003/98/CE, 2013/37/UE, INSPIRE: 2007/2/CE, WFD: 2000/60/CE.

[2] http://www.sandre.eaufrance.fr/

became an integral part of laboratory IS as it has been recognized as an exchange standard.

The French decree establishing the water data national scheme (SNDE[3]) was released in December 2009 and constitutes a major step in the building of a consistency network process: the French water information system (F-WIS). Today, dozens of public institutions, including river basins managers, scientific facilities, professional agencies and expertise groups collaborate to collect data on water topics and to make it more understandable. The French Biodiversity Agency (AFB) coordinates the French WIS, under the responsibility of the French Ministry of the Environment.

Some principles are now commonly acknowledged by the French WIS users and contributors: sharing IT governance and knowledge (a common language, specifications and IT architecture), transparency from technical data to official assessments and statistics (on line working network and partnerships: for instance open data inventories) and pooling the efforts with a coherent level of subsidiarity. A more conceptual vision was built, with the publication of a green book (Lalement & Lagarde, 2005) and the support of research facilities, of which Interoperability for Information System on Water Data cluster (INSIDE[4]) is dedicated to IT interoperability, transparency and innovation.

2 An open data policy to ensure data access

In order to provide the means of an effective societal understanding and involvement, a guarantee of transparency has been identified as a priority (Misuraca & Viscusi, 2016; Toots, 2017).

The early data publication strategy of the F-WIS has been led by two pivotal processes: the first one allows any contributor to disclose trusted datasets on indexed repositories; the second one provides a fully connected infrastructure accessible through online streams thanks to a common vocabulary. Figure 1 shows how the full strategy articulates.

2.1 First step to transparency: a high level of information through repositories

In March 2013, an open data repository (*data.eaufrance.fr*) has been published. More than 250 datasets, designed for data reutilization, produced by 50 contributors, are already online. Where possible and relevant, 'human-readable' format (XLS, ODS, CSV, PDF) are provided. This website is highly synchronized with the national French open data repository[5] thanks to harvesting mechanisms, and allows the AFB to pre-empt future expectations set up with the late *French Digital Republic Law* (voted in October 2016).

The choice of a common open license for data publication[6] indicates a high compatibility with preexisting contracts and initiatives (ODC-BY, CC-BY 2.0, ODbL), a convergence with EC leaders (for instance, Open Government License) and a strong will, to let a wider public invest the field of environmental data and stimulate relevant stakeholders participation as well as digital private actors and communities investment. The reuse of data is a strong will in the PSI Directive (Lakomaa & Kallberg, 2013), and data preparation and explanation can represent a significant cost for public bodies (Johnson et al., 2017). Enhancing the understanding of data production and computations, strengthening the accountability of producers through a greater transparency of each step of

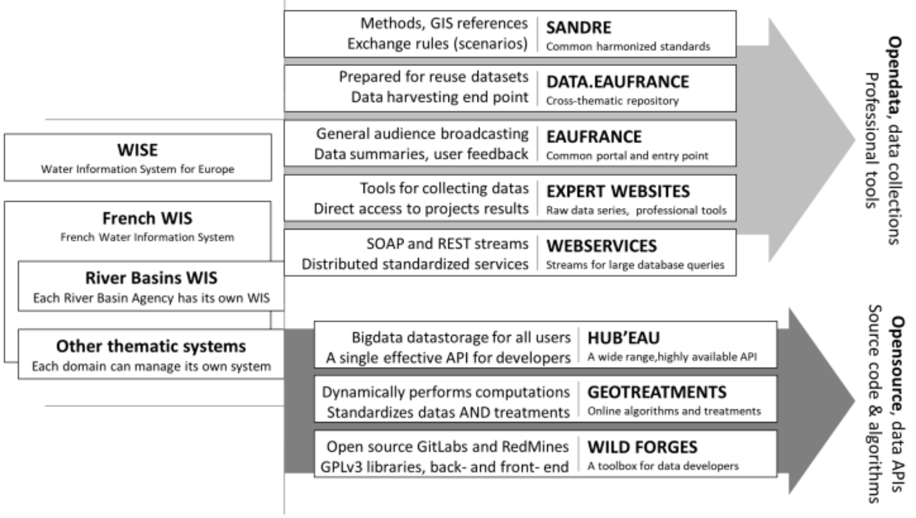

Methods, GIS references
Exchange rules (scenarios)
SANDRE
Common harmonized standards
Prepared for reuse datasets
Data harvesting end point
DATA.EAUFRANCE
Cross-thematic repository
General audience broadcasting
Data summaries, user feedback
EAUFRANCE
Common portal and entry point
Tools for collecting datas
Direct access to projects results
EXPERT WEBSITES
Raw data series, professional tools
SOAP and REST streams
Distributed standardized services
WEBSERVICES
Streams for large database queries
Opendata, data collections
Professional tools
WISE
Water Information System for Europe
French WIS
French Water Information System
River Basins WIS
Each River Basin Agency has its own WIS
Other thematic systems
Each domain can manage its own system
Bigdata datastorage for all users
A single effective API for developers
HUB'EAU
A wide range,highly available API
Dynamically performs computations
Standardizes datas AND treatments
GEOTREATMENTS
Online algorithms and treatments
Open source GitLabs and RedMines
GPLv3 libraries, back- and front- end
WILD FORGES
A toolbox for data developers
Opensource, data APIs
Source code & algorithms

Figure 1. A F-WIS technical overview, with its main components (from left to right): partners databanks used as source, open data collections and tools, open source online processes

[3] https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000022734282

[4] http://www.pole-inside.fr

[5] See data.gouv at http://data.gouv.fr.

[6] More details at https://www.etalab.gouv.fr/licence-ouverte-open-licence and http://www.data.eaufrance.fr/presentation

production have been the main targets of the repositories.

The French WIS takes advantage of a dedicated portal for public opinion and public participation, as well as professional tools: the *eaufrance* portal[7] is the main entry point for public water information. Not only registered websites provide public information for better environmental policies, experts and professionals can find there effective tools and raw data related to field monitoring and technical issues. The scope of data is broad in terms of categories of water bodies (groundwater, coastal and transitional water, lakes, rivers). It includes the state of natural and drinking water (bathymetry, physico-chemistry, hydromorphology, biology, quantity (discharge and low flows), real-time temperatures, economical issues) as well as details about pressures that have an effect on this state (public works, dams, emission of pollutants by human activities) and about measures set up to prevent or mitigate the deterioration of the quality of water (public services efficiency, restoration of rivers...).

2.2 A common language implies common structures (data streams & scenarios)

At Common languages and tools imply common structures and units of analysis (Attard et al., 2015, Toots, 2017). The INSPIRE Directive set up models and facilities; in addition the F- WIS uses standards and methodologies built within expert groups and orchestrated by the SANDRE. Data is not only timely results, it's also modeling and design – and transparency depends on the public ability to re-use data (Geiger & Von Lucke, 2012). Data transaction scenarios were elaborated using XML/XSD technologies, and SOAP services provide direct access to well-formatted data for online applications, known as *Webservices*. These data streams are produced and maintained by the six River Basin Agencies (*Agences de l'Eau*) and recently one oversea Office.

The user feedback is shared and reported by the AFB, using simulations that reproduce the most frequent users behaviors known and collect statistics. This centralized monitoring and guidance is a response to a technical need and takes interest in user experience; although the main effort is deeper. The study of distributed systems designs and architectures, their relationship and the way actors coordinate their work is part of F-WIS governance on the policy side, as part of solving technical issue on the digital side. Dedicated projects and the INSIDE cluster deal with this last side. A recent effort was recently made for URI denomination and attribution.

3 Interoperability and public algorithm through open source approaches

The INSIDE cluster gathers the AFB and the BRGM, with distributed system issues and new means of data acquisition. The structure provides a framework for services orchestration studies and interoperability projects: facilities for distributed services simulations tests and computation, data acquisition from sensors to crowd-sourcing, software developers' rooms (forges, git) for sharing innovation – according to the idea, that transparency isn't only providing output but also committing to creation phases (Reed, 2008; Janssen et al., 2017).

The cluster promotes innovative digital technologies and a strong will is to start a collaborative momentum. Hackathons, workshops and seminaries took place since 2014, knowing a growing success (from barely fifty participants in June 2014 to more than two hundred in April 2017).

3.1 *Hub'eau*, a common entry point for the French WIS data

The INSIDE cluster not only provides ideas and schemes to work together; it also provides prototypes and ready-to-use services. One major example is *Hub'eau*.

3.1.1 From repositories to user-friendly hubs: promoting developers' interface and higher availability

New needs emerged with the increase of data availability, combined to the growth of initiatives for a more meaningful use of data through the web (for instance Linux Foundation's Open API). Previous F-WIS applications, based on datasets broadcasting and master data management, met their limits in regards to new criteria:

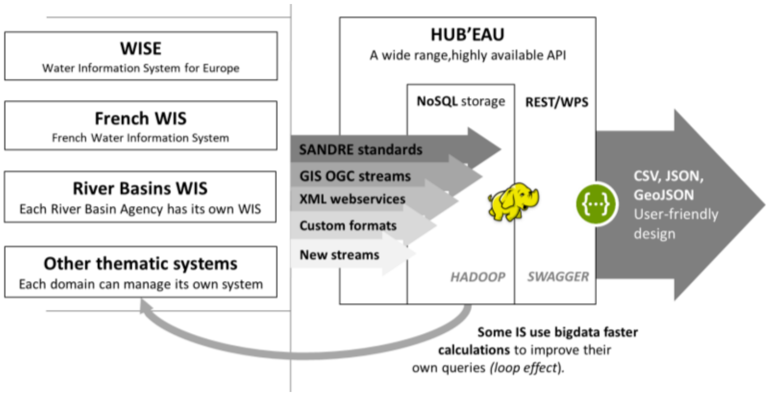
- high availability, even when data quantity is huge, and ready-to-use format,
- data integrity and independence towards format, especially when the source is convoluted, heterogeneous and frequently updated,
- API based design, addressed to web developers' and not requiring a high technical level to explore information.

A technical proof of concept was design by the BRGM: *Hub'eau* (http://hubeau.eaufrance.fr) is the first real big-data datawarehouse for the French WIS (using JAVA and NoSQL-Hadoop, as shown in the Figure 2). It targets a high availability in queries for large scale datasets, such as physical chemistry of lake and river since 1950, groundwater level with high accuracy, times series on fishing communities, economic data on water price for

[7] http://www.eaufrance.fr/ , http://www.data.eaufrance.fr/

each town (…). All these data are accessible through a single entry point, using REST service and producing GeoJSON strands. Developer-friendly documentation and interface, using SWAGGER for instance, tends to guarantee a better and wider use.

Figure 2. Hub'eau is a common hub for the French WIS data. Many various exchange protocols and formats become one single entry point, with GeoJSON or CSV formatting. The API is easily handled using a predictive input help for URL writing.


WISE
Water Information System for Europe
French WIS
French Water Information System
River Basins WIS
Each River Basin Agency has its own WIS
Other thematic systems
Each domain can manage its own system
HUB'EAU
A wide range,highly available API
NoSQL storage
REST/WPS
SANDRE standards
GIS OGC streams
XML webservices
Custom formats
New streams
HADOOP
SWAGGER
CSV, JSON, GeoJSON
User-friendly design
Some IS use bigdata faster calculations to improve their own queries (loop effect).

3.1.2 Multiple use of a powerful tool

The proof of concept was presented at the 95th ASTEE conference to water professionals and scientists throughout a whole coding day (*hackathon*). The feedback was not only positive, but the developers' adoption revealed the use of unexpected software clients: QGIS, office software suite or existing private BI tools (*Tableau software*) were able to deliver fully usable and customized interfaces without the use of any ETL. The scalability of GeoJSON and CSV, and the *self-supporting* of the datasets published (each datasets contains the relevant excerpt of the repositories needed for its understanding) have been noted as practical solutions for apps integration and API dissemination.

Another breakthrough came later directly from scientific communities. Geological survey teams from the BRGM noticed that deep calculous concerning hydrogeological modelling and data preparation were made many times quicker with the use of the big data solution. This loop effect resulted to a new embedded design, were *Hub'eau* not only store data for a better access, but also contributes to the IS it feeds from.

3.2 A toolbox to capitalize on environmental customized source code: WILD[8]

From the idea of tools and service sharing, and its need to gather long-term communities, INSIDE proposed a mutual open source toolbox delivered as an API. The WILD project provides a JAVA library and a strong basis for such developments.

3.2.1 Pooling coding effort, sharing source code and resulting services

The initial perimeter of WILD project is about coping with IT developments in ETL applications, as well as *Hub'eau* renewed F-WIS data engineering. A profound change in IT practices had to result from this search of effort optimization. Indeed, concerned technical and scientific communities came with highly customized needs and a strong "work in silo" habit.
The main challenges are to maintain IT skills networks and to rally open source existing communities, to earn legitimacy and trust of future participants (see also Reeds, 2008). The identification of common needs is the main binder of the initiative: environmental data scope is vast and still being discovered (Armitage et al. 2009). The efficiency of produced tools and their usability is a first selection criterion; the second one is the chance to optimize the solution or to alter some features according to each topic[9].

3.2.2 The open source trend as a redesign tool for engineers and their communities

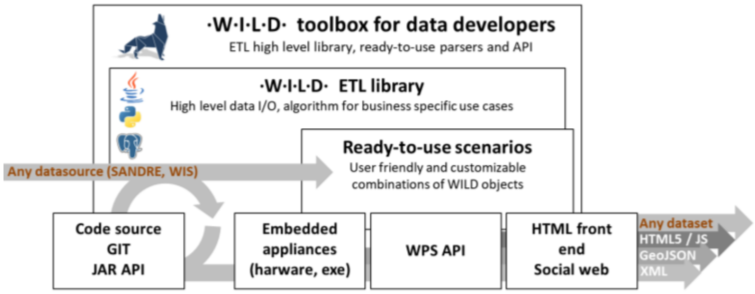
The WILD project targets the advent of an open source strategy to share IT code and capabilities. The deployment of software contributive environments such as *GitLab* and *Redmine* helped to provide a common platform for developers and for webservices. Contracting companies contributing to the code commits use the same tools inner software engineers and developers use. A dynamic way to lead public market, more agile, provides more flexibility and more transparency (Jetzek, 2013, Lakomaa & Kallberg, 2013).
The second goal was to ensure the reuse of the ETL toolbox, in library build-in (see Figure 3). Several business projects already use the WILD library, for F-WIS streams monitoring (OWS WIS project, 2016-2017), ROV sensors data at high speed analysis (Aquadrone, ESIPE & AFB, 2016-2017: Liccardi & Collomb, 2017), for data administration and valorization in multi-stakeholder process.
A long term prototype allows data qualification for the WFD CE reporting, and implies a full back office service and a social designed front office (Liccardi, 2016; Liccardi & Goyen , 2014). More than one hundred officers used this online tool for customized tests and analyses during the

[8] WILD stands for *Wrappers for Interoperability and Lazy Developers*, in recall to its modular computational design to make data processing easier.

[9] More information can be found at http://www.pole-inside.fr/wild.

2016 reporting to the European Commission, ultimately answering the directive which initiated the transparency process on water fields, 16 years before.

Figure 3. The WILD library can be used by developers as a toolbox, and can be embedded in ready-to-use tools for field experts. Additional formats and functionalities can be introduced with scenarios customization and/or code source commit.


·W·I·L·D· toolbox for data developers
ETL high level library, ready-to-use parsers and API
·W·I·L·D· ETL library
High level data I/O, algorithm for business specific use cases
Ready-to-use scenarios
User friendly and customizable combinations of WILD objects
Any datasource (SANDRE, WIS)
Code source GIT JAR API
Embedded appliances (harware, exe)
WPS API
HTML front end Social web
Any dataset
HTML5 / JS
GeoJSON
XML

3.3 The water ecological assessment system (SEEE): making ecological *state* accessible

SEEE was developed by the AFB under the lead of the French ministry in charge of environment. The tool bundles algorithms written for the translation of national methods of assessment, and their publication by open-source services. These algorithms are established by the ministry with the river basin authorities in compliance with the WFD, through a concerted governance process in order to set their indicators, their development and their further integration in the national regulations.

The *ecological state* is a valuable summarized indicator, of which underlying scientific basis can be difficult to fathom for general public and even river authorities (ONEMA, 2016). The assessment system was designed to enhance legibility of indicators and to let stakeholders, and more broadly UE citizens, run their own calculations.

3.3.1 Transparency and scientific opinions: how to reconcile dogmas

Ecological assessment is not an easy matter: large datasets are needed for input and reference ranges, demanding algorithms involving simulations, bootstraps, rankings and partitioning are completed, and expert judgement makes up for the last step of the process. Expert judgement cannot be reproduced by an automated service: experts rely on their deep knowledge of the territories, and can alter results following unstructured information (these exemptions are stored as remarks, in plain text, but are hardly available in an analytical form).

The best that can be done is to maximize transparency by providing data and computational tools, to place the user at the highest available level of information. From professional leaders and project managers' point of view, functional targets are to reconcile the independence, legitimacy and rights of domain specialists and the general audience demand and to allow them work on a same platform (Willinsky, 2005).

Every audience can directly request the same tools scientists and experts should use for their own work. The collation and the verification of input data left to the responsibility of the water manager using SEEE[10].

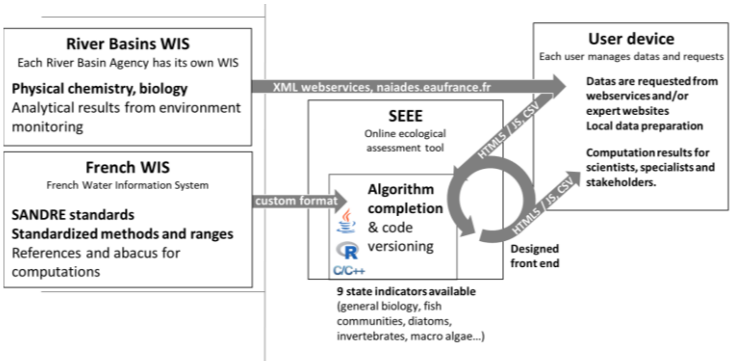
3.3.2 A highly customizable tool fully embedded into the F-WIS architecture

In order to guarantee the integrity and the understanding of results, the proposed design has to answer two more requirements:

- users must be able to add their own algorithms and to modify existing ones, in the simple way of indexing a R script file,
- the whole structure must not seem labyrinthine for engineers, and users must be able to reproduce the computation on a local solution.

The resulting platform is delivered with a highly customizable interface and a versioning of algorithms (see Figure 4).

Figure 4. River Basins webservices and SEEE assessment tool are chained for indicators reporting transparency.


River Basins WIS
Each River Basin Agency has its own WIS
Physical chemistry, biology
Analytical results from environment monitoring
French WIS
French Water Information System
SANDRE standards
Standardized methods and ranges
References and abacus for computations
XML webservices, naiades.eaufrance.fr
custom format
SEEE
Online ecological assessment tool
Algorithm completion & code versioning
C/C++
9 state indicators available (general biology, fish communities, diatoms, invertebrates, macro algae...)
HTML5 / JS, CSV
HTML5 / JS, CSV
Designed front end
User device
Each user manages datas and requests
Datas are requested from webservices and/or expert websites
Local data preparation
Computation results for scientists, specialists and stakeholders.

To facilitate the interconnection between SEEE and databases of district authorities or river managers, the system publishes an API. The underlying service implements REST standards to load input data files (biological, chemical and physical analyses custom results) and performs computations of the WFD ecological state of lake and rivers. SEEE has been used as a common service by water experts (scientific bodies such as IRSTEA and institutional actors such as Water Agencies) for the preparation of the WFD reporting.

[10] Tools are available at http://naiades.eaufrance.fr, http://seee.eaufrance.fr.

4 Recent achievements and outlooks

Current digital trends and researches already brought data management, environmental services profession applications and data broadcasting to an effective and reliable technological level. Many helps were needed, including a new legal paradigm and a substantial financial and human investment. AFB latest developments invests in internet of thing (IoT) thanks to drones and in Semantic Web and RDF ontologies (common modeling following SANDRE productions and INSPIRE obligations paved the way). Public access and process transparency are still the primary objective of the French WIS, as shown by the *closer-to-home data* INSIDE project, recent *eaufrance* reshaping, new datavisualisation services (*Cartograph*) for instance.

Géotraitements is another promising project proposing online algorithms for a better water policy. Using the WPS standard, the related service exploits the potential of complex datasets, such as BD CARTHAGE® (rivers map, provided by IGN) and BD LISA® (underground tables map, provided by BRGM). Chaining computations using hydrological network, 3-dimensionnal cross-matching right under a given position are already available.

Amongst the French WIS principles previously quoted, citizen empowerment tends to emerge as a new duty. Alongside crowdsourcing (rivers water level are planned to be supported by local residents and hikers) and mobile applications (cross-platform front-end solutions for performance of water public service, ecological states are now available), the impulse must be kept through the widening of the community of developers and specialists in a first step, for the lead innovation in the public sector as well as in the private sector. In a second step, the prospect is to rally contributors and users, and to capitalize on their feedback for proactive disclosures and a greater incent to take part in environmental policies.

Acknowledgements

The authors would like to thank François Hissel (AFB) for his insightful comments and suggestions. This paper writing wouldn’t have been possible without the previous theoretical work and the long-term strategy built by the F-WIS contributors, especially René Lalement (AFB) and Pierre Lagarde (BRGM). No data was used in producing this manuscript.

References

Toots, M., McBride, K., Kalvet, T., & Krimmer, R. (2017). *Open Data as Enabler of Public Service Co-creation: Exploring the Drivers and Barriers.* In Proceedings of the 2017 International Conference for E-Democracy and Open Government (CeDEM 2017) (102-112). Krems, Austria: IEEE Computer Society. doi:10.1109/CeDEM.2017.12

Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). *A systematic review of open government data initiatives.* Government Information Quarterly, 32(4), 399–418. doi:10.1016/j.giq.2015.07.006

Armitage, D. R., Plummer, R. , Berkes, F., Arthur, R. I. , Charles, A. T., Davidson-Hunt, I. J. , Diduck, A. P. , Doubleday, N. C., Johnson, D. S., Marschke, M., McConney, P., Pinkerton, E. W., & Wollenberg, E. K. (2009). *Adaptive Co-Management for Social–Ecological Complexity.* Frontiers in Ecology and the Environment7(2):95–102. doi:10.1890/070089.

Lalement, R. & Lagarde, P. (2005). *Architecture du Système d’Information sur l’Eau. Livre vert.* Ministère de l’écologie et du développement durable Printing services, 36 p. La Défense, France.

Misuraca G., & Viscusi G. (2016). *Is Open Data Enough ? E-Governance Challenges for Open Government*. International Journal of Electronic Government Research (IJEGR)10(1), 17p. doi:10.4018/ijegr.2014010102

Lakooma, E., & Kallberg, J. (013). *Open Data as a Foundation for Innovation: The Enabling Effect of free Public Sector Information for Entrepreneurs*. IEEE Access 2010(1): 558–563. doi: 10.1109/access.2013.2279164.

Johnson, P. A., Sieber, R., Scassa, T., Stephens, M., & Robinson, P. (2017). *The Cost(s) of Geospatial Open Data*. Transactions in GIS 21(3):434-445. doi: 10.1111/tgis.12283

Geiger, C. P., & Von Lucke, J. (2012). *Open Government and (Linked)(Open)(Government)(Data).* eJournal of eDemocracy & Open Government 4(2), JeDEM 4(2): 265-278.

Reed, M. S. (2008) *Stakeholder participation for environmental management: a literature review.* Biological Conservation 141(10):2417–243. doi: 10.1016/j.biocon.2008.07.014

Janssen, M., Matheus, R., Longo, J., & Weerakkody, V. (2017) *Transparency-bydesign as a foundation for open government*, Transforming Government: People, Process and Policy, (11/1): pp.2-8. doi: 10.1108/TG-02-2017-0015

Jetzek T. (2013). *The value of Open Government Data.* Perspektiv nr. 23-2013: 47-56.

Liccardi, A. & Collomb, J. (2017). *Aquadrone : geo-tracking and collecting environmental data from an underwater remotely operated vehicle*, In Proceedings of the FOSS4G-Europe 2017, Marne-la-Vallée, France.

Liccardi, A. (2016). *Plateforme de traitement de données environnementales : un projet open source pour les ingénieurs des données.* In Proceedings of the FOSS4G-FR 2016, Marne-la-Vallée, France.

Liccardi, A., & Goyen, J.-P. (2014). *Utilisation de technologies Open Source dans l'administration, la qualification et l'exploitation des données du SIE.* In Proceedings of the FOSS4G-FR 2014, Marne-la-Vallée, France.

ONEMA, Ministère de l'Environnement, de l'Energie et de la Mer, French WIS (2016). *Guide relatif à l’évaluation de l’état des eaux de surface continentales (cours d’eau, canaux, plans d’eau)* Ministère de l'Environnement, de l'Energie et de la Mer Printing services, 107 p. La Défense, France.

Willinsky J. (2005).*The unacknowledged convergence of open source, open access, and open science.* First Monday 2005, 10. doi:10.5210/fm.v10i8.1265.