Data Curation at NISR

Concept note: How it should be done

Introduction:

The word curation comes from the Latin word *curare*, meaning - to take care. Stretching the traditional meaning and juxtaposing it with data life-cycle, 'data curation' as a process means, at least four things. It means to *identify* data (required to be curated). It means to *preserve* data. It means to provide *context* to data. And it means to make data easily *discoverable and accessible*. But it's more than that. Before the arrival of the internet, data curation was mostly confined to the limits of physical location and analog medium. But now, with the internet, both the medium and distribution has changed. Hence, data curation is also, to ensure '*user interactions*' - with the online outlet channels - an extraordinary experience.

Graduate School of Library and Information Science, University of Illinois, defines data curation as the active and ongoing management of data through its lifecycle of interest and usefulness to scholarship, science, and education. Data curation enables data discovery and retrieval, maintains data quality, adds value, and provides for re-use over time through activities including authentication, archiving, management, preservation, and representation[1]. In general, data curation is a term used to indicate management activities required to maintain research data long-term such that it is available for reuse and preservation[2].

Generic Statistical Business Process Model:

From the perspective of NISR, data curation as a process, applies to many 'types' of data flowing through set paths within the organization finally reaching up to the end-users. The Generic Statistical Business Process Model (GSBPM)[3] (Fig. 1.) provides a practical framework also to clearly identify and model the process (of data curation) in NISR.

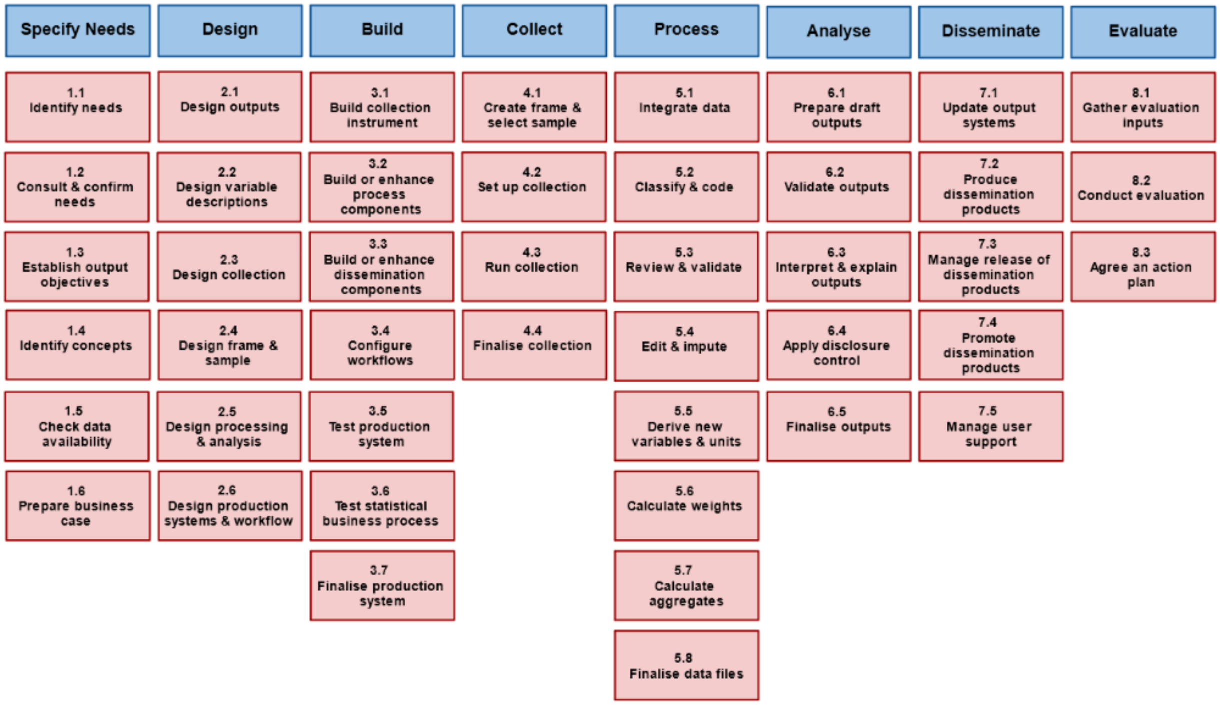
Offering a common language, the sub-processes under each phase (Specify needs, Design, Build, Collect, Process, Analyze, Disseminate and Evaluate) offers the hooks, against which, data curation could take shape for all data types as they progress in the work flow in a given setting.

[1] http://www.lis.illinois.edu/academics/degrees/specializations/data_curation

[2] http://en.wikipedia.org/wiki/Data_curation

[3] http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model

Fig. 1: Generic Statistical Business Process Model



Specify Needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
1.1 Identify needs	2.1 Design outputs	3.1 Build collection instrument	4.1 Create frame & select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
1.2 Consult & confirm needs	2.2 Design variable descriptions	3.2 Build or enhance process components	4.2 Set up collection	5.2 Classify & code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Conduct evaluation
1.3 Establish output objectives	2.3 Design collection	3.3 Build or enhance dissemination components	4.3 Run collection	5.3 Review & validate	6.3 Interpret & explain outputs	7.3 Manage release of dissemination products	8.3 Agree an action plan
1.4 Identify concepts	2.4 Design frame & sample	3.4 Configure workflows	4.4 Finalise collection	5.4 Edit & impute	6.4 Apply disclosure control	7.4 Promote dissemination products	
1.5 Check data availability	2.5 Design processing & analysis	3.5 Test production system		5.5 Derive new variables & units	6.5 Finalise outputs	7.5 Manage user support	
1.6 Prepare business case	2.6 Design production systems & workflow	3.6 Test statistical business process		5.6 Calculate weights			
		3.7 Finalise production system		5.7 Calculate aggregates			
				5.8 Finalise data files			

Considerations:

End-to-end

Specially because of metadata (data about: data or statistics being produced or even the process under taken to obtain the data/statistics), data curation as a process, covers the entire stack of GSBPM, starting from 'Specify Needs' to 'Evaluate' stage. From the time, an act of data capture is ideated (driven by perceived needs), generation of metadata begins and hence 'curation' too should begin along with it.

This however brings up the issue of, 'after the act' vs. 'continuous' curation data/information.

Continuous

Traditionally, the practice has been to document surveys only after the survey is completed and results (in terms of publications and aggregated data) are out. However, data curation must be a continuous process not 'after the act'. As the statistical workflow progresses, data curation as a process, should also progress. This would necessitate that a proper version control is maintained, such that distinguishing between a series of 'drafts' which lead to a final version, which in turn may be subject to further amendments, becomes clear and easy.

Version control

Version control is the management of multiple revisions to the same items, enabling to tell one version from another. It involves a process of naming and distinguishing between a series of 'drafts' providing an audit trail for the revision and update of draft and final versions.

All types of data

Traditionally in NISR, the focus of data curation has been on the survey microdata. But going beyond just the microdata, data curation as a process, must include all types of data produced at the NSO. The data types may include *metadata, microdata, aggregated data, geo-spatial data (map data) and publications (PDF and print)*. The process of curation must apply to all, for their easy discovery, retrieval and re-use.

Job division/Organizational structure

Though the process of data curation is spread across entire organization crisscrossing various departments and sections, the responsibility to curate must lie with a dedicated team. Depending on the type of data to be curated, however, the team may further sub-divide the responsibilities. The processes applicable are illustrated below (Fig. 2).

Fig. 2: Processes and data types

		Data types				
		Metadata	Microdata	Aggregated data (including compilations)	Geo-spatial data	Publications
Processes	Identification (Decide, what to curate?)	▨				
	Preservation (How and where to archive?)	▨				
	Contextualization (Sequencing, policy, use-cases etc.)	▨				
	Discoverability and accessibility (Means to open up/ dissemination)	▨				
	User experience (Observe, empathize, look for gaps)	▨				

The Statistical Methods, Research and Publication Unit at NISR is a natural place for data curation activities. 'Dissemination' as an activity is part of the data curation process. Data curation must also include the One Stop Center and the Information and Communications Technology Unit in the data curation activities.

Data types and tools

Because of variety of data types (including the fact that much of it is digital in nature), the tools required to support the curation process varies. Usually, the tools for managing various data types do make a consideration such that appropriate metadata retain their links with data. In this context, metadata as a data type is a special case, as it traverses across all other data types.

Microdata

The process illustrated below (Fig. 3) by Ms. Lynn Woolfrey, Manager, DataFirst, University of Cape Town, highlights the steps of data curation in case of data type as microdata.

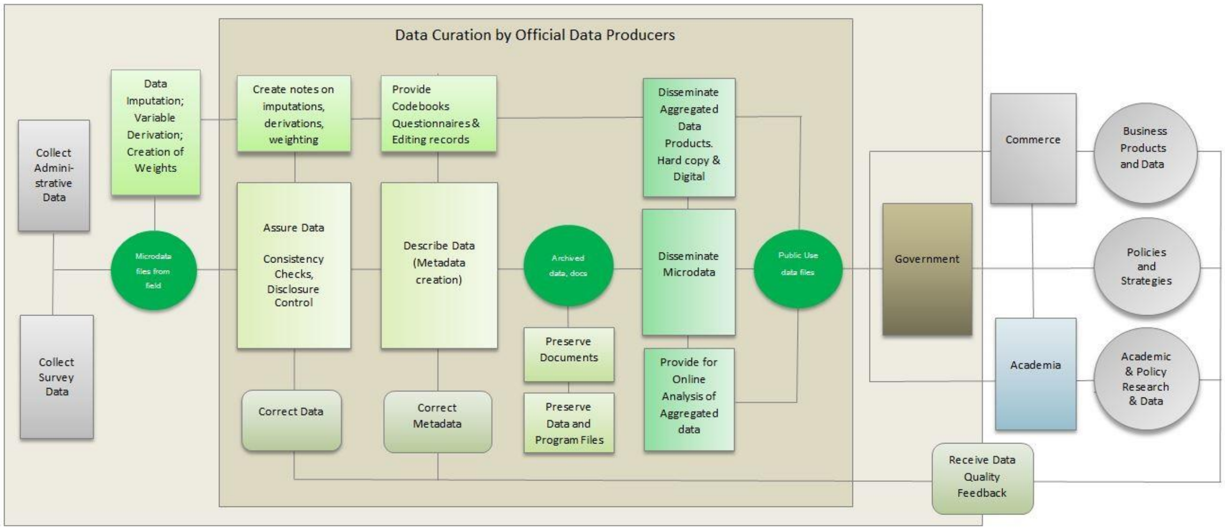

Data Curation by Official Data Producers
Collect Admini-strative Data
Collect Survey Data
Data Imputation; Variable Derivation; Creation of Weights
Microdata files from field
Create notes on imputations, derivations, weighting
Assure Data
Consistency Checks, Disclosure Control
Correct Data
Provide Codebooks Questionnaires & Editing records
Describe Data (Metadata creation)
Correct Metadata
Archived data, docs
Preserve Documents
Preserve Data and Program Files
Disseminate Aggregated Data Products. Hard copy & Digital
Disseminate Microdata
Provide for Online Analysis of Aggregated data
Public Use data files
Government
Commerce
Academia
Business Products and Data
Policies and Strategies
Academic & Policy Research & Data
Receive Data Quality Feedback

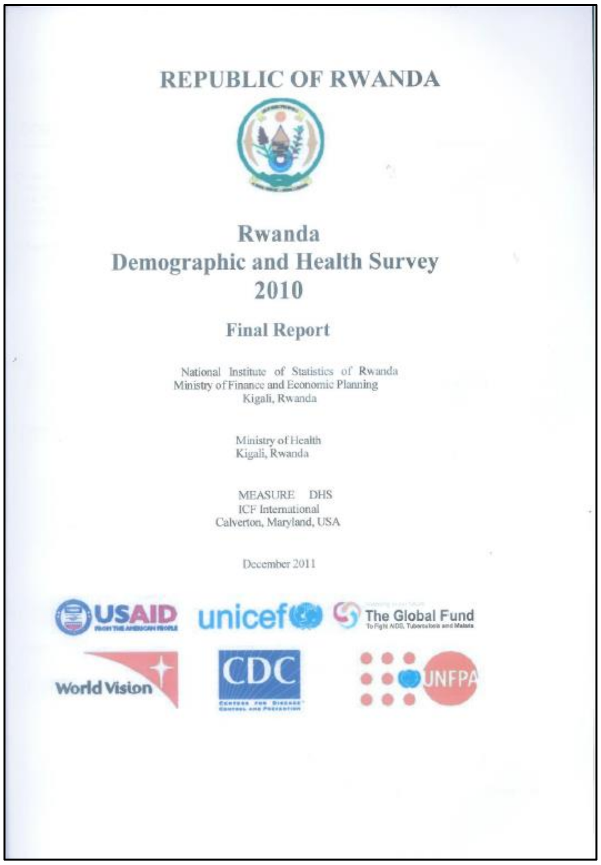
Fig. 3: Woolfrey, L. 2014. Model of Data Curation by Official Data Producers

For microdata, preservation software NESSTAR Publisher (or Microdata management toolkit) could be used. Discovery and dissemination is advised to be managed with National Data Archive (NADA) software.

Aggregated data (including compilations)

The case of DevInfo as a tool of data curation

DevInfo is used at NISR to keep *processed and aggregated* statistical data for dissemination. Data in it therefore, can go in, only after the ***availability*** of aggregated data values (against corresponding indicators). At NISR, *availability* of aggregated data is essentially marked by the *publication* of printed reports containing data in the tables (see the example below of the report front cover from DHS 2010).



REPUBLIC OF RWANDA

Rwanda
Demographic and Health Survey
2010

Final Report

National Institute of Statistics of Rwanda
Ministry of Finance and Economic Planning
Kigali, Rwanda

Ministry of Health
Kigali, Rwanda

MEASURE DHS
ICF International
Calverton, Maryland, USA

December 2011

USAID unicef The Global Fund
World Vision CDC UNFPA

A strategy may be articulated for institutionalizing the process, from the steps currently being followed to keep the database updated.

This is what happens outside DevInfo

1. Identification of the new indicator (depends on the requirements of the 'audience') to be added in the DevInfo database
2. Selection of the source publication
3. Identification of the relevant table in the selected publication
4. **Photocopying and printing of the table**
5. Choosing the required 'disaggregation' from the selected table (as highlighted in the table below from DHS 2010) (or whole table may be selected)

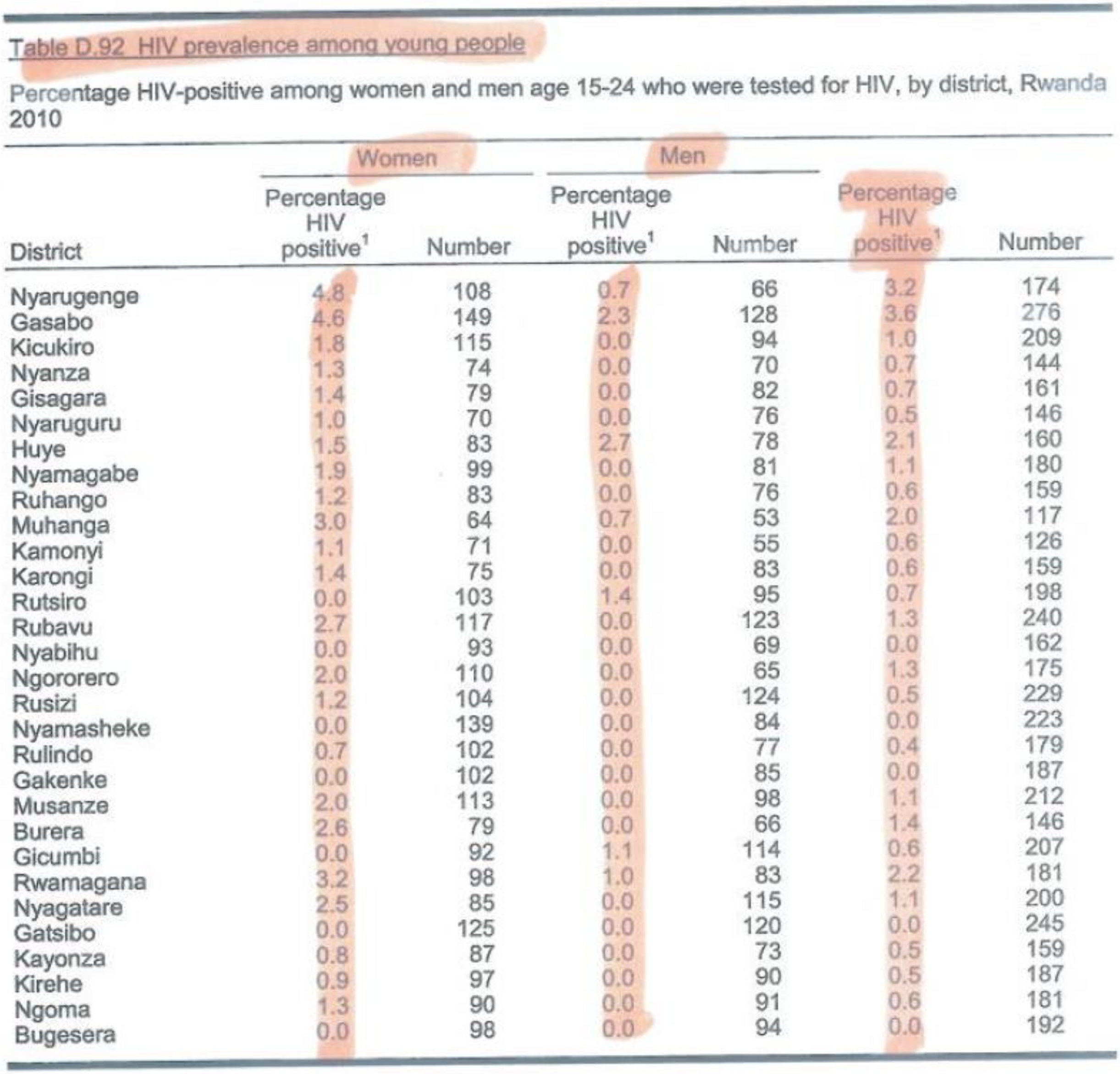


Table D.92 HIV prevalence among young people

Percentage HIV-positive among women and men age 15-24 who were tested for HIV, by district, Rwanda 2010

	Women		Men			
District	Percentage HIV positive[1]	Number	Percentage HIV positive[1]	Number	Percentage HIV positive[1]	Number
Nyarugenge	4.8	108	0.7	66	3.2	174
Gasabo	4.6	149	2.3	128	3.6	276
Kicukiro	1.8	115	0.0	94	1.0	209
Nyanza	1.3	74	0.0	70	0.7	144
Gisagara	1.4	79	0.0	82	0.7	161
Nyaruguru	1.0	70	0.0	76	0.5	146
Huye	1.5	83	2.7	78	2.1	160
Nyamagabe	1.9	99	0.0	81	1.1	180
Ruhango	1.2	83	0.0	76	0.6	159
Muhanga	3.0	64	0.7	53	2.0	117
Kamonyi	1.1	71	0.0	55	0.6	126
Karongi	1.4	75	0.0	83	0.6	159
Rutsiro	0.0	103	1.4	95	0.7	198
Rubavu	2.7	117	0.0	123	1.3	240
Nyabihu	0.0	93	0.0	69	0.0	162
Ngororero	2.0	110	0.0	65	1.3	175
Rusizi	1.2	104	0.0	124	0.5	229
Nyamasheke	0.0	139	0.0	84	0.0	223
Rulindo	0.7	102	0.0	77	0.4	179
Gakenke	0.0	102	0.0	85	0.0	187
Musanze	2.0	113	0.0	98	1.1	212
Burera	2.6	79	0.0	66	1.4	146
Gicumbi	0.0	92	1.1	114	0.6	207
Rwamagana	3.2	98	1.0	83	2.2	181
Nyagatare	2.5	85	0.0	115	1.1	200
Gatsibo	0.0	125	0.0	120	0.0	245
Kayonza	0.8	87	0.0	73	0.5	159
Kirehe	0.9	97	0.0	90	0.5	187
Ngoma	1.3	90	0.0	91	0.6	181
Bugesera	0.0	98	0.0	94	0.0	192

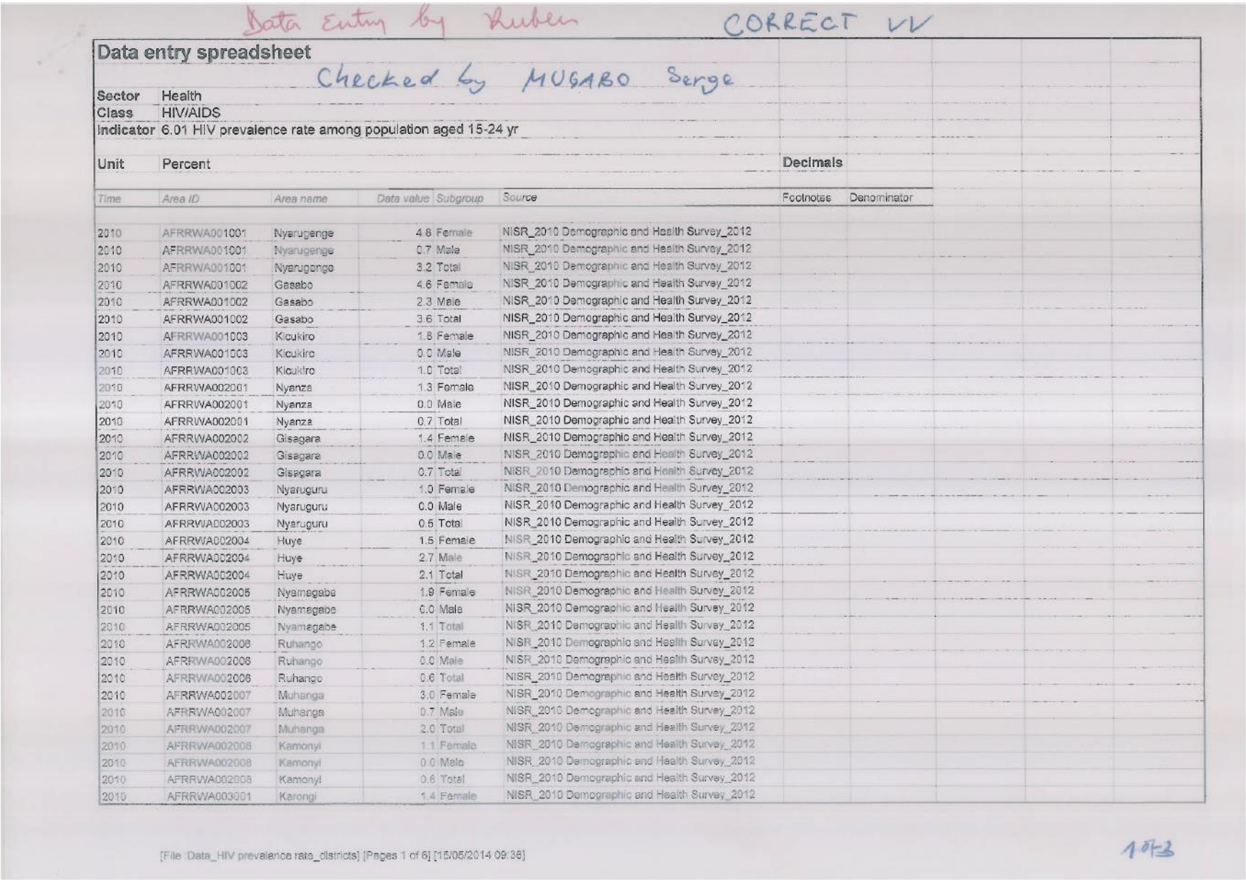
In DevInfo Desktop Admin Application

6. Exporting the blank data-entry template (based of the desired disaggregation) in MS Excel format (This happens after adding the new indicator with its subgroups and associating it with time, area and source in the DevInfo database)

Outside DevInfo

7. Data is entered manually in the exported MS Excel sheet by typing-in
8. Print-out of the filled-in MS Excel sheet is taken
9. The filled-in printed MS Excel sheet is checked (against the printed tables from the publication) and cleared by the 'approver' (other than the person who typed-in the data) for further use (see the example below, filled-in printed MS Excel sheet resulted from the table above and data

entered).



Data Entry by Ruben CORRECT ✓✓

Data entry spreadsheet

Checked by MUGABO Serge

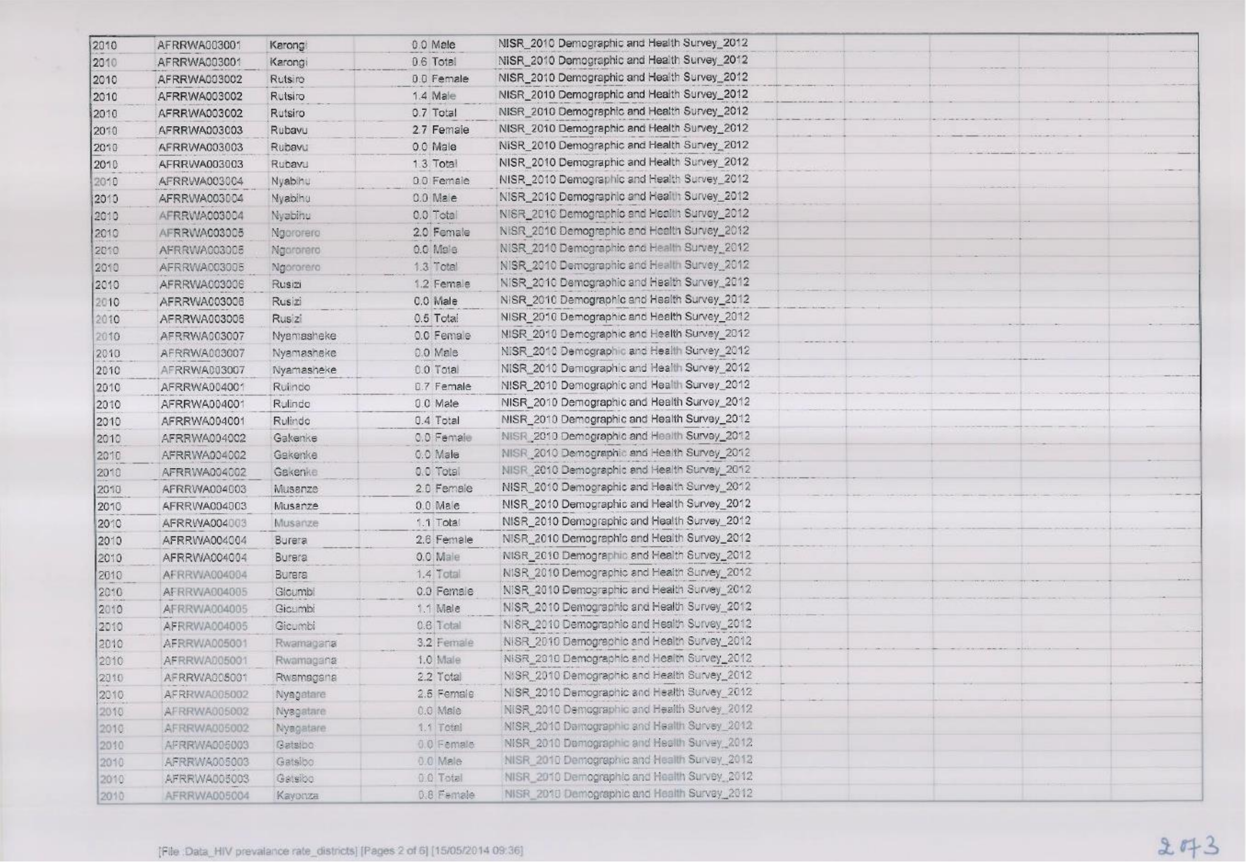
Sector Health
Class HIV/AIDS
Indicator 6.01 HIV prevalence rate among population aged 15-24 yr

Unit Percent **Decimals**

Time	Area ID	Area name	Data value	Subgroup	Source	Footnotes	Denominator
2010	AFRRWA001001	Nyarugenge	4.8	Female	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA001001	Nyarugenge	0.7	Male	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA001001	Nyarugenge	3.2	Total	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA001002	Gasabo	4.6	Female	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA001002	Gasabo	2.3	Male	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA001002	Gasabo	3.6	Total	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA001003	Kicukiro	1.8	Female	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA001003	Kicukiro	0.0	Male	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA001003	Kicukiro	1.0	Total	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA002001	Nyanza	1.3	Female	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA002001	Nyanza	0.0	Male	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA002001	Nyanza	0.7	Total	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA002002	Gisagara	1.4	Female	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA002002	Gisagara	0.0	Male	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA002002	Gisagara	0.7	Total	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA002003	Nyaruguru	1.0	Female	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA002003	Nyaruguru	0.0	Male	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA002003	Nyaruguru	0.5	Total	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA002004	Huye	1.5	Female	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA002004	Huye	2.7	Male	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA002004	Huye	2.1	Total	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA002005	Nyamagabe	1.9	Female	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA002005	Nyamagabe	0.0	Male	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA002005	Nyamagabe	1.1	Total	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA002006	Ruhango	1.2	Female	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA002006	Ruhango	0.0	Male	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA002006	Ruhango	0.6	Total	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA002007	Muhanga	3.0	Female	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA002007	Muhanga	0.7	Male	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA002007	Muhanga	2.0	Total	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA002008	Kamonyi	1.1	Female	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA002008	Kamonyi	0.0	Male	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA002008	Kamonyi	0.6	Total	NISR_2010 Demographic and Health Survey_2012		
2010	AFRRWA003001	Karongi	1.4	Female	NISR_2010 Demographic and Health Survey_2012		

[File :Data_HIV prevalance rate_districts] [Pages 1 of 6] [15/05/2014 09:36]

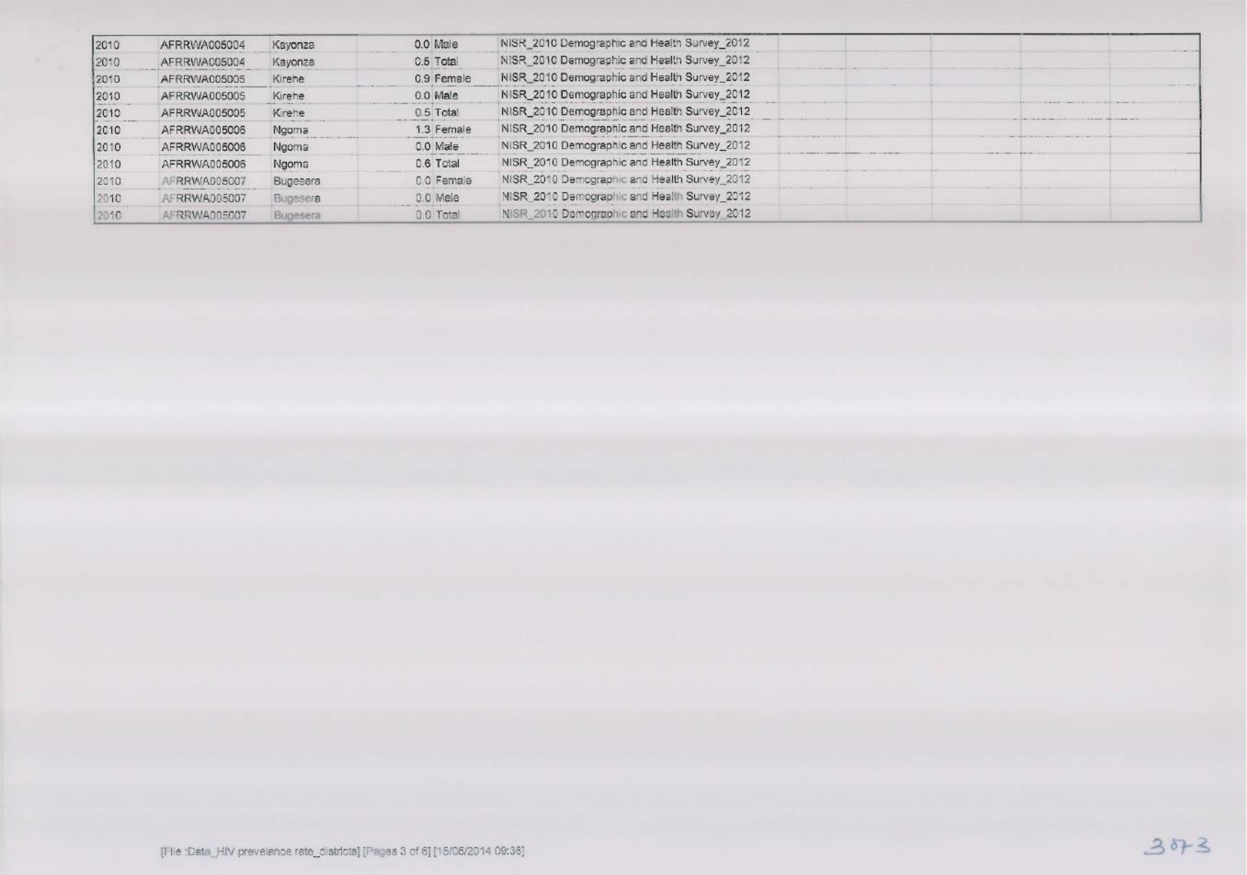
1 of 3



2010	AFRRWA003001	Karongi	0.0	Male	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA003001	Karongi	0.6	Total	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA003002	Rutsiro	0.0	Female	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA003002	Rutsiro	1.4	Male	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA003002	Rutsiro	0.7	Total	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA003003	Rubavu	2.7	Female	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA003003	Rubavu	0.0	Male	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA003003	Rubavu	1.3	Total	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA003004	Nyabihu	0.0	Female	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA003004	Nyabihu	0.0	Male	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA003004	Nyabihu	0.0	Total	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA003005	Ngororero	2.0	Female	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA003005	Ngororero	0.0	Male	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA003005	Ngororero	1.3	Total	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA003006	Rusizi	1.2	Female	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA003006	Rusizi	0.0	Male	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA003006	Rusizi	0.5	Total	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA003007	Nyamasheke	0.0	Female	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA003007	Nyamasheke	0.0	Male	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA003007	Nyamasheke	0.0	Total	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA004001	Rulindo	0.7	Female	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA004001	Rulindo	0.0	Male	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA004001	Rulindo	0.4	Total	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA004002	Gakenke	0.0	Female	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA004002	Gakenke	0.0	Male	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA004002	Gakenke	0.0	Total	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA004003	Musanze	2.0	Female	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA004003	Musanze	0.0	Male	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA004003	Musanze	1.1	Total	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA004004	Burera	2.6	Female	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA004004	Burera	0.0	Male	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA004004	Burera	1.4	Total	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA004005	Gicumbi	0.0	Female	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA004005	Gicumbi	1.1	Male	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA004005	Gicumbi	0.6	Total	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA005001	Rwamagana	3.2	Female	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA005001	Rwamagana	1.0	Male	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA005001	Rwamagana	2.2	Total	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA005002	Nyagatare	2.5	Female	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA005002	Nyagatare	0.0	Male	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA005002	Nyagatare	1.1	Total	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA005003	Gatsibo	0.0	Female	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA005003	Gatsibo	0.0	Male	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA005003	Gatsibo	0.0	Total	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA005004	Kayonza	0.8	Female	NISR_2010 Demographic and Health Survey_2012

[File :Data_HIV prevalance rate_districts] [Pages 2 of 6] [15/05/2014 09:36]

2073



2010	AFRRWA005004	Kayonza	0.0	Male	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA005004	Kayonza	0.5	Total	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA005005	Kirehe	0.9	Female	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA005005	Kirehe	0.0	Male	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA005005	Kirehe	0.5	Total	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA005006	Ngoma	1.3	Female	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA005006	Ngoma	0.0	Male	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA005006	Ngoma	0.6	Total	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA005007	Bugesera	0.0	Female	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA005007	Bugesera	0.0	Male	NISR_2010 Demographic and Health Survey_2012
2010	AFRRWA005007	Bugesera	0.0	Total	NISR_2010 Demographic and Health Survey_2012

[File :Data_HIV prevalance rate_districts] [Pages 3 of 6] [15/05/2014 09:36]

3 of 3

In DevInfo Desktop Admin Application

10. The filled-in and approved MS Excel sheet is imported in DevInfo Desktop (Admin) application

Outside DevInfo Desktop Admin Application (but in DevInfo online deployment)

11. The updated 'mdb' file (the DevInfo database in MS Access) is uploaded online after conversion to MSSQL database

Outside DevInfo

12. 'The photocopied (and printed) tables from the publication' and 'checked & approved printed MS Excel sheet' are stapled together and filed.

Geo-spatial data

The data type relating to the location of geographical features and the relationships between those features are called geo or geo-spatial data. Such data are important not only for contemporary purposes, but are also essential for long-term analyses. Active data management is therefore required to curate, preserve, and provide access to reliable geo-spatial data over time.

This type of data is complex and varies widely in terms of format, size, portability, quality, accuracy and value. The GIS team therefore must implement sound creation practices to ensure that their data are re-usable and sustainable over time, and clearly identify any processing the data have undergone. Effective records management practices should also be implemented to ensure that valuable geospatial data are appraised and selected for long-term curation.

The criteria used to appraise geo-spatial data should reflect both institutional requirements and the potential re-use value of the records for the wider community. SMRP unit can establish such criteria according to institutional priorities and financing availability, and facilitate management of the data for long-term curation. Collaboration between creators (GIS unit) and curators (SMRP unit) will enable sufficient metadata to be preserved with complex data resources so that they can be managed and reliably re-used with authenticity and integrity intact. Due to the often proprietary nature of geospatial data, IT staff should also maintain a technology watch to ensure that data formats do not become obsolete and the data remains accessible.

Publications (digital and print)

This is strictly not a data type but a content type. But even this requires curation. However, this should be treated as per more traditional practice of curation like those followed in traditional libraries.