Health Informatics Hub

An approach to developing National Data Repositories

Fitti Weissglas, MBA, MSc
Technical Director for Global Health Informatics

Michelle Moghadassi, MPH
Informatics Programme Manager

UCSF Institute for Global Health Sciences

Background

National Data Repositories (NDRs) can best be seen as heterogeneous, integrated datasets spanning data from multiple sources within the health sector. An NDR is not just a "dropbox" for data, as it is sometimes perceived, but rather a carefully curated repository with the ability to host data from a wide array of sources, with complex automated data cleaning, management, integration, processing and analysis processes.

Data Integration

Data integration (or linkage) is the ability to compare and use data from two or more disparate data sources, which is a key feature of NDRs but also presents unique challenges, such as using data from different data owners, imperfect or incompatible data not originally designed to be linked, or incompatible data definitions. However, a successful NDR is a powerful platform that allows us to answer questions such as:

- Does my service uptake for a particular drug match the procurement, and is the expenditure within budget?
- Are there any abnormal patterns in diagnoses that cannot be explained by seasonal patterns alone? Are these patterns confirmed by other sources, such as an increase in diagnostic data or zoonotic data? Might these be indicative of a disease outbreak?
- Is my annual health supplies procurement plan in line with the expected disease burden and budget?
- Are we reaching epidemic control with our HIV epidemic, and how well is this supported by the different available data sources?

Data sources

To achieve data integration, careful consideration should be made to the data sources that can be candidates for inclusion. Examples of data sources may include patient data, such as electronic medical record data; surveillance data; research data; programme monitoring data, such as HIV (PEPFAR) indicators; supply chain data, such as stock, procurement and utilisation of medical supplies, including drugs; laboratory and other diagnostics data; disease registry data, such as cancer registries; human resource data, such as training and capacity data on the health workforce; client registration data, such as basic demographics, ideally with some kind of unique identifier; financial data, including budgets and expenditures; and zoonotic data, including data on livestock, animal diseases and abnormal patterns.

Single versus multiple, focused NDRs

Countries may decide they want a single fully integrated NDR, but sometimes the political climate is more conducive for multiple, focused (thematic/programmatic) NDRs. For example, it is possible to develop multiple NDRs, one for HIV, one for cancer, one for surveillance — which can be integrated at a later stage. This approach is sometimes easier leading to faster concrete deliverables. The approach described in this document would apply to each single instance of an NDR.

Key functional requirements for an NDR

Although each NDR has its own set of unique requirements, there are some generic functional requirements applying to most NDRs.

Data governance and confidentiality

First, from a governance perspective, an NDR should guarantee privacy, security and confidentiality of any data. It should also remove, to the maximum extent, personal identifiers and other PHI not immediately needed for analysis. Ideally, an NDR should be backed by policies and governance structures, through, for example, a steering committee. NDR use should be measured through metrics.

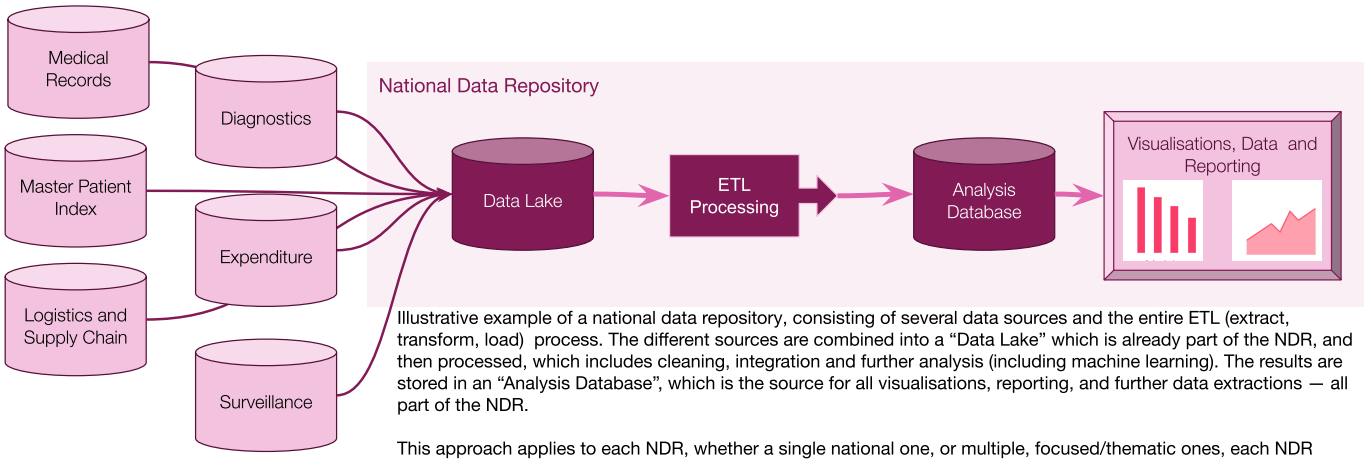
Loading data into the NDR

An NDR should be able to absorb any type of data, regardless of its structure, quality, or update frequency. It should also incorporate this data using any type of input method, including health information exchanges (HIEs), direct imports of files, USB drives, shared folders, etc. Anything that can get the data into the NDR should be accepted.

Processing and analysing data

An NDR should provide routine cleaning, linking, and pre-analysis routines storing intermediate results for faster retrieval, further processing, analysis and reporting (see below). It should also be able to link data across key axes (called dimensions) across datasets, i.e. standardise and link individual

A National Data Repository at a glance



Illustrative example of a national data repository, consisting of several data sources and the entire ETL (extract, transform, load) process. The different sources are combined into a "Data Lake" which is already part of the NDR, and then processed, which includes cleaning, integration and further analysis (including machine learning). The results are stored in an "Analysis Database", which is the source for all visualisations, reporting, and further data extractions — all part of the NDR.

This approach applies to each NDR, whether a single national one, or multiple, focused/thematic ones, each NDR would in essence look like this.

patient data across data sets, link facility data and other geographical data, and time, such as financial years. In most cases, standardising the three W's (Who, Where, When) suffice.

Agile and organic growth

Last, NDRs should "expect the unexpected" and be agile — in some cases, time is of the essence especially in public health emergencies, the ability to quickly absorb, transform and analyse/report that data may in fact avert the course of an epidemic. In addition, NDRs are never finished, they should be able to grow organically, nourished by a team of cross-cutting data managers, data scientists, epidemiologists and public health experts.

Technology considerations

There are many technological options available to implement an NDR. In our experience, it is important to select technologies that:

- Have been around for at least five years, so that they have a proven track record of their usability, support, community and performance.
- Are built around data management industry-standards agnostic of the health domain, so that they can absorb any type of data with ease. There are some technologies developed specifically for the health sector, but they often only address a single area/data source, making them unsuitable to act as a full data repository. Examples of such include DHIS2 (mostly suitable for indicator-based data from health facilities), master patient indexes (MPIs, can only store client demographics) and HAPI FHIR (a database system optimised for storing standardised, coded individual-level patient data). Rather these systems should serve as an *input* into an NDR and not as an NDR themselves.
- Based on platforms with commonly found skillsets, versus something very proprietary/specialised or relatively new. For example, it is much easier to recruit SQL developers than Cypher developers (a graph query language).
- Can stage and prepare data for visualisation, further analysis and reporting, using tools convenient to the end user. For example, if a user wishes to do some advanced analysis in R, the NDR should be able to produce a clean dataset that can easily be imported into R, with already the right linkages between data elements in place.
- Disconnect the storage from visualisation, so that different or multiple visualisation tools can be used, in order to meet the (potentially different) use cases.
- Use at least one visualisation platform that does not require development/software developers to do them, so that end-users and programmatic individuals can build their own visualisations as well, with little turnaround time.

Our recommended technology solutions

UCSF has gained a tremendous amount of experience developing and maintaining data repositories, including in Kenya, Jamaica, Uganda, Mozambique, and Namibia. Based on this experience, we strongly recommend the following:

- A SQL-based relational database, such as PostgreSQL or MS SQL Server, to store the data for an NDR. SQL has proven the test of time, and even after the emergence of non-SQL databases, SQL-based databases are still very popular. Because of this, there is a wealth of SQL developer experience available worldwide and it is not difficult to recruit a SQL developer. SQL-based databases are also allround, they are not limited to a particular type of data, and are extremely efficient in performing common operations for reporting, such as summary statistics.
- A dedicated business intelligence/reporting platform, such as PowerBI or Apache SuperSet, that does not require programming/software development expertise in order to develop visualisations and reports. Ultimately reporting may be a function of programmatic teams rather than software development teams. Advanced tools like PowerBI are not free, but allow for data exploration that free tools do not offer. This data exploration is sometimes critical to meet programmatic needs and discover new findings.

- ETL (extract, transform, load) is the process of extracting data from source databases, transforming it so that it can be used for analysis, and loading the resultset into an analysis database, for which UCSF has two approaches. Historically, UCSF has been using SQL for this, but more recently has also adopted PySpark and Python, an open source platform to quickly manipulate large datasets. The benefit of using PySpark is that it also sets the stage for incorporating machine learning algorithms, as this is also an integral part of Python.

Although open source software has benefits, most notably their low cost, it should be noted that they don't offer all features of paid platforms. For example, PowerBI allows for data exploration and dynamic charts that open source platforms may not offer — thus compromising on the required feature set. National data repositories will require investments either way (hardware, servers, server rooms), the cost of database and BI software (for example, SQL Server) may not be that large in comparison.

Our recommend approaches

An NDR is a complex undertaking, requiring a careful orchestrated process. It should contain a series and mix of workshops, meetings, hackathons, technical working group (TWG) sessions, and individual thinking, conceptualisation and development time. As a guidance, we would recommend the following, generic approach:

1. **Scoping/envisioning** There are many different perceptions of an NDR. It is important to first get stakeholders (MOH, donors, others) together in a room and reach a common understanding of what they are envisioning their NDR to be. This should include technical staff, but also (and especially) programmatic staff, with sufficient programmatic representation from the departments expected to supply the data (i.e. the data owners). The team needs to collectively understand what it means to develop an NDR, what the resource implications would be, agree on data governance, and get a rough sense of what data the NDR would entail, and what outputs are expected. And what should not be expected, to manage expectations early. Scoping can be done in a workshop format.
2. **Planning** The development of an NDR is a large undertaking, and requires commitment from the data owners, the funders, the beneficiaries, and the development/implementation team. This should lead to a plan, with timelines, and budget/resource implications. Of specific importance is dedicated staff time. NDRs are expected to grow organically, and will require continued maintenance and growth. As such, an NDR does not need to be developed in one go, in fact, it is advisable to start small, and then grow it over time. The planning does need to factor in this incremental growth, and provide guidance on priority areas. Planning can be done in a TWG format.
3. **Design and development** NDRs need to be designed, and this may lead to entity-relationship diagrams (ERDs), dimensional models, staging area strategies, and ETL approaches. From there, the NDR can be developed. It is recommended to start with a single, not too large dataset to slowly introduce the team to data warehousing concepts. Design and development can be done in a hackathon followed by a stream of development activities and follow up TA.
4. **Conceptualising and developing visualisations** This is often the hardest part: figuring out what to visualise/report and how to analyse the data. This requires intense involvement from stakeholders, while also bearing in mind that even stakeholders do not always know what they want. Meaningful visualisation of data is as much an art as a science, and requires creativity, flexible data visualisation tools and out of the box thinking. This is especially true where (sometimes for the first time) data from different domains are to be combined. This can be done through a hackathon followed by a stream of development activities.

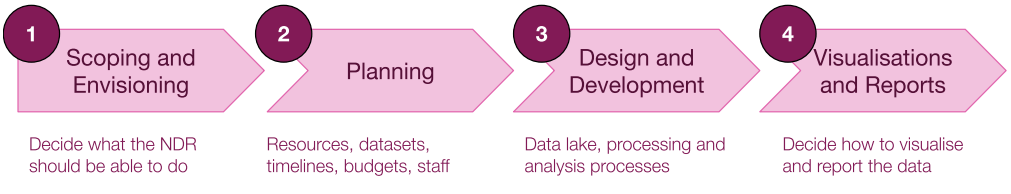
Approaches 3 and 4 are likely to be repeated for each "growth cycle", and it is entirely possible to do this by domain, for example, by focusing on HIV first.

In conclusion: how to start?

This document gives an overview on how to arrive at a national data repository. It shows it is a complex process, both technically and managerially, requiring stakeholders to be aligned and in agreement.

NDRs can be large or small. It is a good idea to start small and expand the NDR over time. All stakeholders should be consulted and informed from the beginning — a wide level of buy-in early in the process is critical to its success. A small TWG with sufficient representation from key stakeholders should be created at the onset, to prepare for Step 1, Scoping and Envisioning, as well as Step 2 (Planning) later in the process.

Planning for an NDR



Planning a NDR requires at least four distinct activities, some of which (especially 3 and 4) may be repeated. Inclusive, programmatic and data owner representation is critical.