# Aix-Marseille University SSH data platforms: Skills to support research in social sciences and humanities (SSH) in the Mediterranean

Sébastien Oliveau, Isabelle Blöss-Widmer, Yoann Doignon, Clément De Belsunce

## ▶ To cite this version:

# Aix-Marseille University SSH data platforms: Skills to support research in social sciences and humanities (SSH) in the Mediterranean

*Les plateformes de données en SHS d'Aix-Marseille Université : des compétences au service des recherches en sciences humaines et sociales en Méditerranée*

منصات بيانات العلوم الاجتماعية والإنسانية في جامعة أيكس مرسيليا: مهارات دعم البحوث في العلوم الاجتماعية والإنسانية في دول البحر الأبيض المتوسط

Sébastien Oliveau, Isabelle Blöss-Widmer, Yoann Doignon and Clément de Belsunce

Sébastien Oliveau, Isabelle Blöss-Widmer,
Yoann Doignon and Clément De Belsunce

# AIX-MARSEILLE UNIVERSITY SSH DATA PLATFORMS: SKILLS TO SUPPORT RESEARCH IN SOCIAL SCIENCES AND HUMANITIES (SSH) IN THE MEDITERRANEAN

SUMMARY

Digital archiving has emerged as a major issue for science at the turn of the 21st century. Within the social sciences, the more specific question of quantitative data is not exempt from this challenge. But just as much as the preservation of data, the question of access to these archives arises. It is from this perspective that we have written this article, proposing a reflection on access to data based on two experiences of support services for researchers in accessing data. The first example takes up the French initiative to archive and disseminate data from official statistics, designed within the CNRS and aimed at archiving and disseminating, and involves the deployment of university platforms. The second is a research project that aims to make data from all Mediterranean countries accessible to the widest possible audience, while respecting the scientific requirements of documentation and rigorous data exploitation. The article therefore discusses the different dimensions of "FAIR" data, with emphasis on the issue of Access.

INTRODUCTION:
THE CHALLENGE OF ACCESSING DATA, PARTICULARLY SSH DATA

The development of computer science significantly transformed scientific practices in the last quarter of the 20th century. Information technology (IT) first made it possible to develop computing power, then information storage capacities. It initially led to an increase in discoveries in the most quantitative sciences as it allowed accelerated calculations, particularly in fields where the data was quantitative and structured (e.g. tables of data in SSH). It then allowed new possibilities to be developed in more qualitative sciences, thanks to the increase in information storage spaces and the development of qualitative or non-explicitly structured data processing (e.g. literary corpora or images in SSH). Interest has gradually shifted from calculation to data.

Today, there is substantial reflection on what IT produces in science in general (largely a renewed questioning of the role of data and the place of reproducibility), and in the sciences more specifically (a new questioning of existing corpora, for example). At the heart of contemporary thinking, the question of data, defined as an elementary description of a reality, is placed at the centre of science and can be defined as an ensemble of systematized knowledge relating to the same field. Data consists of elements resulting from structured knowledge.

In this context, the question of accessing data in order to build knowledge, but also to reproduce reasoning, has become an increasingly important issue in scientific research. The recent development of the global Research Data Alliance network (https://www.rd-alliance.org/), whose objective is the "development of infrastructure and community activities to reduce the social and technical barriers to data sharing and re-use", is testament to this. In France, the creation of the Open Science Committee (https://www.ouvrirlascience.fr) also represents a milestone in the institutional organisation of science and scientific data (ministère de l'Enseignement supérieur et de la recherche 2018).

In this article, we first return in greater detail to the question of accessing data, that has notably led a number of researchers to propose specific computer interfaces. We will give two representative examples in part two. Finally, in the last part of our article, we will tackle a key element of the transformation of research practices, something too often forgotten in favour of technology: the human factor.

In order to remain concise and anchored in practice, we will deliberately restrict our reflection to the field of social sciences in their quantitative dimension.

ACCESS TO DATA: STORAGE AND DISSEMINATION

First of all, we would like to remind that access to data is based on the producer's ability to become a data disseminator (or to entrust it to a third-party disseminator). As well as practical questions, this poses legitimate legal questions, which we will not address here. The task of a disseminator is to make data available, and today this is most often done online. Dissemination must be distinguished from archiving, because the later aims to preserve data with a perspective of long-term storage, while dissemination is driven by the desire to ensure availability.

In France, for example, INSEE (National Institute of Statistics and Economic Studies) produces public statistical data, some of which is general, recent and made available to the public via a website (http://insee.fr). Since 1987, files considered of "historical interest" have been sent by INSEE to the National Archives as part of the "Constance" system (Conchon 1993). Accessing them is then much more complicated, because the researcher must go through a series of authorisations in order to justify their request for access. It is in this context that the "Archives des Données issues de la statistique publique"[1] were created. They document and archive INSEE (and also data from others official sources like ministries) in order to disseminate it to the research community.

However, making data available online does not guarantee that it will be used properly and in accordance with scientific standards. The acronym FAIR represents the qualities necessary for data to be truly usable: Findable; Accessible; Interoperable; Reusable (Wilkinson *et al.* 2016).

These four main principles relate to different dimensions of data archiving and dissemination.

To be "Findable", the data must be indexed on the internet, preferably with a unique and persistent identifier (PID). A persistent identifier is characterised as a unique character string which refers to an object (this may be a physical object, but also a web page or any other document). It is said to be persistent since the identifier will not change even if the designated object is moved. The DOI (http://www.doi.org/) is certainly the most well-known and most widespread of these identifiers, though others are also valid (Cousijn, Kendall *et al.* 2018). Thus, a DOI for a journal article or database will remain the same, even if the article or database is moved to the file tree structure of a website or to another website.

The notion of accessibility is understood here in a narrow sense and refers to the transparency of data recovery protocol. "Open data" is, of course, accessible. Data that requires special protection (confidential data, for example) can be considered accessible if the authentication and authorisation procedures are explicit and visible.

---

1 ADISP – public statistics data archives, http://www.progedo-adisp.fr/.

Interoperability is based on the principle that data is associated with elements that describe it unequivocally, but also on the fact that this set of data/metadata is structured and conceptualised using shared, documented standards.

Finally, to be reusable, the data must be organised in databases using common standards and must be well documented, that is to say with precise and relevant metadata. Metadata consists of sets of elements describing the data made available. These are therefore data notices. For example, the creation date of the data is an important information to know. The metadata can be numerous, with details on all sorts of aspects (who retrieved or constructed the data, where, when, when it was made available, etc.).

While respecting the FAIR principles allows data to be used by the entire scientific community, a growing number of researchers believe this is not enough (see, for example, Boeckhout, Zielhuis and Bredenoord 2018). It is also advisable to provide support for those using this data. The method of facilitating availability is also an issue, whether it is a matter of the ergonomics of the tools offered or even human support for users when it comes to secondary use of this data. A simple file from a spreadsheet on a website, even perfectly documented, is not always enough for a user to be able to retrieve it. To promote this support, a few simple measures can be implemented as a start: promoting the results in descriptive form (graphs, maps) or even referring to earlier analysis (e.g. published results) can help better understand the possible uses. Sometimes this support must be further developed in order to help a user (even an informed one) by means of concrete help in "getting to grips with" the data: advice, training, personalised support.

This position, which maintains the need for mediation between data and researchers, is the one we take (Oliveau 2017; Blöss-Widmer 2019). Mediation can be considered in two complementary ways. The first consists of offering data exploration, visualisation and even analysis tools that let a user enter data without having to retrieve it or having specific software tools. This is the case for the many online mapping interfaces (often based on geographic information systems), but also for an increasing number of data visualisation modules (there are numerous examples, but we may cite the Eurostat Data Browser and its tabular, graphic and cartographic interface for thematic explorations).

The other aspect of mediation is human support. This is based on the involvement of a more experienced user in methodological or technical terms. In the French research system this concerns, for example, the role usually assigned to some of the research assistant found in research laboratories.

## TWO EXAMPLES OF DATA ACCESS TOOLS

To illustrate our point, we will present two data access tools with which the authors have links. These are good examples to show how the FAIR principles

can apply to data dissemination, and how they do or do not include additional elements of accessibility.

The first example is a national research infrastructure that dates back to a reflection at the end of the 20th century. This infrastructure is based on original digital tools and innovative institutional organisation which aim to make access to data effective. To understand the current situation, we need to look back at the way this infrastructure has developed over the last 20 years.

In 1999, the "Social Sciences and their Data" report submitted to the Minister of National Education and Technology proposed the creation of an institute for disseminating data in the social sciences (Silberman 1999). In 2001, the Consultative Committee for Data in Human and Social Sciences (CCDSHS) was created by decree[2]. The mission of the CDDSHS is to define a data policy for the social sciences. On 1 July 2001, the CNRS (National Scientific Research Centre) in partnership with the EHESS (School of Advanced Studies in Social Sciences), the INED (National Institute for Demographic Studies) and the University of Caen, created the "Quetelet Centre"[3]. In 2005, the centre participated in the creation of the Quetelet network (in the form of a Scientific Interest Group). The network includes the INED Survey service, the Sciences Po Sociopolitical Data Centre (CDSP[4]) and the ADISP team[5] (Caporali; Morisset, Legleye *et al.* 2015).

In the 2010s, the tools were gradually simplified. The roadmap for French research infrastructures (ministère de l'Enseignement supérieur et de la recherche, 2008)[6] provided for the existence of a Research Infrastructure called PROGEDO[7]. It was materialised in 2012 with the creation of the "Quetelet PROGEDO" Mixed Unit of Service which took over the activity of the Quetelet network. In 2017, the CNRS decided to assign the ADISP team to Quetelet-PROGEDO. In 2018, Quetelet-PROGEDO was transformed into a Service and Research Unit, called simply "PROGEDO". It fulfils the role formerly carried out by the CCDSHS, namely that of defining a data policy for the humanities and social sciences, as well as the mission of the Quetelet network (data dissemination) through the "Quetelet-PROGEDO-Diffusion"

---

2   Decree no. 2001-139 of 12 February 2001. The decree no. 2015-1469 of 13 November 2015 abolished the CCDSHS.

3   "Mixed Unit of Service T2419", the "Quetelet Centre for archiving and disseminating data in the humanities and social sciences".

4   CDSP is the successor of the sociopolitical data bank established at the Grenoble Institute of Political Sciences, IEP.

5   ADISP inherited the database developed within the Laboratory of Secondary Analysis and Methods Applied to Sociology (LASMAS). LASMAS became the Maurice Halbwachs Centre in 2004

6   Ministry of Higher Education and Research, 2008, p. XIII

7   PROduction and GEstion des DOnnées: Data Production and Management.

internet portal. It is also responsible for archiving quantitative data via the ADISP team[8].

Today, the "Quetelet-PROGEDO-Diffusion" portal (http://quetelet.progedo. fr/) is firmly positioned in the FAIR data perspective, providing access to data according to its level of sensitivity (see below) and documenting it up to the level of questionnaire variables when possible. The portal makes it easy to find data even from a large variety of sources (INED, Sciences Po Paris, INSEE, Ministries, Universities, etc.)[9]. These data are made accessible thanks to protection which varies according to the type of data: open access for data that does not concern individuals; data with reserved access for so-called "production-research" files that do not directly identify individuals but are considered sensitive; data with secure access for individual data. The DDI documentation protocol is used to make this data reusable and also allows interoperability.

Respect for the FAIR principles has undoubtedly contributed to the success of this ambitious data dissemination enterprise, as shown by the results of the survey conducted in 2019[10], with 90% of users declaring themselves satisfied. As we will see, despite positive evaluation, experience shows that it is essential to provide additional support for the least skilled users and to continue efforts to publicise the services offered. This explains why in France, in particular, this tool is not only digital, but is simultaneously based on human resources which can answer to various requests from users. In order to make the data accessible, in the broad sense of the term[11], efforts must be continued in supporting visitors to websites which provide data in order to make the best use of it in their analysis. Two tools we are responsible for have inspired us to illustrate this necessary articulation of digital enhancement tools with personalised support systems using support staff. Namely, the specific interface of the demographic observatory of the Mediterranean, DemoMed, and its tools, designed to simplify the reuse of spatial demographic data for the Mediterranean region, and the Aix-Marseille University data platform.

---

8  PROGEDO's mission is much broader, since it also involves taking charge of the French components of large international surveys as well as more generally supporting surveys in the social sciences and promoting all these data sets.

9  Note that if the data has unique identifiers within the portal, establishing a DOI identifier would constitute interesting progress, in particular to improve the citation of sources.

10  http://www.progedo.fr//app/uploads/2019/07/EnqueteQueteletProgedoDiffusion 2019.pdf.

11  Accessibility is not simply a question of physical or digital availability of an object or service, it also requires skills and know-how. Ensuring availability is the first level of accessibility, but true accessibility may require other features, such as taking into account the diversity of audiences. The most common is the question of disabilities, but it also involves issues related to the education or cultural context of users.

## The DemoMed portal: a multifaceted initiative targeting the Mediterranean region at Aix-Marseille University

The demographic observatory of the Mediterranean is a scientific project created in 2010 within the Maison Méditerranéenne des Sciences de l'Homme (MMSH – Mediterranean House for Human Science) at Aix-Marseille University. The multidisciplinary team of researchers at DemoMed strive to take spatial and temporal dimensions into account simultaneously in demographic analysis conducted on populations in the Mediterranean. This approach offers a new outlook on the interpretation of demographic indicators by revealing the particular spatial organisation of social phenomena. To achieve this, when it was created, the Observatory began a project to equip itself with a digital platform (http://demomed.org) in order to make the data from its research available.

This interface was developed with the aim of making the data accessible. From the outset, this notion of accessibility was designed to function without an intermediary between the user and the data. For this, software development was geared towards the simplest possible access but also taking into account the variety of possible users (Oliveau, Doignon and Blöss 2018). The query interface does not only allow searches via an interactive map at the Mediterranean level, but also through SQL type queries, translated into everyday language. The results of data searches are offered in different forms (tables, maps, graphs), at the user's choice and depending on the data. To simplify access to the data, the interface offers two main starting points. The first is a tool for obtaining data in the form of a table or graph. It leads the user to the data they are looking for (place, date, type of data, theme, phenomena, indicators, etc.) through a classic mode of successive selection via drop-down menus and check boxes. Mandatory items must be completed by the user, however it is possible to refine the search by entering other optional information.

The second way to access the data is a cartographic interface, which allows the database to be represented by exploring the Mediterranean region. The team worked hard to make the use of this cartographic visualisation tool as intuitive as possible for non-cartographers. The imported variables are enriched with as much metadata as possible (definitions, bibliographical references, various information). Moreover, it was also necessary for the interface to take into account the fact that administrative levels of Mediterranean countries are very diverse in terms of area and population. Like NUTS for the EU, administrative levels have been harmonized to improve the comparability of data at the Mediterranean level. The correspondence between administrative levels is based on a correspondence table, which makes it possible to know which level to choose in each country during international mapping, so that it is comparable to other levels chosen in other countries.

The idea guiding this choice of development was that of making the user autonomous in their search for data, whatever their initial level of knowledge, while also satisfying more experienced or demanding users. All the data is

documented by metadata to make it more easily reusable and the interface is intended, in the long term, to function in several languages. The greatest added value of the interface developed by DemoMed at Aix-Marseille University thus essentially lies in the expertise associated with the imported data.


BEYOND THE PORTALS: HUMANS

However, in practice, our team, comprised of teachers and researchers, found that computer mediation is not always sufficient to stimulate the reuse of data. In fact, while digital technology makes it possible to reach a large audience of potential users, it has greater difficulty engaging new audiences and supporting less experienced users. With this in mind, many local initiatives were carried out at a number of university sites before being coordinated and deployed on a national scale by the Large Research Infrastructure PROGEDO, through its University Data Platform (UDP)[12]. The Large Research Infrastructure PROGEDO effectively coordinates a network of research assistants and scientific advisor, located at the heart of universities (in Social Sciences and Humanities faculties). These skill platforms all constitute real skill supports for developing a community of quantitative data users. The UDP of Aix-Marseille University, which we belong to, was officially created in 2018 (http://pud.mmsh.univ-aix.fr/) and, like other UDPs, offers training activities as well as specific support for researching quantitative data in SSH in the Mediterranean.

The history of UDPs dates back to 2001. The idea of locally developing a support unit for researchers first appeared at the University of Lille[13]. This came about in 2003 (Duprez, Cros 2010). Supported by the University of Lille 1 since its creation, the Lille University Data Platform joined the Maison Européenne des Sciences de l'Homme et de la Société (European House for Social and Human Sciences) in 2015. It remained the only initiative of its kind until 2009, when a second platform was created in Lyon[14] on the initiative of the Lyon-Sainte-Etienne House for Human Sciences. Then came the UDPs in Nantes (named PROGEDO Loire) and Caen in 2015. Since 2016 in particular, and under the leadership of the Ministry from 2018 onwards[15], the network of UDPs has expanded to cover all major university sites. At the end of 2020

---

12  http://www.progedo.fr/promouvoir/plates-formes-universitaires-de-donnees/.

13  In an interview in 2009, Jean-Marie Duprez explains that he took the responsibility for it, but it was Emmanuel Lazega who came up with the idea: https://25images.msh-lse.fr/data_shs/video/intervention-de-jean-marie-duprez/fr.

14  The platform is called PANELS: https://www.msh-lse.fr/panels.

15  The Ministry of Higher Educatown, Research and Innovation has financed, under the "Vidal plan for SSH", volunteer universities to help the development of UDPs.

there were 14 UDPs in France: 12 hosted in Houses for Human Sciences outside of Paris and 2 in Paris University and Sciences Po.

These UDPs were created due to a strong need felt by local research teams, primarily to support researchers, but also to support students as they begin research work, i.e. in discovering and using the breadth of data available. More generally, there is a significant need for advice and training on the use of survey data or data from large databases. Faced with this situation, TGIR PROGEDO therefore took the initiative to stimulate the development of these UDPs, which form the necessary link between digital platforms and users. The initiative promotes the use of data from French, European and international research and public statistics (survey and administrative data, etc.) (Chenu, Lesnard 2011). This data is mostly quantitative. For example, the research assistant attached to Aix-Marseille University's UDP provides assistance in searching for data, accessing it, processing and exploiting it. An academic fellow is designated as scientific advisor for the UDP exercises support functions for all researchers, research professors, administrative and technical staff, doctoral students and master's students at Aix-Marseille University.

The additional aim of the Aix-Marseille UDP is to develop expertise on human and social science data for the Mediterranean region. In addition to their primary mission of personalised research support, the research assistant organises training, seminars, study days and summer schools. He visits various teaching and research institutions to publicise the platform and its missions. He also actively participates in exchanges and associated activities in the national network formed by all UDP. Research assistants recruited by the UDPs have a common profile: they have strong statistical skills and teaching motivations. They also must all have a very good knowledge of quantitative disciplines like demographics, statistics, geography, economics or quantitative sociology. This is to ensure that these new research support staff can share their experience in processing quantitative databases and statistical data processing methods with less experienced applicants. They must also demonstrate disciplinary openness to accommodate somewhat diverse requests for help from various backgrounds.

The extremely positive feedback from those using the services offered by the UDPs once again demonstrates that while IT has managed to replace humans in many tasks, there is still a need to maintain close human interaction in many situations. Information technology first supported researchers in their calculations, then in managing their data. Free of these constraints to a large extent, SSH researchers are now asking for support in the choice of available data and the methods for making the best use of it. Today, there is too much data and too many methods for a researcher to master them all. We can rejoice over this, since it allows greater possibilities. Nevertheless, it is all the more necessary to get support in this work, and the UDPs fulfil this mission. "The challenge of making statistical data digitally available would

therefore be to combine training in data sciences and digital sciences" (Blöss-Widmer 2019: 64).

CONCLUSION: SIGNIFICANT YET ESSENTIAL INVESTMENTS

In France, the social science research system is reaching maturity and is structured in the form of collective tools, which notably allow better management of the data produced. We are moving away from the individual observations and interpretations that characterised the system and heading towards a more shared organisation, comparable to that found in other sciences. This method of organisation has a cost, but it is essential for developing and maintaining databases.

One question remains: that of this work's sustainability. The increase in skills due to training by human means is sustainable. It allows those involved in SSH research to be better trained and better supported. This fully justifies investing in university data platforms, as well as in any other human support tool. However, we also know that these investments should not be reduced, at the risk of seeing these gains disappear. With incentives disappearing, researchers could turn away from these objects; technical tools must be maintained at a high level of service so as not to become obsolete; individual trajectories also lead to regular staff turnover.

On the other hand, data management, such as that of IT sites, has revealed itself to be less sustainable than expected. IT tools do indeed become obsolete, and the methods used require regular adaptation. Data formats are evolving, as are website technologies. It is estimated that the long-term lifespan of a website is around 5 years. Some data created ten years ago can no longer be read by current tools.. In addition, the development of documentation and metadata appears to be an absolute necessity to reuse this data (even for the researcher themselves). All of this is time-consuming and expensive. This raises the issue faced by libraries in the 20th century: that of selecting elements to preserve and deleting others. The stakes are high because statistical data "is central to research in the humanities and social sciences and for political decision-making" (Blöss-Widmer 2019: 64).

BIBLIOGRAPHY

Boeckhout M., Zielhuis G., Bredenoord A., 2018, "The FAIR guiding principles for data stewardship: fair enough?", *European Journal of Human Genetics*, vol. 26, no. 7, p. 931-36. URL: https://doi.org/10.1038/s41431-018-0160-0, consulted on 02/12/2020.

Blöss-Widmer I., 2019, "Accéder aux données statistiques en sciences sociales : l'apport des instruments de la révolution numérique", *in* T. Blöss, I. Blöss-Widmer (dir.), *Penser le vieillissement en Méditerranée. Données, processus et liens sociaux*, Paris: Karthala, p. 51-68.

Caporali A., Morisset A., Legleye, S., Richou C., 2015, "Providing Access to Quantitative Surveys for Social Research: The Example of INED", *Population*, vol. 70, no. 3, p. 567-97, URL: https://doi.org/10.3917/popu.1503.0567, consulted on 02/12/2020.

Chenu A., Lesnard L. (dir.) 2011, *La France dans les comparaisons internationales : guide d'accès aux grandes enquêtes statistiques en sciences sociales*, Paris: Presses de la Fondation nationale des sciences politiques.

Conchon M., 1993, "Une girafe est née : l'archivage des fichiers informatiques de l'INSEE aux Archives nationales", *Gazette des archives*, vol. 163, no. 1, p. 324-30. URL: https://doi.org/10.3406/gazar.1993.4215, consulted on 02/12/2020.

Cousijn H., Kenall A. *et al.*, 2018, "A Data Citation Roadmap for Scientific Publishers", *Scientific Data*, vol. 5, no. 180259, p. 11. URL: https://doi.org/10.1038/sdata.2018.259, consulted on 02/12/2020.

Duprez J-M, Cros M., 2010, "Accompagner étudiants et chercheurs dans l'exploitation des sources statistiques. L'expérience de la plateforme universitaire de données de Lille (PUDL)", *Statistique et Enseignement* 1 (1), p. 65-73.

Ministère de l'Enseignement supérieur et de la recherche 2008, *Stratégie nationale des infrastructures de recherche* URL: http://cache.media.enseignementsup-recherche.gouv.fr/file/Infrastructures_de_recherche/62/2/feuille_route_tgir_2008_527622.pdf consulted on 02/12/2020.

Ministère de l'Enseignement supérieur de la recherche et de l'innovation, 2018, *Plan national pour la science ouverte*. URL: http://cache.media.enseignementsup-recherche.gouv.fr/file/Actus/67/2/PLAN_NATIONAL_SCIENCE_OUVERTE_978672.pdf.

Oliveau S., 2017, "Le numérique et les SIG pour présenter et représenter la population", *in* E. Cavalié, F. Clavert, O. Legendre, D. Martin, *Expérimenter les humanités numériques*, Montréal : Les Presses de l'Université de Montréal, p. 145-158. URL: http://www.parcoursnumeriques-pum.ca/le-numerique-et-les-sig-pour-presenter-et-representer-la.

Oliveau S., Doignon Y., Blöss-Widmer I., 2018, "DemoMed – une cartographie interactive des populations en Méditerranée", *M@ppemonde*, vol. 123, URL: http://mappemonde.mgm.fr/123geov5/, consulted on 02/12/2020.

Silberman R., 1999, *Les Sciences sociales et leurs données : rapport au ministre de l'éducation nationale et de la technologie*, Rapport public. URL: http://www.ladocumentationfrancaise.fr/rapports-publics/004000935/index.shtml, consulted on 02/12/2020.

Wilkinson M., Dumontier M. *et al.,* 2016, "The FAIR Guiding Principles for scientific data management and stewardship", *Scientific Data*, vol. 3, no. 160018: 9. URL: https://doi.org/10.1038/sdata.2016.18, consulted on 02/12/2020.