Research data management in the French National Research Center (CNRS)

Joachim Schöpfel, Coline Ferrant, Francis Andre, Renaud Fabre

Research Data Management in the French National Center for Scientific Research (CNRS)

Joachim Schöpfel, Coline Ferrant, Francis André, Renaud Fabre

Abstract

Purpose: The paper presents empirical evidence on the opinion and behaviour of French scientists (senior management level) regarding research data management (RDM).

Approach: The results are part of a nationwide survey on scientific information and documentation with 432 directors of French public research laboratories conducted by the French Research Center CNRS in 2014.

Findings: The paper presents empirical results about data production (types), management (human resources, IT, funding, standards), data sharing and related needs, and highlights significant disciplinary differences. Also, it appears that RDM and data sharing is not directly correlated with commitment to open access. Regarding the FAIR data principles, the paper reveals that 68% of all laboratory directors affirm that their data production and management is compliant with at least one of the FAIR principles. But only 26% are compliant with at least three principles, and less than 7% are compliant with all four FAIR criteria, with laboratories in nuclear physics, SSH and earth sciences and astronomy being in advance of other disciplines, especially concerning the findability and the availability of their data output. The paper concludes with comments about research data service development and recommendations for an institutional RDM policy.

Originality: For the first time, a nationwide survey was conducted with the senior research management level from all scientific disciplines. Surveys on RDM usually assess individual data behaviours, skills and needs. This survey is different insofar as it addresses institutional and collective data practice. The respondents did not report on their own data behaviours and attitudes but were asked to provide information about their laboratory. The response rate was high (>30%), and the results provide good insight into the real support and uptake of research data management by senior research managers who provide both models (examples for good practice) and opinion leadership.

Introduction

In the era of open science, research data management (RDM) is an important though not new challenge for research performing organizations. Not exactly a clearly delimited concept, RDM is an umbrella term for activities related to the creation, organisation, structuring and naming of data; to their backup, storage, conservation and sharing, and to all actions that guarantee data security. It aims to "ensure reliable verification of results, and permits new and innovative research built on existing information" (Whyte & Tedds, 2011). Research data, as one part of scientific output, must be understood in a broad sense, as the "recorded factual material commonly accepted in the scientific community as necessary to validate research findings[1]". Sometimes, they are just generalized as "digital research output" (Pryor et al., 2014, p.VII). But research data are complex objects, dynamic, living, easier to describe than to define, with characteristics changing along with the research process

[1] OMB Circular 110, available at https://www.whitehouse.gov/omb/circulars_a110#36

(André, 2015). Commonly, the term covers laboratory data (spectrographic, genomic sequencing, electron microscopy data etc.), observational data (remote sensing, geospatial, socio-economic data etc.), audio-visual data, images, network-based data, plain or structured text, raw data, statistics, databases, software applications, structured graphics etc.; they are inherently collective and come in sets, as a collation of many individual data (Kowalczyk & Shankar, 2016).

As "researchers have shared data with their peers for centuries" (Klump, 2017), RDM is not a new task for large research performing organizations. These organizations are important data providers, especially because of their large and complex scientific instruments and projects (Large Hadron Collider, Hubble Space Telescope, Human Genome Project, magnetic resonance imaging etc.), and they have a long tradition of best practices in data management. As a result, their information professionals have developed more support activities for the RDM than academic librarians (Martin et al., 2017) where service development is still limited, focused especially on advisory and consultancy services rather than on technical services (Cox et al., 2017). What has changed, however, is the political environment. In the new European strategy towards open science[2], RDM occupies a central place. Open access (OA) to research results, data sharing whenever possible[3] and data management based on the FAIR principles[4] (Wilkinson et al., 2016) become the main objectives of scientific policy, which aim at increasing efficiency and transparency, societal impact and innovation capacity through rapid and unrestricted dissemination of research results – not only by large instruments but also from small scale projects and units. The 2017 European Open Science Cloud (EOSC) Declaration endorses that "All researchers in Europe must enjoy access to an open-by-default, efficient and cross-disciplinary research data environment supported by FAIR data principles"[5]. In this new Open Science policy, research performing organizations must take action: they have a crucial responsibility for research data stewardship and "should play a major role in supporting an open data culture" (The Royal Society, 2012).

In terms of scientific output (articles, citations) and innovation (patents), France is one of the leading Member States of the European Union. In 2015, French scientists published nearly 104,000 articles[6], and France spent 2.3% of its gross domestic product (GDP) on research and innovation (2014)[7]. The French National Center for Scientific Research (CNRS[8]) is the largest fundamental public research organization in Europe. It carries out research in all fields of knowledge, through its ten institutes (life sciences, chemistry, physics etc., figure 1) and 32,500 staff members in more than 1,000 research units (laboratories), most of them run in parallel with universities and/or other research organizations. In 2017, France joined the International Support and Coordination Office (ISCO) set up by Germany and the Netherlands to support the GO FAIR Initiative which aims "to gradually open up existing research data at scientific and academic institutions in all research fields and across national borders" – and is thus a stepping stone towards the realisation of the EOSC mentioned above[9].

[2] *Amsterdam Call for Action on Open Science*, available at http://english.eu2016.nl/documents/reports/2016/04/04/amsterdam-call-for-action-on-open-science

[3] "As open as possible, as closed as necessary."

[4] "All research objects should be findable, accessible, interoperable and reusable, both for machines and for people" (loc.cit.).

[5] http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

[6] Source: http://www.scimagojr.com/

[7] Source: http://www.oecd.org/

[8] Centre National de Recherche Scientifique http://www.cnrs.fr/

[9] https://www.government.nl/latest/news/2017/12/01/progress-towards-the-european-open-science-cloud

Acronym	Full name in French	Discipline(s)	Sample	%
IN2P3	Institut national de physique nucléaire et de physique des particules	Nuclear and particle physics	13	52%
INC	Institut de chimie	Chemistry	55	33%
INEE	Institut écologie et environnement	Ecology and environment	29	34%
INP	Institut de physique	Physics	31	34%
INS2I	Institut des sciences de l'information de leurs interactions	Informatics	33	52%
INSB	Institut des sciences biologiques	Biology	67	27%
INSHS	Institut des sciences humaines et sociales	Social sciences and humanities (with STI)	105	35%
INSIS	Institut des sciences de l'ingénierie et des systèmes	Engineering and systems	42	31%
INSMI	Institut national des sciences mathématiques et de leurs interactions	Mathematics	23	32%
INSU	Institut national des sciences de l'univers	Earth sciences and astronomy	34	33%

Figure 1: The CNRS research institutes (with sample size and representativeness in %)[10]

One of the first signatories of the Berlin Declaration on Open Access, the CNRS is deeply committed to OA. Also, the CNRS supports national and international initiatives, projects and infrastructures fostering OA to research results, with a clear preference for self-archiving of publications and data[11] in open repositories (green road). Recently, the CNRS has confirmed its attachment to the values of Open Science, considering research data as "common goods" that should be shared with the scientific community whenever possible (DIST-CNRS, 2016). But is this enough? Does the CNRS scientific information policy meet the needs and expectations of the scientists? Are its information infrastructures and services in line with the scientists' information and data behaviours? Between 2014 and 2016, the CNRS conducted an internal audit on its information services and policy, including surveys and interviews with scientists, research managers and information professionals from the CNRS, other research organizations and the Higher Education.

One part of the results – a survey on behaviours, needs and attitudes towards open access - has already been published, revealing globally positive opinions towards open access in general and open repositories (green road) in particular and confirming some significant differences between disciplines (Schöpfel et al., 2016). This paper provides additional insight into the field of RDM. Based on a survey with a representative sample of more than 400 laboratory directors, the paper produces empirical elements for a better understanding of data production, curation and preservation, and in particular of attitudes towards data sharing with other scientists. The results are discussed in terms of open data culture, FAIR principles and service development. Because of its intrinsic internationality, the CNRS must comply with the new European Open Science policy; given its central position in the French public research landscape, the CNRS must contribute to a common approach to national RDM in France.

In order to facilitate and foster the deposit of European projects, the CNRS started to develop new services to inform, train and assist the scientists in the field of RDM, via tutorials on open data, DMP and data sharing, a web-based tool to write DMPs with templates and guidance (developed with the

[10] The complete list of the CNRS research laboratories can be consulted at the following address: http://www.cnrs.fr/fr/recherche/labos.htm

[11] The international re3data.org directory contains at least 23 data repositories funded or co-funded by the CNRS

UK Digital Curation Centre)[12], an online helpdesk with an expert network and other customizable services (repositories, TDM…)[13]. Moreover, the CNRS is, via its STI unit INIST[14], the French partner of the DataCite consortium for the DOI assignment to research data[15]. Together with other Higher Education institutions, the CNRS runs HAL, the national repository open for data deposits, and hosts the SSH infrastructure Huma-Num with the NAKALA platform[16] for RDM in the humanities. The CNRS standing committee on research ethics published guidelines on ethical issues of data sharing (COMETS, 2015).

This is ongoing investment, work in progress, and the survey was conducted to guide the CNRS management in further research and development in the field of RDM, to meet new requirements from funding bodies and to improve the excellence of French public research. As with the other institutions and organisations (Aydinoglu et al., 2017; Barsky et al., 2017), the results of this survey will assist the CNRS in making evidence-based decisions about what expertise and which services will be needed to support the laboratories in improving their data management practises.

Methodology

The survey was conducted between July and September 2014 by the CNRS Scientific and Technical Information (STI) Department (DIST)[17]. A questionnaire with 91 items was sent to the directors of the 1,250 CNRS research laboratories representing the whole range of fundamental science. The survey was a component of an internal audit on the CNRS STI policy and service development. Part of the demand analysis, the items' objective was to assess attitudes and needs expressed by research managers regarding four particular functions of scientific information: access to scientific information, publishing of research results, analysis of scientific information (scientometrics), and other research support services, including ethics and legal advice.

432 laboratory directors completed the questionnaire (35%). The respondents are a representative sample of the CNRS research institutes (social sciences and humanities [SSH], life sciences, chemistry, engineering and systems sciences etc.) and of the geographical distribution (Paris, regions).

34 items dealt with RDM, covering production, management and sharing. The raw results were published in March 2015[18]. This paper presents a re-analysis to explore RDM, to evaluate specific needs and demands, and to analyse the differences between scientific disciplines. The findings will be discussed under three different aspects:

- What do scientists think about data sharing and openness? What can be said about their open data culture?
- To what degree are their data behaviours and attitudes supportive of the FAIR principles of RDM?
- Which are the priorities for RDM, and which kind of RDM services do they ask for?

Findings

432 laboratory directors (senior managers) completed the questionnaire. No question was mandatory. The response rates per question range from 0.25 to 0.94 (median 0.84); those questions with lower response rates (<0.5) were sub-questions conditioned by another question.

[12] DMP OPIDoR https://dmp.opidor.fr
[13] DoRANum http://www.doranum.fr/
[14] Institut de l'Information Scientifique et Technique http://www.inist.fr/
[15] http://www.inist.fr/?DOI-Assignment&lang=en
[16] http://www.huma-num.fr/service/nakala
[17] Direction de l'Information Scientifique et Technique http://www.cnrs.fr/dist/
[18] http://www.cnrs.fr/dist/z-outils/documents/Enquête%20DU%20-%20DIST%20mars%202015.pdf

Data production

About 85% of the respondents provided more details on the typology of the research data their laboratories produce and process. For a large and multidisciplinary research organization like the CNRS it is not surprising that the answers cover the whole range of data categories. Figure 2 shows the main primary data types.

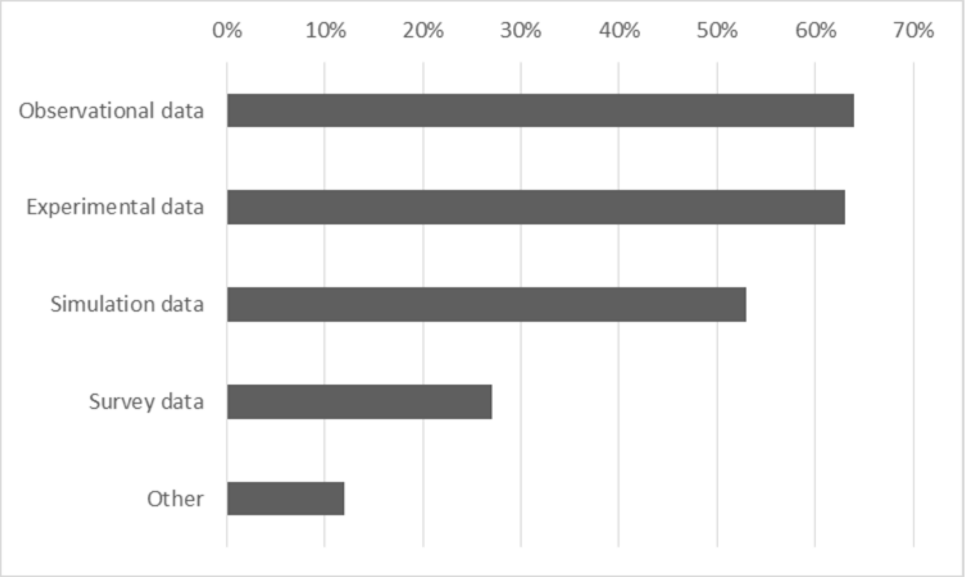


Figure 2: Primary data types, in % (N=367)

Often, a laboratory processes two or more types of data. Most of these data are collected as figures, numbers or statistics (82%) and images (72%), followed by text (59%) and video files (34%); only few data are in audio (sound) files (4%) or another format (6%). 47% of the respondents affirm that their data formats are fully or partly interoperable, but nearly as many (42%) admit that they do not know exactly if the format is open and interoperable or not.

Approximatively half of the laboratories (48%) produce their databases together with other research structures, often with funding from the French National Research Agency ANR[19] (67%) and/or from the European Commission (45%).

Data management

61% of the directors declared that their laboratories' data output need specific RDM. But only one third of them have some kind of tools for monitoring data production, and even less have already established a data management plan (DMP). What can be said about specific resources for RDM?

Human resources: About one third of the laboratories (38%) have specific staff at their disposal, dedicated to RDM. Most of them are permanent staff but more than half of these units (also) hire staff for a limited period to do (or help doing) the job. What is that job? Mainly reformatting and standardization of data (60%), data processing and creation of secondary data (57%) and database production (57%); much less data curation, including metadata (9%). How good are they doing their job? Overall, 75% of senior managers evaluate the RDM skills of their staff as basic (37%), good (30%) or excellent (8%). In four domains (curation, referencing, data security, ethics and law) they rate their staff skills higher, compared to DMPs or data publishing where nearly 40% consider the staff skills as insufficient.

[19] Agence Nationale de Recherche

Financial resources: Only one out of five laboratories (22%) receive specific subsidies for RDM. Most of the time these subsidies are part of a project but in 60% they are (also) recurrent funding.

IT resources: Less than half of the laboratories (46%) use institutional (= nationwide) infrastructures for RDM. Most of the time, their IT resources are local (77%) and/or personal (77%), such as local servers, personal computers and so on. One part of the laboratories have specific software for RDM, because of their instruments or experimental devices; 28% of these information technology (IT) systems are interoperable, while for the other 72% the format is either proprietary or unknown. The main RDM software are spreadsheets (73%) and/or database systems (57%), used for the acquisition, processing and sharing of research data.

Standards: Half of the respondents (52%) answered the question whether they applied standardized procedures for RDM. Most of the respondents did so, especially using standard data formats (86%), describing data in a standardized way (58%) and/or applying standard methods for the data collection (50%). On the other hand, the terminology for curation and indexing is less standardized (40%), and only 30% assign standard permanent identifiers (PID) to their datasets, such as the Digital Object Identifier (DOI).

Data sharing

40% of the laboratory directors state that their research data are published online, often with access restrictions (access on demand or limited to authorized users); only 17% report that their data are freely disseminated on the web, in OA (figure 3).

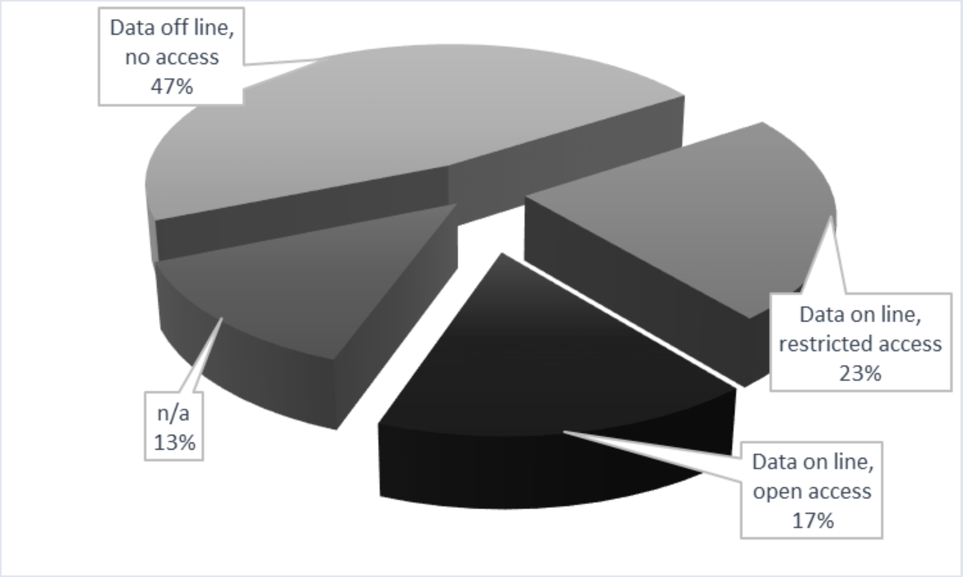


Figure 3: Data sharing (N=432)

The respondents mentioned some data repositories funded or co-funded by the CNRS with data produced by CNRS laboratories, e.g. the *Plasma Physics Data Center* (CDPP)[20] for natural plasmas of the solar system; the *Global Emissions Initiative* (GEIA) database[21] with datasets of surface emissions of atmospheric compounds, and ancillary data, i.e. data required to estimate or quantify surface emissions; the *Open Resources and Tools for Language* (Ortolang)[22] with language data (corpora,

[20] http://cdpp.eu/
[21] http://eccad.sedoo.fr
[22] https://www.ortolang.fr

lexicons, dictionaries etc.); and the French data archives for humanities and social sciences *Réseau Quételet*[23].

Needs

In the field of RDM, collaboration and mutual assistance in the community can be helpful. 59% of the respondents confirm that their laboratories collaborate with other scientists and research units, through shared data tools (84%), workshops (47%), common guidelines (44%) and training sessions (41%).

Yet, this may not be enough. Two thirds of the laboratory directors (65%) are interested in specific information services related to RDM. 45% would need specific tools for the monitoring of their data production, and 50% would like to get online technical assistance and support, such as platforms with RDM tools (81%), user (discussion) forums (52%) or a hot line (44%).

Disciplinarity

Each laboratory is part of one of ten disciplinary CNRS research institutes (see figure 1). Thus, it was possible to compare the survey results among these institutes. Even if the figures must be interpreted with caution, because of the small subsample numbers and also because of undeniable differences between laboratories of the same institute (different size, different instruments and research fields etc.), some general remarks are possible on particular RDM patterns. The following survey results are statistically significant at the 0.001 level (X^2 test[24]).

Awareness: Obviously, the issue of RDM does not have the same relevance and actuality in all research institutes. According to the response rates and patterns, RDM is an issue especially in physics, nuclear and particle physics and in SSH, more than for instance laboratories in chemistry, biology or engineering sciences. This does not mean, of course, that these structures produce less or no data; however, at least in this survey, their senior managers appear less concerned with the issue.

Resources: Dedicated staff (figure 4), specific funding and RDM software are mainly reported from nuclear physics, earth sciences and astronomy and SSH and, to a lesser extent, from ecology and informatics. The same disciplines, together with biology, seem more engaged in collaborative RDM.

Institute	*Sample*	Data staff	No data staff
IN2P3	*12*	58%	42%
INC	*52*	19%	81%
INEE	*28*	43%	57%
INP	*25*	32%	68%
INS2I	*23*	22%	78%
INSB	*63*	29%	71%
INSHS	*93*	57%	43%
INSIS	*36*	22%	78%
INSMI	*17*	12%	88%
INSU	*33*	61%	39%
Total	***382***	**37%**	**63%**

Figure 4: Dedicated staff for RDM in research laboratories (N=382, p<.001)

[23] http://www.reseau-quetelet.cnrs.fr/

[24] Pearson's Chi-Square Test has been performed for all survey questions and research institutes which determine the dominant discipline for each laboratory. Only the most significant differences between institutes are reported here.

Availability: More than half of the laboratory directors in earth sciences and astronomy, informatics, SSH, ecology and nuclear physics declare that their data are available online. This is not the case especially in three other disciplines, i.e. chemistry, physics and mathematics (figure 5).

Institute	*Sample*	Offline	Online	*Online restricted*	*Online open*
IN2P3	*12*	50%	50%	*100%*	*0%*
INC	*51*	71%	29%	*73%*	*27%*
INEE	*27*	48%	52%	*81%*	*19%*
INP	*25*	76%	24%	*89%*	*11%*
INS2I	*22*	41%	59%	*53%*	*47%*
INSB	*62*	56%	44%	*70%*	*30%*
INSHS	*94*	45%	55%	*54%*	*46%*
INSIS	*36*	67%	33%	*57%*	*43%*
INSMI	*13*	92%	8%	*100%*	*0%*
INSU	*32*	19%	81%	*52%*	*48%*
Total	***374***	**54%**	**46%**	***64%***	***36%***

Figure 5: Disciplinary differences of data availability (N=374, p<.001)

Legal issues: Questions on the reuse of research data are reported from nuclear physics laboratories, while privacy issues and questions related to the open data policy (unrestricted dissemination) arouse mainly in the SSH research institute.

On a very general level, we can distinguish three groups: (1) laboratories from nuclear and particle physics and from social sciences and humanities appear globally more advanced regarding RDM than other disciplines; (2) laboratories from the three domains ecology and environment, informatics and earth sciences and astronomy have dedicated resources and make their data available; (3) laboratories in the field of physics appear aware of the challenge.

Discussion

Surveys on RDM usually assess individual data behaviours, skills and needs (e.g. Reilly et al., 2011, Simukovic et al., 2014, Bauer et al., 2015). This survey is different insofar as it addresses institutional and collective data practice. The respondents did not report on their own data behaviours and attitudes but were asked to provide information about their laboratories. So we must be careful when comparing our results with those from other surveys.

On sharing of data and publications

Open Science policy defines openness as OA to scientific results in general, i.e. to publications as well as to research data. As mentioned before, large availability of research data is a crucial element of Open Science, especially to support innovation, transparency and citizen science. But is data and publication sharing the same behaviour? For instance, do the laboratories deposit their publications on the national OA repository HAL[25] and do they make their research data accessible on the Web? Figure 6 shows that the answer is globally negative, apparently there is no strong relationship between both behaviours; but it also shows that there are some differences between disciplines.

Laboratories in mathematics, nuclear physics and computer sciences are generally highly committed to OA publishing on the HAL repository, more than those from chemistry, ecology or biology. But

[25] Hyper Article en Ligne https://hal.archives-ouvertes.fr/

compared to nuclear physics and computer sciences, and even to biology or ecology, the laboratories in mathematics appear less advanced and engaged in making their data available on the web. Another example are research units from earth sciences and astronomy: while they make more effort than others to make their data available, they seem less committed to green OA via HAL. However, more investigation is needed to understand the reasons for these variations.

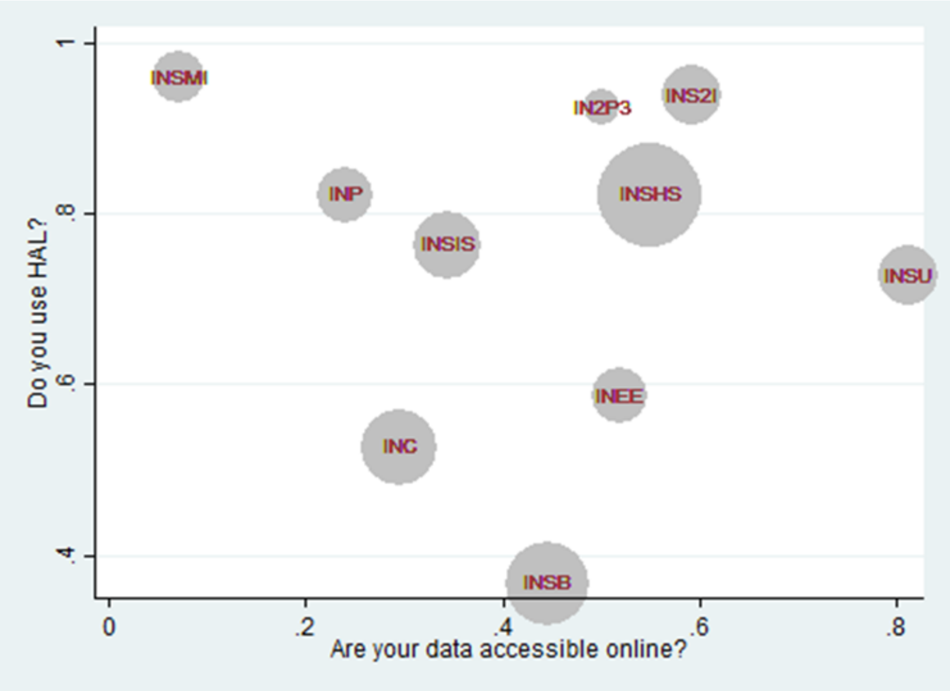


Figure 6: OA publishing and data sharing (N= 371)

The survey revealed that a majority of respondents – 50-70% - are generally supportive of OA and declare actual usage of the French national HAL repository, including the depositing of metadata (records) and documents (full text) while only a small group seem not to be interested in green or gold OA and reluctant to self-archiving and OA publishing (Schöpfel et al. 2016, p.147). Obviously, the impediments to data sharing are more significant, and it is difficult to assess if Open Science policy will change this behaviour, especially as a large majority of these scientists are opposed to mandatory policies. Even if data sharing behaviour is increasing with more favourable views on the practice and the overall movement, this evolution remains a "complex shift, with varying cultures among scientists" (Tenopir et al., 2015).

One example: in a recent study on the Lille SSH campus, about 40% of the respondents expressed a positive opinion about data sharing (Schöpfel & Prost, 2016). But 30% admitted that they were not aware of this possibility and nearly as many (29%) clearly said that they did not share their data in the past and will not do so in the future, for different reasons, e.g. sensitive and confidential data, risk of plagiarism, workload, data illegibility and intellectual property. A survey with Austrian scientists revealed that data archives or repositories are used by less than 15% (Bauer et al., 2015). Also, data sharing and data reuse are largely separate phenomena (Curty et al., 2017), and there are still "perceived risks and barriers that may be slowing the data sharing movement" (Tenopir et al., 2015). Perhaps it is more realistic to speak of "qualified openness" and to acknowledge "legitimate boundaries of openness which must be maintained in order to protect commercial value, privacy, safety and security" (The Royal Society, 2012). The CNRS senior research managers probably comply with the EC open data culture defined as "as open as possible, as closed as necessary". Yet, there should be more institutional guidance about the meaning of “open”, “closed” and “necessary” to frame and inform local and personal decisions on data sharing.

FAIR-ness of research data management

Among the drivers of RDM, such as preservation and research governance, funders’ requirements appear to be encouraging greater engagement with RDM and openness, and “in many HEIs (data sharing) is primarily seen in terms of research funder requirements” (Higman & Pinfield, 2015). Today, governments, research organizations and funding bodies have started to adopt the so-called “FAIR data principles” (Wilkinson et al., 2016), i.e. a set of guiding principles to make data findable, accessible, interoperable, and re-usable.[26] As the 2017 EOSC declaration reminds, “FAIR principles are neither standards nor practices”; they describe four dimensions to assess different levels of FAIR-ness of technical solutions (repositories), workflows, governance etc. They describe “characteristics and aspirations for systems and services to support the creation of valuable research outputs that could then be rigorously evaluated and extensively reused (…) FAIR is not just about humans being able to find, access, reformat and finally reuse data (…) The recognition that computers must be capable of accessing a data publication autonomously, unaided by their human operators, is core to the FAIR Principles” (Mons et al., 2017).

Yet, on a behavioural level, data practices are part of the data culture, and they can be supportive of data FAIR-ness. Therefore we tried to assess if and how the surveyed data behaviours were compliant with these principles. In other words, we wanted to know even if the survey was not specifically designed to assess data services or tools, if its results provide information about skills and practice that may contribute to and increase findability, availability, interoperability and reuse of research data and be helpful, as part of the data culture, for the implementation and transition to FAIR.

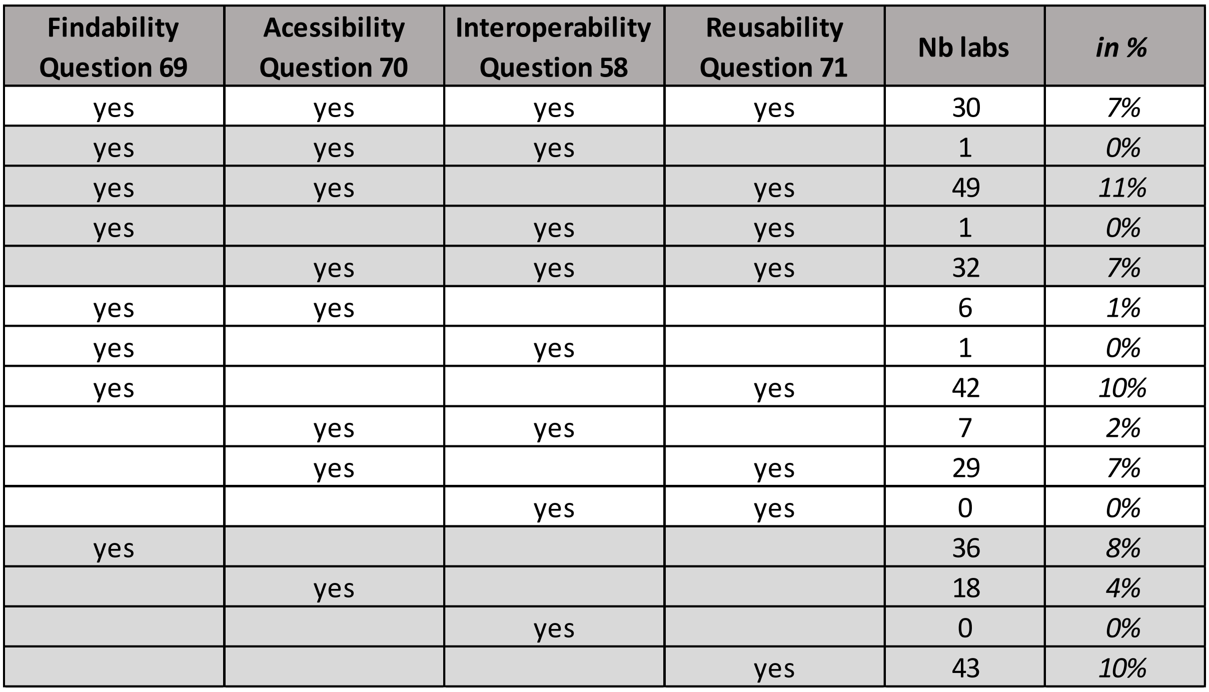
Findable: According to the FAIR guidelines, rich metadata, unique persistent identifiers and searchable resources are necessary to make data findable. In the CNRS survey, 43% of the respondents think that their staff has good or excellent skills related to metadata and persistent identifiers. And 46% report that their data are available online. This means that nearly half of the research units exhibit some skills and practice which could be interpreted as contribution to the “F-principle”.

Available: To increase availability, data should be retrievable by their identifier using a standardized, open protocol which allows, if necessary, for an authentication procedure. In the survey, 53% of the respondents said that their data are available - 19% describe their data as online and open, i.e. freely available, while 34% say that their data are online but available only to authorized users (restricted access). Both access modes appear to reflect a data sharing approach that seems more or less compliant with the “A-principle”. However, this does not necessarily mean that all laboratories apply standardized, open protocols and identifiers.

Interoperable: Nearly half of the respondents (47%) answered that their data were produced and processed in interoperable formats, completely (25%) or partly (22%), and 28% have interoperable, non-proprietary software for their scientific instruments or experimentations. This means that at the time of the survey (2014), less than half of all laboratories were able to produce data that could meet the requirements of the “I-principle”.

[26] For more details see also https://www.force11.org/group/fairgroup/fairprinciples and https://www.dtls.nl/fair-data/go-fair/

Reusable: Reusability, the “ease of using data for legitimate scientific research by one or more communities of research that is produced by other communities of research” (Thanos, 2016) is a multidimensional concept, including legal issues (licensing) and standards. In the CNRS sample, as mentioned above, a more or less important part declares using standards especially for the data format (86%), description (documentation) (58%) and acquisition methods (50%), less for PIDs (30%) and controlled terminology (40%). On the other hand, only 39% evaluate the legal skills of their staff as satisfying, including licensing. With regards to the “R-principle”, these results show a contrasted landscape, with some positive aspects (format) but much progress needed for others (identifiers, legal clearance).



Findability Question 69	Acessibility Question 70	Interoperability Question 58	Reusability Question 71	Nb labs	*in %*
yes	yes	yes	yes	30	*7%*
yes	yes	yes		1	*0%*
yes	yes		yes	49	*11%*
yes		yes	yes	1	*0%*
	yes	yes	yes	32	*7%*
yes	yes			6	*1%*
yes		yes		1	*0%*
yes			yes	42	*10%*
	yes	yes		7	*2%*
	yes		yes	29	*7%*
		yes	yes	0	*0%*
yes				36	*8%*
	yes			18	*4%*
		yes		0	*0%*
			yes	43	*10%*

Figure 7: Levels of FAIR-ness in the sample (explanation in the text)

Even if this survey was not designed as a detailed diagnosis of compliance with the FAIR guiding principles of RDM, some questions help to shed light on supportive practice that contributes to these principles. Figure 7 presents response patterns based on four questions related to

- Findability (#69 online publishing of data?),
- Availability (#70 open access to research data?),
- Interoperability (#58 interoperable data formats?) and
- Reusability of research data (#71 standardized and community-specific procedures?).

Three observations: (1) Only 7% of the laboratory directors confirm that their data practice is compliant with these four criteria: they publish data online, they make at least some of them freely available, they apply interoperable data formats, and they use standardized and community-specific procedures. Of course, this does not mean that their data management tools and infrastructures are FAIR; however, it indicates that they show some data practice that is supportive and helpful for the development and implementation of those technologies. Another 18% answered “Yes” to three of the four questions. The “weak point” appears to be the application of interoperable data formats.

(2) 42% answered in a way that indicates compliance with one (22%) or two (20%) criteria. These criteria are often standards and community-based practice (#71) and, to a lesser degree, online publishing of data (#69).

(3) 32% replied with “No” to all four questions, or did not reply at all. In other words, only two third of the laboratories are to some extent compliant with at least one of these four criteria, in other words, are aware and receptive for data FAIR-ness.

Last comment: this survey was not designed to assess FAIR principles and our results only provide a picture of the situation taken at a given moment and under a specific perspective. These four questions deal with aspects related and supportive to the FAIR guiding principles. However they are neither representative nor exhaustive, they only concern practice and not tools or infrastructures, and in spite of negative answers to these four questions a laboratory may very well contribute in other ways to data FAIR-ness.

Priorities and service development

The survey was designed to provide helpful information for the future development of RDM services. Most studies share the conviction that RDM service development should be “bottom-up”, i.e. it must build on the needs, demands and behaviours of the scientists. Often, RDM is described as a “small-scale service” (Knight, 2015), driven by funder obligations, focused on the research team level (for instance, with embedded research data managers, cf. Schmidt & Dierkes, 2015), and compliant with the requirements that result from the field of application and the data that has to be managed (Curdt & Hoffmeister, 2015) but compliant, too, with the diversity of research methodologies – “data management strategies for a single project may have to include a host of different software packages and file types. Skill development in analytical tools has to be offered broadly; more than just the geographers are working with geospatial data” (Weller & Gulick, 2014).

The specific challenge for a large research performing organization like the CNRS is to satisfy the whole range of needs and requirements, to provide resources (staffing, budget, tools) necessary for local (“bottom”) RDM as well as infrastructures, shared services, and communities “to work collectively on problems” (Awre et al., 2015), on the institutional (disciplinary) and national (multidisciplinary, intra- and inter-organizational) levels (Humphrey et al., 2016). The UK Joint Information Systems Committee (JISC) identified five key areas for action on local and national levels (Brown et al., 2015) which are more or less compliant with the measures recommended by Bauer et al. (2015) for the handling of research data in Austria:

- Policy development and implementation
- Skills and capabilities
- Infrastructure and interoperability
- Incentives for scientists and support stakeholders
- Business case and sustainability

Bauer et al. (2015) insist on specific disciplinary needs, on hiring of information professionals and on the implementation of support services for scientists. So, where do the CNRS research managers set their RDM priorities? On a general level, they are preoccupied more with for instance RDM than with publishing, legal issues, statistics or IT but less than with open access and publication management. RDM, in other words, is clearly identified as an important challenge but it is not at the top of their agenda. They seem less interested in tools for monitoring of data production than in technical assistance especially for the curation and preservation of research data. In particular, they seem concerned with knowledge and skills gaps in their laboratories regarding research data plans, sensitive data and data publishing. Like the Austrian scientists (Bauer et al., 2015), they are interested in technical infrastructure and project-specific support for research data, in legal advice, a general help desk, as well as in training programs; and they probably expect the provision of additional qualified staff, as well as the adoption of guidelines or policies for dealing with research data.

These results are convergent as they confirm a great need for support and basic assistance; in many respects, RDM is not just (only) a technical problem but a “people problem”, e.g. guidance, training, and support (Ward et al., 2011). “Improving tools are not the only steps necessary to overcome barriers. The next steps will likely involve training for scientists, or the ready availability of well-trained data managers to assist with the extra tasks required to describe and share data” (Tenopir et al., 2015). In some situations the best solution may be recruitment of full-time staff with the necessary expertise to work with scientists and develop RDM resources (Knight, 2015); elsewhere the preferred option may be out-sourcing, like in the Yale Open Data Access (YODA) Project’s “trusted intermediary” approach in which an independent partner provides support, accountability, fairness, and transparency (Krumholz & Waldstreicher, 2016).

All these recommendations and initiatives infer additional investment. RDM cannot be done by simple reallocation of existing resources. Key factors of success will be funding requirements (external) and a strong management support (internal), especially because RDM policy will not only introduce new tools and procedures but will also improve existing research practice, from the beginning on of the whole research cycle. Here, again, the crucial role of senior scientists such as the CNRS laboratory directors is evident, especially for the provision of incentives to the scientists and for the development of RDM related policies, tools and practice.

Conclusion

In a recent survey, the development of data sharing and data reuse practices has been described as a “complex shift, with varying cultures among scientists” (Tenopir et al., 2015). “Complex shift” seems an appropriate term to describe the results of the CNRS survey. As a large, multidisciplinary research performing organization, the CNRS has to cope with a complex RDM landscape, with important differences of values and practices (“culture”), tools and skills between laboratories and institutes, and with many different stakeholders, e.g. industry, funders, scientists, citizens, politicians, technical staff, librarians etc., each one with different and sometimes opposed interests. Many laboratories and research teams have more or less experience with RDM, with dedicated staff, software, procedures, budget and cooperation; others are just at the beginning. Also, the situation is anything but static, and international partnerships, national and international funding bodies, technological development and research policy introduce an irresistible dynamic on RDM.

What can be done? What approach should be adopted? There is no evidence, no pattern to follow, and the problem of RDM has already been described as a “wicked problem”, evading easy answers, perhaps even insoluble, at least temporary (Awre et al., 2015). However, as said above, research organizations are expected to play a major role in supporting an open data culture, especially in the new European strategy towards open science which recommends, among others, that they should:

- put in place an institutional data policy that clarifies institutional roles and responsibilities for RDM,
- develop and adopt citation principles for data that include persistent identifiers,
- think actively about what to share and what not to share,
- develop and set standards on privacy,

- and set up and manage local and national e-infrastructures and assist scientists in the selection and use of services[27].

As mentioned at the beginning, the CNRS started to develop a nationwide RDM policy. Our survey provides some guiding principles for the further development of this policy, above all a discipline- and instrument-centred approach, a focus on new requirements from funding bodies (FAIR principles), a coordinated development of infrastructures, RDM tools and training opportunities, and sufficient funding of local and institutional initiatives. Data policy should not make sharing a priority but rather focus on good practice in RDM, compliant with the requirements from the European Commission and other funders.

The senior research managers play a key role in the definition of an institutional RDM strategy (bottom-up) as well as in the implementation of this policy on the local level (top-down). All surveys and case studies confirm that their contribution and support are decisive for the development of a new open data culture in the research communities. To meet the data challenge in good conditions, they need infrastructures (especially for long-term preservation[28]), support from information services[29] and sustainability, e.g. business models and defined responsibilities for maintaining data beyond project-based funding (Knowledge Exchange, 2016). Thus the realistic vision – and the challenge - of an institutional approach to RDM is probably a layered, component-based infrastructure with complementary support functions at various levels and with various types of data services, flexible and compliant with various situations, integrated in the national and international infrastructures, and embedded in an explicit and efficient organizational data strategy (policy), management (coordination) and follow-up (monitoring).

Bibliography

André, F. (2015), "Déluge des données de la recherche? ", in Calderan, L. et al. (Ed.), *Big data: nouvelles partitions de l'information*, De Boeck, Louvain-la-Neuve, pp. 77-95.

Awre, C. et al. (2015), "Research data management as a 'wicked problem'", *Library Review,* Vol. 64 No. 4/5, pp. 356-371.

Aydinoglu, A.U., Dogan, G., Taskin, Z. (2017), "Research Data Management in Turkey: Perceptions and Practices", *Library Hi Tech*, Vol. 35 No. 2.

Barsky, E., Adam, S., Farrar, P., Meredith-Lobay, M., Mitchell, M., Naslund, J-A., Sylka, C. (2017), *Research Data Management Survey, UBC: Humanities and Social Sciences*, Report, Advanced Research Computing, University of British Columbia. http://hdl.handle.net/2429/60639

Bauer, B. et al. (2015), *Researchers and their data. Results of an Austrian survey-report 2015*, e-infrastructures austria, Vienna. http://e-infrastructures.at/das-projekt/deliverables/

Brown, S., Bruce, R. and Kernohan, D. (2015), *Directions for research data management in UK universities*, JISC, Bristol. http://repository.jisc.ac.uk/5951/4/JR0034_RDM_report_200315_v5.pdf

Burgi, P-Y., Blumer, E., Makhlouf-Shabou, B. (2017), "Research data management in Switzerland. National efforts to guarantee the sustainability of research outputs", *IFLA Journal*, Vol. 43 No. 1, pp. 5-21.

[27] *Amsterdam Call for Action on Open Science*, loc.cit.

[28] This is consistent with an ongoing data life cycle management project in Switzerland which puts a major focus on long-term preservation (Burgi et al., 2017).

[29] See the LIBER survey on library-based data services (Tenopir et al. 2016) and the recommendations for research libraries from LIBER (2012) and CLIR (2013)

CLIR (2013), *Research data management: Principles, practices, and prospects*, Council on Library and Information Resources, Washington D.C. https://www.clir.org/pubs/reports/pub160

CNRS-DIST (2016), *Livre blanc : une science ouverte dans une république numérique*, CNRS Direction de l'Information Scientifique et Technique, Paris. http://books.openedition.org/oep/1548

COMETS (2015), *The ethical challenges of the sharing of scientific data*, Comité éthique du CNRS, Paris. http://www.cnrs.fr/comets/IMG/pdf/comets-partagedesdonneesscientifiques-en-2.pdf

Cox, A.M., Kennan, M.A., Lyon, Pinfield, S. (2017), "Developments in research data management in academic libraries: Towards an understanding of research data service maturity", *Journal of the Association for Information Science and Technology*, preprint. http://eprints.whiterose.ac.uk/101389/

Curdt, C. and Hoffmeister, D. (2015), "Research data management services for a multidisciplinary, collaborative research project", *Program,* Vol. 49 No. 4, pp. 494-512.

Curty, R.G., Crowston, K., Specht, A., Grant, B., Dalton, E.W. (2017), *Attitudes and norms affecting scientists' data reuse*, preprint. https://crowston.syr.edu/node/666

Higman, R. and Pinfield, S. (2015), "Research data management and openness", *Program*, Vol. 49 No. 4, pp. 364-381.

Humphrey, C., Shearer, K., Whitehead, M. (2016), "Towards a Collaborative National Research Data Management Network", *International Journal of Digital Curation*, Vol. 11 No. 1.

Klump, J. (2017), "Data as Social Capital and the Gift Culture in Research", *Data Science Journal*, Vol. 16 No. 14, pp. 1-8.

Knight, G. (2015), "Building a research data management service for the London School of Hygiene & Tropical Medicine", *Program*, Vol. 49 No. 4, pp. 424-439.

Knowledge Exchange Research Data Expert Group and Science Europe Working Group on Research Data (2016), *Funding research data management and related infrastructures*, Briefing Paper, Science Europe – Knowledge Exchange, Bristol. http://www.scienceeurope.org/wp-content/uploads/2016/05/SE-KE_Briefing_Paper_Funding_RDM.pdf

Kowalczyk, S. and Shankar, K. (2011), "Data sharing in the sciences", *Annual Review of Information Science and Technology*, Vol. 45 No. 1, pp. 247-294.

Krumholz, H. M. and Waldstreicher, J. (2016), "The Yale Open Data Access (YODA) project — a mechanism for data sharing", *New England Journal of Medicine,* Vol. 375 No. 5, pp. 403-405.

LIBER working group on E-Science / Research Data Management (2012), *Ten recommendations for libraries to get started with research data management*, LIBER Association of European Research Libraries, The Hague. http://libereurope.eu/wp-content/uploads/The%20research%20data%20group%202012%20v7%20final.pdf

Martin, C., Cadiou, C. Jannes-Ober, E. (2017), "Data Management: New Tools, New Organization, and New Skills in a French Research Institute", *LIBER Quaterly*, Vol. 27 No. 1, pp. 73-88.

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O., Wilkinson, M. D. (2017), "Cloudy, increasingly FAIR; revisiting the FAIR data guiding principles for the European Open Science Cloud", *Information Services & Use*, Vol. 37 No. 1, pp. 49-56.

Pryor, G., Jones, S. and Whyte, A. (Eds.) (2014), *Delivering research data management services: fundamentals of good practice*, Facet, London.

Reilly, S., Schallier, W., Schrimpf, S., Smit, E. and Wilkinson, M. (2011), *Report on integration of data and publications*, ODE Opportunities for Data Exchange, The Hague. http://www.stm-assoc.org/2011_12_5_ODE_Report_On_Integration_of_Data_and_Publications.pdf

The Royal Society (2012), *Science as an open enterprise. Summary report*, The Royal Society Science Policy Centre, London. https://royalsociety.org/~/media/policy/projects/sape/2012-06-20-saoe-summary.pdf

Schmidt, B. and Dierkes, J. (2015), "New alliances for research and teaching support: establishing the Göttingen eResearch Alliance", *Program*, Vol. 49 No. 4, pp. 461-474.

Schöpfel, J., Ferrant, C., André, F. and Fabre, R. (2016), "Ready for the future? A survey on open access with scientists from the French National Research Center (CNRS)", *Interlending & Document Supply*, Vol. 44 No. 4, pp. 141-149

Schöpfel, J. and Prost, H. (2016), "Research data management in social sciences and humanities: A survey at the University of Lille 3 (France)", *LIBREAS. Library Ideas*, Vol. 29, pp. 98-112.

Simukovic, E., Kindling, M. and Schirmbacher, P. (2014), "Unveiling research data stocks: A case of Humboldt-Universität zu Berlin", in *iConference*, *4-7 March 2014*, Berlin, pp. 742-748.

Tenopir, C., et al. (2015), "Changes in data sharing and data reuse practices and perceptions among scientists worldwide", *PLoS ONE*, Vol. 10 No. 8, e0134826+.

Tenopir, C., Pollock, D., Allard, S., Hughes, D. (2016), "Research data services in European and North American libraries: Current offerings and plans for the future", *Proceedings of the Association for Information Science and Technology,* Vol. 53 No. 1, pp. 1-6.

Thanos, C. (2016), "Research data reusability: conceptual foundations, impediments and enabling technologies", *Publications,* Vol.4 (forthcoming).

Ward, C., Freiman, L., Jones, S., Molloy, L. and Snow, K. (2011), "Making sense: Talking data management with scientists", *International Journal of Digital Curation*, Vol. 6 No. 2, pp. 265-273.

Weller, T. and Monroe-Gulick, A. (2014), "Understanding methodological and disciplinary differences in the data practices of academic scientists", *Library Hi Tech*, Vol. 32 No. 3, pp. 467-482.

Whyte, A. and Tedds, J. (2011), *Making the case for research data management*, DCC briefing papers, JISC Digital Curation Center, Edinburgh. http://www.dcc.ac.uk/resources/briefing-papers

Wilkinson, M. D. et al. (2016), "The FAIR guiding principles for scientific data management and stewardship", *Scientific Data*, Vol. 3, 160018+.

All web sites visited in January and May 2017.

About the authors

Joachim Schöpfel is senior lecturer, former head of the department of information and document sciences at the University of Lille and researcher at the GERiiCO laboratory. He is also director of the French national reproduction centre for PhD theses (ANRT). He is interested in scientific information, academic publishing, open access, grey literature, usage statistics and service development. He is member of GreyNet, NDLTD and euroCRIS.

Coline Ferrant is a PhD student in the Dual PhD program in Sociology between Northwestern University and Sciences Po. At Sciences Po, she is affiliated to the Observatoire Sociologique du Changement laboratory (Sciences Po / CNRS). She is also an associate fellow at the Alimentation et Sciences Sociales laboratory (INRA).

Francis André is acting as a special advisor for research data at the CNRS Scientific Information Department, based in Paris, working on a variety of projects across scientific data management and open science. He trained in geology and computer sciences, with a PhD from University of Nancy. He initiated and co-chairs the Research Data working group of the French scientific digital library initiative (BSN).

Renaud Fabre is head of the CNRS Scientific Information Department at Paris since 2013. He is professor of economics and former president of the University of Paris VIII Vincennes-St Denis. He graduated in political sciences and holds a PhD in economics.

Appendix – The survey questions

The complete survey (French version) is available online (see footnote 16). This is the list of the 34 items re-analysed for the purpose of our study:

Data production

#47 Does your laboratory's research produce data in need of research data management? (yes/no)

#50 Do you have some idea about the volume of your laboratory's research data volume? (yes/no)

#55 Which kind of data does your laboratory produce? (17 disciplinary choices)

#56 Which kind of raw data does your laboratory collect? (observational, experimental, survey, simulation, other)

#57 Which are the principal raw data formats? (numbers, images, text, video, audio, other)

#58 Are these data formats interoperable or proprietary? Both? Don't know?

#59 Do you think that your laboratory's raw data are not protected by intellectual property? (yes/no)

Data management

#48 Do you have tools to manage your laboratory's research databases?

#53 Does your laboratory's database production receive external funding?

#54 If so, which kind of funding?

#60 Which kind of IT infrastructure do you have for the data collection, processing and sharing? (personal, local, institutional)

#61 Which kind of dedicated software do you use for the data collection, processing and sharing? (database system, spreadsheet, other)

#62 Do you have dedicated software for a specific research instrument or experimentation? (interoperable, proprietary, don't know)

#63 Which part of the research data management is done by special staff? (database integration, reformatting/standardization, creation of secondary/derived data, creation of metadata, other)

#64 Do you receive funding for research data management? (yes/no)

#65 If so, is this funding sustainable (recurrent)? Project-related?

#66 Do you have dedicated staff for research data management? (yes/no)

#67 If so, which kind of staff? (permanent, temporary)

#68 Did you already prepare a data management plan? (yes/no)

#74 Research data management needs specific skills. Please evaluate your laboratory's skills level in the following domains (ignorant, rudimentary, good knowledge, expert):

- Data management plan
- Sensitive data processing
- Data description (metadata, identifiers)
- Data sharing on international platforms
- Referencing and citation
- Protection (security, validity)
- Legal and ethical issues
- Financial and commercial issues

Data sharing

#51 Are your research databases produced with other teams, laboratories or organizations?

#52 If so, which ones?

#69 Are your databases accessible online? (yes/no)

#70 If so, is their access open or restricted?

#71 For at least one database, do you apply standardized and community-specific procedures for

- Data format? (yes/no)
- Data collection (yes/no)
- Data description (yes/no)
- Terminology (yes/no)
- Identifiers (yes/no)

#72 In the field of research data management, do you apply collaborative practices along with your community? (yes/no)

#73 If so, which ones? (shared tools, common guidelines, training, workshops, other)

#87 Have you already had legal problems with the use or reuse of your laboratory's research data?

#89 Have you already had legal problems with the valorisation of your laboratory's research data, in particular with personal data?

#90 Have you already had legal problems with the publishing of your laboratory's research data (open data, public data, administrative data…)?

Needs and demands

#49 Do you need tools to manage your laboratory's research databases? (yes/no)

#75 Do you need online assistance?

#76 If so, which one? (user forum, hotline, platform with software, other)

#84 Which kind of service would you need most? Research data management?