Nicolas de Cordes
(Orange/pilotage de la stratégie d'innovation)
Yves-Alexandre de Montjoye
(Imperial College London)
Zbigniew Smoreda
(Orange Labs)

OPAL: reconciling open innovation and data security

The OPen ALgorithms (OPAL) platform aims to safely and securely provide public access to individual level data collected privately. This project offers an alternative to the relatively inefficient traditional methods of anonymization in terms of protecting a vast quantity of data concerning the private lives of millions of people, such as geolocation tracks or web search history. Instead of irreversibly sharing anonymized data that might be re-identifiable, OPAL allows the original data controller to safely grant access to trusted parties. These parties can then use a series of pre-validated algorithms to perform the necessary analyses on the aggregate private data. OPAL will benefit from extensive privacy and security protections to protect individual level data, Furthermore, to allow transparency and maximise benefits, the OPAL project is open source. OPAL will pilot in developing countries in compliance with their legislation, targeting economic and social development. In the long term, this technology could be applied more widely.

OPAL
bring the code to the data

It is very difficult to access and take advantage of the potential of privately collected data (telecoms carriers, banks, logistics, etc.), whether for the purposes of research, public policy or, in a wider sense, public interest. In spite of the potential demonstrated by hundreds of scientific articles and conferences such as NetMob **[1]**, legal difficulties, competitive risks and threats to client privacy have so far strongly limited access to such sources as mobile phone data. Such complexity has been observed not only in uses during the most severe crises - such as during the Ebola epidemic in 2014 - but also during regular use within the context of public policy.

Orange is one of the major players in the field of open innovation in data sharing, notably following two scientific competitions known as *Data for Development* (D4D), organised by the carrier in 2012 and 2014 **[2]**. These challenges have propelled the company to the forefront of the UN's drive for data mobilisation for sustainable development **[3]**.

However, despite strong interest from the development community, abundant scientific literature on the use of Big Data, to analyse migration, poverty, epidemics, mobility, famines, the effects of climate change, etc., as well as, spectacular research work/initiatives such as Flowminder **[4]** in connection with earthquakes in Haiti or Nepal (to name but a few), the procedures for accessing private data remain excessively complicated, granted only for limited periods and are often very slow which prevents rapid response.

The OPAL project stems from a desire to help solve major social problems – especially in developing countries which lack much of the necessary infrastructure for gathering useful statistics about their populations. OPAL aims to help solve these problems while complying closely with legal, technical and commercial constraints, starting with those associated with the risk of privacy violation as well as fears that data holders will lose control over how that data is used.

How can third parties be given access to detailed data in a way that is both secure and controlled, retaining all the informational value of the data, regardless of the scale? This is exactly the issue OPAL intends to address.

USEFULNESS OF PERSONAL DATA

Personal data, even with identifying information deleted or irreversibly altered (pseudonymized), carries with it the risk of re-identification by cross-referencing against multiple personal databases available elsewhere. Even in a very large database, an individual pattern quickly becomes unique. Once cross-referenced against other information, a user can potentially be re-identified. Traditional solutions for anonymization (also known as de-identification), such as generalisation or the addition of noise, do not sufficiently reduce the risk of re-identification **[5]**. Protecting data through anonymization requires using methods that can be very harmful to the informational value of the data: aggregating data for a number of people taken together, or transforming actual data into summarised, artificial data that reproduces only certain precise, predefined aspects of the initial information.

Although these methods may be sufficient for some specific use cases, they do not generalise. Not only does anonymization greatly reduce data quality and so, the quality of the resulting analysis, but also, performing anonymization properly requires the data publisher to predict all possible future uses of the data before publishing.

There is an inversely proportional relationship between the level of privacy protection and the usefulness of the information for behavioural analysis purposes. At one extreme, the raw data contains all individual information but offers no protection, while at the other, the fully aggregated data (as an overall average) is perfectly anonymous, but is not useful beyond the apparent summary statistics.

Faced with this dilemma, there are two possible options: either we seek a method for transforming the original data (by grouping, sampling, or modelling) which is sufficiently robust to ensure re-identification is impossible (or at least very difficult and costly), or, we do not alter the original data apart from replacing direct identifiers with pseudonyms (thus preserving their value), and then try to provide a secure method of access to this pseudonymised data.

Research into data anonymisation began in 1995 and is now ubiquitous, with each new contribution proposing a variation of an existing method, often for a new type of data. However, many databases that were believed to have been anonymized have been re-identified in recent years, such as web browsing databases in Germany, government medical data in Australia, and public transport data in Latvia **[6]**. These discoveries, together with theoretical analyses showing the intrinsic limitations of anonymization, lead us to doubt the possibility of the development of any sufficiently effective universal method of anonymization in the short term.

We have therefore chosen the second option: Access Control. This option allows detailed pseudonymised data to be retained by the data collector, with strong control over both access to this data and the processed results that are returned to the external user using SafeAnswers, a solution developed by MIT researchers in 2014 **[7]**. This allows the data to remain permanently within the infrastructure of the data controller.

This is the OPAL's ultimate goal: to send the code to the data (rather than the other way around), and to supplement it with open-source code and certification for the associated algorithms. This project was developed in collaboration with London's Imperial College, MIT, the World Economic Forum, Data-Pop Alliance and the Overseas Development Institute. Two mobile operators are participating in OPAL's pilot projects: Orange-Sonatel in Senegal and Telefónica in Colombia.

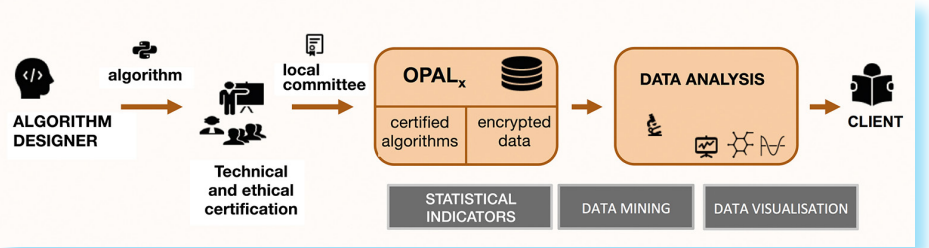
HOW DOES IT WORK?

An examination of the abundant scientific literature on the subject of mobile data confirms our intuition, which is that statistics that are of most interest to researchers tend to be aggregated (graphs showing communication between regions, mobility matrices in or between cities, flows between centres, etc.).
However, this apparent consistency masks the fact that each research question must have access to aggregate data tailored to the problem domain. For example, a mobility matrix for studying the construction of a new bridge will focus on journeys within the city at peak times, while a matrix for studying the spread of malaria will require a detailed analysis of movements between regions of high incidence.

The benefits derived by academic researchers, national statistical institutes and NGOs from gaining access to high quality aggregate statistics provide the justification for the OPAL project; namely, to share algorithms, not data. External researchers can use the algorithms that correspond to their needs, configure it to address their search query precisely and then run it within the operator's pseudonymised database: no data is released or made directly accessible to the researcher.

If currently available algorithms do not meet the user's needs, more specific algorithms dedicated to a particular problem can be developed by a third party and supplied to the data-collecting operator for integration into its library: algorithms to calculate key indicators for national statistics (poverty index, illiteracy rate, etc.), monitoring of population mobility (migrations, city dynamics. etc.), social cohesion (communities, information circulation, etc.), epidemiology (population flows, rates of contact between individuals, etc.), and so on.

Fig. 1: How OPAL works



In order to guarantee the quality of the algorithms and select only those that meet the present needs, algorithms will be published and submitted for review by the scientific community following accepted research principles. Algorithms validated in this way will be close to the state of the art, especially if the scientific community's current level of interest is sustained. Subsequently, other researchers will be able to suggest local improvements or adaptations (e.g. in participating countries) as their work progresses, gradually reducing the cost of algorithm development.

It should be noted that before the new algorithms can be deployed, they will need to undergo technical certification (performance, compliance with privacy requirements, verification of automatic and manual code, testing, etc.).

END USE SUBJECT TO ETHICS COMMITTEE APPROVAL

After the selected algorithm has been technically certified, an ethical use assessment must be carried out. The technical mechanisms implemented on the platform ensure that information about individual persons cannot be extracted. However, this is not enough to guarantee the safe and ethical use of the data, as implied by the emerging concept of «group privacy» **[8]** (protection of collective information) and the potential threats to national security or the data provider's economic interests.

Therefore, as well as implementing technical protection mechanisms, it is vitally important to set up a governance body (at local level, in the country providing the data) which must include not only technicians but also lawyers, members of civil society and also representatives of the data collector responsible for the legal processing. Doing so ensures the representation of all interests relating to the use of the algorithms' results.

It is only at the end of this threefold process (scientific approval of the algorithm, technical certification, identification of ethical risks arising from the end use) that the algorithm can be installed on the OPAL platform in the country in question. Projects accepted by the OPAL platform operator can then access the API (Application Programming Interface) and execute the algorithms for which they have obtained authorisation.

THE OPAL PLATFORM

The OPAL platform is developed entirely in open source by the project's partners, and relies exclusively on existing technologies with open source code. It will be developed on the assumption of use in the telecoms sphere, but also has the potential to be a general «Open-Algorithm / Safe-Answer as a Service» process.

In terms of performance, the infrastructure is built to be able to process call reports and other activity logs from 10 million customers covering a period of several years. The algorithms are written in Python using the bandicoot toolbox **[9]**. For security and privacy reasons, the algorithms are run individually on the pseudonymised data for each client with their pseudonyms being renamed on the fly every time the algorithm is run on client data. Individual results are then aggregated and verified to ensure that the privacy of the operator's clients is protected. The platform is accessed via a REST API, with a controller distributing processes and data efficiently between the machines and a task scheduler ensuring that the most common queries are pre-processed.

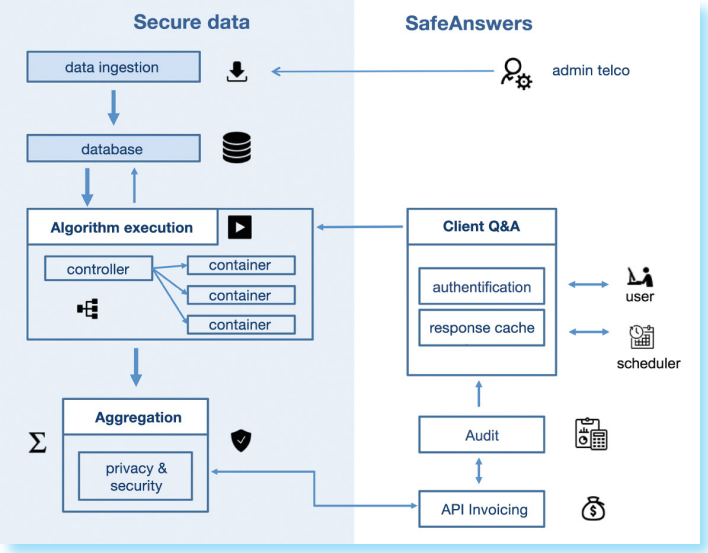
An authentication mechanism verifies all queries submitted to the platform, and metadata from each query is retained for control purposes. As mentioned above, only algorithms that have been technically verified and validated by the governance body and then selected by the OPAL operator can be run, and the processed results are signed by the platform in order to guarantee their source and traceability.

WHAT ARE THE BENEFITS?

The OPAL/SafeAnswers approach allows data collectors to apply their own security procedures and retain control over their clients' personal data, which is never transferred externally. In crisis situations, it can manage access to data using its own rules, selectively granting controlled access (to NGOs, aid and relief organisations) or, alternatively, denying it completely (in the event of political instability, armed conflict, etc.).

The system is developed to comply with data minimisation and Privacy by Default **[10]** principles, and the privacy authority in the country concerned is invited to sit on the OPAL local steering committee to give its opinions on the algorithms used and the risks associated with their use.
Another hallmark of OPAL is response rate. Unless there is a very specific need for an ad hoc development, analysts no longer need to develop and implement their own procedures, as they can use an existing algorithm (on an Algo-Store) which they can fully analyse as all algorithms will be open source.

Fig. 2: OPAL architecture



This also creates an expanded circle of potential users for the data, as specialists from different disciplines rarely have strong data science skills. OPAL will provide them with an analysis system offering many indicator calculation methods which have already been tested and validated by experts.

Thanks to this system, the potential of privately collected data can finally be unlocked and used for the public good. For example, it will be possible to measure the indicators of some of the UN's Sustainable Development Objectives (SDOs) in close to real time, using private data. Furthermore, the transparency built into OPAL allows any potential issues regarding the analytical techniques to be approached openly.

As a corollary, collectors will be able to monetise the information contained in their data by reaching out to interested parties, using the same open-source algorithms. Indeed, various companies have expressed interest in this «Algorithm As A Service» method to provide access to a controlled question-answer mechanism built on their own data: banks, insurance companies, satellite operators, national statistical institutes, DNA analysis, etc.

Lastly, given that this method permits statistics from sensitive data to be made accessible without being disclosed, it could also be of use to an international group needing to perform analyses of the data of its subsidiaries without being obliged to duplicate or move that data. All that remains is to define a series of reference algorithms to obtain comparable operational indicators between subsidiaries: uses, segmentations, churn, ARPU, etc.

CONCLUSION

The OPen ALgorithms project aims to create a flexible system, based on the SafeAnswers mechanism and open source algorithms, that can be implemented within the information systems of the various data providers (private or public), enabling their data to be used without being shared. However, the «Data for Development» approach and our analysis of the social utility of data is prompting us to test this paradigm first on mobile phone data and to deploy it initially in developing countries where there is an urgent need for more reliable and up-to-date statistics. For this reason, the first OPAL deployments will be made between the end of 2017 and the end of 2018 at Sonatel in Senegal and at Telefónica in Colombia, in close collaboration with their respective national statistical institutes.

TO LEARN MORE

[1] [1] http://netmob.org/

[2] We previously covered this subject in the columns of LU&V nos. 49 and 52.

[3] *A World That Counts: Mobilising The Data Revolution for Sustainable Development.* UN IEAG Report, Nov. 2014.

[4] See: http://www.flowminder.org/

[5] See: Y-A de Montjoye (2015) *Metadata and privacy: an unsolvable equation?* LU&V, No. 52

[6] http://www.ndr.de/nachrichten/netzwelt/Nackt-im-Netz-Millionen-Nutzer-ausgespaeht,nacktimnetz100.html; https://www.itnews.com.au/news/health-pulls-medicare-dataset-after-breach-of-doctor-details-438463; http://ieeexplore.ieee.org/abstract/document/7821808/

[7] http://openpds.media.mit.edu/

[8] L.Taylor (2017) Safety in Numbers? Group Privacy and Big Data Analytics in the Developing World, in *Group Privacy: New Challenges of Data Technologies,* Springer.

[9] Y-A. de Montjoye, L. Rocher, A.S. Pentland (2016) Bandicoot: a Python toolbox for mobile phone metadata. *Journal of Machine Learning Research*, 17(175).

[10] Privacy by default or by design consists in putting thought into the protection of the private lives of users of a given technology from its development stage.

Digital Traces and Sustainable Development in Africa

uses and value 57

A newsletter about research in economic and social sciences 11/2017

INTERNAL GROUP

Orange SA au capital de 10 595 541 532 €
78, rue Olivier de Serres
75505 Paris cedex 15
350 129 866 RCS Paris