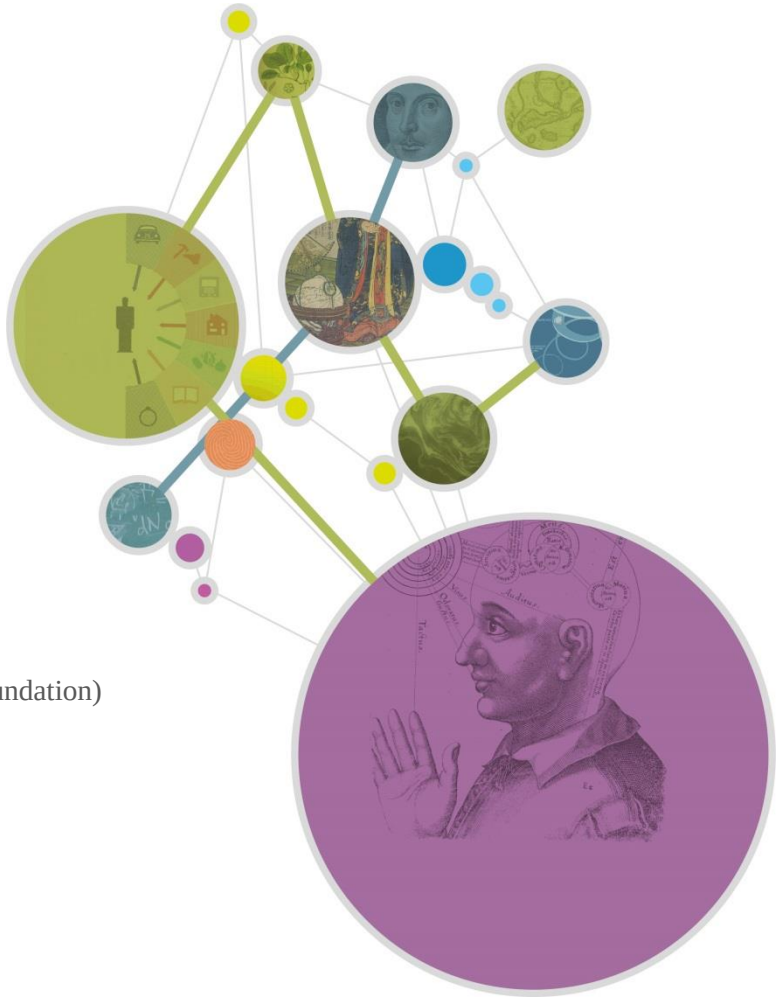
Public Health Research Data Forum

Supporting Capacity and Skills Development for Public Health Data Research Management in Low- and Medium Income Countries

22 November 2013



Gavin Chait (The Open Knowledge Foundation),
Eramangalath Sujith (Frost & Sullivan),
Dominika Grzywinska (Frost & Sullivan),
and Mark Wainwright (The Open Knowledge Foundation)
http://okfn.org and http://www.frost.com

Contents

Executive Summary

Conducting research in public health is both time-consuming and expensive. Ensuring that research data, along with the published findings, are made widely available to the research community will support enhanced discovery and greater efficiency.

The William and Flora Hewlett Foundation, on behalf of the Public Health Research Data Forum, commissioned the Open Knowledge Foundation and Frost & Sullivan to assess labour market dynamics for people who possess relevant expertise, and identify current challenges and support gaps in research data management. The project also seeks to compile an inventory of existing training and capacity building activities relevant to management and sharing of public health research data.

The findings of this study have led to a number of recommendations offered for future development and implementation:

- **Promotion of collaborative research networks:** Many institutions in LMICs lack the scale to employ and support effective research data staff development or to implement research data management systems. The development of collaborative and contributory institutional networks can reduce costs while improving participation and research outcomes.
- **Training and mentorship programs:** While there are numerous schools of public health, as well as courses in systems and database administration, most research data management courses are modules designed for people with existing skills. More effective mentorship and training can act to "join up" the various skillsets required to be a research data manager in public health. Career advice and promotion is also important.
- **Professional certifications or funding of fellowships:** Certifications may be a useful long-term mechanism to build recognition for research data management as a profession, however, at present there are no uniformly agreed standards and many of the leading proponents of research data management are self-taught. Fellowships, as an alternative approach, would be more effective in the short-term for supporting institutional and researcher recognition of the importance of the research data manager role, and of sharing of research data along with published results.
- **Data management as a condition for funding:** Funders can and should ensure that grantees have systems in place to share their research data along with their published findings. Funders will need to ensure that grantees have adequately scoped the lifetime cost of their data management system, or that they are members in good standing of research data sharing networks, and that these systems are audited for delivery. There is also a need to ensure that grantees have mechanisms in place to ensure data managers are supported through training and career development.
- **Data citation as a condition for publication:** It is not sufficient merely to have a system for research data management in place. Grantees and research publishers must demonstrate that they are citing the data in their findings and that such data is available for sharing in a well-managed research data management system. Importantly, data can only be used – and cited – if it is released under an appropriate publication licence.
- **Metadata interoperability:** Different research specialisms support a wide range of different metadata standards which can become a barrier to sharing and reuse of

research data. A mechanism for interoperability of such standards must be developed to ensure that this does not become an excuse for not publishing research data.

- **Open source data publication infrastructure:** If research data sharing and management are to be widespread then the most effective approach is to have software systems that can be easily shared and extended without institutions and publishers having to concern themselves with perpetual license fees or the danger of locking their research into "walled-gardens" and proprietary software. It is important not to replace a problem of unavailable data with that of unavailable systems.

The ultimate objectives and aspirations for ensuring the availability of public health research data to the scientific community will be achieved through collaboration involving funders, research institutions, publishers, and a diversity of service providers.

1. Introduction

In March 2013, Neelie Kroes, Vice-president of the European Commission responsible for the Digital Agenda spoke[1] at the launch of the new global Research Data Alliance[2]. The new organisation aims to accelerate and facilitate research data sharing and exchange.

"Whether it's scientific results," she said, "the data they are based on, the software used for analysis, or the education resources that help us teach and learn, being more open can help, transforming every discipline from astronomy to zoology, and making our lives better."

While not yet widespread in Low and Medium Income Countries, many governments and funders are now promoting greater access to data-sharing repositories.

The UK-based Engineering and Physical Sciences Research Council (EPSRC), for example, has published their expectations of organisations in receipt of funding from them[3]. Amongst the key requirements which research organisations must ensure are that:

- EPSRC-funded research data is securely preserved for a minimum of ten years;
- Effective data curation is provided throughout the full data lifecycle and responsibilities associated with data curation will be clearly allocated within the research organisation;

Ensuring that this happens is not a trivial undertaking involving both skills capacity and financial commitments.

A survey conducted in 2013 by Loughborough University and the Digital Curation Centre[4] received responses from 38 institutions in the UK. They identified data-handling requirements for staff from data librarians, to IT and research support, as well as digital storage facilities. The average cost for this per institution is about £157,000 per year. Larger institutions should be able to afford this; individual research projects are unlikely to.

Within the context of public health research in Low and Medium Income Countries, questions arise about the availability of skilled staff available to perform these roles, as well access to training to support capacity building.

This report offers insight into options for future initiatives to support capacity and skills development in public health data research. Approaches could address the following:

- Skills and systems for data curation and sharing
- Resources for data sharing
- Professional recognition for data publishers
- Citation standards for data reuse

It should also be recognised that while funders are major role players, the success of future capacity and skills development will depend on the value which researchers and institutions experience as a result of these initiatives.

2. Project aims and methodology

The William and Flora Hewlett Foundation, on behalf of the Public Health Research Data Forum, commissioned the Open Knowledge Foundation and Frost & Sullivan to assess labour market dynamics for people who possess relevant expertise, and identify current challenges and support gaps in research data management. The project also seeks to compile an inventory of existing training and capacity building activities relevant to management and sharing of public health research data.

Based on this analysis, we will identify ways in which funding agencies could build on their existing activities to develop and retain key data management skills. They may then improve data management and sharing in research institutions in ways that have the potential to accelerate progress in public health.

Five countries were selected (see Appendix) as representatives of Low and Medium Income Country (LMIC) public health research interests. These are: Brazil, Uganda, South Africa, India, and Vietnam.

Research, both through primary interviews and through secondary research of existing literature, will contribute to our understanding. Our interviewees were categorised and selected as follows:

- Global organisations:
    - Funders who support research institutions at the global level;
    - Research institutions which lead public health research in LMICs;
    - Training organisations which support capacity building for research data management in LMICs;
- In-country:
    - Local research institutions;
    - Training organisations;

In some countries (Uganda and Vietnam), despite our best efforts, it proved difficult to find suitable people to interview.

Opinion presented below is derived over the course of the interviews conducted during the primary research phase and offer insight into how selected organisations work within the industry. It cannot be considered a statistically relevant sample, but does provide a sense-check as to how different entities across the public health research data environment respond to constraints and opportunities.

3. Context for data management and sharing in public health research

3.1 Policy and practice in open data and open access

More than 90 countries now have laws guaranteeing freedom of access to information generated by their national governments[5]. The momentum behind such laws has been growing with the availability of new digital tools for data distribution, as well as the use of open data in a greater number of commercial activities.

Numerous state policies have also built on the Budapest Open Access Initiative in 2002, Bethesda Statement on Open Access Publishing in 2003 and Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities of 2003: that open access works permit users to "copy, use, distribute, transmit and display the work publicly and to make and distribute derivative works, in any digital medium for any responsible purpose, subject to proper attribution of authorship."

In 2004, the UK House of Commons Science and Technology Committee released a report, 'Free for all?'[6], declaring that: "We recommend that the Research Councils and other Government funders mandate their funded researchers to deposit a copy of all their articles in their institution's repository within one month of publication or a reasonable period to be agreed following publication, as a condition of their research grant."

A subsequent report by the Business, Innovation and Skills Committee in 2013[7], noted the following:

"There are 58 UK funder open access policies, all of which have a primary focus on Green, and the largest number of Green mandates in the world, comprising 24 institutional mandates and a further 15 funder mandates. The latest data from the UK Open Access Implementation Group shows that 35% of the UK's total research outputs are freely provided through Green, through an existing network of more than 200 active institutional and disciplinary repositories. In recent years the Government has invested more than £150 million in those repositories."

"Green" open access publication is a mechanism by which institutions are permitted to deposit copies of their published research on their own public websites. The alternative is "Gold" open access, where journals themselves allow free public web access; examples include the Public Library of Science (PLoS) titles. These journals often have author fees (normally met by the research funding bodies) to cover the cost of publishing. PLoS, for example, charges from $1,350 to $2,900 for publication. PeerJ[8], a new open access publisher, intends to disrupt this market with one-off pricing from $99 to $299 per author.

These publications, though, are for the final research papers and not for the raw research data and systems which may have gone in to producing the results. Neither can these systems handle the confidential or restricted-access data that are often part of public health studies. Managing such systems is still for the researcher's own account.

There can be a significant cost advantage to institutions from managing their own publication systems, including the ability to manage confidential data.

The quality and ease-of-use (or accessibility) of the data to be released is as important as ensuring that data is openly available. In May 2013, US President Barack Obama signed an Executive Order making it a default requirement that government data be both open and machine-readable[9].

In 2013, the policy foundation for open access scientific publication is in place in the G8 group of countries[10] (UK, Russian Federation, Germany, Japan, Italy, Canada, France, USA) and the European Union, and Australia.

There are also numerous private initiatives, such as the Panton Principles for Open Data in Science[11], which seek to provide policy guidance for researchers. Private funders are also beginning to expect open access publication in the research they fund.

Research institutions have begun to take the demand seriously and are putting infrastructure in place. Of the 38 institutions responding to the Loughborough survey on research data management policy[4], 82% have a research data management policy in draft or already in place, and 66% are in the process of assessing and setting up research data services.

The expense and complexity of setting up these services has led to the creation of global guidelines which define how a digital data resource and publication platform should work. The leading guide is that from the Consultative Committee for Space Data Systems, the Reference Model for an Open Archival Information System[12].

The cost and complexity of setup, staffing and long-term maintenance of research data publication systems makes it impractical for individually funded research projects to set these up directly.

In addition, public health research – while complex and distinct – shares sufficient requirements with other research specialisms that shared infrastructure and approaches to research data management must be considered.

3.2 Open science in Low and Medium Income Countries

Absent from the initiatives described above are representatives from Low and Medium Income Countries (LMICs).

The Open Knowledge Foundation, along with the OpenUCT Initiative and the International Development Research Centre (IDRC) convened a workshop in Cape Town, South Africa in September 2013. The topic was "Towards a southern-led research agenda on open and collaborative science for development".

Representatives included the University of Sao Paolo, University of Singapore, University of Cape Town, University of the West Indies, Creative Commons, and Universidade Federal do Rio Grande do Sul (UFRGS).

The findings are still to be concluded, but some of the challenges identified include:

- Political systems and cultural norms, including post-colonial legacy and internet filtering, can isolate countries from international engagement;
- Infrastructure and connectivity can limit access;
- Universities are not integrated with society or with similar institutions globally, and sometimes offer little communication of their activities;
- Career development and research communication training are rare;
- There is duplication of research as a result of lack of access to prior research outputs (as differentiated from informed validation and replication of existing research);
- Policy frameworks are required which provide governance and guidance for strategic engagement;
- Language (such as in Francophone and Lusophone regions) mean that research output may be overlooked. E.g. Thomson Reuters published data showing that no research

articles were produced in a certain year by six African countries, because the articles were published in French;

During discussions it was pointed out that many challenges are not specifically southern and relate to science as a whole. Lack of resources and a supportive infrastructure for science is a perennial problem. This is related to an undervaluation of science in many political contexts. And this is before a discussion of open and citizen science begins.

According to Jenny Molloy, coordinator of the Open Science Working Group at the Open Knowledge Foundation, "Open science runs counter to many aspects of the current approach to research and publishing, particularly research assessment, the rewards and incentives that drive researchers, and the intellectual property that may be lucrative for institutions and governments. Both infrastructure and culture need big changes to be considered 'supportive' and these changes may be different to those that would support the current system."

In addition, research infrastructure and national/international research funding structures must deliver the scalability required for larger, team based collaborations rather than small, derivative projects. In addition there are serious tensions within the academic system concerning the combination of prestige versus research relevance.

Many LMICs do now have public access to information policies in place (including Brazil, South Africa, Mexico and India) but there are many which do not. Moreover, even when these policies are in place, they are often difficult to implement.

Molloy believes that Latin America has implemented very successful open science initiatives. Brazil's Scientific Electronic Library Online (SciELO) is highly regarded[13]. In 2011, 43% of Brazilian science articles were free to read on publication (which exceeds the 35% in the UK cited earlier and only 6% in the US that year). Of course, by volume, the absolute number of open access publications does differ.

"What came out in the workshop," says Molloy, "was that developing countries need to learn from each other and generate data/article/other output publishing platforms and research assessment criteria that meet their own needs rather than copying what exists in the north. For example, many research intensive universities in the south reward their researchers solely for publishing in high impact international journals. There's an opportunity for emerging scientific players to leapfrog some of the inertia we're experiencing in mature research institutions."

However, the supporting environment for open data and open access research are nowhere near as developed as in Europe, the US and Australia. Many LMICs feel there is no, or unequal, benefit to sharing data.

Thailand and Indonesia, for example, refused to share virus samples following a H5N1 avian flu outbreak in 2007[14]. Non-government organisations do not routinely share data even with each other and many collect public health data. Linda Raftree, in the Open Development Working Group at the Open Knowledge Foundation, gives the example of the Malaria Atlas Project[15] which had to spend months contacting hundreds of health ministries and other data holders individually to get data sets for global malaria mapping and modelling, and then again to get permission to release under a Creative Commons license.

If LMICs are already concerned that "openness" is not something that developed countries commit to, then it is going to be difficult to convince them to participate in such initiatives.

3.3 Research data management, and research in public health

Public health research data management is a very small component of the much wider context discussed above. The skills required for such professionals are presented in "Sustainable data sharing in public health research: An INDEPTH-COHRED position paper"[16]:

- Ability to design and develop databases that manage longitudinal data efficiently and protect the temporal and data integrity of the information.
- Knowledge of structured query language (SQL) to construct queries for data extraction from relational databases, including experience with automated extract-transform-load (ETL) tools for transformation of non-machine-readable content.
- Familiarity with probability-based record linkage techniques and dealing appropriately with missing data.
- Familiarity with statistical and analytical techniques commonly used in longitudinal data analysis to develop appropriate analytical datasets and communicate effectively with data analysts in satisfying their data needs.
- Knowledge of how to conduct data quality assurance and detect and correct data errors without compromising the integrity of the data.
- Knowledge of metadata standards such as DDI[17]-Codebook and DDI-Lifecycle and the skills to use the available tools to document datasets and surveys.
- Configuration and maintenance of data repositories.
- Data curation.
- Knowledge of data citation mechanisms and their application to shared datasets.

There are few independent ways to gain a combination of skills which cover both system administration and database skills, as well as public health research.

There are a vast number of ways to gain database and data management skills, and numerous schools of public health. Bridging the two qualifications can be done via courses in research data management.

Universities do offer courses in research data management, but these tend to be modules in addition to existing coursework and generic to address institution-wide requirements. Similarly, courses in public health research must address broad requirements ranging from social science studies through to the impact of disease morbidity and mortalities. See the Appendix for more details on these courses.

A person who takes the time to accumulate the skills listed above will be in demand for a range of careers beyond that of public health research. They are known interchangeably as research data managers, data scientists and even data engineers.

The environment can be represented as in Figure 1, where public health research institutions, as well as institutional requirements for research data management, and broader career opportunities all compete for a small pool of appropriately skilled individuals.

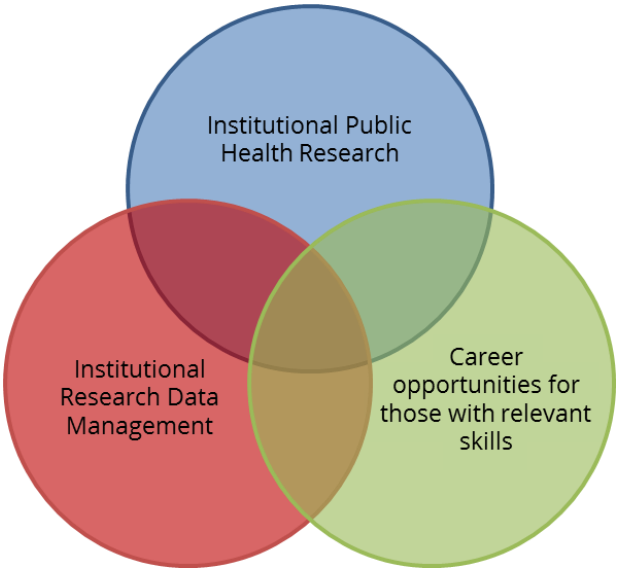


Figure 1: Competing interests for public health research data managers

3.4 Publishing research and data

A set standard for data sharing and role-based training for establishing linkages between multiple databases would improve the skills gap in data sharing.

As discussed in 3.1, there is a growing interest and expectation of open access publication but the tools and services to deliver this are still relatively immature. Technically-astute data managers and researchers have numerous tools available to them for data publication.

On 11-12 June 2013, the Open Economics Working Group of the Open Knowledge Foundation organised the Second Open Economics International Workshop, hosted at the MIT Sloan School of Management. This was the second of two international workshops funded by the Alfred P. Sloan Foundation, aimed at bringing together economists and senior academics, funders, data publishers and data curators to discuss the progress made in the field of publishing open data for economics[18]. The issues raised, while from a different research specialism than public health, present many of the same concerns.

Daniel Feenberg[18], of the National Bureau of Economic Research, a publisher of about a thousand working papers a year, says that requiring data sharing is something that only employers and funders can mandate and there is a limited role for the publisher.

However, in some cases (particularly with the most prestigious or specialised publishing houses), publishers do exert a tremendous amount of leverage and could expect data sharing.

Daniel Goroff[18], of the Alfred P. Sloan Foundation, says that, while some funders may require data management plans and making research outputs entirely open, this is not a simple matter and there are trade-offs involved. The Alfred P. Sloan Foundation has funded and supported the establishment of knowledge public goods, commodities which are non-rivalrous and non-excludable like big open access datasets with large setup costs (e.g. Sloan Digital Sky Survey, Census of Marine Life, Wikipedia, etc.). Public goods, however, are notoriously hard to finance. The involvement of markets and commercial enterprises where,

for example, the data is available openly for free, but value-added services are offered at a charge could be some of the ways to make knowledge public goods useful and sustainable.

"Green" data publication, in which the institution manages its own data, has been supported by JISC[19] in the UK. They have developed a number of approaches to managing and linking research data. The INDEPTH Network, too, has developed an early-stage interface[20] for such research data publication but it is far from a common standard.

Joss Winn, at the University of Lincoln, has presented a paper[21] on using the open-source CKAN data publication platform maintained by the Open Knowledge Foundation for research data publication. Initiatives, such as European Data Watch Extended (EDaWaX)[22] and EUDAT[23], are looking at implementing Winn's proposal but these initiatives are still in development.

Dr Hendrik Bunke[24] at ZBW – part of EDaWaX - describes his requirements from CKAN as being "an abstraction level or framework in CKAN that simplifies the implementation of custom metadata schemes. That would also be a good foundation for a more long-term project: all available metadata schemas (DDI; DataCite, da|ra, etc.) are published in XML (as XML Schema usually)."

The problem of metadata interoperability, as described by Dr Bunke, is a growing concern. Institutions which choose to set up their own infrastructure must ensure it can cater to the widest variety of research.

Various specialisms have their own metadata standards. For example, Ann Clements, at the University of St Andrews, has a requirement for supporting the Common European Research Information Format (CERIF) metadata standard. She has published a paper[25] on matching CERIF to the Marine Environmental Data & Information Network (MEDIN) metadata standard.

Metadata interoperability can be performed, for instance, by producing a set of grammars which relate directly to a Resource Description Framework[26]. A recent initiative from the Open Knowledge Foundation[27] for data catalogue interoperability, based on the US Government's Project Open Data[28], has been launched to develop some of the initial groundwork for this.

That is only one of many requirements for metadata reinterpretation. Once a research organisation commits to data management, open access or not, setting up a self-managed service is a complex undertaking involving decisions around metadata standards required by different research disciplines, interoperability amongst numerous platforms and research teams, software systems, and perpetual hosting costs.

CKAN, obviously, is not the only such software available.

The DataVerse network is a free and open-source service and software to publish, share and reference research data. It is a self-curated platform where authors can upload data and additional documentation, adding additional metadata to make the resource more discoverable. It builds on the incentives of data sharing, giving a persistent identifier, generating automatically a data citation, providing usage statistics and giving attribution to the contributing authors.

A commercial approach to research data management software is FigShare[29], owned by Digital Science, a technology company operated by Macmillan Science & Education. The software was originally created by Mark Hahnel while studying for his PhD in stem cell biology at Imperial College, London.

However, the primary limitation for publishing public health data – beyond metadata standards and interoperability – is the concern about the confidentiality of some datasets and the requirement to protect access and maintain such confidentiality indefinitely.

Unless and until such concerns can be addressed, data collected by agencies will remain locked up and unavailable for further analysis by others.

Such concerns can also become a self-reinforcing excuse for not taking action.

This is already the case in LMICs where data collected by government and regional statistical agencies are used for internal purposes, including planning and policy monitoring. This data remains isolated in files in statistical organisations and is not available for use by researchers, leading to duplication of efforts.

Once software, metadata standards, interoperability and availability are addressed, then there are still issues with the licenses under which data are released.

Peter Desmet, a researcher at the Canadian Research Institute for Nature and Forest, describes how non-standard open access data licenses have made it illegal for him to aggregate 13,297 georeferenced American bullfrog records and place them on a single map[30]. This, despite the data being released as open access on the Global Biodiversity Information Facility (GBIF)[31].

Even when the data are published online poor licensing can still render data inaccessible.

4. The process for undertaking institutionally-funded research in LMICs

The process of producing research in public health requires collaboration and support between a diverse range of institutions and organisations:

- Private institutional funders;
- Local government funders;
- International research institutions;
- Local research institutions;

Given the scale and scope of public health research, consortia in both funding and research are often needed to address the requirements. Beyond the direct research needs, there may also be cultural, legal and political components which can affect the structure of funding and research.

Once funding is approved, then – depending on the nature of the work to be undertaken – research consortia may recruit fieldworkers, data analysts, research data managers and project managers.

Finally, the research outcomes may include research data, publications and data sharing.

This is summarised in Figure 2:

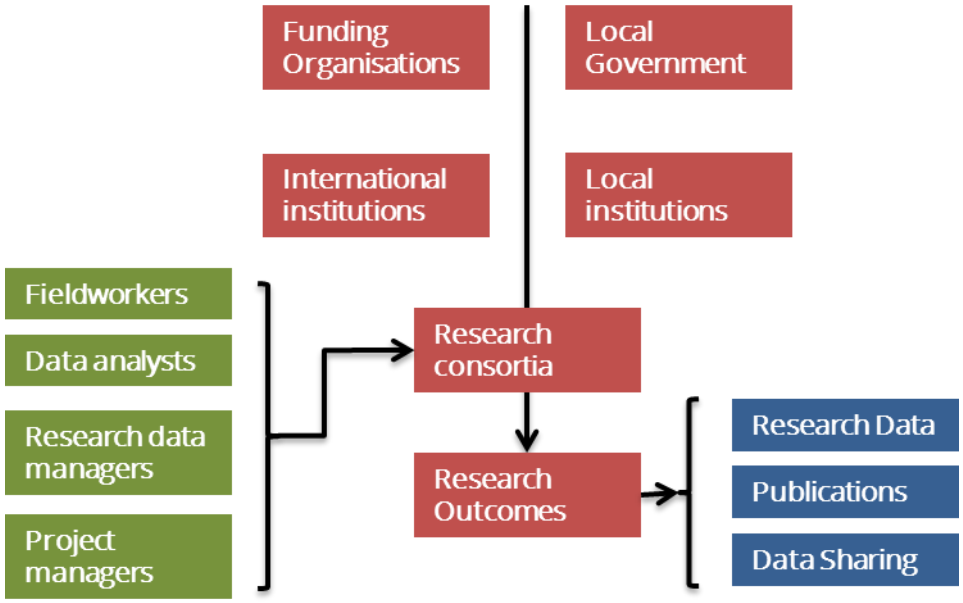


Figure 2: Structuring and implementing public health research

The issues described by the various participants in the study are consistent and portray an industry struggling with growth in the use of data beyond primary collection, and in the systems necessary for data analysis, management and sharing.

The process of producing research data in public health requires that a lead organisation build sufficient competency and capacity to deliver, and then drive, a fund-raising process to finance the research.

Note that structuring funding, research teams and outcomes are diverse and complex. This report is not intended to reflect every case and the case studies below represent a small range of approaches. They are not intended to represent the entire industry.

With that caveat, a series of semi-discrete steps can be identified:

- Capable institutions, with a track-record of high-quality research, develop funding proposals for new research or to continue existing programs, or for evaluation of existing research data;
- They may coordinate a number of other organisations, potentially including ones based in the LMICs, to form a research consortium;
- Funders may also occasionally, and local governments regularly do, call for organisations to tender to perform relevant research;
- Proposals with the best chance of delivering useful research outcomes are financed, either by one funder, or by a consortium of funders;
- Where new primary research is to be conducted, the research consortium may recruit local fieldworkers and ensure that they are trained, as well as putting in place data

processors, research data managers, and infrastructure for the management of research data;
- Once research is complete, then data is processed into publishable content and the source data is (sometimes) prepared for public release;

Public health research studies require several stages, including preparation of questionnaires or queries with inputs from other international studies, data collection and sample collection by field workers, data curation performed manually and using software, data analysis using statistical models, and data sharing to other researchers and external agencies.

There are numerous variations on the above steps. In some projects, no new data is collected and research is based purely on existing resources which have yet to be analysed or processed, or on re-use of existing datasets. That said, this is a reasonably consistent structure upon which to present this report.

4.1 Approaches to funding

One of the most comprehensive non-governmental initiatives designed to improve access to research data is that of the Public Health Research Data Forum (PHRDF). The Joint Statement of the PHRDF[32], which includes signatories from many of the leading public health research funding organisations[33], has the vision of increasing "availability to the scientific community of the research data we fund that is collected from populations for the purpose of health research, and to promote the efficient use of those data to accelerate improvements in public health."

The three organisations identified in this section do not represent the complete range of funding approaches adopted but do provide an indication of the variety which research organisations navigate.

The three different approaches can be described as:

- **Comprehensive:** able to support multi-year, large-scale primary public health research led by complex research consortia [Wellcome Trust];
- **Targeted:** focused on incubating only their own institution's research teams and permitting them to gain scale to attract comprehensive funding at a later date [Emory University];
- **Capacity-building:** greater interest in developing new research capacity through reuse of existing research data than in generating new research [World Health Organisation Research and Training in Tropical Diseases].

4.1.1 Comprehensive funding

According to Dr Jimmy Whitworth[34], of grants funded outside the UK by the Wellcome Trust, around half are now awarded directly to local organisations rather than to global research institutions. This has been driven by changing priorities at the funder, as well as demand from local institutes and government.

The Wellcome Trust funds "quality research" defined in terms of the following:

- Importance of the research question;

- Practicality and conviction that the people and methods will address the research criteria and will succeed;
- Value for money and the value of the result;
- In LMICs, engagement from a host institution and government, and capacity building;

The characteristics of the Wellcome Trust's awards, which fund a diverse range of research objectives from hard-core genomics to health policy research, are:

- Funding is usually for three to five years, and up to a maximum of seven;
- It takes on average around six to nine months to secure an award;
- Track record is a consideration, so if a group has secured funding from Wellcome (or other major funders) in the past and demonstrated they can deliver, this would be a factor in determining future success of applications;
- Core intent is to have a long-term view to working and developing capacity;
- Funding can be from £100,000 to £5 million; 95% of grants are £30,000 to £1 million a year;
- Have 5,000 current active grants per year, of which LMIC are £70-£100 million per year;

Few private institutional funders have the resources to finance such **comprehensive** large-scale research projects on their own. Amongst them are the Wellcome Trust, Bill and Melinda Gates Foundation, and Hewlett Foundation.

4.1.2 Targeted funding

A more **targeted** approach is led by Dr Jeffrey Koplan[35], at Emory University, who administers two large grants received from the Bill and Melinda Gates Foundation which are then offered via institution-wide faculty-level support for projects. There are many models and many projects, including: diabetes prevention in Chennai, India; public health in Mexico; environmental health, sanitation, chronic diseases and infant malnutrition in Chennai and Delhi; tobacco control in China.

Each faculty is aware that they can apply for funding from the program. This permits researchers to leverage initial studies to secure larger funding support. Emory acts as an incubator, permitting research teams to get started, and then apply for additional funding to gain scale. The process is also one of integration since a comprehensive view of institutional funding and support needs is known. After receiving the initial pilot investment, the researchers are expected to deal with comprehensive funders themselves.

Emory looks for a 14:1 return, approximately. For example, $750,000 provided for a three-year diabetes program which, five years later, raised $7 million from other organisations.

"It is a platform for broader, long-term work," says Dr Koplan. Emory, as with most funders, is not interested in one-off studies, going for the long term and the potential to secure greater revenue later.

4.1.3 Capacity-building funding

Both the comprehensive and targeted approaches start with the assumption that the applicants have the capacity to deliver quality research. **Capacity-building** programmes are less

concerned about producing new primary research and more about ensuring that funding recipients are trained to become future research leaders. That does not mean that there are no research outcomes.

The World Health Organisation has begun a new model of funding via their Research and Training in Tropical Diseases (WHO TDR)[36]. Dr Garry Aslanyan, the WHO TDR manager, has a different funding focus on providing training for public health organisations to undertake analysis on the data they already have. They provide $20,000 per person and each person gets assigned a mentor. The mentor may be from anywhere in the world but they do their best to ensure appropriate matching (such as South-South pairing). More details on the training can be found in Section 6.2 where the WHO TDR relationship with The Union is described.

The need for such interventions to promote capacity with existing data collection is clear in South Africa. Despite oncology incidence and outcomes rates being collected, the collation of such data is now seven years out of date[37]. Dr Elvira Singh, a public health specialist at the National Health Laboratory Service which manages the register, provides a summary of the constraints: there are capacity constraints in staffing and IT infrastructure.

4.1.4 Co-funding with local government

In addition to the above single funder approaches, there are also collaboratively funded projects. Many funders now require that local governments accept their responsibility for public health research through co-funding requirements. In some ways, this is a branch of **capacity-building** as it ensures that local government has the capacity to engage and consider their requirements and priorities for public health research.

For example, Brazil and India have research projects on longitudinal cohort studies of ageing, which are funded by their Ministries of Health, for tackling issues related to ageing populations[38]. Brazil has project-Epigene, for understanding the genetics of the entire spectrum of population, which is co-funded by the Wellcome Trust and the Brazilian government.

The World Bank and other international agencies, such as OECD, co-fund projects in numerous African countries, including in South Africa, Uganda, Namibia and Zambia. Public health research projects, such as Arcade RCH and Arcade HSFR, are conducted through a collaboration of around 7-12 partners across Asia and Europe, including South Africa and Vietnam[39].

4.2 Research organisation structures

As with funding, research organisations have a wide range of approaches and structures. Two approaches are described here:

- **Consortium-based:** a consortium of research agencies collaborate to deliver on a single project over a single period of time [Institute of Development Studies];
- **Network-based:** a network of research organisations acting independently work collaboratively to produce research which can then be aligned and assembled within a single framework [INDEPTH Network];

Researchers intend making careers in research and so long-term relationships with institutions and funders is critical to future success.

4.2.1 Consortium-based research

Knowledge Services at the Institute for Development Studies specialises in the mobilisation of research knowledge and the development of institutional architecture at different levels of government[40].

"We harness evidence to shape decision-making and mobilisation of resources, including advocacy influencing skills development amongst our client groups," says Tom Barker, who has a special focus on health and nutrition.

They structure the type and range of projects they undertake according to the funders they approach. They propose, in partnership, to funders like the UK's Department for International Development (DFID) and form a consortium with other international organisations. More recently, local governments will require that a local partner be included in the initial high-level consortium.

Their proposal stage builds on existing networks to partner on the expression of interest and full proposal.

Implementation starts with the inception phase (up to nine months) to put bilateral relationships in place (groundwork, landscape and evidence review). This may also lead to stronger representation by local partners. Funders are placing increasing emphasis on capacity building. These requirements may be very explicit and formal.

Projects tend to be from one to six years and a number of projects are clustered in the same country. Long-term capacity building is often secondary. They have an immediate requirement for skilled staff so that projects can begin and don't necessarily have the time to wait for the outcomes of skills development. Conversely, federal governments and local partners see capacity building as crucial.

This has resulted in tension between complex and occasionally contradictory requirements.

There is a further concern raised: many local institutions lack experience or capacity to develop research projects independently and prefer to be junior collaborators or consultants to funded international consortia-led projects[40].

This has implications for their recognition within academia and in driving policy debate with local government. The lack of experience at senior levels has implications for junior academics who don't gain field experience. Apprentice training under experienced mentors reinforces education in each subsequent generation. This is often absent in LMICs[35].

While consulting to internationally-led projects does permit local organisations to gain experience, it also means that they are often not part of project scoping conversations with local authorities. This becomes self-perpetuating with a lack of engagement between local academics and local policy-makers.

This also implies that local institutions don't apply for funding, for lack of capacity and experience to do so, and may remain unknown to funders.

4.2.2 Network-based research

INDEPTH Network is a continuous project but taking a different approach to collaboration. 48 member centres around the world run health and demographic surveillance in which people in a specific area are enumerated and then observed longitudinally.

In total, INDEPTH Network centres collect data on 3.5 million people, representing 40 million person-years of observation. They have various long-term funders, including the Wellcome Trust.

"The INDEPTH Network promotes the harmonisation and sharing of data from its member's centres through data use agreements with them that allow the Network to publish and share data on the INDEPTH Data Repository and INDEPTHStats," says Kobus Herbst, the Deputy Director of the Africa Centre, an independent institution which is a member of the INDEPTH Network[41].

Herbst is also the Principal Investigator of the iSHARE2 INDEPTH project, which falls wholly under the auspices of INDEPTH.

The Africa Centre for Health and Population Studies, at the University of Kwazulu Natal is primarily funded by the Wellcome Trust as one of its Major Overseas Programmes. They have been conducting demographic surveillance on approximately 90,000 individuals in rural KwaZulu-Natal since 2000. As a long-term project, they conduct capacity building and offer a South African Qualifications Authority (SAQA) accredited training programme to their fieldworkers. They are also a member of other network-based research collaborations, like the ALPHA network.

"The ALPHA network aims to maximise the usefulness of data generated in community-based longitudinal HIV studies in sub-Saharan Africa for national and international agencies involved in designing or monitoring interventions and epidemiological forecasting," according to the London School of Hygiene and Tropical Medicine, which coordinates the network.

This is funded by both the Wellcome Trust and the Bill & Melinda Gates Foundation. There are ten different centres based in Africa.

The iSHARE2 project is a continuation of the INDEPTH iSHARE initiative to train, equip and support INDEPTH member centres to harmonise, quality assure, document and share their research datasets. This initiative has had various funders, including Hewlett Foundation, Sida, Bill & Melinda Gates Foundation, Wellcome Trust and the International Development Research Centre.

iSHARE2 uses tools developed by others (e.g. NESSTAR Publisher, NADA, Kettle, etc.) to support their work. They have developed a 'research data management appliance', the 'Centre-in-a-Box', to bring these tools together in a device that can be deployed ('plug and play' fashion) into the diverse and often under-resourced IT infrastructure of research centres in LMICs and can be remotely managed and supported.

Networks can use consortia too. Knowledge Services at the Institute for Development Studies has worked with the ALPHA Network to produce a series of briefing reports on studies on HIV in African Communities.

Networks can act to support individual research consortia or centres through provision of shared infrastructure, capacity-building and engagement on long-term projects. **Consortia** often come together for a single project and then disband, whereas networks can collaborate indefinitely.

5. Public health research project implementation in LMICs

Public health research studies involve several stages, including preparation of questionnaires with inputs from other international studies, data collection and sample collection from the region under study, data curation performed manually and using software, data analysis using statistical models, and data sharing with other researchers and external agencies.

For new projects, once funding has been approved, research organisations begin to recruit for implementation. Where fieldwork is required to collect primary data, semi-skilled local people are employed on short-term contracts. Teams may also require research data managers and other skilled positions to support research outcomes.

Any capacity-building exercise must start with an assessment of the skills resources required.

5.1 Research data collection

The vast majority of data collection efforts are funded by, and the majority of data is generated for, local government[39 42]. All LMICs surveyed have developed capacity in field work.

Field work for data collection is often performed by graduate and postdoctoral students, especially if the project is affiliated to universities in a region. For example, in Brazil, public health projects have partnered with Fiocruz University and University of Sao Paulo, where they have a resource base of around 50 PhD students who are supervised by 15 tenured faculty advisors. However, it is also common to recruit fieldworkers and, occasionally, outsource the data collection to external agencies such as Instituto Brasileiro de Geografia e Estatística (IBGE), a third-party service provider[38].

Latin America has greater scope for large scale collaboration and movement of workers over, say Africa or South East Asia, because of a more homogenous language and culture than other regions. In India, fieldworkers are recruited from the region to be surveyed, as expertise in local language and culture helps in effective data collection[43].

An area highlighted for improvement is the organization of the collected data, and database management. A major constraint faced by field workers is the need for translation of the questionnaires from English to the local language and then converting the collected data back to English. Data loss or incomplete data may be generated due to this process.

This occurs in research conducted in Brazil where surveys are mostly designed in Portuguese and later translated to English for harmonisation with other international studies[39].

5.2 Data curation and analysis

Database management and data curation are performed by data managers who are often PhD students associated with the universities, researchers with relevant experience in public health, and officials from the government statistical offices. While Brazil and South Africa engage students for data analysis, India and Vietnam usually recruit experienced researchers with a background in public health for the purpose[44].

Most public health projects in LMICs have the majority of their efforts focused on data collection and, although there are data managers, they often do not place emphasis on data curation which affects the quality of the output information.

As a result of this focus on immediate strategic data objectives, data managers are usually engaged on a short-term contract basis. The lack of a focus on continuous data curation not only affects data availability beyond the duration of the immediate project, but also the prospects for a career path for data managers.

That leads to a lack of research data management capacity for data quality assessment, sharing, and packaging of the data for redistribution or publication. There are capacity problems for generating consistent, timely data and then the ability to analyse that data. Those with talent and ability at the junior level often lack support or guidance at the senior level[35].

The result of this lack of both initiative and experience is that, where research data is collected through government projects, the data goes unused or isn't effectively captured. Where projects are led by international research organisations, the work experience gained by local research participants is not effectively communicated and shared within their organisations.

Even where skills do exist – in the case of data scientists – it is difficult to attract such people into research because of opportunities in other industries. Data scientists are not recognised academics, are not often cited or referenced in research as providers or curators of the data used, and so find raising their own profile difficult.

5.3 Data preparation and publication

Practical difficulties in data sharing are often underestimated. The burden of making data shareable is most often entrusted to those who generate the data.

Many researchers and data managers share data informally as no industry-wide, standard procedures for sharing have been adopted. The cost and time implications associated with data sharing are high, as significant effort has to be put towards documentation and addressing queries on the data with other researchers or organisations.

There is also concern among researchers that they are able to have an exclusive access period to the data they have generated before it should be shared with others, in order to justify the effort spent by them to collect the data.

Once research data has been collated there are no formal or standardised publication and citation mechanisms for sharing that data, although numerous efforts to address this are underway (as detailed in 3.4). Research by the Digital Curation Centre (DCC), commissioned by the Wellcome Trust, has documented approaches to public health research citation[45].

Micah Altman, of MIT Libraries, Brookings Institution, says that data citations should be treated as first-class objects of publication and that reproducibility policies should be developed to support publishing replications and registering studies. He emphasised that policies are often not self-enforcing or self-sustaining and compliance with data availability policies, even in some of the best journals, is very low[18].

Brian Hole, of Ubiquity Press, offers the suggestion that dedicated data journals or encouraging data sharing and improving data citations through the publication of data and methodology in data papers, are also an option. He says that the publication of a data paper where the data is stored in a repository with a DOI and linked with a short data paper which describes the methodology of creating the dataset could be a way to incentivise individual researchers to share their data as it builds up their career record of publications[18]. Additional benefits of having data journals is having a metadata platform where data from different disciplines can be collected and mashed up producing new research.

Given that data may need to be available indefinitely beyond the funding of any particular project, ensuring the long-term support for publication is critical.

Once funders and legislation insist on research data publication, the next step is verification of compliance.

The Wellcome Trust requires that applicants submit a data management and sharing plan as part of grant applications for those grants which are likely to generate datasets of value to others[46]. Many other funders (e.g. NIH, MRC, etc.) have similar requirements.

The principle is that data sharing plans, and the costs associated with them (which might include costs needed to support data managers), are reviewed and funded as an integral part of the grant. This is the same for LMIC researchers and UK.

"However", says David Carr of the Wellcome Trust, "it is almost certainly fair to say that the extent to which this happens is variable. It is felt that plans do not sometimes get the level of scrutiny that they should particularly for subject areas where there aren't well established community-norms for data sharing and my personal perception is that the costs and resources aren't always planned for appropriately."

At present there is very little structured follow up to check on the implementation of data sharing plans for many of the grants supported by the Wellcome Trust, nor are there clear consequences for non-compliance with the data sharing plan.

The Wellcome Trust are actively considering what they may do to improve this and ensure that implementation is tracked in a proportionate way. The question of whether to introduce explicit sanctions is one they are considering but there is a question on the extent to which they feel they can legitimately sanction researchers when, for many areas, there are legitimate constraints in terms of infrastructure and capacity.

EPSRC has taken a slightly different approach to the majority of UK-based Research Councils in that, rather than considering research data management on a grant by grant basis, they place responsibility on the institution itself to have an overarching approach for data sharing (which covers all the EPSRC grants funded there).

Given the cost and complexity of managing any system for research data management and sharing, increasing the number of stakeholders also increases the value while reducing the cost of implementation and long-term management.

6. Support and training for public health research data management in LMICs

Many LMICs are showing progress in initiating research projects for public health, although skills shortages in data management, curation, and analysis, limit projects.

Many existing research projects provide basic training modules with some offering specialised courses in association with the World Bank, OECD and other agencies. Training is mostly performed on an on-the-job basis with experiential learning serving to solve problems during the course of work.

It is striking that training initiatives are, for the most part, very new, with the most mature identified in this study that of the International Union Against Tuberculosis and Lung Disease (The Union) only starting in 2009[47].

Emory University has developed a formal mentorship program for research project leaders with funding for four years. They partner a bright, young, mid-level person with similar people, but usually older and more senior. The mentor offers management, research and technical science experience support. Mentorship lasts two to three years, including telephonic- and email support, as well as short trips to meet in person. They have six of these relationships in total[35].

They also do direct training. Four years ago, in Uganda, an institute had difficulty raising funds. Emory performed one week of training in how to set up a research project and how to raise funds. The group was subsequently able to raise £1 million.

IDS provides training for their teams through one of their partners (Oxford Policy Management) on information literacy[40].

However, it is The Union which has taken the most ambitious steps in training, measuring their outcomes based on an 80:80:80 requirement: 80% of those trained should submit papers for publication, with 80% of these being published (with a preference for publication in open access journals), and 80% of papers then resulting in meaningful policy change[47].

6.1 Research data collection

The majority of LMICs involved in public health research projects have developed data collection skills through performing field work, although the next levels of data processing skills are often unsupported.

Increased local government involvement brings with it the experience of doing similar research studies in other industries.

In India, field workers are trained though a ten module program and receive hands-on training in IT skills, using notebook computers to enter data collected through fieldwork[43]. Similarly, in South Africa and Uganda, the majority of time and effort is invested in data collection. As a result the capacity for data collection has a solid foundation and supplementary courses, workshops and online refresher course-work may serve to address skills shortfalls.

There is a skills gap in the area of basic epidemiology, and a lack of short term courses in data collection which results in poor data collection methodologies at community level and at the university level. Training programs, workshops and specialised courses on raw data collection given to project staff can significantly improve the quality of data collected.

The following types of training were described by interviewees in the countries selected for this study:

Training Need in Data Collection	Preferred Method of Training	Preferred Training Provider
Questionnaire Design for Field Work Public Health Data Requisites Basic Epidemiology Basic Data Collection	Workshops Online Course Material Refresher Courses Hands on training for IT skills	In-house Exchange programs with international universities/ agencies Third Party Provider

Training for data collection on introductory epidemiology can be performed through in-house classroom-based training. However, if the project requires a specific skill set, involvement of a third-party provider (such as a visiting professor, or Oxford Policy Management in the case of IDS) can help with specific competencies.

This is the model taken by INDEPTH, Emory University and IDS, amongst others.

6.2 Data curation and analysis

Herbst describes his greatest concern as being the lack of adequate availability of research data managers and data scientists[41].

Neither is this unique to public health research. McKinsey, in a report on global trends in "big data", declares: "There will be a shortage of talent necessary for organizations to take advantage of big data. By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions." [48]

While the nebulous "big data" industry is only tangentially related to research data management, they require similar core skills. A shortage of this nature will drive wages up and pull those sufficiently skilled into industries willing to pay. The research industry will have to compete.

Even in entry-level positions, LMICs lag in developing the skills required for data management, data curation and data analysis. The risk is that, unless steps are taken to address this critical issue, much of the data generated will be rendered useless.

With respect to data curation and analysis, Lynn Woolfrey, at the University of Cape Town's DataFirst, believes a pure lecture training method may not be effective, as the data managers should be able to apply the skills in their work environment[49]. Similarly, using only online training will not be effective as there are constraints with respect to IT infrastructure and bandwidth, especially in parts of Brazil and remote areas in Africa. In addition, the motivation and engagement levels of participants in online statistical analysis courses are poor.

Feedback regarding data curation needs for the various LMIC respondents are as follows:

Training Need in Data Collection	**Preferred Method of Training**	**Preferred Training Provider**
Basic Data Curation Data Analysis Database Linkage	Combination of lecture training, on-the-job training, workshops and exchange programs with international universities or researchers.	In-house capacity development, with involvement of third party providers for specific skills. After initial training by third party provider, the in-house department has to take over to sustain interest in the training.

Note that interviewees are likely to request training to support needs which address their delivery requirements. Where researchers are not expected, or expecting, to perform data curation and long-term data publication they will not request it.

There are fewer skilled workers for data management in Vietnam and Africa as few universities offer relevant graduate courses[44].

The University of Cape Town has initiated a master's course in data management which started in 2013. Another South African University, the University of the Witwatersrand (Wits), will be offering a research data management post-graduate course for public health from 2014[49]. These are the only formal courses offering such skills in Africa we identified during the course of this study. Universities in North America and Europe have been offering similar – albeit less focused – courses for more than 15 years.

That said, and as described previously, research data management is offered as informal courses for students undertaking graduate studies, rather than as discrete courses offered to general students.

Neither India nor Brazil have specialised courses in data management, even though there are data managers available in the regions. India has been a data management hub for several years, and Brazil has access to principal investigators who have completed their MS and PhDs from universities in the US and Europe.

Some statistical offices and funded projects do provide basic training for staff, but there is significant staff turnover as these research studies are not focused on staff retention[39]. Trained researchers prefer to move out to the private sector from their current positions in government, especially in Africa and in India, where financial remuneration in research cannot compete with that of a corporate career.

Course examples from DataFirst include Saudi Arabia and Ghana, where training participants are provided with licensed data which they can use to apply skills they learn from the training workshop. Similarly, in Uganda, trainers spend time with data managers in statistical offices to understand work dynamics and operate together on questionnaire design, data derivation and data usage[39].

Blended training courses of workshops and on-the-job training, which provide hands-on experience on curation and analysis, followed by online refresher courses, appear beneficial to data managers.

Certification and accreditation are important for graduates, but The Union's model provides meaningful recognition. Anthony Harries, Senior Advisor for the course, describes its structure as follows:

- **Module 1:** research questions, protocol development and ethics (5 - 6 days)
    - Principles of operational research;
    - Teach how to write a protocol;
    - Then a day on ethics plus fill in ethics forms;
    - Protocols are done over five days;
    - How to source references, e.g. PubMed;
- **Module 2:** Data management and data analysis (5 – 6 days)
    - EpiData, CDC source for data
    - Use protocol (six-page document for defining a project – aims, objectives, methodology, variables to collect) – sources, registers, analytics
    - In EpiData have:
        - Quest file coded
        - Check file to ensure no errors
        - Results compiled
- **Module 3:** Paper writing, peer review and policy implications (5 – 6 days)
    - In the 8 months up to module 3, the project is implemented and data are collected and analysed in the field;
    - Everyone gets together in module 3 and writes the paper;
    - Within a month of course the paper must be submitted for publication or the student will fail;

From their courses, they identify promising people and provide one-year fellowships for operational research and leadership in government. The training sessions are divided between lectures, mentor-group sessions, and plenary sessions for presentations by participants.

The value to participants is as follows:

- Accredited by WHO and participants will have written and published a paper which provides immediate and lasting credibility;

- There is a plan to engage with universities to get the course recognised for credit for a Masters in Public Health ; 60 credits required for degree, this would count 10 credits;

As of end of June 2013, they had 95 students who completed the nine courses; 90% completed submissions, 75% have been published but they expect this to increase as publication is a slow process.

They do not stream students; everyone does the same course, but do have a mentorship program and match appropriate support to each student. The cost is about $7,500 per participant.

The Union and Medecins Sans Frontieres (MSF) run these courses now in collaboration with WHO-TDR branded as "SORT IT" (Structured Operational Research Training Initiative) – this means they have to follow a certain structure and are bound to the 80-80-80 targets[36].

The Union's course certainly doesn't cover the full ambit of requirements for research data management. The most appropriate course we have discovered (as far as its stated aims are concerned) is Wits University's Master of Science in Epidemiology in Research Data Management[50], as mentioned above. However we have not managed to contact them regarding the program which is due to start in 2014.

Harries describes the following concerns in offering training, which encapsulates many of the issues raised by others[47]:

- Courses need to take place in countries which permit visa issues to be circumvented (not everyone can travel easily). E.g. Nepal courses for Indian, Pakistan and Indonesian students; Kenya or Tanzania for Africa – Europe is harder since everyone needs a Schengen visa;
- Skills expectations amongst course participants have an effect on outcomes; out of 12 delegates there will be six good, four no problem, and two who are weak – mainly people nominated by government / public sector where pressure is exerted;
- Graduates can be head-hunted to the private sector even though there is a requirement for them to remain in public sector;

In the case of The Union, they also require participants to produce research papers of adequate quality for publication, and following this up is a concern.

The other difficulty is in scaling courses to accommodate the demand. Developing and providing course content is expensive and there are gaps in areas such as research data publication, as well as curation of such resources.

6.3 Data preparation and publication

As described in 3.4, research data publication is a long-term responsibility with high associated costs. There are no standardised products available at the time of writing. Work on methods for data citation as well as the interoperability of various metadata standards are still underway.

Research organisations which commit to data publication must take on obligations to provide support as well as complex technical assessments on their present and future data

requirements. It is understandable then, that many organisations consider data publication and sharing as having a secondary priority as they wait for others to take the lead.

Funders, though, are increasingly expecting every research project to have a plan for data management and data sharing. Interviewees report that their limitations for doing so are adequate training and infrastructure in which to publish such data.

They summarise their training needs as follows:

Training Need in Data Collection	Preferred Method of Training	Preferred Training Provider
Data Sharing Standards Linking Databases	Combination of lecture training, on-the-job training, workshops and refresher courses.	Combination of in-house training and sessions by specialised third party providers.

The selection of data-sharing infrastructure is another difficulty entirely and one that, potentially, research institutions shouldn't have to confront on their own. The choice is one they will make infrequently (potentially only once) and hiring or contracting in the skills they need to commission such software can be prohibitive.

7. Conclusions on issues limiting capacity and skills development in LMICs

7.1 Institutional scale and experience

Many of the issues raised throughout this report have to do with institutional scale and capacity.

Small research institutions cannot afford high-level experienced research directors nor does the volume of their research justify a complex hierarchy. However, without this in place, talented junior researchers lack the guidance they need to mature and are often forced by dint of professional ambition out of research.

Large and diversified research institutions and networks (including the larger world-class universities, and networks like INDEPTH) have the scale and research integration to justify investment in infrastructure and support.

The research organisations with that hierarchy of experience either already have processes in place to improve their skills deficits and improve their technical infrastructure or are working towards it, and vice versa.

Individual, consortia-based research projects – even those that operate over a number of years and are well-funded – frequently have immediate requirements for research generation and delivery of results. They also lack the capacity or time to develop skills amongst junior team members. They need to hire appropriate people for the task at hand.

Institutions that lack such a hierarchy of experience will also then be inexperienced at developing and writing funding proposals for public health research. They are less likely to lead such research, and then unlikely to develop a comprehensive and consistent research "culture" able to provide capacity building and mentorship to junior staff.

Support, then, is not only about putting training or infrastructure in place, it is about supporting collaborative support mechanisms. Once these are in place then the remaining technical issues will have the necessary champions receptive to addressing the concerns raised below.

7.2 Skills and systems for research data management training

Scale in diversified research institutions (i.e. organisations handling multiple projects on a range of topics simultaneously) can limit internal resources for training. The same is true of single-project research consortia.

In this case, support from a network or from third-party training services can support skills and systems development. Guidance on appropriate courses would reduce the complexity of selecting training as well.

7.3 Skills and systems for data sharing

Data sharing requires both staff experienced and skilled in preparing data for publication, as well as in responding to requests for data, and for the curation and long-term management of data resources.

LMICs experience a shortage of sufficient skilled research data managers and data scientists, and a consequent curtailed ability to manage, analyse, produce insight from and publish research based on their data.

There is also an issue (and by no means restricted to LMICs) that plans for data management and sharing are not developed or costed appropriately. Funders need to implement their policies more effectively through providing better guidance, assessing and tracking data plans, and ensuring costs are included into grants. Researchers and institutions also need to place more emphasis on planning.

That, however, is only one part of the problem. The software infrastructure for publishing research is still immature. There is no standardised mechanism or platform for sharing research data. Metadata systems are still largely independent of each other and interoperability is limited. Similarly, research data are released under a plethora of licenses which can severely limit downstream use of the data.

While a research organisation may be able to hire the skills to manage research data and publish it, setting up a research data publication system is cost-prohibitive and extremely complex. Choices regarding metadata standards, data security, archiving and the long-term support of the systems all require experienced data research managers working in well-integrated research environments. It is essential that any system work not just for immediate research needs but also for future projects without extensive modification for each project.

7.4 Professional recognition and citation standards for data publishers

There is limited recognition of research data managers through citation, references or acknowledgement. This has implications for their recognition as researchers, and their ability to develop suitable career progression.

Requiring professional certification has consequences, however. A similar recommendation for cybersecurity (unrelated skills, but a similarly new data-intensive technical industry) resulted in a report from the US National Academy of Sciences which declared that cybersecurity is an occupation, not a profession[51]. They describe the difference as follows:

"For example, formal education or certification could be helpful to employers looking to evaluate the skills and knowledge of a given applicant, but it takes time to develop curriculum and reach a consensus on what core knowledge and skills should be assessed in order to award any such certification."

"Once a certification is issued, the previously mentioned barriers start to emerge. The standards used to award certifications will run the risk of becoming obsolete. Furthermore, workers may not have incentives to update their skills in order to remain current. Again, this issue is seen in the industry today, as some professionals chose to let their certifications lapse rather than renew them or try and collect the required CPE credits."

"But the largest barrier is that some of the most talented individuals in cybersecurity are self-taught. So the requirement of formal education or training may, as mentioned, deter potential employees from entering the field at a time when they are needed the most. So while professionalization may be a useful tool in some circumstances it shouldn't be used as a proxy for "better.""

This introduces a new level of technical jargon, but the implications are clear. Certifications may be a useful long-term mechanism to build recognition for research data management as a profession, however, at present there are no uniformly agreed standards and many of the leading proponents of research data management are self-taught. Such certification can disadvantage self-taught research data managers by creating an expensive new barrier to entry.

This does not mean it is a terrible idea or that it should never happen, but that it should be adopted with care.

Funders can encourage institutions to enhance career progression but it isn't possible to mandate this. Raising the status of data managers is an important concern and mechanisms for raising their institutional profile, such as giving data managers general responsibility for contributing to training for research staff on data expertise, could support this.

Presenting research data management as a career choice to undergraduate students would also build awareness and recognition for the emerging profession.

8. Recommendations for supporting research data management

Any constructive approach to address the issues should be constrained by the following:

- A demand-driven solution addressing needs expressed by the end-users but which support the objectives of the funders;
- Offering incentives via the grant process, or enforcing existing requirements (regarding, for instance, data sharing) more clearly;

8.1 Promotion of collaborative research networks

Many research institutions in LMICs are too small to employ and support effective research data staff development or to implement research data management systems. The development of collaborative and contributory institutional networks can reduce costs while improving participation and research outcomes.

The strength of collaborative networks, such as INDEPTH, is in the international and comprehensive range of additional services that can be delivered to its members. The network permits individual organisations to specialise, or for specialist services to be affordably developed and provided.

Just as researchers can mentor other researchers, successful universities, such as the University of Cape Town or University of Sao Paolo, could mentor universities in LMICs through peer relationships across management and program level, rather than only at the research level.

Some universities could form regional hubs to support smaller institutions in their own country. In countries where local research capacity is limited collaborating globally would be beneficial.

Networks like INDEPTH or ALPHA are organised around specific public health research interests. A new series of collaborative networks focused on research data management training and infrastructure would similarly be beneficial. Members would have to participate and not just receive services.

Funders could insist that research organisations demonstrate capacity in research data management, including in long-term data curation and sharing, or become active members of a future shared research data management and infrastructure network.

Where institutions and networks become sufficiently experienced and capable, funders could support the institution's ability to incubate new research projects. This would permit appropriate research institutions in LMICs to set up funding models similar to that employed by Emory University (having a pool of funds to direct towards their own research priorities), as well as to ensure local awareness of the mechanisms for applying for funding the type of research already being performed in their countries.

8.2 Training and mentorship programs

While there are numerous schools of public health, as well as courses in systems and database administration, most research data management courses are modules designed for people with existing skills. More effective mentorship and training can act to "join up" the various skillsets required to be a research data manager in public health. In addition, specific careers advice at universities that such a career exists would also be useful.

Funders could commission training programs and offer these to their funded institutions, or – in support for the collaborative model in 8.1 – establish training hubs at one or more of the LMIC institutions they support.

Similarly, funders can support recognition of capable data research managers and librarians by arranging for them to act as consultants and trainers in LMICs and at LMIC training hubs.

Online training courses offer a lower cost and wider distribution of resources. This permits global standards in training to emerge as well as enabling trainers to circulate to support training but without having to be permanently available. In isolation, however, purely online training does suffer from poor completion rates and needs to be integrated with further support.

Online training requires meaningful research data in an online format which can then link such courses to the use of shared data and the direct production of new research insight (i.e. new data from existing sources).

A way of joining all of these together, and scaling the existing Union training course, would be to link shared research data publication to a mix of online- and in-person training with the requirement for publication of new research as a condition for graduation.

A partnership between iSHARE2 and The Union / WHO TDR would provide global data produced by iSHARE2 to become the source data used in training. Given the global nature of INDEPTH research data, this would permit similar scale in training opportunities for participants to work with research meaningful to themselves.

8.3 Professional certifications or funding of fellowships

Recognizing or promoting formal professional qualifications or accreditation has been suggested as helping to raise the profile of data managers. This is worth exploring but should also be managed with care for the reasons raised in 7.4.

Fellowships, as an alternative approach, would be more effective in the short-term for supporting institutional and researcher recognition of the importance of the research data manager role, and of sharing of research data along with published results.

Research institutions can be supported into developing a dedicated fellowship funding scheme (or Masters scheme) for ‘data scientists’, or data managers and librarians. This would be a path on the way towards professional certification, or simply a mechanism by which formal training and recognition can be achieved. Course materials can be shared through a network to ensure that consistent approaches to data management are achieved.

8.4 Data management as a condition for funding

Funders have their greatest leverage through their funding relationship and agreements with grantees.

Funders can and should ensure that grantees have systems in place to share their research data along with their published findings. Funders will need to ensure that grantees have adequately scoped the lifetime cost of their data management system, or that they are members in good standing of research data sharing networks, and that these systems are audited for delivery.

As part of their research data management system, institutions should have a clear strategy in place for data management and sharing, including career paths for data managers.

This needs to be at the institutional or network level, rather than requiring that individual research projects offer this capacity. Research data sharing would then be embedded within institutional support.

This will challenge research agencies to ensure that they have such staff available, and help to create a viable market for research data publication software. Institutional – rather than project-level – research data support will also reduce the proliferation of competing standards.

Institutions need not purchase software or resources for each project, but can share resources or take advantage of open-source software systems. Similarly, they need not set up entire research data management programs from scratch but could join recognised networks which support this.

New mechanisms for auditing need to be established to ensure that funders are able to track compliance. Future funding for institutions should depend on the degree to which individual research projects comply with these requirements.

It is also important that compliance not create a new cost and logistical burden for both funders and grantees. A third-party audit report to a common set of standards should be sufficient for any funder without each requiring their own due diligence.

8.5 Data citation as a condition for publication

It is not sufficient merely to have a system for research data management in place. Any data used in a research publication, whether from primary sources developed by themselves, or from the use of already published data from others, must be cited in that final research. Grantees and research publishers must demonstrate not only that they are citing the data in their findings but that such data is available for sharing in a well-managed research data management system.

Grantees should also report, or be encouraged to report, data citation metrics as an indication of the value of their research.

Data citation is separate from data management and sharing. Just because the data are shared does not imply that they are cited. Citation cannot happen where the data are unavailable (i.e. are not shared via a research data management system) and such sharing becomes meaningful and valuable when the data are cited. Both are required.

Finally, data must also be released under a progressive license permitting easy reuse without having to chase individual publishers for permission. An equivalent license to the Creative Commons Attribution, Share-alike and Redistribution classification would be most beneficial[52]. Training, as in 8.2, would be severely inhibited if students had to request permission from each of hundreds of publishers in order to produce simple analysis.

8.6 Metadata interoperability

Different research specialisms support a wide range of different metadata standards which can become a barrier to sharing and reuse of research data. A mechanism for interoperability of such standards must be developed to ensure that this does not become an excuse for not publishing research data.

A research data management software system can serve the broadest possible research interests within an institution if it can offer support for a wide range of metadata standards. The infrastructure costs can then be borne by a larger community of researchers. Promoting interoperable research data metadata standards also ensures that research does not become "locked" into particular standards after publication and that future changes can be incorporated.

Metadata interoperability can be performed, for instance, by producing a set of grammars which relate directly to a Resource Description Framework (RDM)[53]. It will be necessary to build a library of mappings from each metadata standard through to a common RDM.

8.7 Open source data publication infrastructure

If research data sharing and management are to be widespread, then the most effective approach is to have software systems that can be easily shared and extended. Institutions and publishers should not be concerned that they are exposing themselves to perpetual license fees or the danger of locking their research into "walled-garden" or proprietary software.

It is important not to replace a problem of unavailable data with that of unavailable systems.

Over time, as demand for research data publication and data scientists rises, it would be worthwhile supporting development of an open-source and common approach to data publication. This would reduce individual institutional costs as well as software vendor lock-in. Interoperable standards will similarly permit data to be accessible across different research sectors.

The combination of open source software along with open standards can lead to wide adoption and an increase in the number of applications for such systems.

A shared research data management and infrastructure network can become responsible for the development and maintenance of such a platform. There are numerous examples of how this can work, which includes open source software projects like the Apache suite of software, the Python programming language, and even the Open Knowledge Foundation's CKAN data publication platform.

8.8 An implementation matrix

While it is difficult to specify within the constraints of this report exactly how each of these proposals could be implemented, we have presented a table indicating the degree of difficulty and time-frame for each of the recommendations:

	Difficulty	Time-frame	Funder led
8.1 Promotion of collaborative research networks	Medium	Multi-year	Partially
8.2 Training and mentorship programs	Low	Short-term	Yes
8.3 Professional certifications	High	Long-term	No
or funding of fellowships	Low	Short-term	Yes
8.4 Data management as a condition for funding	Low	Short-term	Yes
8.5 Data citation as a condition for publication	Medium	Multi-year	Partially
8.6 Metadata interoperability	Medium	Short-term	Partially
8.7 Open source data publication infrastructure	Medium	Multi-year	Partially

In addition, these objectives can be related to the goals espoused in the joint statement by funders as part of the Public Health Research Data Forum:

Immediate goals

- Data management standards support data sharing [8.1, 8.2, 8.6, 8.7]
- Standards of data management are developed, promoted and entrenched so that research data can be shared routinely, and re-used effectively. [all]
- Data sharing is recognized as a professional achievement [8.3]
- Funders and employers of researchers recognize data management and sharing of well-managed datasets as an important professional indicator of success in research. [8.4, 8.5]
- Secondary data users respect the rights of producers and add value to the data they use [8.5]
- Researchers creating data sets for secondary analysis from shared primary data are expected to share those data sets and act with integrity and in line with good practice - giving due acknowledgement to the generators of the original data. [8.4, 8.7]

Longer-term aspirations

- Well documented data sets are available for secondary analysis [8.1]
- Data collected for health research are made available to the scientific community for analysis which adds value to existing knowledge and which leads to improvements in health. [all]
- Capacity to manage and analyse data is strengthened [8.1, 8.2]
- The research community, particularly those collecting data in developing countries, develop the capacity to manage and analyse those data locally, as well as contributing to international analysis efforts. [8.1, 8.7]
- Published work and data are linked and archived [8.4, 8.6, 8.7]
- To the extent possible, datasets underpinning research papers in peer-reviewed journals are archived and made available to other researchers in a clear and transparent manner. [8.4, 8.6, 8.7]

- Data sharing is sustainably resourced for the long term [8.1, 8.3]
- The human and technical resources and infrastructures needed to support data management, archiving and access are developed and supported for long-term sustainability. [8.1, 8.4, 8.6, 8.7]

No doubt, new opportunities to support the objectives of the Public Health Research Data Forum will emerge as these initial proposals are developed.

The ultimate objectives and aspirations for ensuring the availability of public health research data to the scientific community will be achieved through collaboration involving funders, research institutions, publishers, and a diversity of service providers.

We hope that this report will serve as a helpful foundational step towards achieving these ends.

Appendix: University courses for research data management in public health

Research data management for public health, with a focus on scientific research and publishing, requires skills that link data management to public health and research. As far as training goes, though, those wishing to learn often have to study each stream (public health research, system and data administration, and research data management) independently of each other.

System and data administration

Just about every university and vocational training provider offer courses in computing and IT. Importantly, any suitably ambitious individual can gain software-related skills through independent study, unlike public health research which will require training through a suitable institution.

Online courses are available through **The Open University** (http://www3.open.ac.uk/study/undergraduate/qualification/q62.htm) or even the **O'Reilly School of Technology** (http://www.oreillyschool.com/individual-courses/).

There are also numerous informal mechanisms, such as **Stack Overflow** (http://stackoverflow.com/), where people can gain information.

Schools of public health

In the United States, the **Association of Schools and Programs of Public Health (ASPPH)** maintains a list of Council on Education for Public Health (CEPH)-accredited schools and programs for public health (http://www.aspph.org/members/cephaccreditedmembers.cfm).

In Europe, the Association of **Schools of Public Health in the European Region (ASPHER)** serves a similar role (http://aspher.org/pg/pages/view/78/aspher-members).

A small sample of schools offering graduate degrees in public health are listed below (descriptions are taken from the relevant course outlines):

Oxford University, MSc in Global Health Science:

http://www.dph.ox.ac.uk/courses/gradstu/globalhealth/specs

The course aims to promote advanced study of the challenges of global health and their potential solutions by in-depth study of a range of scientific disciplines, so that students may understand and integrate medical, epidemiological, social and economic aspects of ill-health in developing countries.

The course will aim to develop students':

- knowledge of the major global health problems and their potential solutions
- knowledge and skills in techniques of analysis of global health problems, particularly principles of epidemiology and statistics, health policy and public health, and

international development with opportunities for training in additional specialist fields
- capacity to critically appraise evidence in global health
- skills and practical experience in researching specific health problems

Cambridge University, Institute of Public Health, MPhil in Public Health:

http://www.phpc.cam.ac.uk/graduate-studies/mphil-in-public-health/

- A strong foundation in epidemiology and biostatistics
- Modules on major disease/exposure clusters from research leaders in these fields (including genetic epidemiology and public health genetics
- A major module on public health assessment methods – including metrics increasingly used in international public health assessments
- Strong coverage of communicable disease epidemiology and control
- Coverage of standard professional curricula – including the UK Faculty of Public Health
- Teaching examples drawn from faculty experience in a range of national and international settings – including UK, Australia, Bulgaria, Poland… and the EU more generally
- Major emphasis on thesis work – with potential supervision for a wide range of topics – from epidemiology to public health assessments to qualitative studies

Sheffield University, Master of Public Health (MPH):

http://www.shef.ac.uk/scharr/prospective_students/masters/mph

The MPH seeks to provide students with an in depth understanding of issues in public health principles and practice, and apply this to the specific challenges in delivery, planning and management of health services in their national context. The learning outcomes are:

- Thorough understanding of global and national public health issues
- Insight into the global drivers of reform in health systems and their potential impacts on future public health policy directions in both developed and developing nations.
- Comprehensive understanding of the tools available to systematically assess and evaluate health needs
- Critical awareness of how the research process may be applied in the study of public health

Boston University School of Public Health, Master of Public Health:

http://sph.bu.edu/Academic-InformationsDegrees-a-Programs/degrees-a-programs/menu-id-617074.html

The Master of Public Health, with a concentration in Epidemiology, provides training in the principles and methodology of epidemiological research and practice. Students in this program explore the theories and methodologies underlying the science, and learn how to design, conduct, analyze, and interpret research studies in such areas as cancer epidemiology,

reproductive epidemiology, and infectious disease epidemiology. Graduates pursue advanced degrees or research or management careers in the public, private, or academic sectors.

Brown University School of Public Health, Master of Public Health

http://brown.edu/academics/public-health/undergraduate/graduate-programs/mph-program-about-us

The mission of the MPH program is to preserve and enhance the health and wellbeing of human populations by preparing graduates in the knowledge, skill, and analytic capabilities required to 1) advance the principles and practice of public health; 2) enter public health careers at the local, state, and national levels with the skills necessary to assume leadership roles; and 3) uphold and foster an ethic of social responsibility which recognizes the value of equal opportunity for health and wellbeing among all and which respects individual, family and community values.

University of California, Los Angeles, Master in Public Health

http://ph.ucla.edu/degrees-and-academics/degree-programs/mph-degree-programs

The Master of Public Health is a professional degree that will prepare you to solve public health problems by applying professional disciplinary approaches and methods in professional environments such as local, state or national public health agencies and health care organizations.

The MPH is a School wide degree, allowing you to gain broad training in public health, but you'll also specialize in one department so that you can concurrently gain focused knowledge in a particular area. Students specialize in one of the School's five departments: Biostatistics, Community Health Sciences, Epidemiology, Environmental Health Sciences or Health Policy and Management.

Harvard University, School of Public Health, Master of Public Health:

http://www.hsph.harvard.edu/master-of-public-health-program/program/

The MPH is a demanding, interdisciplinary program emphasizing active, student-directed learning, problem solving, and the acquisition of skills essential to the practice of public health. MPH applicants must hold an MD, DO, DMD, JD, or health-related doctoral or prior master's degree plus experience.

Johns Hopkins, Bloomberg School of Public Health, Master of Public Health:

http://www.jhsph.edu/academics/degree-programs/master-of-public-health/

The Hopkins MPH degree prepares students through multidisciplinary approaches that apply the latest scientific knowledge, common sense and teamwork to solve important health problems. Students in the program will obtain a population-based perspective on health, along with rigorous training in a school wide curriculum focused on the core disciplines of epidemiology, biostatistics, management sciences and the environmental, biological, behavioral and social factors that influence the health of populations and communities.

Hopkins MPH students apply their knowledge and skills to practical problems and integrate their competencies in a culminating capstone project.

General training in research data management

Universities do offer research data management courses to their graduate students, however, these courses are extracurricular and often not part of the students' registered courses. A student on an epidemiology, public health, or similar health related course could access this RDM training. Such courses are not necessarily specific to healthcare. There is also the assumption that you already have a background in research.

The Digital Curation Centre maintains a list of courses associated with research data management (http://www.dcc.ac.uk/training/data-management-courses-and-training). The University of Glasgow maintains another list (http://www.gla.ac.uk/services/datamanagement/training/).

Possibly the most interesting and valuable is that the online university, **Coursera**, now offers a free course (provided by **Vanderbilt University**) entitled **Data Management for Clinical Research:**

https://www.coursera.org/course/datamanagement

This course is designed to teach important concepts related to research data planning, collection, storage and dissemination. Instructors will offer information and best-practice guidelines for 1) investigator-initiated & sponsored research studies, 2) single- & multi-centre studies, and 3) prospective data collection & secondary-reuse of clinical data for purposes of research. The curriculum will balance theoretical guidelines with the use of practical tools designed to assist in planning and conducting research. Real-world research examples, problem solving exercises and hands-on training will ensure students are comfortable with all concepts.

JiscMRD supported a UK-based project called 'Datum for health' and lead by Professor Julie McLeod, Northumbria University, School of Computing, Engineering & Information Sciences[54].

The materials have been released as a series of training materials[55] from the DATUM for Health project page[56]. The materials are described as follows:

"It aims to provide PGR students with the knowledge to manage their research data at every stage in the data lifecycle, from its creation to its final storage or destruction. Students learn how to use their data more effectively and efficiently, how to store and destroy it securely, and how to make it available to a wider audience to increase its use, value and impact."

It comprises three sessions:

Session 1: Introduction to research data management
Session 2: Data curation lifecycle
Session 3: Problems and practical strategies and solutions

JISC's Digital Curation Centre specifically recommends the DATUM materials for health students[57].

Other online examples include:

University of Edinburgh, Research Data Management Training

http://datalib.edina.ac.uk/mantra/

MANTRA is a free, non-assessed course with guidelines to help you understand and reflect on how to manage the data you collect throughout your research. The course is particularly appropriate for those who work with digital data.

On completion of this course you will:

- Be aware of the risk of data loss and data protection requirements.
- Know how to store and transport your data safely and securely (backup and encryption).
- Have experience in using data in software packages such as R, SPSS, NVivo, or ArcGIS.
- Recognise the importance of good research data management practice in your own context.
- Be able to devise a research data management plan and apply it throughout the projects life.
- Be able to organise and document your data efficiently during the course of your project.
- Understand the benefits of sharing data and how to do it legally and ethically.

MANTRA is maintained by Data Library staff in Information Services, University of Edinburgh. It was originally developed in collaboration with the Institute for Academic Development as part of a JISC-funded Managing Research Data project (2010). The content was developed based on a needs assessment with three postgraduate training programmes at the University of Edinburgh in the fields of geosciences, social and political sciences and clinical psychology.

University of Lincoln, Research data documentation and training materials:

The following suite of training materials have been developed at the University of Lincoln by the Orbital project up to March 2013. They provide an overview of research data management concepts and practices, links to useful external resources, and specific information on using data tools & services at Lincoln.

https://orbital.lincoln.ac.uk/training-introduction

University of Oxford, Research Data Management:

http://www.admin.ox.ac.uk/rdm/

If you are starting a new research project, then you will need to consider issues relating to the management of research data. Many of these issues will be relevant whether or not the research is funded by an external sponsor. By managing your data you will ensure:

- Funding and regulatory body requirements are met
- Research data remains accurate, authentic, reliable and complete.
- Duplication of effort is kept to a minimum
- Research data keeps its integrity and research results may be replicated.
- Data security is enhanced, thus minimising the risk of data loss

University of Cambridge, Support for Managing Research Data:

http://www.lib.cam.ac.uk/dataman/

This data management website is a product of the Incremental project, a collaboration between the Cambridge University Library and the University of Glasgow's Humanities Advanced Technology & Information Institute, funded by JISC. It is intended to provide researchers, computing officers, and administrators with guidance and tools to manage, re-use, and preserve electronic resources as easily as possible.

Examples of offline resources include:

University of Minnesota, Data Management Course

https://www.lib.umn.edu/datamanagement/workshops

The University Libraries' course on data management is designed for graduate students at the University of Minnesota who seek to prepare themselves as “data information literate" scientists in the digital research environment. Detailed videos and in-class workshop activities will help you prepare for the specific and long-term needs of managing your research data. Experts in digital curation will describe current sharing expectations of federal funding agencies (like NSF, NIH) and give advice on how to ethically share and preserve research data for long-term access and reuse

Students will get out of this course:

- Five workshop sessions to gain hands-on skills for managing digital research data accompanied with 3-9 min video lessons that you can watch anytime online or download to your device.
- A Data Management Plan (DMP) template with tips on how to complete each section. Your completed DMP can be used in grant applications or put into practice as a protocol for handling data individually or within your research group or lab.
- (Optional) Feedback and consultation on your completed DMP by research data curators in your field.

Participants that attend all 5 data management workshops in this series will receive a Certificate for their UMN training records.

University of Bath, Managing your research data

http://www.bath.ac.uk/learningandteaching/rdu/courses/pgskills/modules/RP00076.htm

This session will introduce the subject of managing research data: how it should be done, why it is necessary, and what options and strategies are available for successfully managing, storing, archiving and retrieving information.

After this session, you will be able to:

- Explain what research data are and to whom they belong,
- Keep your data safe, secure and organised,
- Explain who has responsibility for different aspects of data management,
- Find and cite existing data to reuse for your own research,
- Archive and share your research data and understand why you might want to do this.

Research data management networks

These organisations promote research data management and may be able to provide support to research institutions or provide a model for a future research data management support network for LMICs:

American Society for Information Science and Technology (ASIS&T)

http://www.asist.org/

This group sponsors annual meetings, workshops, and symposia, including the 2010 Research Data Access and Preservation Summit in Phoenix, Arizona. Special interest groups provide avenues for professional development, including the Bioinformatics, Digital Libraries, and Scientific and Technical Information groups.

Association of European Research Libraries (Ligue Des Bibliothèques Européennes De Recherche, LIBER)

http://www.libereurope.eu/

A working group of this association is focusing on the topics of e-science, research data, and workflows for the 2010-2012 biennium.

Association of Research Libraries (ARL)

http://www.arl.org/

ARL, which includes libraries from North America, is concerned with the changing roles of research libraries, including the curation of research data. The association provides links to member activities, surveys, policy guidelines, reports, outreach resources, and training.

Australian National Data Service (ANDS)

http://ands.org.au/

ANDS is an organization for all higher education providers and publicly funded research organizations in Australia. It investigates and develops policies, guidelines, and examples of research data ownership and access, including curation policies and tools for collecting and managing data.

Australian Partnership for Sustainable Repositories (APSR)

http://www.apsr.edu.au/

This Australian organization provides outreach and education, and funds collaborative research and development projects for digital collections, including digital research data.

Coalition for Networked Information (CNI)

http://www.cni.org/

CNI, sponsored by Educause and the Association of Research Libraries, supports projects, meetings, and conferences to build systems, standards, practices, and capacity for networked information. Task force meeting presentations and links to sponsored programs are available on the site.

Commerce, Energy, NASA, Defense Information Managers Group (CENDI)

http://www.cendi.gov/index.html

Fourteen U.S. Federal Agencies make up this working group. Its goal is to improve efficiency in the areas of scientific and technical information capabilities. Interest areas in CENDI include metadata, taxonomies, preservation, and virtual libraries, and the group has provided a workshop on managing scientific data for its members.

Committee on Data for Science and Technology (CODATA)

http://www.codata.org/

This committee of the International Council for Science (ICSU) works to improve accessibility and quality of scientific data sets, through working groups, workshops, publications, and conferences.

DataCite

http://datacite.org/

This international consortium works to facilitate access to scientific research data through data registries and the promotion of research data as citable materials in the scientific record.

Digital Curation Centre (DCC)

http://www.dcc.ac.uk/

Created and funded by the United Kingdom's JISC (Joint Information Systems Committee), this organization supports and funds a large number of projects in data curation research, policies, tools and systems for UK higher education institutions.

Digital Preservation Coalition (DPC)

http://www.dpconline.org/

This not-for-profit coalition of both private/commercial and public organizations and individuals of the UK provides support for the adoption of digital preservation policies and practices.

Dublin Core Metadata Initiative (DCMI)

http://dublincore.org/

This open membership organization is dedicated to the development of metadata standards that are vital to describing scientific data.

International Association of Scientific and Technical University Libraries (IATUL)

http://www.iatul.org/

This association sponsors an annual conference with opportunities to network with librarians from around the world. The 2010 meeting included a variety of presentations on digital data curation (http://docs.lib.purdue.edu/iatul2010/).

International Council for Scientific and Technical Information (ICSTI)

http://www.icsti.org/

An international organization which sponsors scientific and technical projects, such as the integration of data citation with text, which culminated in the DataCite consortium.

Joint Information Systems Committee (JISC)

http://www.jisc.ac.uk/

Funded by UK higher education funding bodies, JISC manages an extensive list of projects, programs, and services for information technology innovation. These include data services and collections, digital repositories, open technologies, standards, and infrastructure.

Research Data Strategy Working Group

http://rds-sdr.cisti-icist.nrc-cnrc.gc.ca/eng/index.html

A component of Research Data Canada, this working group draws its membership from libraries, research institutions, agencies and individuals, and is focused on solutions for managing Canadian research data.

Research Information Network (RIN)

http://www.rin.ac.uk/

The Research Information Network is funded by UK higher education and national libraries. RIN does open science case studies, research on data centers, and develops principles and guidelines for data stewardship.

UK Research Data Service (UKRDS)

http://www.ukrds.ac.uk/

This joint project funded by Joint Information Systems Committee (JISC) and the Higher Education Funding Council for England (HEFCE) is developing a planning and costing model for a national shared digital research data service for UK higher education.

Capacity building for research data management and public health in LMICs

There are a small number of initiatives catering to public health research data management training in LMICs:

The **University of the Witwatersrand is to offer a Master of Science in Epidemiology (MSc) In Research Data Management**[58] from 2014.

http://www.wits.ac.za/10568/

This appears to be an epidemiology MSc, but with a module specifically on research data management. In many ways, the course does not appear to be much different from those listed under the Masters in Public Health programs above, but it is new for Africa:

The Division of Epidemiology and Biostatistics of WSPH offers a Master of Science in Epidemiology to provide training in the principles and practice underlying the discipline of epidemiology. There are currently three fields of study within the MSc programme:

- Epidemiology and Biostatistics
- Population-based Field Epidemiology
- Infectious Disease Epidemiology

Epidemiology is fundamental to clinical and community medicine and public health. In combination with basic medical science and clinical research, epidemiology provides the tools with which we can learn more about the aetiology and effects of disease, the opportunities for prevention, the cost and effectiveness of various diagnostic and therapeutic approaches, and the health status and risks of individuals and populations.

University of Cape Town, DataFirst

http://www.datafirst.uct.ac.za/home/

DataFirst's Research Data Centre (RDC) provides a secure setting for improved access to national census and survey microdata for research purposes. Trained staff in the RDC assists

data analysts with data and software queries. The RDC is located in the School of Economics building on the Middle Campus at UCT. An extensive collection of books devoted to survey design and analysis is available for use in the RDC.

DataFirst does provide workshops and training but has no standardised generally available research data management courses at this time.

Another project is Africa Build[59], which "aims to use information technology to improve capacity for health research and education in Africa. It seeks to provide innovative learning and research opportunities to individuals and institutions throughout the continent.":

AFRICA BUILD is a Coordination Action aiming to support and develop advanced Centres of Excellence in health care, education and research in the African countries, through Information Technologies. This project is supported by the EU's Seventh Framework Programme (FP7-ICT). AFRICA BUILD started 1st August 2011 and will run for a period of 36 months.

This support will come, in the initial phases, from established research centres located at EU countries, including WHO, the leading medical organization at a worldwide level. For later phases, a significant challenge of this CA is to generate virtual communities of African researchers that can continue these initial efforts by creating, developing and exchanging, collaboratively, new knowledge, methods, informatics tools and data. The AFRICA BUILD vision aims to address fundamental problems in health research and education in a low income geographical area like Africa, providing innovative solutions by optimizing and sharing resources through the use of novel technologies.

From a conceptual perspective AFRICA BUILD will provide the scientific, technological and financial support for developing centres of excellence in health education and research in Africa. These activities will foster the capacities and scientific excellence in the African centres and will be the starting point for future collaborative developments that will ensure sustainability once the CA ends. Experiences of the consortium in previous successful projects and initiatives will be used for the exchanges and initiatives proposed in this Coordination Action.

A symposium on 8 October 2013 aimed to present the outcomes of the research phase[60]. Unfortunately, updates from the project appear to have stopped being disseminated since September 2013, so we are unable to detail their findings.

Appendix: Study interviewees and acknowledgements

The complete list of people interviewed as part of the study (in alphabetical order) are:

- Anthony D Harries, Senior Advisor, International Union Against Tuberculosis and Lung Disease (The Union)
- Cesar Victora, Universidade Federal de Pelotas, Brazil
- Dr Cesar De Oliveira, University College of London, Institute of Epidemiology and Healthcare
- Dr Garry Aslanyan, Manager, Partnerships and Governance for Special Programme for Research and Training in Tropical Diseases, World Health Organization
- Dr Jimmy Whitworth, Head of International Activities, Wellcome Trust
- Dr Tran Huu Bich, Vice Dean, Hanoi School of Public Health, Vietnam
- Jeffrey P. Koplan, Vice President for Global Health, Emory Global Health Institute
- Kobus Herbst, Deputy-Director at Africa Centre for Health and Population Studies, INDEPTH Network, iSHARE2
- Lynn Woolfrey, Manager, DataFirst, University of Cape Town
- Salle Atkins, Scientific Co-ordinator, Department of Public Health Sciences, Division of Global Health (IHCAR), Karolinska Institutet, Stockholm, Sweden
- Tathagata Bhattacharjee, Senior Data Manager - iSHARE2, INDEPTH Network, Asia Node
- Tom Barker, Senior Nutrition Convenor, Institute of Development Studies

Other contributors who provided time and insight:

- Dr Hendrik Bunke, ZBW
- Joss Winn, University of Lincoln
- Michelle Brook, Sander van der Waal, Lucy Chambers, Marcus Dapp, Jenny Molloy, Open Knowledge Foundation
- David Carr, The Wellcome Trust
- Ruth Levine and Kristen Stelljes, The Hewlett Foundation

Appendix: Country assessment criteria for study inclusion

Following the general secondary-research overview, a set of five low-medium income countries (LMICs) were to be selected for analysis of in-country research data management experiences in existing projects.

The assessment criteria for inclusion of a country in the primary research phase are:

1. Health data released by public sector:

    - Does the department of Health or national statistics office release regular disease mortality and morbidity data?
    - Does the government appear to track major public health issues and emerging infectious disease threats?
    - Is the government an active participant in the Global Health Observatory http://www.who.int/gho/en/

2. Overall economic health:

    - Are international training organisations able to operate locally (e.g. any local branches of foreign universities)?
    - Is the national statistics office a statutory body, independent of government interference?
    - Foreign Direct Investment as a percent of GDP;
    - Foreign Direct Investment as a percent of global total;
    - Corruption perceptions index score;
    - Economic Freedom index score;
    - Doing Business index score;

3. Public health interviewees landscape:

    - How many public health research bodies are active in the country?
    - How many healthcare teaching hospitals and universities in the country?
    - How many contacts do we already have in hand?

A basic "traffic light" (red, orange, green) overview of the results is detailed below:

	Health data released by public sector	Overall economic health	Interviewees landscape	Selected (Y)
Latin America				
Brazil				Y
Mexico				
Africa				
Egypt				
Kenya				
Morocco				
Senegal				
South Africa				Y

Uganda				Y
Asia				
India				Y
Indonesia				
Thailand				
Vietnam				Y

The final selection is:

- Brazil (Portuguese)
- Uganda (English)
- South Africa (English)
- India (English)
- Vietnam (Vietnamese)

Appendix: Africa Centre, member of INDEPTH Network

The Africa Centre for Health and Population Studies, UKZN is:

- Primarily funded by the Wellcome Trust as one of its Major Overseas Programmes
- Has been conducting demographic surveillance on approximately 90 000 individuals in rural KwaZulu-Natal since 2000
- Offers a SAQA accredited training programme to its fieldworkers
- Is a member of the INDEPTH Network (but also of other collaborations, like the ALPHA network)

The INDEPTH Network:

- Is a network of 48 Centres like the Africa Centre
- Has various funders, including the Wellcome Trust, as far as I know WT is a long term funder of the network but probably not the dominant one
- Promotes the harmonisation and sharing of data from its members Centres through data use agreements with them that allows the Network to publish and share data on the INDEPTH Data Repository and INDEPTHStats

The iSHARE2 project:

- Is a continuation of the INDEPTH iSHARE initiative which is aimed supporting the last point above, this initiative has had various funders, including Hewlett Foundation, Sida, Bill & Melinda Gates Foundation, Wellcome Trust & IDRC
- Is funded through a strategic award from the Wellcome Trust
- Trains, equip and support INDEPTH member Centres to harmonise, quality assure, document and share their research datasets
- It uses tools developed by others, e.g. NESSTAR Publisher, NADA, Kettle, etc to do this
- It has developed a 'research data management appliance', the 'Centre-in-a-Box' to bring these tools together in a device that can be deployed ('plug and play' fashion) into the diverse and often under resourced IT infrastructure of research centres in LMIC and can be remotely managed and supported
- Runs a helpdesk that users can email to get technical support

Africa Centre[41] recruits a number of general employees to operate their projects in KwaZulu / Natal and the approach below is relatively typical for their projects globally.

Fieldworkers and data capture/process:

- High-school graduates, must be from local community;
- There are about 1,000 high-school graduates per year in the area and 30-50 are recruited for fieldwork;
- Fieldworkers and data-capturers receive training from the Africa Centre for the job and this is South African Qualifications Authority (SAQA) registered;
- Some parts of course are standardised (e.g. verbal autopsy to WHO standard) but still have variability due to different health priorities;

- The trainer for the fieldworkers can be from specific specialisms (e.g. blood measures), from Medical Research Council, external expertise, or contract workers;
- Pay for workers depends on who the nominal employer is (many are with University Kwazulu/Natal, some permanent, some not);
- Even with the uncertain employment conditions staff tend to remain in the area since there are no other jobs (5,000 applications for every 5 jobs); 10% staff turnover;

Data processing:

- Large and complex longitudinal datasets in large and complex data available on a database;
- Recruits come from computer or statistics background;
- Data processing skills – research data manager – translates what scientists ask into SQL calls;
- Never have enough and struggle to recruit;

Pay:

- $7,000 p.a. per fieldworker;
- $10,000 p.a. per senior fieldworker;
- $18,000 p.a. for coordinator;
- $60,000 p.a. for a data scientist;

Appendix: Careers in Healthcare Information Management

Healthcare information management (HIM) is large field that is relatively different from that defined for this study. One classification[61] gives about 130 job types, 30 of them specifically in the area of data analysis.

The World Health Organisation classifies healthcare information management professionals as people who "plan information systems, develop health policy, and identify current and future information needs. In addition, they may apply the science of informatics to the collection, storage, use, and transmission of information to meet the legal, professional, ethical and administrative records-keeping requirements of health care delivery. They work with clinical, epidemiological, demographic, financial, reference, and coded healthcare data."

While there is certainly a requirement for research data management in public health research, the greater employment demand is across the entire healthcare industry, from hospitals to medical device and pharmaceutical manufacturers. Research data management in public health can be considered a highly specialised subset of the health information management industry but data managers would perform similar roles throughout.

The information presented below gives some context to the professional milieu in which research data managers may find themselves.

Health providers collect a huge amount of data - on patient symptoms, diagnoses, interventions, outcomes, and cost of treatments. From this they need, among other things, to measure and report on performance, evaluate treatments and cost effectiveness, check that they are meeting targets and any applicable laws and regulations. Manufacturers similarly track disease morbidity and mortality to assess demand for new products and treatments.

This requires that the raw data is captured in information systems, coded according to standard classifications, and a battery of analyses and presentations run. Data collection, coding and analysis are three important areas of HIM. Another is clinical audit, working with health professionals to use data to improve outcomes and clinical effectiveness.

In order for all these tasks to be possible, data collection systems and databases must be designed, implemented and maintained, a role for more senior staff. The analysis is only as good as the data that feeds it, so a data quality manager might be in place to review the information systems and ensure the data is of a consistent quality. Finally the whole array of data systems and procedures will be overseen by dedicated managers.

From a public health perspective, this entire chain then reports into research institutions and statutory bodies. For instance, reporting on maternal and child mortality starts at the local clinic with appropriate data collection and classification.

Careers in Health Information Management

In the US, there is a professional body specifically for HIM practitioners, the American Health Information Management Association (AHIMA)[62]. AHIMA runs a Health Information Careers website where a Career Map[63] contains information on over 60 jobs in the field, graded from 'entry' to 'master' level. These cover a wide range of subfields (IT infrastructure, administration, etc.), but even in the narrow area of informatics and data analysis, the map

lists 10 jobs, an indication of how important the field has become. This is a specialised field: none of the 10 are entry-level jobs (though some clearly-related jobs are). The lowest listed in the field is 'Clinical Data Analyst', a post whose holder will need to use database software such as SAS, MS Access, and SQL, as well as medical terminology and classifications.

In the UK, the National Health Service, Department of Health, and others have collaborated on the still more comprehensive Health Informatics Career Framework (HICF)[61] abovewith around 130 roles. They are classified into nine levels, from entry level to the most senior management roles, and across seven subfields. One of these, Information Management, includes 30 types of job related to health data analysis. The Framework is intended for various purposes including helping practitioners plan their own careers, and includes very helpful notes on types- and levels of qualifications and experience needed, as well as the types of work involved. The NHS's own careers website has a similar list with 12 jobs[64], and a page specifically on entry requirements for HIM jobs[65].

A range of current job advertisements can be found at HealthJobsUK[66], which gives an idea of salary ranges. In practice the range of job titles advertised is much more various. E.g. a data analyst might be called a 'Senior information manager and data services officer'. Despite the large proliferation of roles and titles, it's possible to distinguish four main types of job in the area: data entry, data analysis, data systems planning, and senior management. Some very rough salary indications and sample job titles from the HICF classification are listed below for each category. The US salary ranges given by AHIMA's Careers Map are mostly comparable.

- Clerical and data entry ($32,000 p.a.)
    - information management apprentice
    - clinical audit assistant
- Data analysis ($30-45,000 p.a.)
    - data analyst
    - coder
    - clinical audit officer
- Data systems planning ($60-90,000 p.a.)
    - informatics co-ordinator
    - clinical audit manager
    - data quality manager
- Senior management ($65-150,000 p.a.)
    - planning and performance manager
    - public health intelligence manager

Data entry type jobs don't have many specific pre-requisites, besides general numeracy and IT literacy. The knowledge health-related terminology and classifications they require can be learnt on the job, as the NHS's apprenticeship scheme makes explicit. At the top end, people in senior management roles will usually be those with an in-depth knowledge gained from a number of years work in HIM as well as post-graduate degrees in health economics and similar qualifications.

Qualifications and experience

General IT skills and experience required vary, but as a rough guide:

Data entry	Basic IT - office applications, spreadsheets, etc.
Data analysis	Data analysis & presentation - e.g. SPSS Database use - SAS, SQL, Access
Systems planning	Database design Programming & system administration

In addition, work in HIM requires varying knowledge or experience of health-specific areas, specifically of medical and procedural terminology and coding according to standard classifications. These may vary by region. For example, in both the US and the UK, diseases are classified according to the International Classification of Diseases (ICD)[67], published by the World Health Organization. However the coding used for medical procedures vary: the AMA's Current Procedural Terminology list (CPT)[68] in the US, but in the UK, the classification of interventions and procedures known as OPCS-4[69]. Jobs in coding are likely to require degree-level education in a biological or related subject. Practitioners in clinical audit will require familiarity with concepts such as clinical effectiveness.

To address this fairly wide spread of skills, it is possible to take degree courses in Health Informatics most commonly at masters level, or various specific qualifications available while working. A list of universities and colleges worldwide offering HIM programmes is available[70].

The main skills of a HIM practitioner - recording, coding, analysing data, designing and implementing databases and information policies, etc. - are obviously extremely transferable. From a cursory look at general job listings, the salaries at different levels quoted above appear to be very broadly in line with the practitioners' 'market' rates in similar jobs elsewhere.

References

[1] Launch of the Research Data Alliance, Neelie Kroes, European Commission, 18 March 2013, http://europa.eu/rapid/press-release_SPEECH-13-236_en.htm

[2] https://rd-alliance.org/about.html

[3] Engineering and Physical Sciences Research Council, Policy Framework on Research Data, Expectations, http://www.epsrc.ac.uk/about/standards/researchdata/Pages/expectations.aspx

[4] UK Higher Education Research Data Management Survey - Raw Responses (Redacted). Martin Hamilton, Sue Manuel. figshare. http://dx.doi.org/10.6084/m9.figshare.817926

[5] Access to Information Laws: Overview and statutory goals, Right2Info, http://right2info.org/access-to-information-laws/access-to-information-laws-overview-and-statutory

[6] Scientific Publications: Free for all?, House of Commons Science and Technology Committee, 7 July 2004, http://www.publications.parliament.uk/pa/cm200304/cmselect/cmsctech/399/399.pdf

[7] Business, Innovation and Skills Committee – Fifth Report on Open Access, 3 September 2013, http://www.publications.parliament.uk/pa/cm201314/cmselect/cmbis/99/9902.htm

[8] https://peerj.com/pricing/

[9] Executive Order -- Making Open and Machine Readable the New Default for Government Information, http://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-

[10] G8 Science Ministers Statement, 13 June 2013, https://www.gov.uk/government/news/g8-science-ministers-statement

[11] Panton Principles for Open Data in Science, http://pantonprinciples.org/

[12] Reference Model for an Open Archival Information System (OAIS), Consultative Committee for Space Data Systems, June 2012, http://public.ccsds.org/publications/archive/650x0m2.pdf

[13] Scientific Electronic Library Online, http://www.scielo.br/

[14] Thailand Joins Indonesia in Withholding Bird Flu Data, Bloomberg News, 22 March 2007, http://www.bloomberg.com/apps/news?pid=newsarchive&sid=aSm9ZP9Guu9U

[15] http://www.map.ox.ac.uk/

[16] Osman, S. & IJsselmuiden, C. (2011). Sustainable data sharing in public health research: An INDEPTH-COHRED position paper. Retrieved October 9, 2012 from http://www.indepth-network.org/index.php?option=com_content&task=view&id=1262&Itemid=595

[17] http://www.ddialliance.org/

[18] Second Open Economics International Workshop Recap, Open Economics at the Open Knowledge Foundation, 5 July 2013, http://openeconomics.net/2013/07/05/second-open-economics-international-workshop-recap/

[19] http://www.jisc.ac.uk/whatwedo/programmes/mrd/clip.aspx

[20] http://www.indepth-ishare.org/indepthstats/

[21] Open Data and the Academy: An Evaluation of CKAN for Research Data Management, Joss Winn, University of Lincoln, http://eprints.lincoln.ac.uk/9778/1/CKANEvaluation.pdf
[22] http://www.edawax.de/2013/09/adapting-ckan-for-open-research-data/
[23] http://www.eudat.eu/
[24] Dr Hendrik Bunke, ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Innovative Informations- und Publikationstechnologien (IIPT)
[25] CERIF for Datasets, Scott Brander, Anna Clements, et al. 2011, http://cit.unis4ne-system.net/downloads/files/TPDL_C4D_FINAL.pdf
[26] http://en.wikipedia.org/wiki/Resource_Description_Framework
[27] Simple Data Catalogue Interoperability Proposal, Open Knowledge Foundation, https://docs.google.com/a/okfn.org/document/d/1JFgk-U7so4V0aecut53JS2GPtI2tgFMShUvLoINyKQY/edit#heading=h.c7k2dl806f2f
[28] http://project-open-data.github.io/schema/
[29] http://figshare.com/
[30] Showing you this map of aggregated bullfrog occurrences would be illegal, Peter Desmet, 17 October 2013, http://peterdesmet.com/posts/illegal-bullfrogs.html
[31] http://www.gbif.org/
[32] Sharing research data to improve public health, Mark Walport, Paul Brest, The Lancet - 12 February 2011 (Vol. 377, Issue 9765, Pages 537-539), DOI: 10.1016/S0140-6736(10)62234-9
[33] Signatories to the Joint Statement, http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-and-epidemiology/Signatories-to-the-joint-statement/index.htm
[34] Dr Jimmy Whitworth, Head of International Activities, Wellcome Trust - Interviewed 31 July 2013
[35] Jeffrey P. Koplan, Vice President for Global Health, Emory Global Health Institute – Interviewed 17 July 2013
[36] Dr Garry Aslanyan, Manager, Partnerships and Governance for Special Programme for Research and Training in Tropical Diseases, World Health Organization – Interviewed 23 July 2013
[37] Cancer record is useless, Times Online, 28 August 2013 http://www.timeslive.co.za/thetimes/2013/08/28/cancer-record-is-useless
[38] Dr Cesar De Oliveira, University College of London, Institute of Epidemiology and Healthcare – Interviewed 13 August 2013
[39] Salla Atkins, Scientific Co-ordinator, Department of Public Health Sciences, Division of Global Health (IHCAR), Karolinska Institutet, Stockholm, Sweden – Interviewed 9 August 2013
[40] Tom Barker, Senior Nutrition Convenor, Institute of Development Studies – Interviewed 5 July 2013
[41] Kobus Herbst, Deputy-Director at Africa Centre for Health and Population Studies, INDEPTH Network – Interviewed 18 July 2013
[42] Cesar Victora, Universidade Federal de Pelotas, Brazil – Interviewed 13 August 2013

[43] Tathagata Bhattacharjee, Senior Data Manager - iSHARE2, INDEPTH Network, Asia Node – Interviewed 25 July 2013
[44] Dr Tran Huu Bich, Vice Dean, Hanoi School of Public Health, Vietnam – Interviewed August 19 2013
[45] Enabling the citation of datasets generated through public health research, Jonathan Rans, Monica Duke and Alex Ball, DCC, http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtp051762.PDF
[46] David Carr, Policy Advisor, Wellcome Trust - Correspondence, 29 October 2013
[47] Anthony D Harries, Senior Advisor, International Union Against Tuberculosis and Lung Disease (The Union) – Interviewed 27 August 2013
[48] Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute, May 2011 – http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation
[49] Lynn Woolfrey, Manager, DataFirst, University of Cape Town – Interviewed 15 August 2013
[50] http://www.wits.ac.za/10568/
[51] National Research Council. Professionalizing the Nation's Cybersecurity Workforce?: Criteria for Decision-Making. Washington, DC: The National Academies Press, 2013. http://www.nap.edu/catalog.php?record_id=18446
[52] http://creativecommons.org/licenses/by-sa/3.0/deed.en_GB
[53] http://en.wikipedia.org/wiki/Resource_Description_Framework
[54] http://www.jisc.ac.uk/whatwedo/programmes/mrd/rdmtrain/datum.aspx
[55]http://www.northumbria.ac.uk/sd/academic/ee/work/research/clis/information_records_management/rmarea/datum/health/materials/?view=Standard
[56]http://www.northumbria.ac.uk/sd/academic/ee/work/research/clis/information_records_management/rmarea/datum/health/
[57] http://www.dcc.ac.uk/training/train-trainer/disciplinary-rdm-training/disciplinary-rdm-training
[58] http://www.wits.ac.za/10568/
[59] http://africabuild.eu/
[60] http://africabuild.eu/symposium-kenya/
[61] https://www.hicf.org.uk/
[62] http://www.ahima.org/
[63] http://hicareers.com/CareerMap/
[64] http://www.nhscareers.nhs.uk/explore-by-career/health-informatics/careers-in-health-informatics/information-management-staff/
[65] http://www.nhscareers.nhs.uk/explore-by-career/health-informatics/entry-requirements/
[66] http://www.healthjobsuk.com/
[67] http://www.who.int/classifications/icd/en/

[68] http://www.ama-assn.org/ama/pub/physician-resources/solutions-managing-your-practice/coding-billing-insurance/cpt.page
[69] http://systems.hscic.gov.uk/data/clinicalcoding/codingstandards/opcs4
[70] http://www.healthinformaticsforum.com/health-informatics-degrees-and-certificates