DEPARTMENT OF GLOBAL HEALTH
UNIVERSITY of WASHINGTON

**PhD in Global Health**

Metrics Track



# Data Hybridation approaches in African HMIS

Grégoire Lurton

IHME
Institute for Health Metrics and Evaluation

I-TECH
International Training & Education Center for Health

BLUE SQUARE .ORG

**Abstract** Here the abstract will come

# Contents

# List of Figures

# List of Tables

# Acronyms

**EMR** Electronic Medical Record. 3, 4

**HIS** Health Information System. 1

# 1 Introduction

If a literary form had to be chosen to write or talk about Health Information System (HIS), the complaint would probably be everyone's favorite pick. Be it complaints on the burden of work involved in collecting, managing and analyzing data in health systems, or laments on the inexistence of good quality data in most developing countries health systems, HIS are usually described as a non performing burden of health systems, that can only be improved [HMN citation]. This frustration has multiple causes, and is only matched by the expectations placed in HIS and their widely recognized importance, some authors calling HIS "the foundation of public health"[**?** ]. Collecting and analyzing information on activities and results of health systems and on the populations served is indeed essential to guide strategic decision making and to inform health policies.

National HIS are complex, in the sense that they are expected to provide a wide variety of information, acquired from a wide variety of data sources. Producing this information requires the contribution of a multitude of actors, and is a huge organizational and methodological challenge. Data used to produce relevant health information may come from administrative records, organizational documents or population surveys, and are produced by a variety of actors and organizations, with differing cultures and approaches.

Moreover, HIS have to be able to adapt rapidly to changing epidemiological, organizational or political situations. They have to be able to produce relevant information on emerging health issues, and to adapt to the entry of new actors or to a modification in the mode of management of health systems, or to new priorities or questions.

In richer countries, the issue surrounding health information is often one of regulation and standardization. The existence of well performing and well established data sources on populations, and the relative ease of collecting massive amounts of data on individuals pose questions that are mainly related to the protection of privacy, and to the definition of standards for interoperability. The definition of what information should or can be produced is usually a legislative and matter, handled by dedicated entities.

In developing countries, where population data is scarce and data collection can be much more of a challenge, the issues are much different. Health policies depend heavily on the financial, political and technical support of international actors, with differing orientations and priorities. As a consequence, developing countries HIS strengthening programs can be classified into two great families : technology based solutions and institutional reforms. The first family is mainly targeted towards improvement of data collection and relies on solutions relying on the revolution in data collection capabilities. The second family relies on the adequation of health information produces to policy makers needs, and usually promotes a top-down and normative approach to Information Systems.

If each of these approaches has its benefits and some success stories, they appear to miss an important part of what makes information systems work. They put their emphasis to the extremes of the information production chains (cf. Figure 2) and undervalue the middle tiers of these systems. This proposal designs research program oriented towards exploring ways in which this middle tier can be mobilized to improve the value of information produced by HMIS.

This document will start by an introduction of the conceptual framework surrounding the proposed research. We will define Health Management Information Systems, using two classic schematic approaches. We will then define the objectives we will pursue in our research, and will finally describe our different aims and the methods we want to use to complete them.

# 2 Conceptual framework

We will first look at a representation of HMIS through the goals they are meant to fulfill. Once we will have understood why HMIS are used, we will look at a representation of how they are organized.

## 2.1 HMIS definitions

### 2.1.1 Goal approach

A first approach to HMIS is a consideration of the stated goals of these information systems. Figure 1 shows what these main goals are.

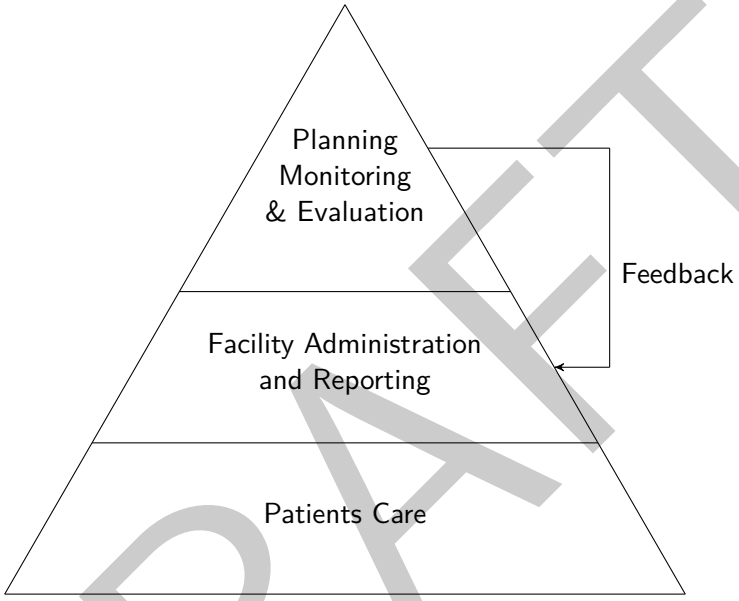


Figure 1: Information needs for HMIS



**Patients Care** Taking care of patients is the primary goal of a health facility. To do so, it is necessary to collect data on these patient, data that will be transmitted (to other services), stored and reused during further follow-ups.

**Facility Administration and Reporting** At facility level, HMIS data is used in daily activities to quantify and forecast needs in health inputs, and to create reports for higher levels of the health system. @spiegel

**Planning, Monitoring & Evaluation** People in charge of the administration of health systems at local or national also need data to monitor activities in the health system, to evaluate the results of interventions, to report to funders or to plan later interventions.

The pyramidal representation of these needs is used to show that these goals fill data needs at different levels of health systems. The different needs for information can roughly thought as being the needs of different type of actors of the health system. Meanwhile, this understanding is not fully true, as at local level, actors will often hold multiple roles and will thus have to use information in different situations. For example, a physician may also be in charge of managing his health facility, and will thus need to plan activities and report on them.

### 2.1.2 Functional approach

A first way to approach HMIS is to describe the four principal functions that are necessary to have a HMIS to run. Figure 2 presents a simplified sketch of the principal functions that are to be filled in order for HMIS to produce useful information.

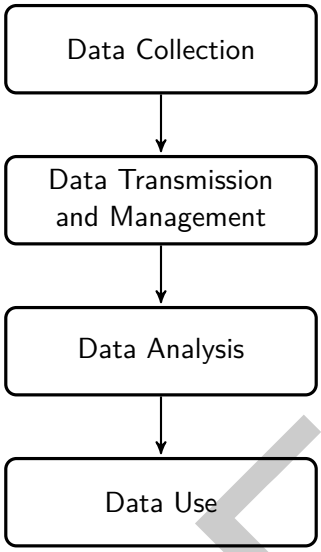


Figure 2: Different functions inside the Health Information Systems

**Data Collection** Primary data collection is essential to the production of any information system. In the case of HMIS, data collection happens in health facilities, and is made by health professionals. Data collected in health facilities can be individual patient data collected in patients files or cards. It can also be a first level of aggregation of this data, as for indicators that are reported on a regular basis by facilities to higher levels of the health system. This reporting usually happens through standardized reports, that are then transmitted by successive aggregation to the top of the health pyramid.

**Data Management** Data collected in health facilities has to be stored and archived, to be later accessed and reused. Data management work can encompass managing paper data, or managing computerized data. Individual patient data will be computerized in Electronic Medical Record (EMR) whereas aggregated indicators are stored in data-warehouses, like the DHIS2 software.

**Data Analysis** Data that is collected and stored in HMIS can then be analyzed. Analysis can be defined as the transformation of data into information. The results of data analysis can be varied, from collection of graphs and maps that can constitute dashboards, to the results of complex models that provide evidences of causality.

**Data Usage** Information is the end product of the HMIS, and is used by decision makers or health workers to achieve their tasks. For example, a nurse in a health post may need the monthly consumption of a health product to place an order. A District Health Officer may consider the evolution of monthly number of cases of malaria in his district to plan malaria prevention activities. At national level, a worker at the Ministry of Health may use the number of patients tested positive for HIV to design grant applications for the Global Fund.

Even though the schematic representation of this functions is linear, it should be noted that this linearity is not true in practice. Even once data has been analyzed the results of this analysis has to be archived, and transmitted to the information end users. In some situations, data can be

used in its raw form. For example, a physician may use a biological result that he has received in a raw form. Finally, for some data, data collection won't happen inside the health facility. For example, population survey results may be used to plan and target some interventions, but primary data collection will have happened outside of the health system, and the first function to be used in the HMIS will be the data management function.

The way a program considers and plans each of these functions will define how a specific HMIS will work. We will now present three common approaches to HMIS.

## 2.2 HMIS strengthening strategies

Depending on the HMIS model that is used, programs will implement different type of HMIS strengthening approaches. Programs who privilegy a jigsaw puzzle approach to HMIS will tend to focus on standardizing procedures and methods for data collection and data analysis. Meanwhile, programs who privilegy a pixel approach to HMIS will tend to favor solutions geared towards the implementation of new and performing data collection tools.

**The institutional approach** operates under the assumption that all functions of information systems should be geared towards and submitted to the end information users. This approach tends to be extremely normative as any activity in the information system has to be oriented towards one main predefined goal. In doing so, this approach undervalues the benefits of both the integration of external data, and the positive externalities data collection and analysis may have on multiple users.

**The technological approach** relies on the assumption that collecting data and making it available is a sufficient enabler for all other functions of information systems to operate. In this sense, a direct link is made between an information need and a data gap. This approach comes at a cost, and provides only limited benefits if it is not supported by improved data analysis. These solutions tend to provide highly specialized and siloed data collection systems.

We argue these two approaches focus on the most expensive ways to strengthening HMIS (data collection and systemic reforms), and are emphasizing the design of systems and tools that are specific to precisely defined data needs, thus limiting the possibility to implement secondary data usage and the positive externalities of their interventions. The archetype of these pitfalls are the well known parallel and siloed data systems present in many developing countries health systems.

Meanwhile, some of the most significant successes in the strengthening of health information systems in developing countries have been reached precisely through the strengthening of their middle tier. The District Health Information System (DHIS2) project has become a pervasive system to store and organize data collected in developing countries health systems. The DHIS2 approach to health information is based on the organization and storage of multiple data types and sources in a generic data warehousing approach. Its versatility and its ability to adapt to different contexts and data has made it increasingly used in multiple context, thus arguably improving the storage and the availability of health data in low resources countries. Other approaches geared towards the promotion of interoperability of different dimensions of data systems, such as the Open Health Information Exchange framework are also gaining traction.

If these approach have provided efficient solutions to organize and access HMIS data, there is still a lack of solutions to analyze and use HMIS data. Indeed, the high dimensionality of HMIS data and its average low quality make it essentially hard to analyze using standard methods available in developing countries health systems. We will now describe the research project developed to explore ways to analyze this data.

We want to ancher hospital data in a multiplicity of data sources that are available to health professionals to create evidence and make decisions. This data can be metadata that is routinely generated during data collection exercises. It can also be data generated by other organisations of the community. Finally, it can be similar data generated by different actors.

We will explore and propose ways to use these different types of data. First, we will

## 2.3 Approach and research questions

This project aims at developing new tools to use heterogeneous data sources to provide information for health systems deciders. We will thus explore methods for *data hybridization*

The generic question we ask is : How can data currently routinely used data sources be used to provide actionable information for decision makers ?

More precisely, we will ask three main data analysis questions :

- How can metadata collected in an EMR be used in EMR data analysis ?
- How can different non standardized HMIS sources be mapped and jointly analyzed ?
- How can multiple data source be integrated to HMIS and analyzed to provide information at local level ?
- How do decision makers in health systems consider HMIS generated data, and how does it influence the way HMIS are engineered ?

Our method is taken in a decision theoretic framework. We do not aim at providing substantial knowledge with the results of our analysis, but rather we aim at providing information that will emopower decision makers to take decisions.

Cadre bayésien

We operate in decision theoretic framework, our end objective being to inform decision making for the

Approche Probabilistique des données

Control Charts

Each of these questions explores a different aspect of how standard HMIS data analysis can be expanded to produce useful information with HMIS data. Metadata like the times of creation and savings of EMR forms are indeed seldom used. Meanwhile they provide useful information on how data is collected, and on the working patterns inside health facilities. Our second question explore a different problem, which is often taken as a question of interoperability. Indeed the multiplicity of programs and actors working in many health systems generates a multiplicity of indicators used, that are often related but not identical. There is a need for simple and effective methods to map and conjointly analyze data from different HMIS systems. Finally, HMIS should be useful for local level analysis and decision making. Meanwhile, HMIS data is seldom useful on its own and has to be integrated in larger analysis frameworks to produce interesting information. This integrating can be difficult at national level, but it is even more complicated at local levels, has mapping precisely different data becomes more and more of a challenge at small scale.

We also aim at providing a critical evaluation of the way HMIS are thought of in developing countries societies. Authors like Alain Desrosières and Ted Porter have shown how statistics and computation have come from and generated different cultures of public action in modern societies. The ambivalence of numbers as descriptors or norms has an influence on how information systems are thought of as top-down normative systems instead of knowledge systems. We aim at interrogating how this perception has its roots in long term local historical trends as well as in the tradition and methods of quantitative public health.

To answer these question, we will conduct four distinct research projects.

**Aim 1** Evaluation of the benefits of improved data collection for HIV patient care in Kenya.

**Aim 2** Test of multiple semantic approaches to interoperability in Bénin.

**Aim 3** Definition and test of a local malaria elimination metric in Namibia.

**Aim 4** Analyze of the theory and practices of Health Information Systems for national decision makers in Mali.

These aims have been designed to provide insights on the problematic posed for each HMIS goal and function. We will now describe each of these aims in more details.

# 3 EMR and individual health[1]

A first aim will be to understand how data collection itself impacts quality of care. As we postulate that data collection is not a neutral activity, we want to look into how primary data collected in HIV care setting can impact the outcome of care and organizational capabilities of HIV services. The case we will explore for this project is provided through a project implemented by ITECH in Kenya.

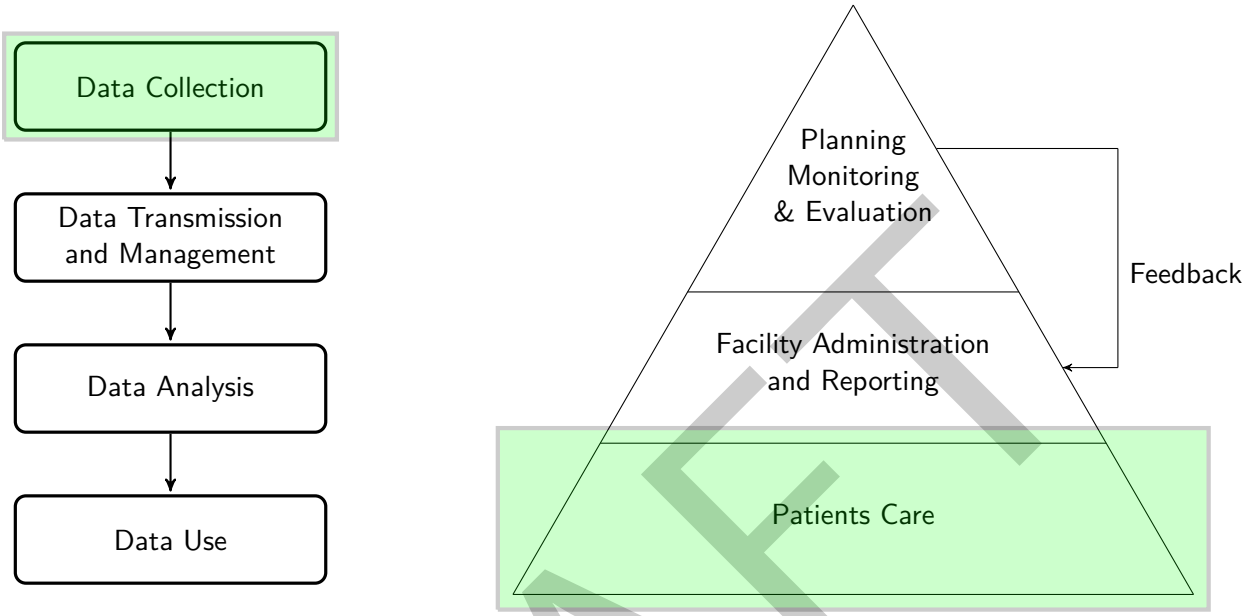


Figure 3: Objective one definition

## 3.1 Setting

In Kenya, I-TECH has implemented an EMR for HIV care, called KenyaEMR, in 341 facilities. The evaluation of this program is currently being carried out. One objective of this evaluation is to assess the effectiveness of KenyaEMR implementation. This effectiveness will be evaluated on two dimensions:

1. Improvement of reporting quality in facilities after KenyaEMR implementation
2. Improvement of quality of care metrics after KenyaEMR implementation

## 3.2 Data

Kenya's legal framework for protection of confidentiality of personal health information prohibit transfer of individual patient-level data from any health care facility, even if the data is de-identified. For this reason, the data we will use for this evaluation will be indicators of quality of care, aggregated monthly at facility level, and used for Continuous Quality Improvement (CQI) (see section 3.5). These indicators will be aggregated on site in Kenya and transmitted for data analysis.

To monitor the maturity of implementation of KenyaEMR (see section 3.3), we will measure the delay in data entry using metadata stored with KenyaEMR forms, with time stamps for form creation. We will also trace utilization of reporting features of KenyaEMR by using time stamps linked to the use of reports generation. All this data will be extracted and transmitted in raw form for analysis.

To measure the quality of the reports produced for different periods (see section 3.4), we will consider counts of number of forms entered for a given period, and mean completeness of entered

[1] This section is heavily based on KenyaEMR evaluation protocol.

forms. These will be aggregated on site and transmitted for data analysis. We will also use results from Routine Data Quality Assessments (RDQA) that have been conducted in different sites with KenyaEMR implementation. Data for these RDQA are collected in Excel format, and will be used as an external measure of the quality of data entered in KenyaEMR. In the remaining of this document, we will thus use the following terms:

- Patient data refers to the data collected by health workers during patients' visits. They are stored in paper patients' files, or entered in KenyaEMR forms. We will thus refer to paper patient data or to electronic patient data. This data will not be directly used for analysis in this project.
- CQI indicators refers to aggregated indicators used to measure quality of care.
- CQI Report refers to a set of CQI indicators computed for a specific month for a specific facility.
- DHIS Report refers to the MOH 731 and MOH 711 reports. We will differentiate between paper reports for which the data and the computation of indicators have been made without any digitalization of patients' data, and electronic reports for which patients' data has been digitalized. We will be able to use the paper reports as they have been entered in DHIS2 or other data collected by health districts administrations.
- Patient Forms Metadata refers to the metadata generated by KenyaEMR when patient forms are entered. The metadata used should mainly be timestamps related to time of data entry.
- Reporting Metadata refers to timestamps generated by KenyaEMR when different types of report are generated.
- RDQA Data refers to raw data collected during RDQA exercises.

## 3.3 Implementation maturity

We distinguish three different periods in the implementation of the EMR. Each of these periods is characterized by different ways the data is collected, entered, analyzed or used. For each of these periods, we will also have access to different types of data. We will describe the characteristics of each of these periods, and present a strategy to categorize the available data in each of these periods, using DHIS and CQI reports and metadata.

### 3.3.1 Paper Based

In the first period, no patient data entry is made in the facility. Patient data is collected in paper files, and reports are computed manually using these files. In the meantime, health workers can only use paper data to follow their patients.

The data we will be able to access from this period is:

- The patient data that will have been retrospectively entered in KenyaEMR
- The paper reports that will have been entered in DHIS2 or other reports available from health districts administrations

### 3.3.2 Retrospective Data Entry

In a second period, data entry has been implemented in the facility. The backlog of paper patient data has to be retrospectively entered, and current patient data is also entered in KenyaEMR after a delay. During this period, health workers will still refer to paper data to follow their patients. This is the Routine Data Entry phase (RDE).

The data we will be able to access from this period is:

- The patient data entered in KenyaEMR
- The metadata for patient data entered in KenyaEMR
- The CQI and DHIS reports computed from this data
- The reporting data metadata
- Evaluations of data quality from RDQA

### 3.3.3 Point of Care

In a third period, the patient data is entered either by the health worker or by a specialized data clerk based on patient data collected on paper by the health worker, in quasi real time with the medical consultation. We call this phase the Point of Care (POC) phase.

The data we will be able to access from this period is:

- The patient data retrospectively entered in KenyaEMR
- The metadata for patient data entered in KenyaEMR
- The CQI and DHIS reports computed from this data
- The reporting data metadata
- Evaluations of data quality from RDQA

### 3.3.4 Transition periods

There may be some overlaps between different periods. For example, the limit between the paper collection and routine data entry may not be clear cut, as some facilities may have tried to start entering more recent patient forms during the RDE period, to be on top of the work quickly. Similarly, some facilities may have been at the same time doing retrospective data entry for some forms, and POC data entry for some others, depending on the organization of care.

To take this into account, we will need to consider overlapping periods for different aspects of the data.

- Data quality: the process to collect and enumerate patient data is identical in paper based period and RDE period. Meanwhile, in the POC period, data is possibly directly entered in KenyaEMR, without using a paper form. Also, rapid data entry may allow to go back to the HW to complete missing data, or to correct unclear information.
- Report computation quality: Once the data entered in KenyaEMR, the reports can be computed automatically. Thus, the quality of computation of reports will be identical in POC and RDE, but will likely differ from the Paper Based period (see Section 3.4 for more details on Reporting Quality).
- Quality of care: in the paper based period as in the RDE period, HW can only access patient data through paper files. They thus can't use automated reminders, or summary information offered by KenyaEMR. Meanwhile, starting in the RDE period, some reports can be edited through Kenya EMR that would allow health worker to better track late and defaulting patients, and thus would allow them to pass reminders calls, or plan lab tests. We would thus anticipate to see a slightly improved quality of care for RDE period compared to Paper Based period, and to see an additional improvement for POC period compared to RDE period.

Based on this periodization, we will want to test three main hypothesis:

1. Observed data quality is similar in paper and RDE period, and better in POC period.
2. Computation quality is bad in paper-based period but then improved in RDE and POC periods.
3. Quality of care is worst in paper-based period, improves in RDE and is best in POC period.

### 3.3.5 Methods for periodization

For each facility included in this analysis, we will have to define when they enter or exit each of these periods. To do this, we will use programmatic data collected by I-TECH staff to monitor the implementation of KenyaEMR, and time stamps associated with forms entered in KenyaEMR, and Building on the characteristics of the different periods, we will categorize the different dimensions of the data collection and use separately:

1. Data Quality: The passage between stage 1 and stage 2 of data collection will be tracked looking at the delay of data entry of forms. Looking at the distribution of this delay, and using I-TECH monitoring data for confirmation, we will define a threshold to define stage 2 data entry. We will also use comparison of data completeness between different periods.
2. Report Computation: The passage between stage 1 and stage 2 of report computation will be tracked looking at the source of the reports available for the facility. Existence of reports from DHIS2 or similar source that were not produced using KenyaEMR computation will lead to the categorization of the stage of report computation as stage 1. Reports computed with KenyaEMR will lead to a categorization of the period as a stage 2 for report computation. The categorization will be validated with data from I-TECH monitoring, and by a comparison of results reported in DHIS2 and results computed for the same month from KenyaEMR.
3. Data usage: The passage between stage 1 and stage 2 for data usage will be used considering metadata from different reports, and delay of data entry. A different threshold as the one used for data quality will be used to categorize a facility as stage 1 or 2 for quality of care.

Using available data to explore this different dimensions, we will be able to categorize each facilities' reports into its corresponding period. As we anticipate some exceptions due to unclear transition periods, we will design a continuous index of maturity of implementation of KenyaEMR, to be included in latter stages of the analysis. Depending on the results of the exploratory work, we will use a continuous index or a discrete periodization of the intervention.

## 3.4 Reporting quality

To estimate the impact of KenyaEMR on the quality of reporting, we will compare aggregated monthly reports on HIV activities in facilities produced before and after implementation of KenyaEMR. Evolution of reporting quality involves two evolutions: amelioration of primary data quality, and amelioration of report computation quality.

We will measure data quality by looking at specific metrics:

- Proportion of data fields used to compute reports that have contain valid data
- Mean monthly number of visits by active patients

We will also use RDQA data to evaluate the quality of the data. Using RDQA results as training results, we will explore systematic classification of data quality based on reports indicators and patient forms metadata distribution.

We will then measure the evolution of data quality between RDE and POC data in KenyaEMR and we will perform simple comparisons to evaluate changes in data quality when entering data directly in computerized form.

Also, we expect computation quality to have multiple measurable impacts:

- Greater coherence of indicators involving longitudinal data analysis,
- Greater coherence of indicators involving multiple data sources
- Greater coherence of indicators evolution in time, as computerized computation will be exactly the same in time
- Greater coherence of indicators between facilities, as computerized computation will be exactly the same in all facilities.

We will compare reports generated for the same facilities and same months, in Period 1 and Period 2, and we will perform simple comparisons to evaluate changes when using standardized computation methods.

Based on these two dimensions of reporting quality, we will finally design an index of reporting quality that will be used in subsequent analysis. Quality of reporting will then be modelled using, using facility characteristics as covariate, and the index of maturity of implementation. The coefficient associated to maturity of implementation will be considered as the measure of the impact of KenyaEMR on reporting quality (see section Quality of Care and patient health outcome4 for presentation of the modeling strategy).

## 3.5 Quality of Care and patient health outcome

Using existing aggregate-level longitudinal data from KenyaEMR sites, we will retrospectively compare quality of care and patient health outcome indicators during each period of the EMR transition. The specific quality of care and patient health outcome indicators to be examined will be determined in collaboration with CDC and the MOH, based on commonly-used indicators within Kenya and globally. A list of these indicators can be found in Annex C.

To model the association between using KenyaEMR and the level of each quality of care and patient health outcome indicators, we will use Generalized Estimating Equations (GEE) that will allow us to take into account the temporal correlation of observations. Covariates that we will introduce in this model include:

- Facility type
- KenyaEMR implementation maturity index
- Reporting quality index
- Number of patients followed for HIV in the facility
- Number of HW involved in HIV care in the facility
- Time trend

The coefficient estimated for KenyaEMR implementation maturity index in this model will be considered as the measure of the impact of KenyaEMR on the quality of care and the health outputs of HIV patients. The index will be introduced in continuous form or in dichotomized form. Alternative proxy of KenyaEMR implementation will also be tested such as period of implementation as defined for quality of care in table 1.

## 3.6 Timeline

Figure 4 presents a timeline for the realization of this objective. Even though the data collection process could be a sort of blackbox, we expect this paper to be finished by February 2017.

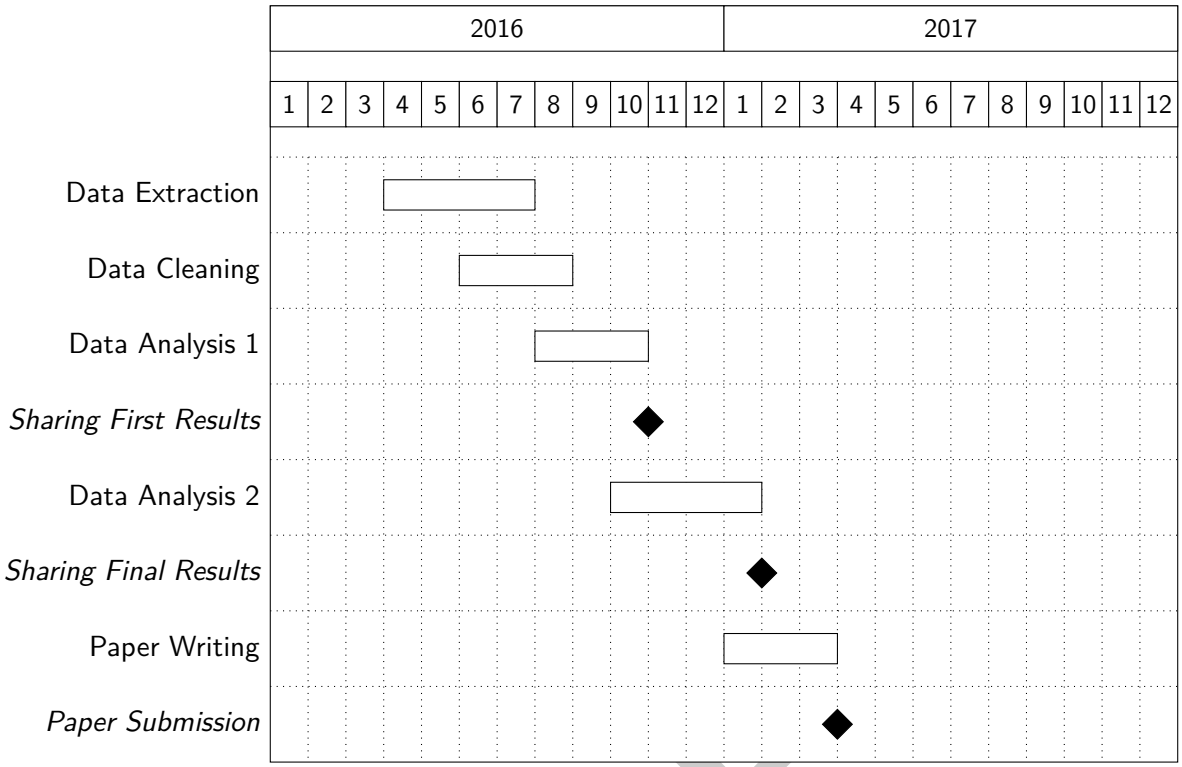


Figure 4: Gantt Chart for Paper 1

# 4 Using divers readily available data sources to map population in the Sahel

Knowing where people live is essential to the design and the evaluation of policies or interventions targetting populations. Population maps is thus an important tool for policy makers, or for actors of social policy or relief actions. Meanwhile, in data poor settings, the quest for precise and exact mapping of population has long proved an illusory goal, for reasons that touch both on the paucity of data, and on technical mapping issues.

This project is exploring an innovative approach to mapping of populations in resource-limited settings. Using voters registration lists, a data source seldom used in demography, we will endeavour to offer a map of population in Niger that will be easily callable and usable by different actors intervening in Niger. To do this, we will challenge current practice in population mapping, and will aim at producing a point map, more adapted to daily usage than raster surfaces.

In the rest of this documents, we will first take a quick historical detour to better understand how our work is justified by an under-investment in population data sources in former French colonies, and how it is also situated in a long standing tradition of population mapping methods. We will then present the data sources we are using, and will discuss the methods use. Finally, we will describe the outputs we hope to produce by the end of the quarter.

## 4.1 Mapping populations in sub-Saharan : a historical challenge

### 4.1.1 Data issues

The first population map can be traced to Crome's 1785 *Groessen-Karte von Europa* [**?** ], in the time at which Foucault dates the "discovery of population"[**?** ], thus hinting to how the concept of a governed population is consusbstantial to the localization of this population. Meanwhile, the production of these maps necessitates the availability and the processing of amounts of data that grow exponentially with the desired level of details. In countries that were formerly colonized by the

French powers, the investment necessary to produce this geolocalized data on population has seldom been done. In the first movements of colonization, maps were mainly produced by the military, and in 1902, Paul Pelet's very first *Atlas des colonies française*[**?** ] did not include a lot on information outside of topographic data[**?** ]. Additionally, there was a radical choice made to use spellings for places in colonies that were adapted to metropolitan French rather than to local languages[**?** ] The second period of colonisation was more interested in economic and commercial exploitation, and thus, the 1922's *Exposition Coloniale* showed some progress in the mapping of geologic and natural resources, but still no population mapping was produced[**?** ]. In 1929, a first map of the population density was produced for Sénégal, using indirect methods to estimate the size of communities, using the amount collected for personnal taxes by local authorities[**?** ]. This undervaluation of human data was prolunged in the structures that were left after the decolonizations[**?** ], and reinforced by structural underinvestment in statistical systems in sub-saharan Africa[**?** ]. As a result, the main source of data for population size estimation in most sub-Saharan countries are census, that are performed at best every ten years. In countries with dynamic demographies, this means that most population estimations used are either outdated, or heavily modelled for projection purposes. There is a blatant need to find new data sources to estimate the size of local communities, that are updated more frequently, and are valid at local level.

### 4.1.2 Methodological questions

Mapping populations is a tricky task. In contrast to geological data or natural features, population is non continuous in space, and is changing very quickly. In 1935, Fawcett discerned three facts one may want to describe in a population map[**?** ] :

1. The actual number of the people within given areas
2. The density of the population
3. The grouping, or arrangement, of the population.

Each of these facts require a different type of mapping. The number of people in an area can be communicated through discretization of space, and association of a number of inhabitants in each subzone. Density mapping necessitates a more continuous approach to mapping, in terms of scale and of color scale. Arrangement of population can be shown through the plotting of points representing settlements, and with size representing the number of inhabitants. Each of these require different amounts and nature of primary data, and different computational approaches.

The latest advances in population mapping are geared towards the production of density surfaces, presenting a continuous description of where populations live on a territory[**?** ]. This new approach is made possible by the availability of large aggregate datasets for land use, and other usable co-variates[**?** ], and the computational ability to interpolate these different data sources for popualtion distribution[**?** ].

Meanwhile, this approach presents two main problems that make its result of little use in country like Niger. First, the output format if these maps, a raster of population distribution, is of little use at a local level, where actors typically use places name and not GPS coordinates. Second, in countries with little urbanization, and poor population data, these rasters end up displaying an overlay of covariate layers more than they present a credible distribution of populations.

Moreover, if this top-down approach is useful to provide macro-level perspectives on population distribution, it is not useful for local level use. Local actors typically think in terms of places names, more than in terms of GPS coordinates. Additionaly, attractivity of a urban center for rural population is hard to model through macro level covariates, as it often depends on such factors as tradition, habits or administrative border drawing, as is evident in Figure 1.

In order to offer more useful maps for local actors, we will use an approach supported by very minimal modelling of primary population data distribution, and geared towards the anchoring of population in callable localisation names.

Figure 5: Mapping of Niger population from AFRIPOP

## 4.2 Project concept

### 4.2.1 Data sources

**Voters list as a demographic Datasource** A data source that is, to our knowledge, seldom used to inform population mapping for public health purposes, is voters registration lists. There is meanwhile a case to be made for the use of voters' registration data to estimate size of populations. By definition, voters' registration should aim at being as complete as possible a register of adults in the nation. Moreover, in most democracies, some form of national elections are held at least every five years, leading to an update at least partial of voters' registrations. In sub-Saharan Africa, between the years 2015 and 2016, 27 countries were supposed to hold national elections, leading to a theoretical registration of more than half of the adult population of the continent. Finally, for transparency and accountability reasons, electors registries are usually supposed to be easily accessible.

Due to the sensitive and political use of these data, the quality of voters registries are often described as not being trustworthy. On other hand, for the same sensitivity reasons, voters registries are receiving a very high level of scrutiny from many different actors, and are audited sometimes multiple times before validation. This level of scrutiny before validation is much higher than the attention given to a lot of studies or other often used data sources.

**The Niger 2016 elections voters registry** In Niger, presidential and parlementary elections were held in February 2016. Voters lists were updated during the second half of the year 2015, under the supervision and control of a mission of the Office International de la Francophonie (OIF). The operations for registration of voters were conducted during the third quarter of 2015[2]. A first version of the voters list was published on December 21, 2015, tallying 7,569,172 voters, out of 8,569,309

[2] http://www.ceni-niger.org/article-region/more-24

that were expected based on the 2012 census[3]

Final lists were validated in early January 2016 after being corrected for some incoherencies noted by the supervisory bodyhttp://www.nigerinter.com/2016/01/le-fichier-electoral-du-niger-valable-sous-reserves/. A final report on these lists was published in may 2016[4]. The Comission Electorale Nationale Independante (CENI) later made these lists fully available on its website, from which we extracted, anonymized and formatted the lists.

**RENALOC and RENACOM** The *Répertoire National des Localités* (RENALOC) is a geolocalized repertory of all localities in Niger. The 2012 version was downloaded as a pdf file from the *Institut National de la Statistique* (INS) website. The tables were extracted in bulk from this file using the Tabula Package, and then processed in Python to recompose the geographic structure of the document. The final data consists in 34507 localities, for which the INS provides the number of inhabitants, by gender, as well as the number of households, and the number of agricultural households. For most of the localities, a GPS coordinate is recorded, as well as the type of locality (neighborhood, village, camp, water well, hamlet).

The 2006 version, named RENACOM, was retrieved in tabular format directly from the INS website.

**OpenStreetMap** Additional geolocalization method will be extracted from OpenStreetMap (OSM), using the python api for OSM.

### 4.2.2 Objectives and Methods

This project has three distinct but complementary objectives :

**Population estimation** We model Niger population, using the voters list by precinct as a primary covariate. Moreover, we model age structure of the population at local level using the age of registered voters. To do this, we combine traditionnal demography and Machine Learning (ML) methods, to get a proper estimation of uncertainty at locality level.

**Name Matching** We already noted that different spelling of locality names are in use in Niger. We have no reason to privilegy on spelling over another for our project. Moreover, we want users of our map to be able to use whichever writing they prefer. Using our four data sources, we will offer a matching of different spellings of same names. This matching will be made using a combination of direct correction, unsupervised and supervised learning approaches. The output of this work will be, at the same time, a database, and a trained prediction algorithm able to offer a suitable guess to users using unknown spellings of localities.

**Locality mapping** The three data sources for localization (RENALOC, RENACOM, OSM) have GPS coordinates for different subsets of localities in Niger, with some common coverage. It appears that RENALOC GPS coordinates are biased, and that OSM coordinates are sometimes rough estimates. We will try and correct GPS coordinates when needed, and map localities based on these coordinates.

Based on these objectives, the output of the project for the quarter will be to produce a simple dashboard server, allowing a simple and usable display of our results. This dashboard will have the following feature :

1. An interactive map of Niger localities, selectable by clicking, or panning for multiple selection
2. An estimation of the population in the localities currently selected on the map

[3] http://www.iinanews.org/page/public/news$_d etails.aspx?id = 98929NL = True$

[4] http://www.nigerinter.com/2016/05/remise-officielle-du-rapport-du-fichier-electoral-au-ministre-detat-a-linterieur-par-le-cfeb/

3. An histogram representing the age structure of the population in the localities currently selected on the map

4. A search box through which the user will be able to search for a given localiy. Unknown localities will return a list of probable similar places.

Additional features of this dashboard may be, in no particular order : the display of uncertainty in population and age structure, the differentiation of recorded voters and extrapolated population, by age group, the ability to explore different modelling options.

# 5 Semantic approaches to HMIS interoperability for external data validation

The second paper of this dissertation regards the management of data collected in hospitals in developing countries. The lack of standardization of indicators computed in different places or by different HMIS makes the use of HMIS data a complicated task. This issue is often handled as an issue of interoperability between systems.

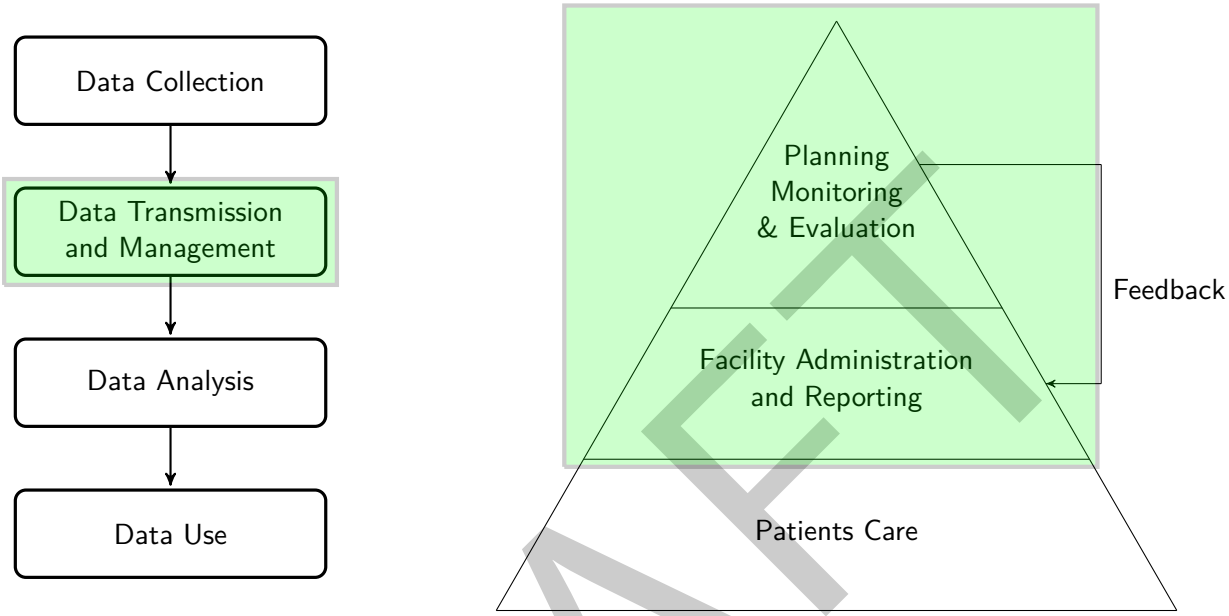


Figure 6: Objective two definition



## 5.1 Introduction to Interoperability work

There are multiple reasons why interoperability between HMIS datasets is a great asset :

**Comparison** Being able to compare the results of different health systems is essential to be able to benchmark these results, or to make different systems benefit from each others' experiences.

**Validation / Completion** When multiple systems are operating in the same place, one can wish to compare results from different system in order to validate the data, or to fill missing data from another system.

**Co-analysis** Finally, pooling results from different systems provides higher power for analysis that can be made on different subjects.

It should be noted that the conditions for interoperability can be seen in different ways for each of these uses. If comparison of results necessitates that measured indicators have quasi identical definitions and methods, it is less the case for validation and completion, where a set of indicators can be used as a proxy to check the coherence or impute values of another data set. Finally, co-analysis may or may not require an exact mapping of indicators from different systems, depending of the subject matter.

There are multiple levels at which interoperability of data systems has to be enforced [**?** ]. At the Syntactic-Technical level, protocols have to be designed and implemented, to ensure that different data-systems can communicate. At the Organizational level, processes have to be implemented to allow the exchange and to define the condition of usage and aggregation of different data systems. Finally, at the Semantic level, qualitative meaning and understanding of the nature of the data being exchanged and compared has to be enforced.

I am interested in considering the semantic level of interoperability. Indeed, I will make the hypothesis that perfect standardization of HMIS indicators across contexts and platforms is an elusive goal, and may not even be desirable, as it is a factor of rigidity for HMIS, which should be able to evolve rapidly. I am thus be interested in defining methods to ensure ex-post semantic interoperability between different HMIS systems, in order to map indicators and with an objective of data validation and missing data imputation.

## 5.2 Research questions

This project is in the framework of a larger project, defining an interoperability framework between different systems. As part of this project, a tagging system has been defined that allows mapping of indicators from different standard hospital indicators sets.

Meanwhile, this tagging approach, if it is effective, has an important entry cost for users. Tagging a set of multiple dozens of indicators can indeed been an harrowing task in a field where quick fixes are the gold standard. We are thus working on two ways to improve this uptake. One is the definition of automatized learning methods that could help users in the tagging task. Another approach is to provide users with sufficiently strong incentives that the tagging work will be worth going through. As actors working towards interoperability of systems are likely to be middle tier actors in the information system, we think empowering them to use the benefits of interoperability for validation of data, correction of faulty data and completion of missing data would be a strong incentive.

I will thus explore methods to validate, correct and complete data sets from routine HMIS, and measure the benefits of indicators matching between different dataset to improve validation, correction and completion. This will be considered looking at different intermediary questions :

1. What is a good metric to assert data quality and completeness of a given HMIS data set ?

2. What is the performance of different approaches to HMIS data validation, correction and completion ?

3. What is the benefit of using mapped datasets on this last metric ?

Our main aim will be to gauge the credibility of a data value or a dataset. Additionally, we want to provide some evidence as of what the source of error or missingness in a dataset is. There are three main situations we want to be able to differentiate between :

**simple outlier** are situations when an isolated data value in a dataset is wrong, for one facility once. These are the situations that are the most commonly recognized as outliers.

**outlying report** are situations when all values of one report appear to be off. This may be due to an update in the tools or methods for some indicators, or to the training of a new Health Worker in the facility who does not fully comply with usual ways to compute indicators. To identify this type of issue, it is necessary to compare

**outlying facility** are situations when one facility is consistently reporting numbers that are different from surrounding facilities. This may happen when structural conditions in one facility are leading to discrepancies in data collection, or in data computation.

Our general approach will be to attach to each data value a probabilistic value for its credibility. The combination of these credibility measures for a dataset will give an overall estimation of the credibility of this dataset, and the type of error it may suffer from. In a third and last step, we will explore methods for imputation of data values with low credibility or for methods with low credibility.

## 5.3 Data Validation / Method

We will test two approaches to do so :

**Error prediction** Using the validation dataset from OpenRBF, we will try and predict wrong data values using a simple predictive approach and bagging different Machine Learning Classification methods. Result of this approach will be a probabilistic assessment of data quality for each indicator value.

**Variable imputation** Using all available data, we will impute all data value and get a posterior estimated distribution of the value.

### 5.3.1 Error Prediction

For a each indicator value, we will model the probability that the value is right. We will use a logistic model specified as :

Using *Forward model selection* and training models on data on verified values (cf. 5.6), we will find the best models to predict data value errors. We will then use this model out of sample to compute the probability a value is true on each value.

### 5.3.2 Variable Imputation

For a given indicator $X_{i*}$ measured at time $t*$ in facility $f*$, we will compute an imputed distribution based on all other information available in our data. A general representation of this approach could be written as:

$$\widetilde{X}_{i^*,t^*,f^*} \sim \mathrm{D}\left(f\left(X_{i,t,f}\right)_{(i,t,f)\neq(i^*,t^*,f*)}\right)$$

where $\mathrm{D}$ is an unspecified distribution derived from a function of all available data excluding the data point $\widetilde{X}_{i^*,t^*,f^*}$.

Using verified data as our validation set, we will define a threshold for credibility of data that we will later use as a decision rule for considering data as regular or outlying (see 5.4).

We will test different approaches for this model.

## 5.4 Credibility Pattern Screening

The validity probability of each data value will be pooled at report, facility and district levels. An analysis of the distribution of credibility at each level of aggregation will be made to characterize the type of data error pattern in one of four categories : No error, Single outliers, Outlying report or Outlying facility.

**No error** will be situations in which no data values will be over the fixed threshold of acceptable credibility.

**Single Outliers** will be situations in which less than 10% of data values in a single report for a facility will be under the fixed threshold of acceptable credibility.

**Outlying report** will be situations in which more than 10% of data values in a single report for a given facility will be under the fixed threshold of acceptable credibility.

**Outlying facilities** will be situations in which more than 10% of data values will be under the fixed threshold of acceptable credibility in more than two consecutive reports for the same facility.

## 5.5 Performance Metric

Validation of these methods will be made using cross-validation. We will first evaluate the data validation framework independently of the indicators mapping framework. We will then test the performance of the combination of different indicator mapping methods and data validation methods.

## 5.6 Data

This work will be conducted in collaboration with the Belgian startup Bluesquare. Bluesquare has developed a data system for management and validation of Results Based Financing indicators, called OpenRBF. This solution has been implemented in Bénin in XX facilities in YY départements. Indicators are collected on a monthly basis in OpenRBF, and a data validation system is in place, to check the accuracy of reported indicators.

In the meantime Bénin has been implementing and using DHIS2 nationwide since AAAA. There is considerable interest in mapping indicators from the two systems, and using the two systems as validation and completion solutions.

## 5.7 Timeline

Figure 7 will present a timeline for the realization of this objective.

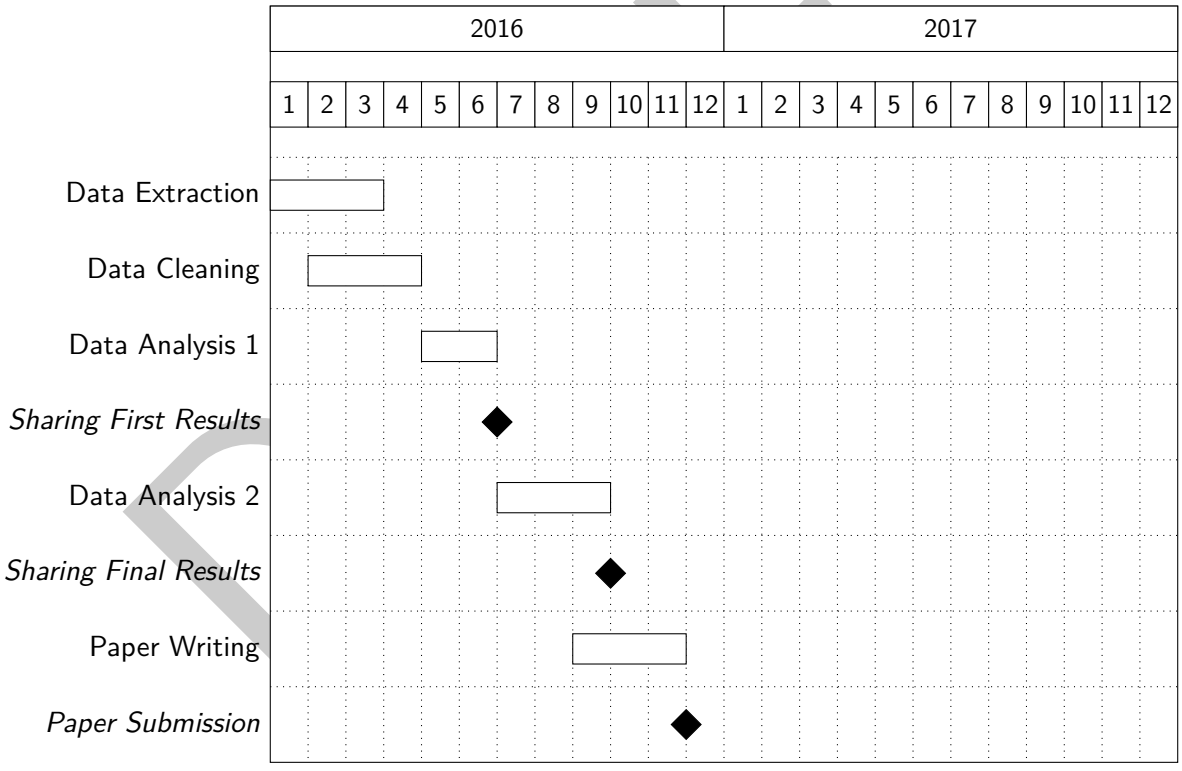


Figure 7: Gantt Chart for Paper 2

# 6 Data Use

construction d'un espace politique d'équivalence et de codage

"Les outils statistiques permettent de découvrir ou de créer des êtres sur lesquels prendre appui pour décrire le monde et agir sur lui"

## 6.1 Information in Health Systems

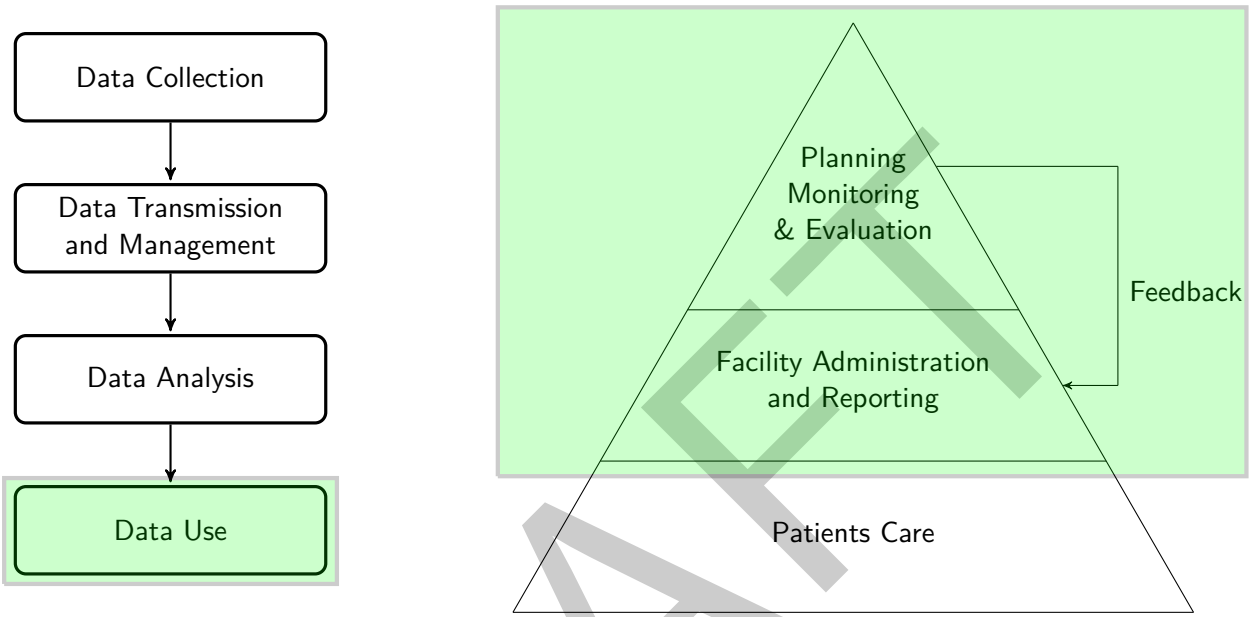


Figure 8: Objective four definition

The use of statistical information for the management of complex organizations has evolved since the beginning of the XIXth century. Since the invention of population by XVIIIth century demographers[DESROSIERES], and the integration of numbers in the political and administrative language in the second XIXth century [PORTER], multiple types of information have been used for the orientation of public policies and the administration of public services. Meanwhile, the rise of epidemiology and the critalization of a body of knowledge around the institutions in charge of the defense of Public Health helped creating a specific Public Health oriented spin on quantitative information for health systems.

The use of data for policy making is a combination of data sources, statistical methods, and political or social norms, that will define the conditions of utilisation of statistical evidence for policy making. Finding the proper data source, being able to analyze it and incorporating the results of this analysis in a political process is essential to the proper use and utilization of information systems. In this regard, Alain Desrosières has shown how two traditions have been cohabiting in the early ages of the production of social statistics[**?** ].

> The first tradition is administrative, and is based on political science and the law, on the German Staatenkunde, from the time of Conring and Achenwall. It is more taxonomic than metrological: it is designed to classify facts systematically rather than measure them, which is the essence of the other tradition, the "English" tradition. The latter, inspired more by the natural sciences and by progress made in measurement and probability theories, is a distant relation of the English political arithmetic of Graunt and Petty.

Desrosières later shows how these two traditions have bee reconciled in the modern figure of the statistician, at the same time administrator and scientist. It is useful to keep considering this tension

when thinking about maturing statistical systems like HMIS. Being able to distinguish between situations when actors of HMIS are acting as administrators, and when the position is that of a metrician is essential to understand HMIS issues and offer informed solutions.

This distinction is essential at many levels. The whole debate around the level of uncertainty that is bearable around a measurement is not only important for statisticians. Choosing a given approach will have an impact on how primary data will be collected, how it will be analyzed, and how it will be used. In many usages of HMIS, complete enumeration is deemed necessary, but this can be discussed. What is the level of confidence one can bear around the estimation of a stock of drugs ?

In sub-Saharan Africa, this tension is reinforced by a political tradition that has been structured around the colonial state. The structures and political traditions coming from this specific have complex relationships with the notions of uncertainty and control. Moreover, these structure are reenacting the colonial culture of exogeneous power structure, through that international actors take in the governance of African country.

This last paper will aim at understanding how some program managers in Mali the data available in HMIS, and how it impacts the way they think, design and implement HMIS programs. We will interrogate the notions of uncertainty, sampling, control and norms for this managers, and their appreciation and use of numerical evidence.

UNFINISHED

Reflections on social conditions of HMIS data usage / politics of administrative statistics.

Data is not produced to create knowledge, but to implement disciplinary monitoring. Thinking mainly in terms of indicators.

Figure 9 will present a timeline for the realization of this objective.

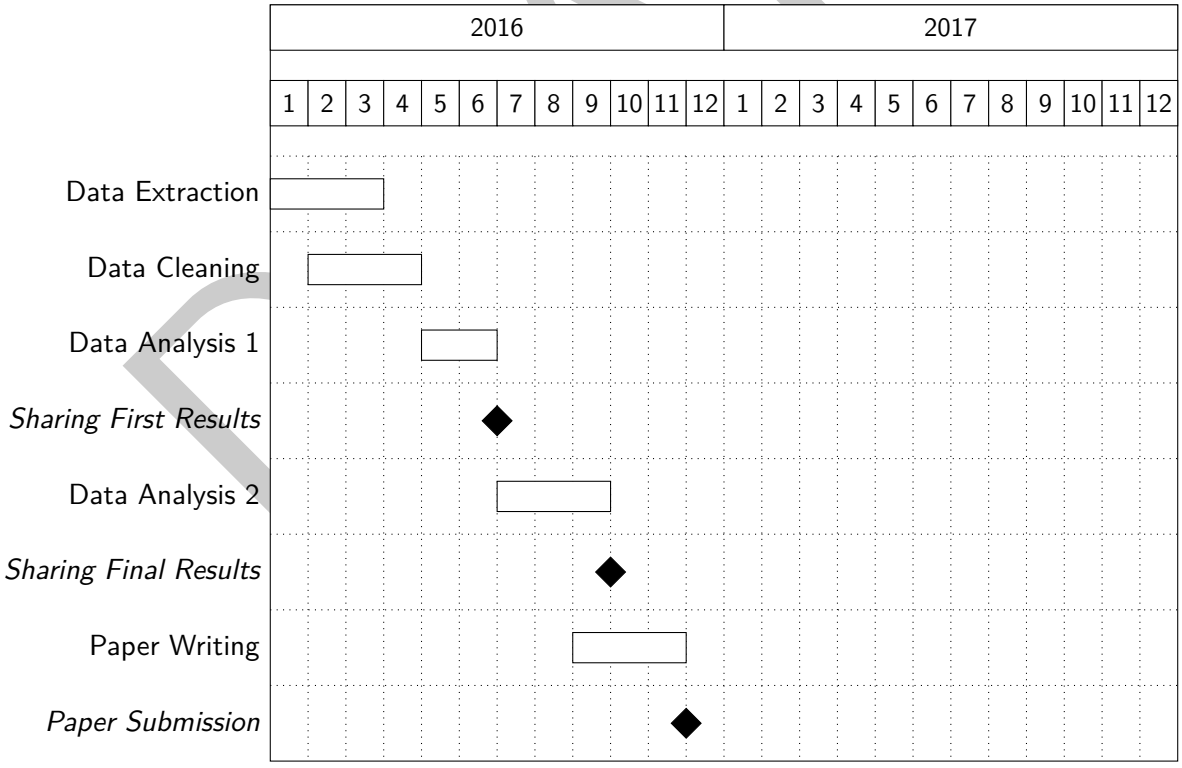


Figure 9: Gantt Chart for Paper 4

# References