





Workshop Report

Global Health Research & Development: Mapping funding flows – working towards a common approach

October 20, 2014, Charité CrossOver (CCO), Berlin, Germany

Workshop Conclusions

- Consensus among attendees of the need, importance and benefits of funding agencies improving the quality, consistency and public access to their funding data
- Current challenges to access and provision of quality data can be diverse and plentiful, however
 it would be relatively tractable for agencies to provide a basic, regular view of their funding data
 within a specified format
- Significant interest by all present to move beyond manual classification of research grants to more automated text mining solutions where the accuracy can be assured. This involves further exploring how automated harvesting of data and text mining can help agencies more easily participate in R & D mapping exercises (e.g. G-Finder)
- The funders involved were keen to further explore and identify:
 - The minimum data set, lowest common denominator or set of 'minimum necessary elements' that would describe a current research project adequately to enable useful analysis (not necessarily for adoption of a global standard but that could be available for ex-ante standardization)
 - Standards for publishing that data on the internet that facilitates free access and stimulates innovation in the analysis of this data
 - The appetite among other funders including those who participate in the Heads of International Research Organisation (HIROs) grouping of major public and charitable research funders to adopt a common approach
 - o The scope to extend the discussions to include commercial sector R&D funders
- General support that an R&D observatory hosted by WHO might be one platform / home for such a portal

Workshop Background

Since the first World Health Organization (WHO) meeting on this topic which took place at the Wellcome Trust in 2013¹, there has been progress in the area of mapping and automation of research classification. Funders, globally, are increasingly improving their transparency and reporting of R&D funding. New tools are available to help collate and analyse data, reducing our dependency

¹ http://www.who.<u>int/phi/workshop14022013/en/</u>

on costly and time consuming surveys. In time, a comprehensive picture of health R&D could be generated from these available electronic sources. This would be one major contribution towards the development of a global health R&D observatory as called for recently by the Member States of the WHO as part of the WHA 66.22 resolution of May 2013². The R&D observatory is expected to demonstrate proof of principle by May 2016.

Workshop Aims

- Discuss the role health R&D funders can play in facilitating greater transparency in the reporting of the research they are funding
- Better understand the challenges that are faced by both the generators of R&D data (funders) and third-parties who also use the data (harvesters) in the generation of analysis
- Understand the benefits to all stakeholders in creating this map from the perspective of priority setting, strategy development, benchmarking, monitoring and evaluation
- Identify the minimum data elements that would be needed to create a useful map of health R&D, including funding flows
- Consent on practical next steps to move this agenda forward

Workshop Structure and Outputs

The open-workshop was hosted by The Charité medical school as a 2014 World Health Summit (WHS) satellite session in Berlin, Germany on October 20 2014. The full day workshop was an interactive and informal session that saw the attendance and active participation of a global-mix of representatives from both public and privately financed funders, a diverse selection of the current users or (harvesters) of funding data as well as academics and representatives from the general public and civil society.

The workshop was split into three main sessions: the first was a technical session in which 4 data users presented their tools and experiences of obtaining and using funding flow data. The second session sought to obtain the funder perspective on what their requirements and constraints would be in terms of sharing of grant information. The final session pulled together the day by establishing areas of commonality for where both data providers (funders) and harvesters (users) could realise mutual benefits.

A one-page summary of the main challenges faced, and benefits forseen by, both the providers of data (funders) and current harvesters/analysts of data follow on pages 3 and 4. The appendices provide a more detailed log of the discussions had: Appendix 1 on the discussion around a 'minimum data set'; Appendix 2 on the 'challenges faced' by both sides and Appendix 3 on list of participants.

Next steps

Action item Timings -Workshop report to be reviewed and revised by workshop sponsors early Nov Finalised workshop report, of key themes and action items, list of Nov 2014 attendees and slides circulated to participants Nov-Dec Incorporate issue onto agenda for forthcoming HIRO's meeting to seek 2014 'in principle agreement' to a common and consistent approach to sharing data; as a starting point for wider discussion with other R&D funders. Small Working Group to be established / convened for the further Jan-Feb 2015 development of a set of 'minimum elements' (see provisional discussion in Appendix 1) and to contribute to move agenda forward

² http://www.who.int/phi/documents/CEWG-WP/en/

Benefits to....

Data generators (funders)

<u>NB:</u> While there was largely a consensus of participating funders of the benefits – it was acknowledged that this might not be the case for all funders. Will be important to get funders to 'feel the benefits' for them and not just about 'seeing the benefits for others'.

Strategic and practical benefits:

- Know who [else] is funding in a particular area
- Identification of possible collaborators to improve linkages / networks
- Increase efficiencies of funding processes
- Promote evidence-driven funding decisions: improved priority-setting
- Matching funding to health burden/needs
- Facilitating the identification of research gaps
- Identification of technologies that could assist in realization of transparency goals
- Broader / collateral benefits i.e. identification of peer-reviewers, niche expertise
- To enhance evidence-informed global coordination of future funding commitments to priority research areas
- To assist in creating a 'needs case' of funding in neglected areas to member states
- Need for International comparison data for global benchmarking and positioning
- Identification of domestic strengths
- Where are domestic weaknesses where a partnership/supporting role would make more sense?
- Information on networks: which are most efficient, which models work, best-practice identification and sharing
- Assists with accountability: the 'so what' of research funding
- To assist in monitoring/evaluation of spend
- Interested in outcomes / impact / products of their money

Data harvesters/analysts

Recently surveyed by G-Finder (commissioned by WHO)³ to identify what users/stakeholder needs are:

- Stakeholders want to have a comprehensive centralized reliable source with standardised definitions, from a trusted organisation
- They want to see at a glance what they should do (analysis)
- Preference among high level policy makers of low-granularity 'big picture' information while researchers, product developers were more interested in the ability to drill-down into more detailed data
- Simple, fast, easy to use search portal
- The data harvesters see huge benefits in an open-access resource as it would 'free them' enabling them to focus on [where they see] their real added-value - the analysis and provision of data. Their current role in 'data handling' is merely a means to an end. They do it as it is not being done elsewhere.
- Providing raw data (open access) will lead to innovation in research analysis; the data will be used!
- For small organisations (charities) or countries
 with low-level funding (SSA) or limited history of
 data sharing (China): provision of a basic set of
 minimum standards would assist International
 organisations in meeting their needs (through
 technical support). The advantages to them that
 they do not have to grapple with the 'legacy
 effect'; can start from scratch, building on the
 learnings of others.
- Necessary to empower and inform civil society, advocacy and civil society organisations put data into a context to approach stakeholders and decision makers to lobby for increased resource etc

³ http://www.who.int/phi/implementation/stakeholder_ReportG-FINDER_primer.pdf?ua=1_

Challenges for....

Data generators (funders)

Obstacles to providing useful data

- Providing open, public access to raw data can be constrained by a number of factors:
 - Political: fear of data end-use, galvanization of research community and benchmarking
 - Socio-cultural: variability in 'data sharing culture' across countries, sectors and funders
 - Economic: maintenance of up-to-date, quality, grant data in public domain very time and resource intensive
 - Legal: data protection limitations can apply throughout funding process i.e. from grant application portals to use of outputs (e.g. ResearchFish reluctance to share)
 - Technical: barriers to public data release in specified formats can be higher for those with long-established, incrementally evolved, internal datacapture systems or legacy databases. Most funders have IT systems originally designed for processing grants and do not readily support analytic capabilities.
- Pioneering countries and organisations of data sharing culture can serve as an example/tool for internal advocacy within reluctant or slow adopters
- Workshop discussions acknowledged to be within a broader movement, expectation and legal framework towards increased transparency
- 4. Despite all funders stating that they have met their legal obligations with regards data disclosure:
 - Publication does not necessarily guarantee the data were accessible, usable or useful
 - Many ignorant of whether/how these data are accessed by users
- Huge inconsistencies were acknowledged across and within databases with regards to how data structures, formats, nomenclatures, definitions, classification approaches are used and why
- 6. Acknowledged that funders are a

Data harvesters/analysts

Obstacles in obtaining and using data

- Continue to face many data access challenges
- Most data not stored/available in a structured, standardized, easy-to-collect way and from a multitude of funders. Data coding⁴ is the major challenge, however:
 - Global consolidation of main funders provides an opportunity to build classification consensus (12 org's = 75% of NTD funding)
 - Some guides to coding / data dictionaries are available from participants and can be built upon
- 3. Ideally, funders should present data in formats able to be easily aggregated with other funders' data. It is a major challenge for users/data analytics organizations to attempt to aggregate data in variable formats, states of completeness and quality. The primary interest of data analytics is in adding value through analysis/expertise (not recoding and checking quality).
- 4. Legal tools are available to facilitate data access i.e. '[especially commercially] sensitive data', however even when legal framework in place: does not always ensure disclosure i.e. private sector and clinical trials data
- 5. Plea for a 'move to the middle' where coding burden is removed from the users and shared with the providers:
 - Key fields or minimum elements ideally would <u>not</u> be user defined
 - Desire to work towards standardization to reduce transaction costs
- 6. Challenges of different data collection approaches:
 - Self-reporting/surveying, in theory, shifts coding (and resource) burden to providers
 - Reality: manual coding frequently undertaken ex post
 - Automated approaches involve the data harvesters assuming coding (resource) burden
 - Reality: have tried to encourage data providers to ex-ante adopt greater standardization (with limited success)
- 7. Automation felt to be only a part-solution:
 - Would ease data collection, enable ex-ante coding and enable provision of a macro overview BUT
 - The lowest level of agreement between reporting funders (regarding coding) was considered as likely being insufficient to enable fruitful/in-depth analysis of the data

⁴ Data coding is the term used – throughout this document – to refer to the processes of data cleaning, consistent classification, harmonization, disambiguation etc all the steps required to make data 'usable'

Data generators (funders)	Data harvesters/analysts	
heterogeneous group with different histories, mandates, motivations and data needs	8. Face legal challenges either in the republishing of data i.e. non commercial reuse clauses (Japan) or in collection because harvester is a commercial entity i.e DFG of Germany	
	9. Low capacity in low and middle income countries to collect and report on this data	

Appendix 1: Provisional discussion on minimum data set

Objective of agreeing a minimum data set is not to seek approval and adoption of a 'global standard', but to agree to put data out there that can subsequently be standardized.

Criteria for minimum data set:

- Bringing to lowest common denominator. What is the minimum that is sufficient?
- Every additional data field will increase resistance to provision must be kept as simple but meaningful as possible
- Need for providers and users to meet in the middle
- Standardization contentious, necessary to bring down transaction costs but data anyway will be retrospective
- If you want to build up a platform with little effort you have to set high standards of your datainput (high manual effort) but there might be less compliance and you have to bring down your expectations

Provisional list of 'must have fields/elements'

- Individual / Person/ Name holder of grant: ... only resistance to this was a possible legal restriction on the European Commission (EC) which considers this confidential information
- Location of beneficiary / Institution: institution and country department considered too granular
- Time scale (start/end date/extension): for further discussion grant period or grant disbursement? Commitments or realized commitments (disbursements)
- Funds: grant value and duration (yearly or multi-year)
- Grantee: name of recipient organization
- Title of grant/research proposal
- Abstract / overview of research proposal with sufficient depth to enable textual mining of content.
- Unique grant identifier/code some challenges were raised with this (i.e. re-use of funder references). Discussion on using existing codifying systems such as those logically derivable or global norms such as the digital object identifier DOI (paper or dataset). To be further discussed if this is a 'must' or a 'nice' to have.

Provisional list of 'nice to have fields/elements'

- Classification standard: HRCS, MESH (not made for reporting, terms not yet in MESH), title (sentence), abstract (200 words), key words.
- Methodology of research
- Co-applicants and collaborators of grant

<u>Appendix 2:</u> Full, detailed, notes from workshop documenting 'challenges faced' by data generators and users present

	_
Data generators (funders)	Data harvesters / analysts
Obstacles to providing useful data	Obstacles in obtaining and using data
Political:	Political:
Outputs / impacts tend to be more politically	Funders can be hesitant to provide raw data
complicated, hence meeting and starting point for	which will later be published publicly
discussion to focus on inputs	Is it an unwillingness or inability to provide
 Providing open, public, access to raw data can be 	data?
constrained or complicated by considerations of:	Discussions focused on top-down
The inputs themselves:	approaches (data provider = the funding
 Consideration for 'not wanting to upset the 	agencies), bottom-up approaches (data
research community'	provider = the beneficiary/researcher)
 Nervous about revealing deficiencies = risks of 	might mitigate political resistance and
galvanization	surmount the commitment vs.
Awareness that transparency comes with risks	disbursement issue example provided from
Outputs (how the data will be used):	Brazilian initiative LATICE platform / US NIH
 Fear of 'misrepresentation' 	• Don't want to be data coders ⁵ – consider
Uncertainty and suspicion of end-use	the responsibility for these databases a
 Fear of benchmarking/comparison i.e. between legal jurisdictions (Canadian 	public role. Interest in adding value through
provinces) and within an organization but	analysis/expertise
acknowledged that benchmarking can be	Ideal would be open-access to raw/primary data
powerful for leverage	data
Political constraints to providing access was	Face data access challenges: heightened when data request is from a required.
acknowledged to differ between countries and sectors	when data request is from – or when data provider – is a commercial company
(see 'cultural') but not necessarily agreement on for	(safeguards can alleviate these challenges
whom barriers highest	i.e. disclosure under confidentiality
Global movements/initiatives/visibility of issue help to	agreement and aggregate reporting)
put pressure on ministries to adopt greater openness	 Desire to incorporate private sector data, in
Willingness to share can be higher when internal	time, however sensitive and acknowledge it
systems are being updated – many present indicated	will be slow and incremental process
that that time is now (window of opportunity)	will be slow and incremental process
Pioneers of transparency (organizations or countries)	Economic:
can help 'bring along the rest'	Plea for a 'move to the middle' where
Public disclosure can be a driver for improved data	alignment of coding burden is removed
quality	from the users and shared also with the
Publishing only minimum level of detail helps orgs with	providers i.e. key fields or minimum
less detailed information save face	elements ideally would not be user defined
Acknowledged a growing public and government	Self-reporting/surveying, in theory, shifts
movement/expectation of transparency:	coding (therefore resource) burden to data
 For charity/private foundation sector not 	providers
necessarily a history of this – starting from	 Reality: manual coding frequently
scratch	undertaken <i>ex post</i>
 For public sector, accountability is driver and 	 Survey fatigue –and diminishing
political focus and pressure (on many	responsivity - a downside of surveying
agencies) remains on demonstrating impact	approach as funders faced with
	increasing data requests
Economic:	Automated approaches involve the data
Hugely time and resource intensive to maintain up-to-	harvesters assuming the burden of data
date, quality, grant data in public domain	coding

⁵ Data coding is the term used – throughout this document – to refer to the processes of data cleaning, classification, harmonization, disambiguation etc all the steps required to make data 'usable'

Data generators (funders)

- Every additional data field / requirement will increase resistance to provision
- Legacy databases: for many organizations the internal systems were built to get funds out of the door – they are historically determined and evolve organically, slowly and incrementally
- Often internal systems /coding are constrained by the structure and legal coverage of the 'first entry point' the grant application databases (which were considered 'unchangable').
- Internal systems were not originally built for 'fund and follow' or analytical capabilities
- Acknowledged a need and desire to 'move to the middle' where funders facilitate users use of data.
 However provision of 'basic minimum elements' was seen as more feasible than 'change to internal systems'

Social/Cultural:

- Different cultures of research have produced different terminology i.e. development vs funding institutions, basic researchers vs clinical
- Complexity of the funding / researcher habitat is not linear (as often represented) by researcher – grant – output schematics
 - Funder making a funding decision uses data from different sources (content consolidation)
 - Interrelationship of people, products and institutions → valuable to be able to represent true complexity
 - No longer a situation of one researcher increasingly trans-national networks as beneficiary

Legal:

Many of the public bodies present noted legal constraints to data handing for example:

- Grant data (and in the case of Canada, all evaluations) must be in the public domain
- Many countries have 'freedom of information (FOIA) acts' making the funders – hypothetically – subject to FOI requests
- Data protection limitations can apply throughout process i.e.:
 - Third-party data publishing portals i.e.
 ResearchFish has its own terms of use
 - Through the initial point of data input (frequently the grant application process) whose terms of use are designed to protect the applicant/researcher
- Despite all funders stating that they have met their legal obligations with regards data disclosure:
 - Publication does not necessarily guarantee the data were accessible, usable or useful
 - Many ignorant of whether/how these data are accessed by users

Technical:

Data availability

Consensual desire expressed to avoid duplication and

Data harvesters / analysts

- Reality: have tried to encourage data providers to ex-ante adopt greater standardization (with limited success)
- Automation was felt to be only a partsolution:
 - Would ease data collection, enable ex-ante coding and strength remains in providing macro snapshot overview BUT
 - The lowest level of agreement between reporting funders (regarding coding) was considered as likely being insufficient to enable fruitful/in-depth analysis of the data

Social/cultural:

- Some countries have a greater willingness / openness to share than others – depending on their adoption of the 'sharing economy'
 - Canada reported very positive response to 'open access policy'
 - Users agreed on the challenges faced with obtaining Chinese data
- Willingness to share also dependent on sector (private, public, philanthropic, charity)
 - Smaller bodies or those with less history of transparency – can be slower adopters, more cautious and see fewer benefits. However also potentially easier as unconstrained by 'legacy database' effect
- Willingness to share dependent on institution and individual researcher level aswell ('personal responsibility' for issue)
- Limitations/constraints of self-reporting was raised as were the potential benefits of 'triangulation' approaches for data verification

Legal:

- Challenges occur in the republishing of data i.e. non commercial reuse causes (Japan)
- Some countries do not make data available to commercial entities i.e. DFG of Germany, Japan
- It was noted that the private sector discloses less than public bodies regarding clinical trials (despite a legal mandate to do so), so laws do not necessarily assure disclosure
- Legal protection in the form of confidentiality agreements can be a way of obtaining access to 'sensitive data' by providing reassurance to providers on how the data will be treated and published (private sector especially).

Data generators (funders)

- recreating the wheel: Are there different ways of using instruments we already have?
- Formats html (web)-based, SQL or excel were considered the most usable (.xls as a min.). PDF's and others were considered unusable
- Language currently English-centric, automatic translation would mean quality reduction

Data structure

Classification:

- Huge inconsistencies across and within databases
- Many different coding nomenclatures were mentioned by participants as the basis of their classifications: ODA (OECD), MeSH, natural language processing (text form), ORCHID, UMLS, HRCS however non of the vendors present used these
- Discussion around the feasibility of adopting standards and norms across all funders (unified classifications/definitions):
 - Perhaps too late and / or unworkable
 - Would have to be applied retrospectively to existing data
- Alternative suggestion:
 - Possible to translate existing standards into a new database?
 - Determine minimum meaningful data elements (see Appendix 1 for provisional discussion of this)

Definitions:

- Inconsistencies arise throughout spectrum of research terms, i.e. diseases, research types, financial years, definition of grant or project (repeated over years)
- Inconsistencies and how to avoid or disambiguate (often their own institutes will have >20 spelling variations)
- Not even harmonized Nationally, how to standardize (or at least understand differences)?
- Terminology can be culturally-determined (nationally or by different disciplines) and is a moving target (evolves)

Outputs:

Considered beyond the scope of workshop, however:

- Linking funding data to output data (such as product development, publications) is challenging and compounded with time-lag issue
- Funding flows do not follow same time as outputs
- Challenges of defining then measuring output/impact.
 Differences between outcome, output, results and impact.
- Inconsistent reporting on research progress
- Network analysis seen as having limitations (time-lag from grant award and H-index limitations) but also valuable for input/output-analysis and identifying where there are still gaps and key 'bridging or linking researchers' in a given field/topic i.e. Ebola

Data harvesters / analysts

Technical:

Data availability

- Very limited health R&D data especially in LMIC
- Huge variability in availability in western countries:
 - Between countries and different sectors
 - In how/what data are stored
 - o In quality/details of data available

Data structure

Comparability:

- Internal storage systems/processes: most data not stored/available in a structured, standardized, easy-to-collect way
- Inconsistent classification and definitions across databases and challenge of inconsistent entry (need for disambiguation) within databases
- However global dominance / consolidation of main funders provides an opportunity to build classification consensus (12 org's = 75% of NTD funding)
- Desire to work towards standardization to reduce transaction costs
- Most private sector funding is not in the form of 'grants'
- Some 'guides to coding' / data dictionaries are available from participants:
 - G-Finder
 - o Uber research provided already
 - UK MRC said they could provide from GateWay research portal

Appendix 3: list of participants

	First Name	Last Name	Institution	Email
Sponsors &	Taghreed	Adam	WHO	adamt@who.int
participants	Liz	Allen	Wellcome Trust	l.allen@wellcome.ac.uk
	Robert Fraser	Terry	WHO / TDR	terryr@who.int
	Peter	Tinnemann	Charité	peter.tinnemann@charite.de
	David	Carr	Wellcome Trust	d.carr@wellcome.ac.uk
	Suzanne	Edwards	TU-Berlin	suzanne.edwards@tu-berlin.de
	Gary	Aslanyan	WHO / TDR	aslanyang@who.int
	Alain	Beaudet	CIHR, Canada	Alain.Beaudet@cihr-irsc.gc.ca
	Detlef	Boecking	PT_DLR / BMBF	detlef.boecking@dlr.de
	Max	Bender	Charité	max.bender@gmail.com
	Marlon	Cerf	SAMRC	Marlon.Cerf@mrc.ac.za
	Nick	Chapman	PolicyCures/G-Finder	nchapman@policycures.org
	Christian	Herzog	ÜberResearch	christian@uberresearch.com
	Enno	Hoffmann-Dose	Lundbeck Foundation	ehd@lundbeckfonden.com
	Paul	Houston	Clinical Data Interchange Standards Consortium	phouston@cdisc.org
	Ghassan	Karam	WHO	karamg@who.int
	Stefan	Kohler	Charité	Stefan.Kohler@charite.de
	Judith	de Kroon	NOW - WOTRO	j.dekroon@nwo.nl
	Jeff	Mphahlele	SAMRC	Jeffrey.mphahlele@mvc.ac.za
	Jonathan	Nagle	CIHR, Canada	Jonathan.Nagle@cihr-irsc.gc.ca
	David	Peckham	CIHR, Canada	David.Peckham@cihr.gc.ca
	José	Salm	Stela Institute, Brazil	salm@stela.org.br
	Beverley	Sherbon	MRC UK	Beverley.Sherbon@headoffice.mrc.ac.uk
	Christian	Wagner	Health Action International	cw@bukopharma.de
Students/	Peter	Philipsborn	Student for Protocol	peter@philipsborn.eu
rapporteurs	Katharina	Last	Student for Protocol	katharina_last@posteo.de
	Anna-Lisa	Behnke	Student for Protocol	anna-lisa.behnke@charite.de