comets

CNRS Ethical committee

OPINION n° 2015-30

THE ETHICAL CHALLENGES OF THE SHARING OF SCIENTIFIC DATA

Opinion no. 2015-30 approved May 7, 2015

I. SUMMARY

The massive development of IT tools for data collection, measurement and processing has changed the role of data in the production of scientific work. The scientific data sharing movement provided for by international mechanisms such as the Berlin Declaration in 2003 is a response to the need to share results as quickly as possible and overcome legal and technical barriers to the circulation of this data. Similarly open data policies from governments and the European Union have for some years been aimed at the wide dissemination of data acquired with public funding. However not all scientific communities are subject to the same constraints in this respect. The general guidelines can appear to be in conflict with legal restrictions regarding privacy (data protection), copyright law and the obligation of secrecy, or security. Given the complexity of the obligations encountered by researchers, this opinion is intended to reaffirm the need for rational sharing of data and include new requirements for data availability in the assessment of scientific work. The data issue, whether in terms of obstacles to be overcome or limits to openness, has become crucial in the definition of science policy.

CONTENTS

II. SELF-REFERRAL

Data has a central role in scientific production in all disciplines. Researchers increasingly need large amounts of data, alongside more modest amounts to explore, view and compare results, validate assumptions or formulate new ones or even to automate the building of new insights through machine learning. Large infrastructures and common platforms building on recent advances in digital technologies are continuously developed for archiving, storing or processing information. Movements promoting open access are therefore becoming critical. Public and private research laboratories increasingly need to come together to coordinate their efforts and reuse data acquired by others.

The stance regarding data sharing and openness differs greatly according to data types and disciplines. In astrophysics or genomics for example, data reconciliation and comparison is clearly a source of new discoveries; any impediment to the flow of results is contrary to fundamental principles of the generalised pooling of insights. For other disciplines, particularly in human sciences, or when major industrial challenges appear, data is often collected individually or under restricted sharing conditions. Depending on the subject of research, data can be shared only with the same embargo as on the publication of results. We should add that raw data is not always stored once it has been processed because this may take up too much space.

The scientific data sharing movement must adjust to the latest government policies on open data which have for some years been aiming at the wide dissemination of data acquired with public funding, with significantly differing objectives and legal and ethical constraints. Public data and scientific data however partially overlap. The HORIZON 2020 European program also enshrines the principle of free access to scientific publications and data. All these general guidelines can appear to be in conflict with legal restrictions regarding privacy (data protection), copyright and the obligation of secrecy, or security. Pressures in respect of research value or duty of confidentiality may also play a role in restricting the dissemination of data. While a large number of researchers support the principle of open data, many feel helpless in the face of constraints that can appear contradictory. This opinion aims to inform researchers on their obligations and the impacts of their choices with respect to the data they collect, share or recycle, and suggest what response scientific institutions should make in respect of these new obligations.

III. ANALYSIS

A. Nature of data and strategic context for openness

Scientific data as considered here applies to all data collected for scientific research[1], namely – empirical, observed, measured – *primary data*, some of which is not intended to be stored let alone shared; *secondary data* derived from primary data, annotated, enriched and interpreted, adding value to the original data and possibly involving other actors; *metadata* that structures, manages and facilitates the access to primary and secondary data and provides information on the conditions of data sharing. This data can consist of digital streams from sensors, or text, graphics, pictorial and multimedia documents. The gap between the status of data and that of publications is moreover tending to narrow with the concept of open science, i.e. dissemination of the data and insights used and developed during the process of elaboration and writing of scientific publications.

Successive agreements and charters have marked the history of the scientific data sharing movement. In 1996 for the first time, researchers involved in sequencing the human genome, signed a set of agreements providing for the basis of a system for data sharing as of its production. Then the first definition of *open data* was given by the international declaration on open access in Budapest on 14 February, 2002, known under the acronym BOAI (Budapest Open Access Initiative). Note that the clarification of the meaning of various terms that appear here was established gradually. Today the term *open data* is used for the free availability and use of data obtained using public funding in general. The term data sharing is usually reserved for the movement initiated by research communities affected by the making available of research data on which this opinion focuses. Many other initiatives have followed, such as the 2003 Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, which was strengthened in 2005. Most scientific bodies including the CNRS (the French national centre for scientific research) have signed these declarations and legitimised this open access culture. The need for reflection on openness and knowledge sharing is included in the 2014 CNRS DIST (scientific and technical information department) strategic plan. We should add that free access to scientific data is now included in the objectives for public research, with this proviso: access does not extend to the dissemination of confidential data, which falls under other legal regimes. Some data used by researchers to conduct studies is sent to them by third parties subject to confidentiality, pursuant to a contractual clause or specific regulations. These conditions can be very restrictive, as in the case of tax data, which was made available for research by the Law of 22 July 2013.

[1] H.Tjalsma & J. Rombouts, *Selection of research data- Guidelines for appraising and selecting research data,* The Hague: Data archiving and Networked services 2011, p.13,14 at http://www.dans.knaw.nl

B. Very rapid expansion of the volume of data, diversification of its use in research

Scientific activity relies increasingly on the creation and use of shared multi-source and multi-purpose data infrastructures linked to three types of changes: the development of measurement and raw data capture tools, considerably higher IT capacity for storage and archiving, and collaborative internet networking that allows direct enrichment of online databases and platforms by many contributors.

Currently, a lot of data collected by experimenters is lost. It is estimated that publications provide access to about 10% thereof, the remainder being available on computer hard drives but not used. In some disciplines, valid and important results remain unpublished and a lot of data is underutilised or lost (this is particularly true of data from negative results, which is ignored). For that collected by large tools, the raw data gathered is so massive that it is processed directly online without being stored, such as, for example, the data provided by spatial observation. It is then necessary to indicate the origin of the data constructed from the raw data, even where such data has disappeared. It is also important to recognise the valuable work of staff (researchers, engineers, technicians) who have contributed to the processing of raw data to make it available in a usable form for the rest of their community. This often thankless work is not always valued to the extent of the considerable effort it generally requires.

Large volumes of data may improve our understanding and help in predicting phenomena through the use of machine learning techniques in all scientific fields, particularly in the health sector. However, due to the huge volumes of information involved and, above all, the dimensions of descriptions, such massive processing of data does not generate causal relationships, but correlations, which suggest the existence of causalities, without proving them. New practices are being developed, resulting in a true epistemological revolution often referred to as 'data driven research': starting from already established datasets, algorithms automatically explore fields of assumptions and detect irregularities or unexpected phenomena that differ from known laws. The use of these practices is developing, especially in areas such as seismology or large-scale biology (genomics).

Another advantage of the availability of large amounts of data, obtained thanks to the fantastic possibilities of the internet, is to allow access to the genesis and the various stages of the development of research, showing partial interdependent and scalable results.

Cognitive interactions are manifested in so-called 'knowledge hubs' where several layers of more or less developed knowledge coexist as the dynamic result of a continuous exchange between researchers. All these developments have a heuristic effect on traditional scientific processes.

Access to primary data then becomes crucial to enable the checking of its quality and also to judge the resulting methodology and interpretation. The issue of tracking and scientific evidence therefore also has to be raised in the light of these new practices.

C. Making public data available and scientific data sharing

Unlike the data sharing movement developed by researchers themselves, public open data policies were born outside the scientific community. In Europe, following the Directive on reuse of public sector information, and the Directive creating an infrastructure for spatial information, the majority of countries have opted for a policy aiming to promote public open data. States are important providers of data produced, reproduced, collected, distributed or redistributed by governments as part of their institutional missions. These include demographics, geographical, weather, economic, financial, cultural and tourism data, which are intended to ensure the quality and continuity of public services. In France the remit of ETALAB, a service that manages public open data under the authority of the Prime Minister, is to disseminate data produced with public funds and make it available (almost) free in order to facilitate its reuse in the least restrictive way possible. One of the other objectives of *open data* is to allow the valuation or the monetisation of this data by creating wealth for companies that use it. Finally, a more collaborative aspect concerns the data made available for communities of citizens or civil society. These policies all aim to create accessible and sharable data repositories and also to promote transparency of knowledge production. In the framework of public *open data* measures, the principle is statutory and applies legally to all public officials, including employees working in public research.

Policies that promote public open data do not have the same objectives as those of scientific data sharing. To clarify the applicable regimes, it is important to differentiate scientific and public data. Scientific data produced with public funds is, with some exceptions, done so with a view to being made public. Public data is meant to be used as scientific data when it concerns the environment, climate, the state of society or health. Researchers should therefore be able to use *Open Data* in line with state policies. One example is health insurance which has developed the world's largest database on health: SNIIRAM, which has for decades been supplied with information generated by the management of care provision and hospitalisation in France (20 billion lines of services). Responding to the demand for this data repository, the Caisse Nationale d'Assurance Maladie (National Health Insurance Fund) intends to develop its policy of openness through public interest criteria (including research), depending on the quality of protocols, the need for data access, safety procedures and merits of applicants.

D. Constraints on the processing of personal data

The previous example illustrates the legal constraints surrounding the availability of public data. Health information is sensitive because it can lead to the identification of individuals. The data processing model that was part of in the Computing and Liberties (*Informatique et Libertés*) law of 6 January 1978, amended in 2004, no longer corresponds to the current modalities of algorithmic processing. The features of the law are no longer adapted to the new contexts of massive data, as reported by the researchers interviewed. Indeed the rapid, open flow of data between researchers upsets the order of the proceedings and makes the stream of data relatively autonomous in relation to their sources or authors. It often becomes impossible to respect the principle of research objectives when assumptions are not developed a priori, the principle of proportionality if the necessary data is not known prior to use, or even the principle of non-conservation, because you cannot destroy the data used at the end of research because of its open access and reuse character.

In many cases the use of data relating to individuals is subject to constraints. Thus viewing using computers, which aims to automatically recognise visual scenes, is subject to image rights. If this relates to an identified or identifiable person, the image of a person is considered as personal data. Computer processing of this data (scanning, diffusion from a website, etc.) must be carried out in compliance with the '*Informatique et libertés*' law. The *Commission nationale de l'informatique et des libertés* (CNIL - National Commission on Computing and Liberties) approves research on face recognition for research purposes, provided this data is not retained beyond the end of the project, except where a request for extension is made. This therefore leads to the paradox whereby experiments with other systems on the same data to compare performance are banned, although this would be the normal scientific process. It seems inconsistent to limit the use of these images in this way, while the goal is to allow the development of the most effective possible operational algorithms.

It has become difficult to apply in all cases the basic principles for the processing of personal data, such as informing people about the future and the use of data, or obtaining their consent. Sometimes the researcher's approach requires obtaining information without the knowledge of the person who is the subject of the investigation. Principles therefore must be put into place to be followed where consent has not been given, such as an undertaking to inform this person post-research. Likewise the issue of consent arises when research focuses on information from data mining of social networks. Though publicly available, this data is considered by the CNIL as personal data. Finally, note that the future European regulation on the protection of individuals with regard to the processing of personal data provides for exemptions from the requirement of consent in research in cases where it serves the overriding public interest and cannot be completed in any other way. For the new digital bill, initial consultations have set priorities defining a legal framework for data mining.

E. Making research data available and the Science Commons movement

In recent years, the birth of data science and the explosion of massive data has led to awareness of the legal and technological barriers that impede the free flow of data, even when scientists aspire to it. Either large databases may be subject to access rights, or the data available in closed formats, requiring proprietary software. Initiatives have encouraged some scientific communities to get together to assert new principles of openness and data availability. In 2005 a community of researchers, aware of the resistance encountered during the implementation of open data policies, launched a global initiative to create Science Commons, with tools and methods (access platforms, model author contracts, etc.) to speed up the flow of results and enable the reuse of data on which they are based. In general today, data to be shared is subject to a minimum questionnaire on the identification of the researcher and the use of the data.

These shareable research platforms facilitate the development of new services such as *reuse of research data* through policies and tools that help individuals and organisations to make their production available; *immediate access to tools* (online calculation) through standard contracts for duplicating, verifying and expanding research and also facilitating scientific peer review; the *integration of fragmented sources of information* through a common, standardised, machine translatable language.

The general movement of openness and data sharing has been facilitated by open archiving policies developed within scientific institutions (ArXiv, 1991). In France, HAL (Hyper Articles en Ligne), established in 2000, was based on 'the model of direct communication between researchers' of their pre-published articles. Its management and tasks, currently being revised, are still to be defined vis-à-vis the archiving of scientific data; in particular the embargo period and open license link, which are decided on by researchers when registering their work, will be included directly on the platform.

A successful example of open networking initiated by researchers is an initiative dating back to 2013, in the field of human biology: the '*Global Alliance for genomics and health'*. This is a unified movement of 285 member institutions from 30 countries that have decided to contribute to facilitate the sharing of scientific data by establishing standards, sharing experience and best practices and establishing a framework for responsible data sharing. Similarly another international initiative initiated and piloted in France in the field of biology (BRIF) promotes sharing through better recognition of shared resources and their authors. Disciplines such as Earth and Space Sciences have insisted on other imperatives which also require archiving and the free release of data: perennial observation of natural phenomena involves processes whose time constants can be large compared to human life; the data from these measures are of their essence non-reproducible and are the basis of our knowledge of the world around us, its developments and risks to our societies. As nature is common property, the community of researchers involved and organisations that employ them have an obligation vis-à-vis the public in this respect.

Other types of data platforms sometimes develop outside researchers' initiatives. This is the case in disciplines such as biology and medicine, where publishers require researchers to make their data set available to check the reproducibility of experiments or processes covered by the publication to monitor the results to be published by challenging the data and detecting fraud or errors that may lead to withdrawal. Once collected, if it is not in public databases but remains exclusively in the hands of publishers, this data may constitute for publishers a closed data set that is independent of publications, even though it is only requested in order to monitor results. To prevent this eventuality, access to data relating to a publication required by publishers must not be restricted by copyright imposed by journals. The data must be provided with the article in an associated file but must remain available to researchers for repeat analysis and publication in any journal.

Sharing policies involve informing researchers on the limits of sharing. The data concerned may be unavailable in non-anonymised personal data format, or may be subject to special regimes such as national security and confidentiality, to restrictive contractual clauses or various business interests. In addition, researchers must remain rights holders of the data they have produced or analysed as they are of their publications, if they want to share it or allow reuse. In this case, researchers are strongly advised to put their protectable data under a free license like *Creative Commons* in order to inform future users of the status of the data. Researchers need to be alerted to the consequences of their choices when they make over their exclusive rights to other parties.

F. Researcher responsibility

Today, despite encouragement at European level and through the CNRS, researchers do not all follow the same practices or suffer the same constraints vis-à-vis their data, as highlighted in the survey conducted at the CNRS with the directors of its laboratories. An embargo of between six months to a year after publication looks to be a minimum time period for making primary data available in the human sciences. In chemistry, probably because of industrial development issues, communication of results and data does not precede the acceptance of publication. Physicists however usually put their articles in open access archives, possibly with additional data, as of submission or even earlier. For the use of processed data that is produced by major physics or astronomy equipment, there is a delay before this data is made available to the whole community. This delay is fixed in advance (between one and two years), giving preference for a limited period to the researchers who have contributed to this equipment.

In general, public researchers are encouraged to pursue the ideal of sharing and peer exchange and participate in the dissemination of data obtained with public funds, provided they abide by the exceptions defined by contractual commitments. Conversely, consortium agreement models involving public and private partners (particularly in competitiveness clusters) are often very restrictive in terms of open data: agreements will now have to be negotiated upstream by public researchers to avoid improper confiscation of untapped data by private partners. Another concern involves decisions regarding the terms of benefits sharing, to prevent powerful institutions or private companies deriving exclusive benefits from data tapping. We have observed in fact that teams funded by public funds or the European Union now disclose their data allowing large private groups to exploit it for their own benefit, irrespective of reciprocity vis-à-vis public researchers.

Researchers are becoming aware that open data - but also software, ontologies and metadata that allow its exploitation - implies a new level of responsibility. They now need to be particularly concerned about the quality of information and data that they offer, as well as the clarity of the accompanying documentation. To allow others to replicate or reuse data, the integrated and interoperable nature of the data must be checked, identifying sources, dates of collection and processing and a detailed examination of the different steps in the creation of data repositories: collection, classification, standardisation, provision, reuse, conservation, destruction or archiving. Thus the organisation and maintenance of interoperable data is becoming fundamental for ensuring the integrity of scientific data in the digital era. These new tasks create new responsibilities among researchers. The implications of these policies in respect of the ethics of research need to be assessed on a case-by-case basis.

In short, faced with this dynamics of the movement of data sharing supported by their supervisory authorities and communities, researchers should be aware of their individual responsibility, deontology and ethics towards the community to which they belong, be aware of international undertakings of the institutions which employ them, know the limits of the exploitation techniques they use regarding the volumes of data they handle and interpretation problems that might result. It is also up to researchers to participate in the definition of best practices specific to their disciplines in terms of data sharing. For these reasons, research institutions must develop new skills that meet the information needs of researchers in terms of their data and create ethics committees on research data by discipline or institution.

May 7th 2015.