WASHCost Data Organization and Coding Protocol

June 2010

By: Jeske Verhoeven, Catarina Fonseca, Kwaku Adjei, Dr. Charles Batchelor, Dr. Kristof Bostoen, Dr. Nagarajan Jayakumar, Dr. Amah Klutse, Arjan Naafs, Boudewijn Meijs, Vemula Suryanarayana Murthy

General information Word Document

Reference number*	P1
Country	WASHCost the Netherlands
Contact person	Jeske Verhoeven
E-mail contact person	Verhoeven@irc.nl

Content document

Short description*	Data protocol, specifying agreements on the organisation, structure, coding and sharing of the information collected in the WASHCost project.
Objective	WC04 Data Collection
Remarks/other	

Status document

Final – can be shared outside the project team

Version history

Date of Publishing	Status	Version
1/04/2010	Draft – shared for review	v1.0
21/06/2010	Final draft – shared for review	v2.0
27/07/2010	Edited by Peter McIntyre	v3.0
10/08/2010	Final	v4.0

Storage

Linked to other sources of information?*	• WASHCost Research protocol version 7 • WASHCost Data Organisation and Coding proposal version 2 (20100219) • Data Organisation and Coding meeting report version 1 (201003) • WASHCost Indicators and Formulae version 2 (20100331) Common Excel sheets version 1 (20100401)
Link to folder where it is stored?*	• Wiki workspace WASHCost archive • Wiki workspace research

Acknowledgements

Revised by: Dr. Kirstin Komives, Dr. Richard Franceys, Dr. Ratna Reddy, Dr. Christelle Pezon and Boudewijn Meijs

Abbreviations

GIS	Geographic Information System
RQ	Research Question

Table of Contents

Introduction and background

WASHCost, a five-year initiative, is focused on exploring and sharing an understanding of the true costs of sustainable services. Since 2008, WASHCost has developed new methodologies to better understand and use the costs of providing water, sanitation and hygiene services to rural and peri-urban communities in Ghana, Burkina-Faso, Mozambique and India (Andhra Pradesh).

The WASHCost Data Organisation and Coding protocol describes the agreements reached on how we organise, code, store and share our research data in the WASHCost project. These agreements were made during the WASHCost Data Organisation and Coding meeting that took place in The Hague, the Netherlands from the 25th of February till the 2nd of March 2010. In preparation for this meeting, a proposal[1] was prepared based upon an assessment of current ways of working by the different country teams and experts opinions of the International Advisory Group and IRC staff members.

The main objective of WASHCost data organisation and storage system is to enable the WASHCost team and outsiders to track the source of each data point if and when required. It is to be expected that the research results presented by the WASHCost project will be questioned and scrutinised. The information collected by the WASHCost project therefore needs to be organised and stored in a transparent and accountable manner that enables us to track and show the information on which we base our results.

This protocol builds on existing ways of working in the WASHCost project[2]. Each country team has been organising, coding and storing the information that they have collected during the pilot testing of the research protocol in 2009. The agreed system is simple and transparent and works within the different country contexts and existing ways of working. The aim has been to link the different systems together without causing much extra work or making it unnecessarily complex, rather than imposing the same structure on each country team. Harmonization in this respect is not about being the same, but being compatible.

The scope of this protocol is to organise the quantitative and qualitative information that is collected by the WASHCost project on costs, technologies, service levels and contextual information. Together this information answers to the four key research questions of WASHCost. The four key research questions (RQ) are:

- RQ 1: What is the current, actual magnitude and relative magnitude of different cost components per technology? (per capita, per m3, etc.)
- RQ 2: What is the current, actual magnitude and relative magnitude of different cost components per service level? (per capita, per m3, per village, per district, etc.)
- RQ 3: How do service levels received by poor and non-poor households differ?
- RQ 4: What are the main cost drivers?

The agreed system enables the conduct of any form of analysis, including a cross country analysis answering the four key research questions at international level. It is also flexible enough to allow the export of the research data in the future to a still undefined Decision Support Tool (DST).

[1] WASHCost Data Organisation and Coding proposal v2.0 (20100219).

[2] Current ways of working in the WASHCost project are described in the WASHCost Data Organisation and Coding proposal (20100219) and in the report of the WASHCost Data Organisation and Coding meeting report.

The information collected is often referred to as “the research data” or just “data”. Within the context of the WASHCost project, data is both numerical and descriptive. The scope of this protocol entails the organisation of both types of data.

Another objective of this protocol is to stimulate the sharing of research information within the project. A system to organise and store information is more than a database: it is a mindset or shared culture. This is reflected in how this protocol and the Data Organisation and Coding meeting served as a platform for cross country sharing and learning. That is illustrated in this protocol by the tips, tricks and recommendations that were shared in the Data Organisation and Coding meeting[3].

This protocol has been divided into two parts. The first part of describes the general overview of the agreed structure for organising and storing data. In addition, the first part discusses the quality control and reliability procedures and information sharing. The second part goes into greater detail on data entry and includes tips and tricks to allow all team members to work comfortably with the agreed system and format in Excel.

[3] A detailed report of the WASHCost Data Organisation and Coding meeting is available that provides an overview of the discussions at the meeting.

Part 1: General overview

1. How we organise and store our data in WASHCost

This chapter describes the agreed structure for organising and storing the information that is collected in the WASHCost project.

The core of the system is to store the information that is collected by each country team on costs, technologies, service levels and contextual information in several in several Excel sheets that have a common format between countries and are kept together in a single Excel file in each country. The agreed structure of this Excel file is:

1. **Cover sheet:** general information such as country, data owner and version number with an index of the sheets in the Excel file.
2. **Country dictionary:** an overview of research locations, surveyed technologies and formulae used in each country.
3. **Common indicators:** raw data needed to answer the 4 key research questions.
4. **Sources of information:** an index of sources with a classification of the type of source.

The information stored in this Excel file with 'common Excel sheets' can, when analysed, provide answers to the four key research questions per country and collectively for the international level.

Each country team can choose to:
A. Insert the common Excel sheets as the first sheets in their own country raw data Excel files.
B. Keep the common Excel sheets as a separate file next to their own country Excel files. This requires that country teams copy and paste required information into the Common Excel sheets.

As part of the ongoing reporting system, every country sends their 'common Excel sheets' to the WASHCost - the Netherlands team[4]. The country Excel files are then compiled into one file. The team in the Netherlands can use the common raw data in this file to conduct analysis, including cross country analysis.

1.1 Cover sheet

The first of part of the shared structure is the cover sheet. The cover sheet can be compared with the cover of a report or book as it gives general information on the content of the Excel file, such as the country, the history of the file and the owner of the data. The cover sheet is illustrated in **Annex A: Cover sheet Excel files**.

The cover sheet should enable the recipient of the information, whether a member of the same country team or the WASHCost - the Netherlands team , to know the content of the file and its status. The cover sheet is inserted as the first sheet **in all WASHCost Excel files of all countries**.

The cover sheets makes it easier to share information in the project and to trace information to its source as it specifies a contact person who can respond to questions on the file. The cover sheet also clearly marks the status of the file and the type of data it contains (raw, treated, analysed or modelled). The

[4] To verhoeven@irc.nl and fonseca@irc.nl

recipient will be able to tell, for example, that the file contains raw data from Mozambique and that information from 15 out of 32 villages has been entered.

Please note that **a cover sheet is also inserted at the beginning of all Word documents**. The cover sheet for Word documents contains general information, including the author(s) of the document, the version history and the status. The proposed cover sheet for Word documents is included as **Annex B: Cover sheet Word documents**.

The key information in the cover sheet for Word documents can be stored in the Excel sheet 'sources of information'. In this way an index is generated of what information is stored where and information from different sources can be linked. Still under discussion (1 August 2010) is whether to use automatically generated text fields in the Word cover sheet. This would make it possible to select all field information from all Word files and export them in one go to the Excel sheet, sources of information[5].

1.2 Country dictionary

The key to the success of the agreed structure for organising, coding and storing the information lies in having a clear and precise dictionary for each country. The country dictionary allows people in and outside of the WASHCost project to 'read' the identification number (discussed in next section) and to trace each data point to its source. An example of the first draft of the country dictionary is included as **Annex C: Country dictionary Excel sheet**.

The country dictionary gives an overview for each country of:

i. The research locations under each governance level for each country and its specific codes.
ii. The surveyed water technologies and their specific codes.
iii. The surveyed sanitation technologies and their specific codes.
iv. An overview of the variables each country is using to measure the common indicators.
v. An overview with a description of the different formulae used in each country to calculate the common indicators.

1.3 Identification number

The information collected by the WASHCost project needs to be organised and stored in a transparent and accountable manner that enables us to track and show on what we base our results. The key measure to ensure the source of each data point can be tracked is the agreed identification number.

The identification number is composed of unique country identifiers for different governance levels and technologies. Each governance level at which data collection takes place and all surveyed water and sanitation "technologies" or "infrastructures" are given country specific codes and need to be numbered. When collecting information, all teams document, in consecutive order, at which governance level and to which water and sanitation infrastructure(s) the information relates. Where relevant, this code includes a household identifier.

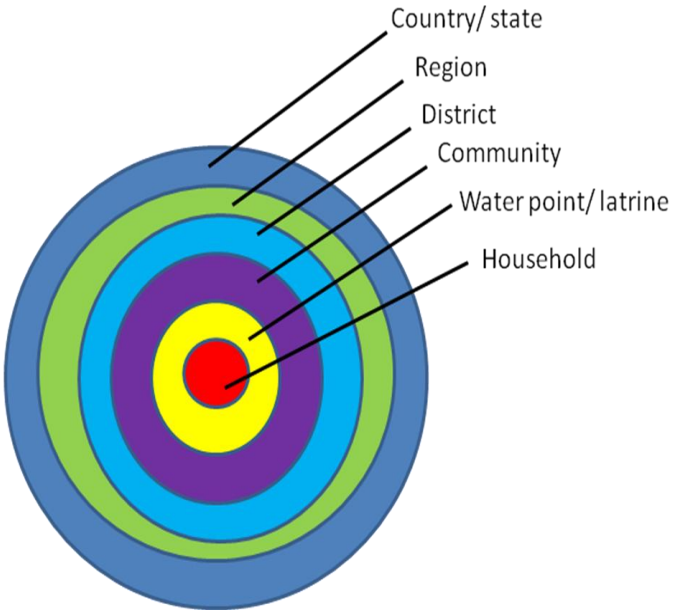
It is important that next to giving a country specific code to all governance levels and types of water and sanitation technologies, that each surveyed water point and latrine is numbered as well as each household. **The number of the specific water point, latrine or household needs to be added in the identification number, after the code for the type of water or sanitation technology or household.**

[5] Jeske Verhoeven is looking into generating automatic text fields in Word, if you know more on how this works please contact Jeske.

Otherwise it is not possible to differentiate between the costs of two different hand pumps in the same village or to separate the costs for households which are poor or non poor within one village.

The different codes together form the identification number. The identification number then allows anyone to follow the trail back to its source. Figure 1 Schematic overview showing how an identification number is made up.

Figure 1 Schematic overview showing how an identification number is constructed



Governance levels differ between the four WASHCost countries. Table 1 provides an overview of the governance levels at which WASHCost is collecting information in the four countries with their equivalents in the other countries. The international description will be used by the WASHCost - the Netherlands team. Table 1 forms the basis of the identification number for each country and will be used as the first variables for each raw data entry Excel sheet by all country teams. In order for others to be able to 'read' the identification number, each country has to document governance levels and water and sanitation technologies with their codes in their country dictionary.

Table 1 Variables of the identification number

Level	International	Ghana	Burkina Faso	Mozambique	India
Level 1	Country/State	Country	Country	Country	State
Level 2	Region	Region	Region	Province	Zone
Level 3	District	District	Province	District	District
Level 4	Multi-district				Mandal (multi)
Level 5	Sub-district		Communes	Posto Administrativo	Mandal
Level 6	Multi-village	Multi village		Town (part of a)	Mandal (part of)
Level 7	Community	Community/Village/ Town	Secteurs/ Village	Community	Habitation/ ward
Level 8	Water point or sanitation infrastructure	Water point or sanitation infrastructure	Point d'eau/ assainissement	Water point or sanitation infrastructure	Water point or sanitation infrastructure
Level 9	Household	Household	Menages	Household	Household

* = Level does not exist in country

This system of identification numbering also enables us to link different sources of information, which is required to answer the research questions. For instance when calculating the cost per technology and per service level, we need to be able to calculate the costs for a specific water point. We are collecting the costs for that specific water point from different sources such as households, surveys at the water point, from the district, region etc. However, since each governance level and surveyed water and sanitation technology is given a country specific code and each surveyed water and sanitation point structure is numbered, we can identify all the information for a specific location collected per technology and analyze it.

Each country team is free to decide what code they give to their governance levels and water and sanitation technologies. It is recommended to use the same codes as the country government for all geographical locations at which we are collecting information. Harmonising coding with national statistic institutes and government systems is a key strategy for embedding research results within government structures. However, each team needs to indicate in the dictionary if the code is being created by the team or if the code comes from the national statistic institutes or government. If a code from the national statistic institutes or government does not exist, teams are free to invent their own codes. However all country teams need to indicate it clearly in their country dictionary when they are using a government code or a code only related to the WASHCost project. For example you can put a certain letter, as W, in front of the code to indicate that the code is only related to the WASHCost project.

Recommendation: Code your questionnaires with the identification number before going into the field to prevent as much as possible problems with data entry.

1.4 Common variables and indicators

The common variables and indicators are the agreed quantitative and qualitative information needed to be collected at a minimum on costs, technologies, service levels and contextual information to answer the four key research questions. The list of common indicators has been divided into the following categories;

1. Contextual information
2. Technologies and infrastructure
3. Water - cost components
4. Sanitation - cost components
5. Water - Service levels indicators
6. Sanitation - Service levels indicators
7. Cost drivers indicators
8. Currency and financial indicators

The list of common indicators will be the same for all countries. The variables can be different for each country. To clarify, an indicator is the “thing” you want to measure and variables are the way you are going to measure it. For example, the way you measure household income can vary from country to country, however each country needs to measure household income. The country dictionary will contain an overview of the variables each country is using to measure the common indicators.

The draft list of in total seven hundred variables and indicators that are common for all four countries has been generated upon review of all country questionnaires and discussions in the Data organisation and Coding meeting. This list will evolve throughout 2010 (and probably start being smaller)[6]. **Each country**

[6] The list of common indicators was shared for review on 1 April 2010 as the WASHCost Indicators and Formulas v2.0 (20100331).

team now needs to define which variables for their country fall under each indicator and check if some indicators are still valid after the data collection.

The common raw Excel sheets will contain the raw data for each of the variables that fall under the common indicators. We suggest that the common raw excel sheets are organised in the same way as the countries collect the data – listed either according to the research tool used or the catgegory of information being collected. The organisational structure of the raw data Excel sheets in the common Excel file will therefore be composed in the following order:

- Household surveys
- Technology survey
- Focus group discussion
- District and regional information
- Contextual information, including GIS, population etc
- Currency and financial information
- Cost drivers

Of course more than one Excel sheet can be used to store the information collection per research tool.

The first variables in all common raw data Excel sheets are the combination of variables forming the identification number for each country. The other columns will contain the data itself for the variables and indicators being collected by the country teams.

1.5 Sources of information

The common way of working in WASHCost is that all country teams currently use Excel 2007 and Word 2003 to organise and store the information that is collected[7]. Additionally, each country has a number of documents that are only available in hard copy or as a picture or PDF.

In order to link these different sources of information, all Word documents, pictures, PDFs and hard copy files are given a reference number. Each country team is free to generate this reference number in its own manner. However it is recommended to use the same system of identification coding as for the Excel sheets to generate a reference number for Word documents, pictures and hard copy files.

The reference number together with the title of the data or information and the place of storage needs to be added in the Excel sheet 'sources of information'. In this way an index is generated of all sources of information. An example is included as **Annex D: Sources of information**.

All focus group discussions, semi structured interviews and questionnaires also need to be given a reference number[8]. This reference number together with some general information is stored in the Excel sheet 'sources of information'.

All hard copy files (for example, hard-to-get government cost documents) have to be scanned. In order to allow anyone to retrieve the information we are referring to (and to track information with the monitoring tool Infolution), we need to make that information available in soft copy. As official government

[7] For data analysis other software can be used. Country teams can decide what best suits their needs.

[8] Organisations such as UNICEF also use this system and create a reference number for all their focus group discussions. UNICEF inserts a small box at the beginning of their Word files with some general information about the focus group discussion. They do this in order to link their Word documents with the information stored in Excel.

documents are sometimes very lengthy with up to a hundred pages and more, it is agreed (at a minimum) to scan the page(s) with the information that is being referred to and the front page of the document. Each scanned document is numbered and this reference number together with some general information is stored in the Excel sheet 'sources of information'.

Note: be careful with copyright and other restrictions when you scan and copy documents. [9] The information should be traceable within our own system, but we have to be careful with references in documents we make public.

Each country team is responsible for respecting copyright on the information that they are collecting. **By June 2010, each country team needs to examine what copyright exists on the information that they are collecting and define and implement procedures to ensure that copyright is respected.** Any restrictions (such as copyright) on information collected can be stored in the 'sources of information' Excel sheet together with the identification number and a general description of the source.

To summarise, basic principles are:

- For collecting and storing data and information we work in Excel 2007 and Word 2003 (we sometimes use Acrobat or image editing programmes).
- A cover sheet is inserted into all Excel files and Word documents.
- Each country develops and implements its own identification number system.
- The variables forming the identification number are the first variables of each raw data entry Excel sheet.
- In order to be able to 'read' the identification number, a country dictionary is made available by each country which gives an overview of research locations and surveyed water and sanitation technologies and their code.
- A list of formulae used to calculate the common variables and indicators is part of the country dictionary.
- The raw data of the common variables and indicators that are needed to answer the 4 key research questions is stored in the common raw data Excel sheet(s).
- The list of common indicators will be the same for all countries. The variables can be different for each country. This information is stored in the common raw data Excel sheet(s).
- An index of sources of information with a classification of the type of source is generated by each country in the Excel sheet 'Sources of information'.
- The index needs to be regularly updated.
- As part of the ongoing reporting system, every country sends their Excel file of 'common Excel sheets' to the WASHCost - the Netherlands team as a key "support document".
- Each country team is responsible for respecting the copyright of information that they are collecting, and alerting other teams when sharing copyright information.
- By June 2010, each country team needs to examine what copyright exists on the information that they are collecting and define and implement procedures necessary to ensure that copyright is respected.
- All country teams need to document the procedures for respecting copyright. This information from each country will be included as an annex to this protocol.
- Any restrictions (such as copyright) on information collected can be stored in the 'sources of information' Excel sheet.

[9] Example: in the literature review being done in the Netherlands (the top 10 documents), some sources of information are articles only available in pay-to-view websites. In the WASHCost website, we only include the abstract of the document but not the full document itself (which is storied in our filing system). In general, copying a few relevant pages and the cover sheets will be covered by "fair use" or "fair dealing" allowances in Copyright regulations that allow use for commentary, criticism, news reporting, research, teaching or scholarship.

As long as it is well documented with an up-to-date index and dictionary:

- Each country can use its own names for variables.
- Each country can use its own codes for questions.
- Each country can use its own country specific area codes.
- Each country can work in its own language.

2. Data quality control and reliability procedures

We expect that the research results presented by the WASHCost project will be questioned and scrutinised. A good data quality control system that shows reliability is therefore essential for the WASHCost project. This chapter specifies agreed data quality control and reliability procedures in WASHCost.

The main principle of the WASHCost data quality control and reliability procedures is to be accurate. In WASHCost, accuracy is the degree of closeness of a measured or calculated quantity to its actual (true) value. The main aim is therefore not to be *precise* but to be *accurate* (you can be precise without being accurate). The sample size[10] proves the precision of the project, the data quality control and reliability procedures we implement will ensure the accuracy. The relationship between accuracy and precision is explained in detail in Annex E.

It was agreed that each country team can have its own data quality control and reliability procedures. These procedures have been developed by each country team in 2009 during three rounds of pilot testing and revision which has refined our system of data collection and the research formats. This has created a robust system which aims at a continuous process of improving the reliability of information and has installed a culture of checks and balances at the different levels of data collection, data entry and analysis. The core of the system for each country consists of triangulation of information, filtering for outliers, and manual checking of data for logic by different experts. The step by step data quality control and reliability procedures specified per country are included as **Annex F: Country data quality control and** reliability procedures .

In the WASHCost project we have agreed to implement a **data entry error margin of 5%**. This means that when records are checked the error rate must be less than 5% (one in 20 entries) or the whole batch must be cross-checked and re-entered. Country teams must build into their procedures a step to check 5% of records at random, and then ensure that fewer that 1 in 20 of these records is incorrect. Please note that we would expect our data entry to be much more accurate than that (one country team estimates their error rate at less than 0.2%). Large mistakes should be exposed during data cleansing and triangulation. For example, you can filter for a decimal point in the wrong place or a where a letter has been placed when there should be a number. Double data entry (data is entered twice by two different people and the entries are cross-checked) is considered a good practice but not a requirement. . All discrepancies between the entries are checked and corrected.

2.1 Sources of information

As the source of information determines for an important part the reliability of the information, each country team will create an index of the sources of information they are using. The index of sources of information can be stored in the last part of the common Excel sheets, 'sources of information'.

As well as listing the source of information with a general description, each country will categorise the type of source. In total, nine different categories of sources can be defined in the WASHCost project. These are:

1. Official government document
2. Official publication other than government

[10] The sampling strategy for each country is documented as part of the WASHCost research protocol.

3. Grey literature
4. Other reports (including media)
5. Interview
7. Household survey
8. Focus group discussion
9. WASHCost team member
10. Photo

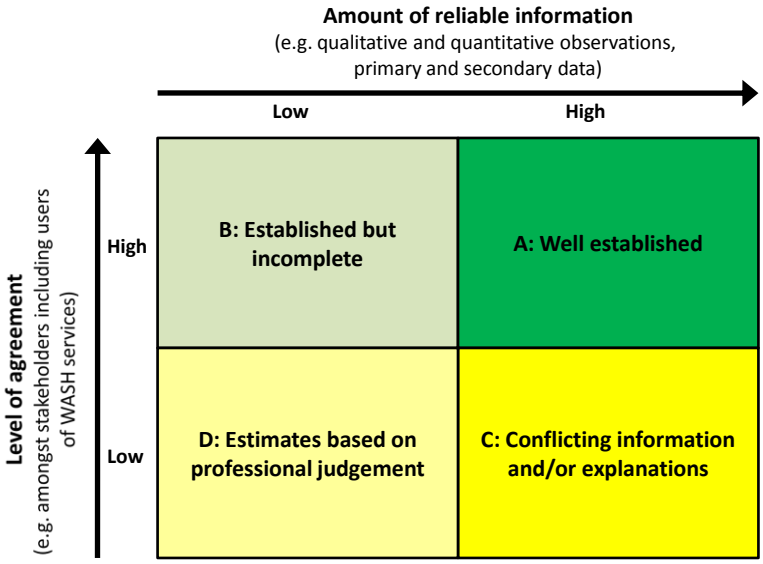
In addition, each country team will add if the source of information is a primary or secondary source. In the WASHCost project a primary source of information is defined as a person with direct knowledge of a situation, or a document created by such a person. A secondary source is person or a document that discusses or relates to information originally presented elsewhere.

The source of information together with some general information and the category of source are added in the common Excel sheet, 'sources of information'.

2.2 Information-reliability classification system

In order to ensure that users of shared information (e.g. in analysis, models, reports etc) are aware of the reliability and limitations of information, the WASHCost project has developed an information-reliability classification system (see Figure 2). In this system information is labelled from A to D, with A being well established information and therefore the most reliable and D estimates based on professional judgement. As more information is collected, discussed and validated by Learning Alliances we expect this to change to well established.

Figure 2 WASHCost information-reliability classification[11]



[11] This classification system is based loosely on the one used by IWMI in the Comprehensive Assessment of Water management in Agriculture (2007).

The main principle of this system is that country teams have a good understanding of the accuracy, variability[12] and uncertainty of information that they have collected. Country teams can therefore label the information that they collect. This label together (with the source of information) gives users of shared information (e.g. in analysis, models, reports etc) an understanding of the reliability and limitations of this information. The information-reliability classification system also shows the level of agreement that exists regarding this information amongst stakeholders.

The label A to D is added in the common Excel sheet, 'sources of information' (along with the source, general information and a category of source).

The different levels in the information reliability classification system can be explained as follows:

Rating	Title	Description
A	Well established	This rating is achieved if: 1) A body of accurate and precise information has been built up 2) This information is not contested amongst stakeholders (including users); 3) The sample size is large enough to support statistical analysis of variability and analysis of the main causes of this variability; and 4) An understanding has built up on levels and root causes of uncertainty in this information.
B	Established but incomplete	This rating is achieved if: 1) A limited body of accurate and precise information has been collected 2) This information is not contested amongst stakeholders (including users).
C	Conflicting evidence and/or explanations	This rating is achieved if: 1) A body of accurate and precise information has been built up 2) The sample size is large enough to support statistical analysis of variability and analysis of the main causes of this variability; and 3) An understanding has built up on levels and root causes of uncertainty in this information. However, the information or interpretations of this information is contested.
D	Estimates based on professional judgement	Initially, all the information collected and quality controlled by the project is likely to have this rating. However, we expect this to change as more evidence is accumulated and consensus is reached in the learning alliances and, if relevant, triangulated with users.

To summarise, the basic principles are:

- The main principle of the WASHCost data quality control and reliability procedures is to be accurate.
- Country teams can use their own quality control system.
- Implementing a system of double data entry is considered good practice.
- The data entry error margin is 5%
- Country teams classify the source of information.

[12] The precise definitions of accuracy, variability in the WASHCost project are available at Annex E.

- Country teams label information from A to D to classify the reliability of the collected information, with A being well established, most reliable information, and D being estimates based on professional judgement.
- All sources of information with their reference numbers, a general description, the classification of the type of source and an information reliability classification are entered in the 'sources of information' common Excel Sheet.

3. Data sharing – who, what, when and how?

This protocol should help facilitate the sharing of information that is collected by each country team. In order to be able to share information, it helps to define what, when, how and who needs to share information.

3.1 What needs to be shared?

The information that needs to be shared at a minimum is the quantitative and qualitative information that is collected by the WASHCost project on costs, technologies, service levels and contextual information, which together will provide answers to the four key research questions.

In practical terms this means:

- The common Excel Sheets with the data on the agreed list of 'common indicators'.
- A 'Sources of information' Excel sheet, with a country index giving an overview of which information is stored where, including Word, PDF, pictures and hard copy files.
- A country dictionary that includes an overview of research locations under each governance level, all surveyed water and sanitation technologies with their specific codes, an overview of the variables each country is using to measure the common indicators and a description of the different formulae used in each country to calculate the common indicators.
- Country questionnaires
- The Word files and PDF documents

3.2 When to share?

As every country is collecting large amounts of information in 2010 and the agreed system is new, it has been agreed to share the agreed research information **every month**. Any problems can then quickly be defined, discussed and resolved.

The information (at a minimum) needs to be shared by storing it on the wiki[13] and sharing it with Jeske Verhoeven (verhoeven@irc.nl).

At a later stage, when the system is functioning well and the majority of the information has been collected and entered, we will share the agreed research information **every four months,** together with the regular reporting.

3.3 How to share?

The following main ways to share information have been agreed:

E-mail: One of the main ways to share information in the WASHCost project is e-mail. The WASHCost contact booklet will be updated to include all e-mail addresses of persons involved in data management. At the same time, the e-mail address of the contact person for each Excel file or Word document is stored on the cover sheet of each file.

[13] The exact location for storing research information on the wiki still needs to be defined. This information will be part of the next version of this protocol.

Phone: For more detailed or sensitive discussions calling by phone is often the easiest and most efficient way to share information. The phone numbers for all WASHCost team members are stored in the WASHCost contact booklet.

Skype: WASHCost team members working with research data can install the free application, Skype, for online free communication via phone and chat. Version three of Skype (and above) also includes the possibility of screen sharing so you show your computer screen to some else via the Internet and see their screen. This can be very handy when discussing research data. Skype can also be used to transfer files. The Skype ID of all WASHCost team members can be found in the WASHCost contact booklet.

TeamViewer: 'TeamViewer' is a free application which gives you the possibility to share your screen and remotely control a desktop via the Internet. The application 'TeamViewer' connects any PC or server around the world via the Internet within a few seconds. TeamViewer also has an integrated file transfer system that allows you to copy files and folders to and from a remote partner.

Wiki: The WASHCost wiki (www.mywash.net) is an online repository and shared platform which is accessible to WASHCost team members via the Internet. The wiki is a safe place to regularly store backups of your data as it is an online archive which can be accessed via the Internet and is only accessible by team members. In case of a fire in the office or a laptop being stolen, information that has been posted on the wiki will be safe.

3.4 Who shares?

One of the main principles in the WASHCost project is that each country team is responsible for managing, storing and sharing the information that it collects in the WASHCost project. In order to facilitate the sharing of information each country has a dedicated person responsible for managing, storing and sharing the research information. **However the lead researcher of each country team needs to clear all research information/data before it is shared.**

Lead researchers are the following:

- WASHCost Ghana: Dr. Kwabena Nyarko (nyark10@yahoo.com)
- WASHCost Mozambique: Arjen Naafs (arjen.washcost@gmail.com)
- WASHCost Burkina: Dr. Amah Klutse (amahklutse@yahoo.fr)
- WASHCost India: Prof. Ratna Reddy (vratnareddy@lnrmi.ac.in)
- WASHCost the Netherlands: Catarina Fonseca with backup from Jeske Verhoeven (verhoeven@irc.nl or fonseca@irc.nl)

3.5 Version control and naming convention

When sharing information it is often difficult to tell the status of a document or if you have the correct document. A system of version control has been agreed in order to keep track of the status of a document and to ensure that when working with shared information all team members work with the correct and latest version of a document.

It has been agreed that all WASHCost team members will use a naming convention for their files and include the following information in the file name:

- Version
- Name of the document
- Date (YYYYMMDD)
- Country

In order to keep track of the version of a document and its status all WASHCost team members will:

- Inserting a cover sheet with the version history of the document, author/contact person, the status (draft, final etc) and if there are plans to update the document in the future[14].
- Keep track of the date in the file name with the data convention: YYYYMMDD.
- Do not use spaces and dots '.' in the file name
- When naming a document, think of a name that will makes sense to other people not a name that only makes sense for you.
- Fill in document properties in Word Documents (in the File, Properties menu).
- Save as 'final', when a document is final.

3.6 Basic principles to prevent data loss

In order to prevent the loss of data, some basic principles have been agreed:

- Excel 2007 and Word 2003 are used to store collected information. If required, this information can be exported to any other programme for specific analysis as the Statistical Package for the Social Sciences (SPSS).
- A free file format converter (FileFormatConverters.exe) can be downloaded from www.microsoft.com to open Word 2007 documents in Word 2003.
- Raw data is always stored in a separate file. After data entry is complete, all raw data files are locked (make them read only) and stored separately from analysed and modelled data.
- Every WASHCost team member needs to have a legal version of Word 2003 and Excel 2007 installed on his/her computer.
- All research data must be stored in several places.
- All research data needs to be stored on the wiki[15].
- A secure server should be used as the main storage area for research data (not individual computers).
- Back-ups must be made on external drives which are kept outside the WASHCost offices.
- Old style folders with paper copies are also part of a good backup strategy.
- Make a PDF of a final version of a document and only share the PDF.
- Make links in an Excel file but do not make links between different Excel files as the link will break when the name of the document changes.

3.7 Guaranteeing anonymity

The WASHCost project guarantees our respondents anonymity. We therefore need to organise, store and share the information we collect in such a way that we are guaranteeing anonymity. This means that names must not be attached to shared information. It has been agreed that the names of respondents are not part of the common indicators. Names of respondents are to be stored separately from the other raw data, preferably in a WinZip file that is password protected and are not to be shared (inside or outside the project).

The Geographic Information System (GIS) data that each country is collecting is also sensitive. We are collecting the GIS coordinates of the dwelling of each household and with this information a household can be traced back exactly. It has therefore been agreed that GIS information must be stored separately from the other raw data in a locked file that is password protected. If possible a country team can;

[14] For an example of a first draft of a cover sheet for Excel and Word files please see Annex A and B.

[15] The exact location for storing research information on the wiki still needs to be defined. This information will be part of the next version of this protocol.

A. Remove or randomise the last digit (or last two digits) of the GPS coordinates. The decision if you need to remove one or two digits depends on the level to which the data needs to be made inaccurate and this correlates with the housing density.
B. If precision is important and you do not want to remove or randomise the last two digits of your GPS coordinates. Instead you can remove all data that is not relevant to the analysis and use an identification number to link the GIS information with the rest of the collected data.

To summarize, basic principles are:

- The common Excel sheets are shared every month. At a later stage, the common Excel sheets will be shared every four months at the same time as the regular reporting.
- At a minimum the information needs to be shared by storing it on the wiki and sharing it with Jeske Verhoeven (Verhoeven@irc.nl).
- Each country has a dedicated person responsible for managing, storing and sharing the research information.
- The lead researcher needs to clear all research information/data before it is shared.
- A system of version control will be implemented by all WASHCost team members, which includes a naming convention and a convention for the date.
- Raw data is always stored in a separate file. After data entry is complete, all raw data files are locked (make them read only) and stored separately from analysed and modelled data.
- All teams must have a legal version of Word 2003 and Excel 2007 installed on their computer.
- All country teams are responsible for ensuring that the anonymity of respondents is guaranteed.

Part 2: Going into detail

4. Data entry

In order to be able to compile the 'common Excel sheets' of each country to one Excel file that can be used for cross country analysis, we need to enter the raw data according to some shared principles. The general shared principles for data entry in Excel 2007 are described in this chapter[16].

Some general recommendations on data entry that were discussed in the WASHCost Data Organisation and Coding meeting are:

- Let field enumerators experience how it is to enter data into Excel so they understand what possible problems can arise with entering data.
- Make a reasonable effort to reduce mistakes in data entry, realise that even a motivated PhD student will make errors when entering large amounts of data.
- Make a special Excel sheet for questions where you do not know how many answers you will get. In this way you do not waste space in your data base and it is easier to analyse. For example for questions as; 'How many members are there in your household (list each member with age and gender)? / How many times has this water point been repaired (list the repairs)? / How many water sources, informal and formal, do you use?'
- In general, when a cell should be empty in Excel do not enter a number that could be in the answer range. You can enter a very high number as 999, put a dot '.' or enter no NR (No response), NA (Not Applicable) or DK (Don't Know) like the Ghana team. Negative numbers as "-2" will create problems when trying to export data from Excel to other software for data analysis as the Statistical Package for the Social Sciences (SPSS).

4.1 Answer categories

Most countries have started with coding the questions in their questionnaires to facilitate easy data entry in Excel. It is recommended to try to format Excel sheets as much as possible with defining a response category for your questions and add these as an answer category in Excel. This can make data entry easier. However do not give the possible answers when asking a question to a respondent as this could influence their answer.

Example: In Ghana, each question has been labelled with a code and this code is written next to each question on the questionnaire. For some questions the Ghana team has also categorized the possible answers and labelled them with a number as for example payment modes when asking how people pay for water:

(1) Pay-as-you fetch
(2) Weekly
(3) Monthly
(4) Yearly
(5) Other, please specify

[16] The shared principles for data entry that are discussed in this section have been developed together with Boudewijn Meijs of IcTonic, the Netherlands and were practised during the WASHCost Data Organisation and Coding Meeting.

These categories with their label (1, 2, etc) can be inserted as a default list of standard answers into Excel. When entering the data into Excel, fewer mistakes can be made as the person entering the data can only choose out of a predetermined set of options.

When categorizing answers, the Likert scale is most commonly used in questionnaires. The Likert scale constitutes of a five to seven point scale, as for example:
(1) Strongly disagree
(2) Disagree
(3) Neither agree nor disagree
(4) Agree
(5) Strongly agree

This gives the respondent a possibility of a more precise answer than using a 3 point scale of:
(1) Bad
(2) Good
(3) Very good

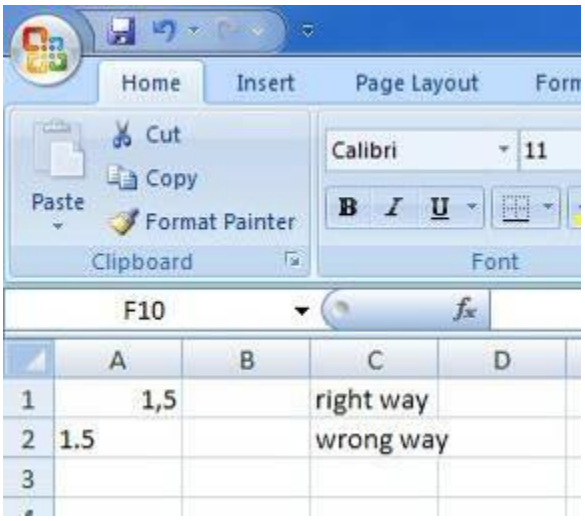
Tip: "Other" is an important option as an answer to a question.

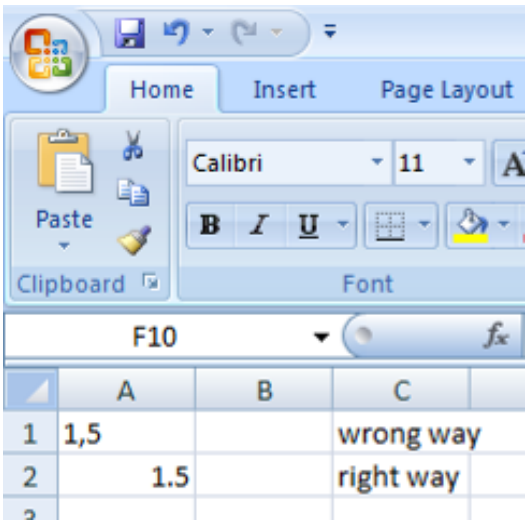
Tip from WASHCost Ghana: Have a de-brief session every evening when collecting data in the field with the team of data collectors. You can discuss the answers given by respondents and see which answers that were put into the 'other' category can fit in one of the pre-defined categories.

4.2 Decimal Separator

Insert all numbers in Excel without a thousand separator as Excel will convert the numbers to the correct format with the correct decimal separator.

! Check: if a number in a cell is aligned to the left side of a cell, then it has been inserted as text. This means that you've used the wrong decimal separator. If the content of a cell is aligned to the right side of a cell then it has been inserted as a number and it will be converted the right way. Off course to check: de-activate any alignment settings.

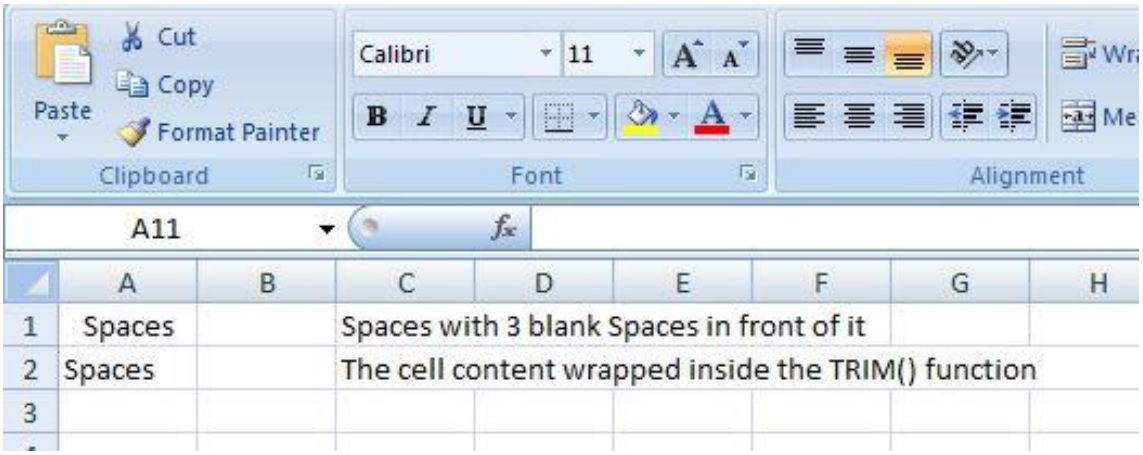




4.3 Trim

Always be careful that you do not add any blank spaces behind or in front of text you enter in Excel. In many excel functions you want to refer to a cell with text contents. If they don't exactly match, the formula could give an error or even worse: give the wrong answer. If you are referring to text, it could always be useful to wrap this cell referral inside a TRIM-function.

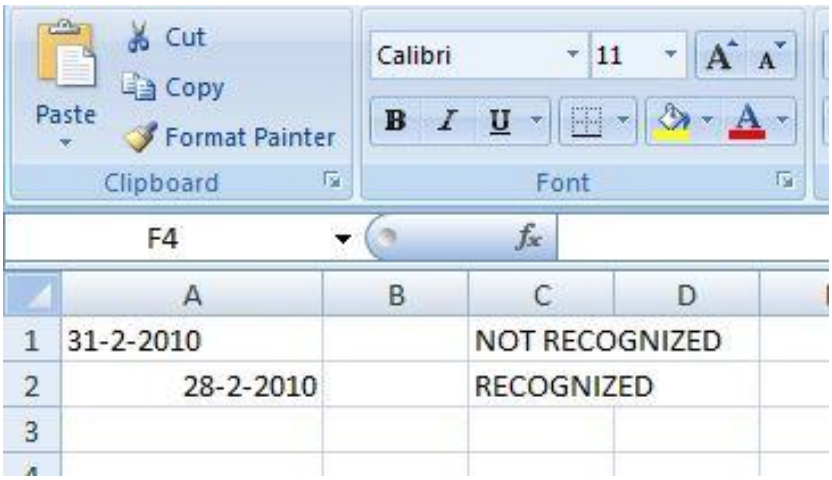
Example: The first 3 characters of cell A1 are spaces. After the 3 spaces comes the word "Spaces". If you use the TRIM() function, you'll see that the first 3 spaces are removed (the same if there are any spaces at the end of the cell content). This will give you the exact content of a cell. Very useful if you're working with e.g.: VLOOKUP(), HLOOKUP(), MATCH() etc..



***Formula cell A2 is: "=TRIM(A1)"**

4.4 Date Recognition

The same as with the decimal separators, if you're entering a date, always check if it is entered as a valid date in Excel. An invalid date will be aligned to the left side of a cell. A valid date will be aligned to the right side of a cell.



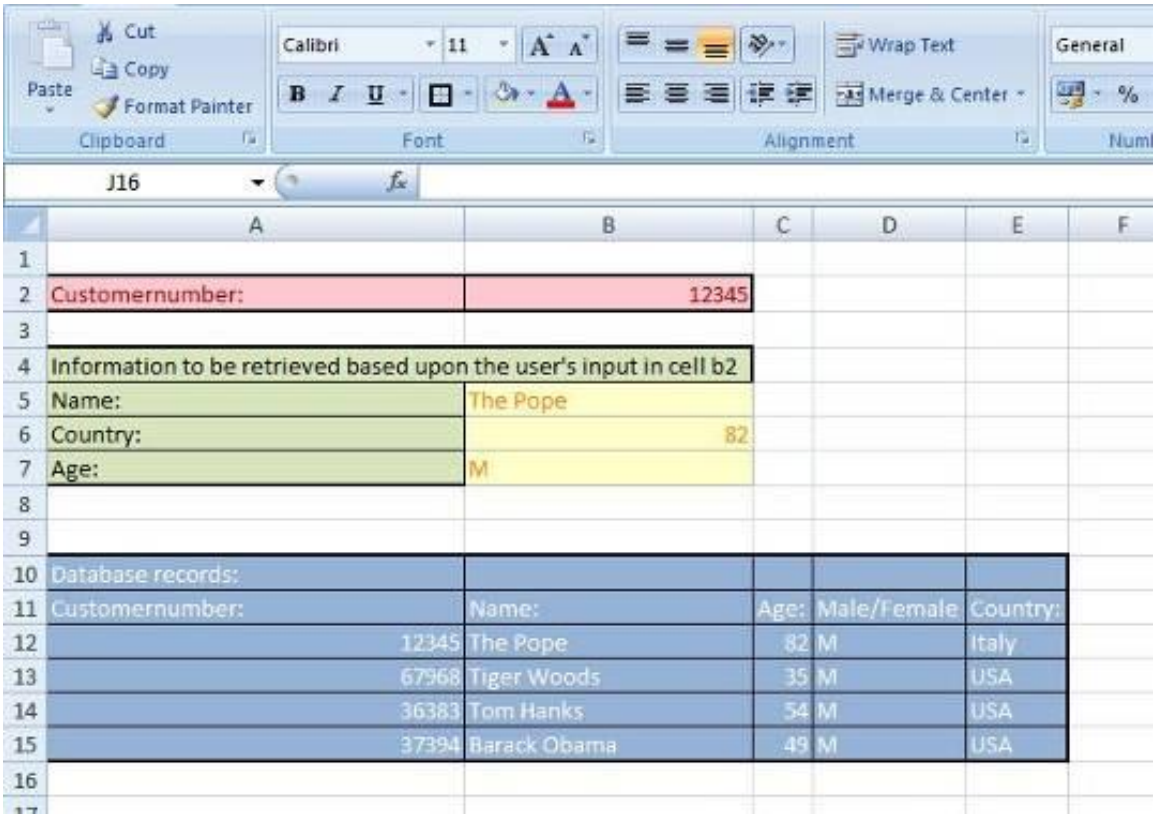
4.5 Vlookup()

This function is used often in an Excel worksheet. We can use the VLOOKUP() or HLOOKUP() to find information based upon one specific value. In the example below we're trying to find information about a person.

1. The name of the person,
2. The country in which the person is living,
3. The age of the person.

We want to find this information based upon his "Customer number". Procedure:

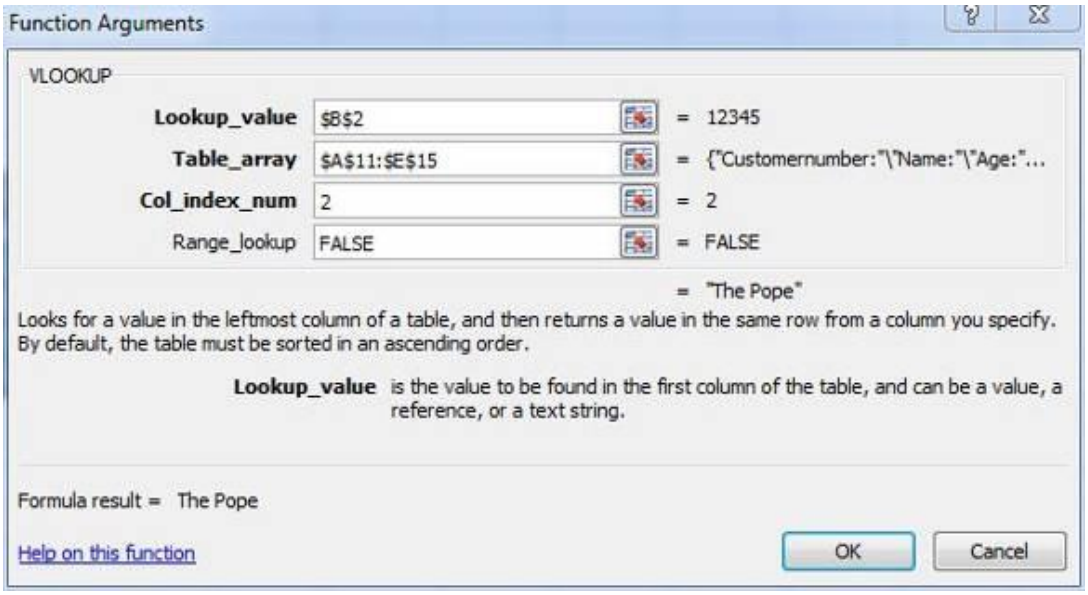
1. Insert in cell b2 the customer number.
2. In cells, b5,b6 and b7 will appear the correct date. The date is extracted from the table in blue.



	A	B	C	D	E	F
1						
2	Customernumber:	12345				
3						
4	Information to be retrieved based upon the user's input in cell b2					
5	Name:	The Pope				
6	Country:	82				
7	Age:	M				
8						
9						
10	Database records:					
11	Customernumber:	Name:	Age:	Male/Female	Country:	
12	12345	The Pope	82	M	Italy	
13	67968	Tiger Woods	35	M	USA	
14	36383	Tom Hanks	54	M	USA	
15	37394	Barack Obama	49	M	USA	
16						

In the VLOOKUP () function are four arguments used.

1. The lookup-value (in this case: cell b2)
2. The array in which we have stored all our information (the table in blue (A11:E15), but off course this could be a table on a different sheet),
3. The column in which we can find the asked information (Name is in the second column of our array as defined in argument 2),
4. Should there be an exact match (are the numbers or text equal) or not. If an exact match is required the argument is FALSE, otherwise it's TRUE. In most of the case the argument will be FALSE.



- The formula: "=VLOOKUP($B$2;$A$11:$E$15;2;FALSE) " would give us the name as a result.
- The formula: "=VLOOKUP($B$2;$A$11:$E$15;3;FALSE) " would give us the age as a result.
- The formula: "=VLOOKUP($B$2;$A$11:$E$15;4;FALSE) " would give us the sex as a result.
- The formula: "=VLOOKUP($B$2;$A$11:$E$15;5;FALSE) " would give us the country as a result.

4.6 Different type of measurement (for example Kilo or KG)

When filling in the data sheets always use the same type of measurement. All information that you would like to export to other software for example for data analysis such the Statistical Package for the Social Sciences (SPSS) needs to be numeric. If you add a measurement in the same case as the number, for example 30km, Excel will read it as text and not as numeric data. Unless you declare a number format in Cell properties/number/custom

Example: do not use Gallons and Litres in the same column. Use Litres or Gallons.

4.7 Pivot tables

Pivot tables are dynamic tables, of which you can easily change the layout and change the format of the data presented. When you start to work on Pivot tables make sure of two important things:

1. Make sure that your list has column headers. They should not be merged cells or with duplicate names.
2. Make sure that in your range of data there are no complete empty rows or empty columns.

How to create the pivot table?

1. Select one of the cells of your list of data.
2. Go to the insert ribbon,
3. Click on pivot table,
4. Check in the window that appears if the range of your data is correct. If not, select the right region of data.
5. Choose for creating the pivot table on a new sheet (it's best to choose for this option, do not create it on the same sheet as where your data is).
6. Once the new sheet is visible, you have to decide what kind of data you want to see.
7. Drag from the list with all your column headers, the items you want to see in your pivot table. Per item you want to categorise, place its name-item in the Row- or Column area.

8. Place the name-item of which you want to do the calculation in the values area.
9. If you want to see how the pivot table came to a specific number in the data region: simply double-click on the number and all the related data will appear on a new sheet (it's an extraction of the original data).

4.8 Conditional formatting

Conditional formatting can be used to emphasize numbers or data that meet a specific condition.

Example: if a number is below 5, give it a red background colour. If the number is above 5 give it a blue background colour. Steps:

1. Select all the data on which you want to place the conditional formatting.
2. On the insert ribbon, click on the conditional formatting button.
3. Select "New rule",
4. Select a rule type: "Format only cells that contain"
5. Enter the rule description and set the format.
6. This way you can add multiple rules.

More useful examples can be found on: http://www.contextures.com/xlCondFormat01.html

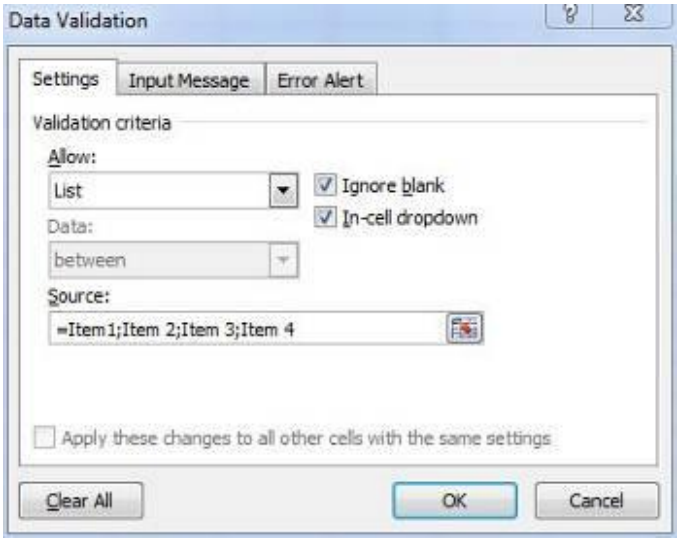
4.9 Data Validation

Data validation can be used to force users of the excel sheets to only enter data in a specific format inside a cell. For instance, the content of a cell should be a date, a number, free text, or an item out of a list.

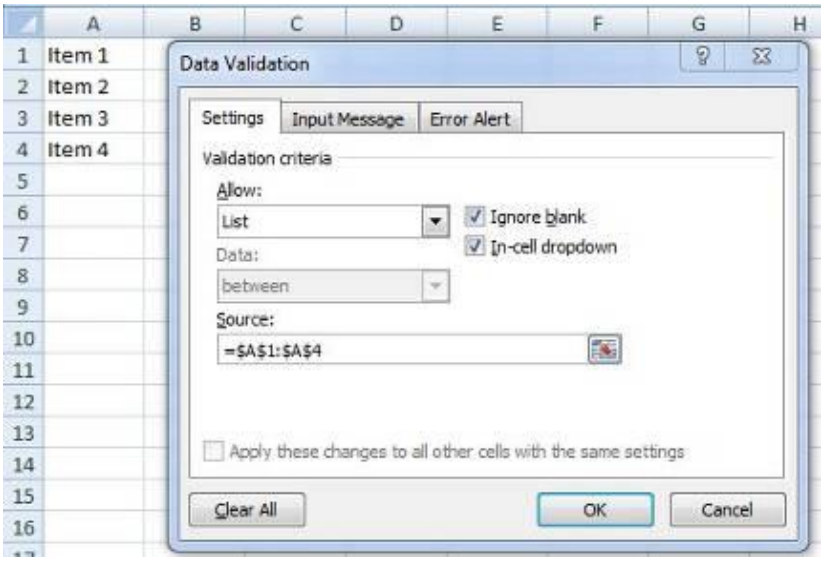
Example for an item out of a list:

1. On the Data Ribbon, go to Data Validation.
2. On the settings tab, choose "list" from the "allow" dropdown box.
3. Insert the source you're using:
   a) Type the list items in the source box,
   b) Refer to a range on your actual excel sheet,
   c) Refer to a named range (if the list you're referring to is on a different sheet).
4. On the "Input message" tab, insert the title and message settings.
5. On the "Error Alert" tab, insert the title and message settings.
6. Click OK and select the cell.

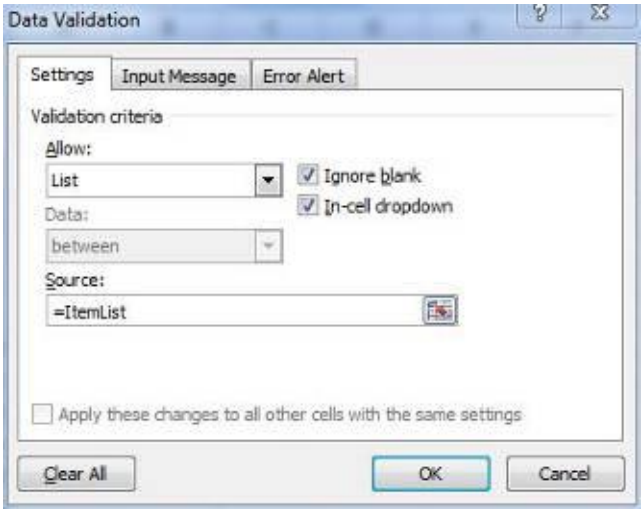
Picture 3a:

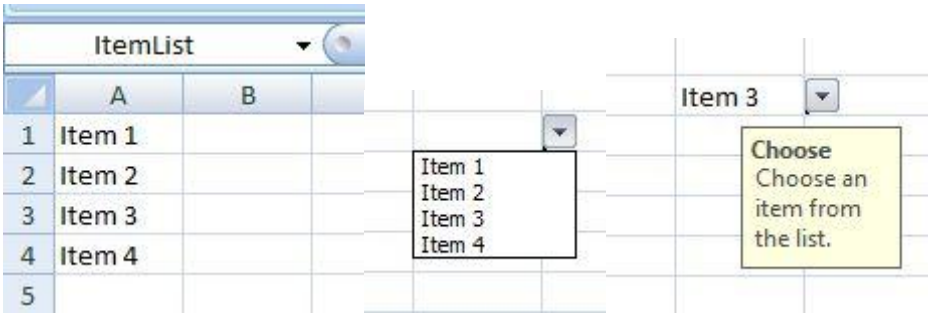


Picture 3b:



Picture 3c:



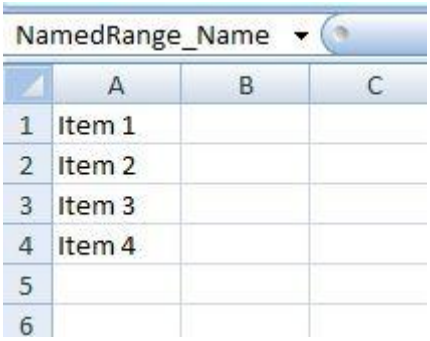


4.10 Named ranges

In step 3c of the data validation example we entered a named range. A named range is a number of cells referred to with a name. This could be very useful in long formulae. You don't refer to a range (E.g.: sheet2!A8;H9), but to a name (E.g.: data2010). In step 3c of data validation the source was referred to as a named range "ListItems".

How to create a named range:

1. Select all the cells within your range.
2. Go to the cell address box and type a name (above cell a1).
3. Then press "Enter".

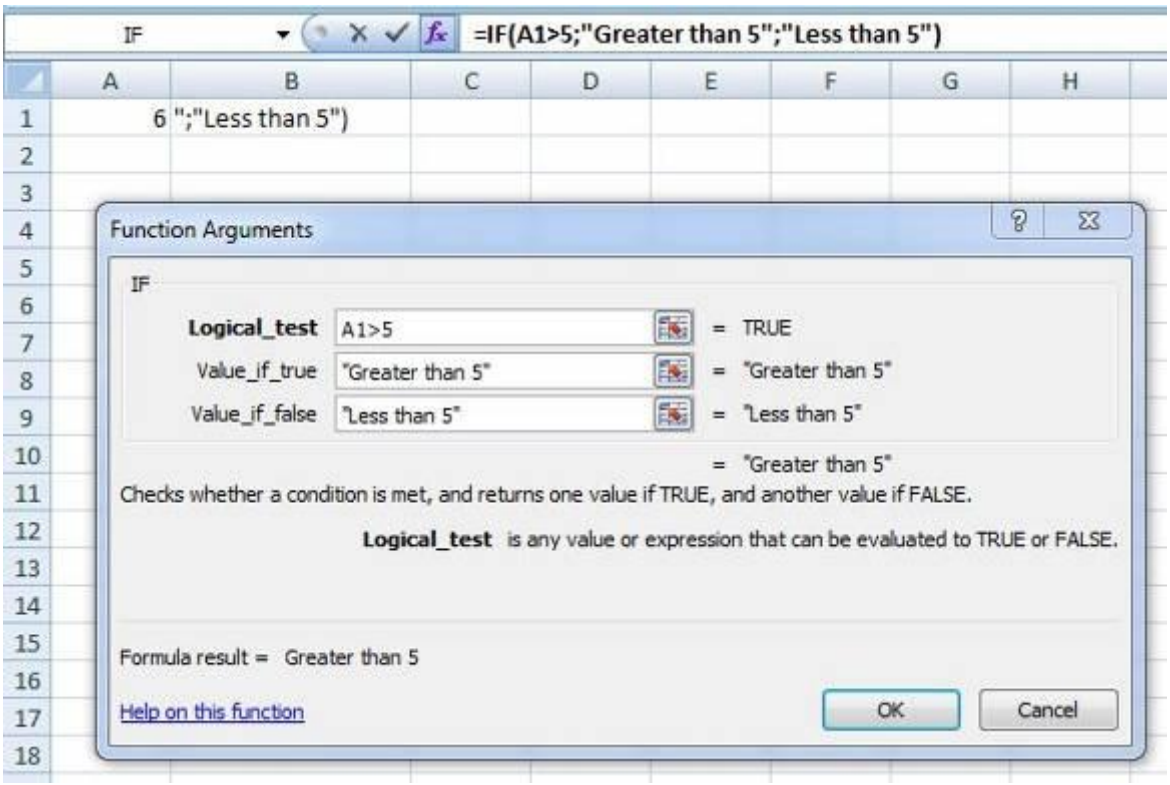


4.11 The IF() function

With this function, you can examine of a specific condition (logical test) is true or false. If the condition is true then the "Value if true" appears, otherwise the "Value if false" appears. E.g.: if the number in a cell is greater than 5 then in the cell next to it the text "Greater than 5" will appear. In case it's less than 5, the text "Less than 5" will appear.

Steps:

1. Select the "if" function.
2. In the logical test set the condition,
3. Insert the values (formulae, numbers or text) in Value if true/Value if false



4.12 IsError()

The output of the IsError function is “True” or “False”. It indicates if the cell content contains an error. As an example, see pictures below.

	A	B
1	text	ABCD
2	number	100
3	text * number	#VALUE!
4		
5	IsError	TRUE
6		
7	If - IsError	Error, please correct!

In cell B3, the formula is B1*B2, which gives an error as a result. In cell B5 the IsError function determines that it has an error in the calculation. If you enter the IsError function inside an IF function, you can rewrite the “#Value!” message to something more meaningful. E.g.: “Error, please correct!”

	A	B
1	text	ABCD
2	number	100
3	text * number	=B1*B2
4		
5	IsError	=ISERROR(B3)
6		
7	If - IsError	=IF(ISERROR(B3);"Error, please correct!";B3)

4.13 Indirect

This function transforms a text string into a usable argument in a formula.

	A	B
1	Which sheet?	Sheet3
2	Which cell?	a2
3		
4	Value of the cell:	Test data of sheet 3
5		
6	Formula used:	=INDIRECT(B1&"!"&B2)
7		
8	Normal way:	=Sheet2!A2
9		

In the example above, we can type the name of the sheet in cell b1, also we can type the address of the cell of which we want to see the data of. Instead of making the formula like the content of cell b8, we can make the formula more dynamic.

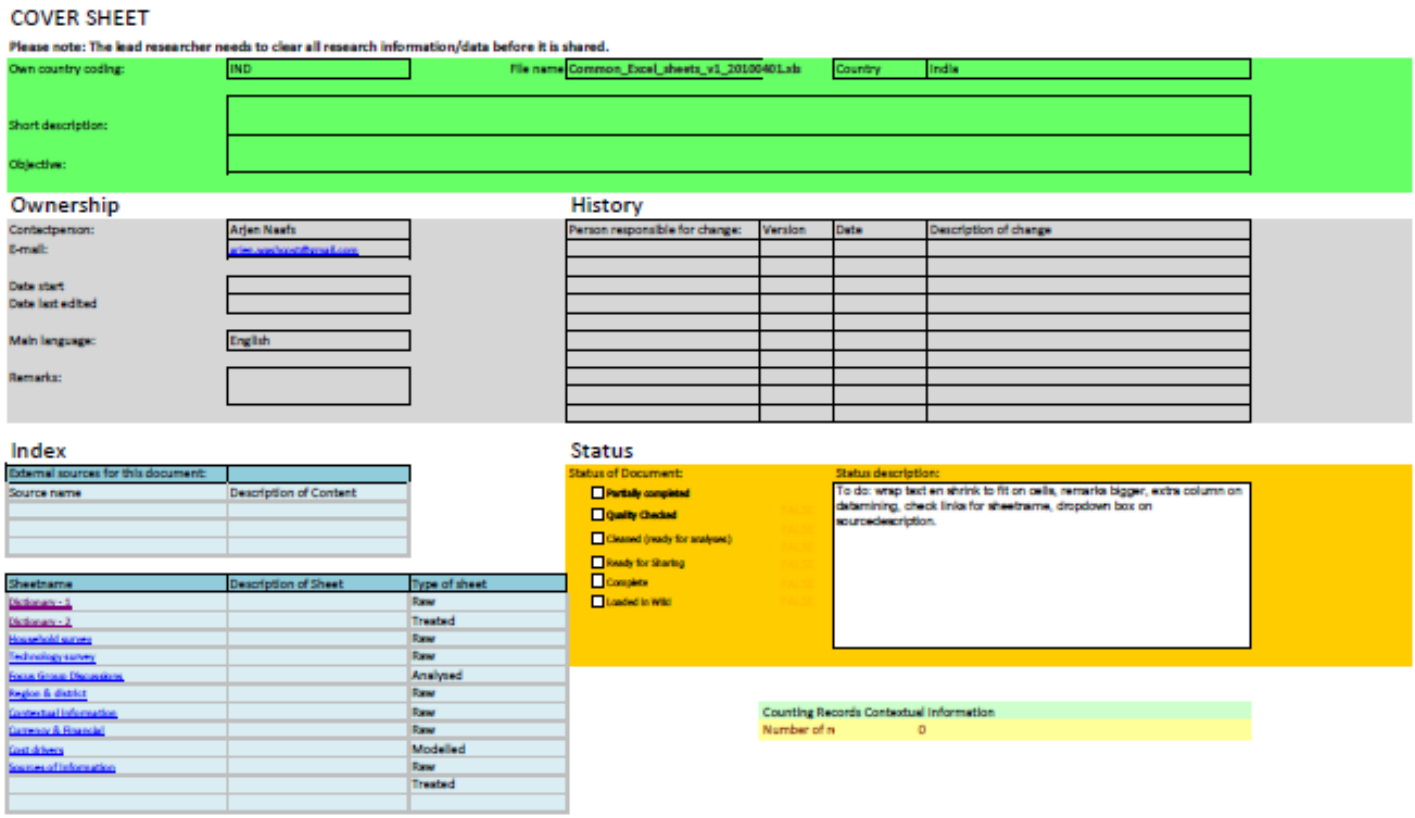
To summarize, basic principles are:

- Try to format Excel sheets as much as possible with defining a response category for your questions.
- Insert all numbers in Excel without a thousand separator.
- Always be careful that you do not add any blank spaces behind or in front of text you enter in Excel.
- If you're entering a date, always check if it is entered as a valid date in Excel.
- When filling in the data sheets always use the same type of measurement.

Finally:
We look forward to compile the one Excel sheet with all the data!
Thank you all for your inputs in making this data protocol ready.

Annex A: Cover sheet Excel files

For a copy of the Excel file with the common Excel sheets please contact Jeske Verhoeven (verhoeven@irc.nl)



COVER SHEET

Please note: The lead researcher needs to clear all research information/data before it is shared.

Own country coding:	IND	File name	Common_Excel_sheets_v1_20100401.xls	Country	India
Short description:					
Objective:					

Ownership

Contactperson:	Arjen Naafs
E-mail:	
Date start	
Date last edited	
Main language:	English
Remarks:	

History

Person responsible for change:	Version	Date	Description of change

Index

External sources for this document:	
Source name	Description of Content

Sheetname	Description of Sheet	Type of sheet
Dictionary - 1		Raw
Dictionary - 2		Treated
Household survey		Raw
Technology survey		Raw
Focus Group Discussions		Analysed
Region & district		Raw
Contextual information		Raw
Currency & Financial		Raw
Cost drivers		Modelled
Sources of information		Raw
		Treated

Status

Status of Document:

- Partially completed
- Quality Checked
- Cleaned (ready for analyses)
- Ready for Sharing
- Complete
- Loaded to Wiki

Status description:

To do: wrap text en shrink to fit on cells, remarks bigger, extra column on datamining, check links for sheetname, dropdown box on sourcedescription.

Counting Records Contextual Information	
Number of n	0

Annex B: Cover sheet Word documents

General information Word Document

Reference number*	
Country	
Contact person	
E-mail contact person	
Date of data collection	
Date of data entry	

Content document

Short description*	
Objective	
Remarks/other	

Status document

Version history

Date of publishing	Status	Version

Storage

Linked to other sources of information?*	
Link to folder where it is stored?*	

*This information needs to be stored in the Excel sheet 'sources of information'.

Annex C: Country dictionary Excel sheet

Please see Common Excel sheets version 1 (20100401) for latest version

Code	Indicator group	Sub indicator group	Indicator	International	Description
CG1	Contextual information	General information	Level 1	Country/state	
CG2	Contextual information	General information	Level 2	Region	
CG3	Contextual information	General information	Level 3	District	
CG4	Contextual information	General information	Level 4		
CG5	Contextual information	General information	Level 5		
CG6	Contextual information	General information	Level 6		
CG7	Contextual information	General information	Level 7		
CG8	Contextual information	General information	Level 8	Community	
CG9	Contextual information	General information	Level 9	Household	
CG10	Contextual information	General information	Level 10	Water point / Latrine	
TW1	Technologies and infrastructure	Water infrastructure	Technology water 1	Bottled water	
TW2	Technologies and infrastructure	Water infrastructure	Technology water 2	Unprotected shallow well	
TW3	Technologies and infrastructure	Water infrastructure	Technology water 3	River/ stream/ lake/pond	
TW4	Technologies and infrastructure	Water infrastructure	Technology water 4	Rain water harvesting	
TW5	Technologies and infrastructure	Water infrastructure	Technology water 5	Shallow well with bucket and rope	
TW6	Technologies and infrastructure	Water infrastructure	Technology water 6	Shallow well with handpump	
TW7	Technologies and infrastructure	Water infrastructure	Technology water 7	Manual drilled borehole with handpump	
TW8	Technologies and infrastructure	Water infrastructure	Technology water 8	Mechanically drilled borehole with handpump	
TW9	Technologies and infrastructure	Water infrastructure	Technology water 9	Mini system with borehole	Max 5 standpipes, no HH connection
TW10	Technologies and infrastructure	Water infrastructure	Technology water 10	Small piped system	Up to 500 HH connections
TW11	Technologies and infrastructure	Water infrastructure	Technology water 11	Medium piped system	500 to 2000 HH connections
TW12	Technologies and infrastructure	Water infrastructure	Technology water 12	Large urban system	More than 2000 HH connections
TW13	Technologies and infrastructure	Water infrastructure	Technology water 13		
TW14	Technologies and infrastructure	Water infrastructure	Technology water 14		
TW15	Technologies and infrastructure	Water infrastructure	Technology water 15		
TW16	Technologies and infrastructure	Water infrastructure	Technology water 16		
TS1	Technologies and infrastructure	Sanitation infrastructure	Technology sanitation 1	Open defecation	
TS2	Technologies and infrastructure	Sanitation infrastructure	Technology sanitation 2	Traditional latrine	
TS3	Technologies and infrastructure	Sanitation infrastructure	Technology sanitation 3	Slab latrine	
TS4	Technologies and infrastructure	Sanitation infrastructure	Technology sanitation 4	VIP Latrine	
TS5	Technologies and infrastructure	Sanitation infrastructure	Technology sanitation 5	Pour flush latrine	
TS6	Technologies and infrastructure	Sanitation infrastructure	Technology sanitation 6	Pour flush latrine with septic tank	
TS7	Technologies and infrastructure	Sanitation infrastructure	Technology sanitation 7	Public latrine	
TS8	Technologies and infrastructure	Sanitation infrastructure	Technology sanitation 8	Toilet with sewage	
TS9	Technologies and infrastructure	Sanitation infrastructure	Technology sanitation 9	Toilet with septic tank	
TS10	Technologies and infrastructure	Sanitation infrastructure	Technology sanitation 10		
TS11	Technologies and infrastructure	Sanitation infrastructure	Technology sanitation 11		

Annex D: Sources of information

Please see Common Excel sheets version 1 (20100401) for latest version

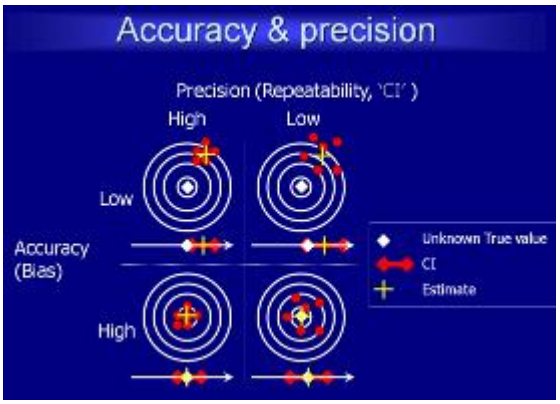
Item - Unique number	Type of source	Reviewer
CBKF-D01-HH05-TW-TTHP	Government report	jv
CBKF-D01-HH08-TW-TTHP	PDF-file	cf

Name of document	Indicator	Place of storage	Hard/Soft Copy	Short Description
New ways on WashCost		Library	H	report on how to do things....
just any title			S	

Annex E: Important definitions

We should take care when using statistical terms (e.g. accuracy, uncertainty, variability etc) as using them incorrectly will lead to confusion. "Accuracy", "confidence" and "uncertainty", have specific statistical meaning. "Reliability" is not a statistical term and it is word that is in everyday use, therefore it has become the title of the WASHCost classification system. Figure 3 Relationship between accuracy and precision, might help as you can be precise without being accurate, the misconception that small repeatability (CI) indicates accuracy.

Figure 3 Relationship between accuracy and precision



Accuracy (and precision) In the fields of science, engineering, industry and statistics, accuracy is the degree of closeness of a measured or calculated quantity to its actual (true) value. Accuracy is closely related to precision which is the degree to which further measurements or calculations show the same or similar results. Accuracy indicates proximity to the true value, precision to the repeatability or reproducibility of the measurement. The results of calculations or a measurement can be accurate but not precise, precise but not accurate, neither, or both. A measurement system or computational method is valid if it is both accurate and precise

Variability The degree to which differences exist amongst, for example, the costs of the same goods or services. As variability is an inherent characteristic, it cannot be reduced by collecting additional information or by carrying out additional research. However, WASHCost should have the target of explaining and characterising important causes of variability.

Uncertainty The degree to which a value is unknown. In the context of WASHCost, uncertainty could arise from an imperfect understanding of the nature of disaggregated life-cycle costs and/or from differences of opinion over what is known or even knowable. One aim of WASHCost is to reduce this uncertainty.

Standard deviation In statistics, standard deviation is a simple measure of the variability or dispersion of a data set. In mathematical terms it is equal to the square root of the arithmetic average of the squares of the deviations from the mean in a frequency distribution. A low standard deviation indicates that all of the data points are very close to the same value (the mean), while high standard deviation indicates that the data are “spread out” over a large range of values.

Annex F: Country data quality control and reliability procedures

The data quality control and reliability procedures in the four WASHCost countries include the following:

WASHCost Burkina Faso

1. The system of data collection was designed in three rounds of pilot testing and revision of the research formats.
2. The research tools have been reviewed and revised by a team of external experts different fields (mathematics/modelling, 2x WASH expert, 1x GIS expert, 2x economist).
3. The system of data collection was validated by the most important stakeholders.
4. The name of the enumerator or field investigator who collected the information and the supervisor are stated on each questionnaire, format or survey.
5. During data collection in the field enumerators check each answer for logic and if unclear verify the answer.
6. During data collection in the field both the field investigator/enumerator and the field supervisor check all the collected research formats for inconsistencies and gaps. If consistencies or gaps are found the field investigator/enumerator tries to fill the gaps or verifies the information.
7. The data is entered by experienced data entry staff.
8. The data base manager supervises data entry. The database manager visits the field teams regularly.
9. The database manager checks data entry. Out of a 100 records at least 5 records are checked at random. If more than 5 records are incorrect all data needs to be cross checked and re-entered.
10. The database manager cleans the database by filtering and screening for outliners with a mixture of cross tabulation and basic statistics. Errors are corrected and information is verified when needed.
11. The database is now *clean* and aggregated by the data manager.
12. The aggregated figures are checked by a 2 different experts for consistency. All inconsistencies are verified. After the approval of these 2 experts the database is send to the lead researcher for analysis.
13. The lead researcher filters the database for inconsistencies before he starts analysis.

WASHCost Ghana

1. The system of data collection was designed in three rounds of pilot testing and revision of the research formats.
2. The research tools have been reviewed and revised a consultant.
3. Field enumerators and supervisors have entered the data collected during the pilot testing in order to understand the set up of the database and the data entry process.
4. The system of data collection was validated by the most important stakeholders.
5. The name of the enumerator or field investigator who collected the information and the supervisor are stated on each questionnaire, format or survey.
6. During data collection in the field enumerators check each answer for logic and if unclear verify the answer.
7. During data collection in the field both the field investigator/enumerator and the field supervisor check all the collected research formats for inconsistencies and gaps. If consistencies or gaps are found the field investigator/enumerator tries to fill the gaps or verifies the information.
8. The data is entered by experienced data entry staff.
9. All formats (other than household questionnaires) are entered twice (double data entry) by two different people and differences are compared and corrected.

10. The data base manager supervises data entry. The database manager visits the field teams regularly.
11. The database manager checks data entry. Out of a 100 records at least 5 records are checked at random. If more than 5 records are incorrect all data needs to be cross checked and re-entered.
12. The database manager cleans the database by filtering and screening for outliners with a mixture of cross tabulation and basic statistics. Errors are corrected and information is verified when needed.
13. The database is now *clean* and aggregated by the data manager.
14. The aggregated figures are checked by the lead researcher for consistency. All inconsistencies are verified.
15. The lead researcher filters the database for inconsistencies before he starts analysis.

WASHCost Mozambique

1. The system of data collection was designed in three rounds of pilot testing and revision of the research formats.
2. The system of data collection was validated by the most important stakeholders.
3. The name of the enumerator or field investigator who collected the information and the supervisor are stated on each questionnaire, format or survey.
4. During data collection in the field enumerators check each answer for logic and if unclear verify the answer.
5. During data collection in the field both the field investigator/enumerator and the field supervisor check all the collected research formats for inconsistencies and gaps. If consistencies or gaps are found the field investigator/enumerator tries to fill the gaps (by returning to the field) or verifies the information.
6. Both the field investigator/enumerator and the field supervisor sign the research format after checking the format for inconsistencies and gaps, stating that the information is correct.
7. The data is entered by experienced data entry staff. Data entry is restricted by software to minimise data entry mistakes.
8. All household questionnaires are entered twice (double data entry) by two different people and differences are compared and corrected.
9. The data base manager supervises data entry. The database manager visits the field teams regularly.
10. The database analyst checks data entry. Out of a 100 records at least 5 records are checked at random. If more than 5 records are incorrect all data needs to be cross checked and re-entered.
11. The database analyst cleans the database by filtering and screening for outliners with a mixture of cross tabulation and basic statistics. Errors are corrected and information is verified when needed.
12. The database is now *clean* and aggregated by the data analyst.
13. The aggregated figures are checked by one expert for consistency. All inconsistencies are verified. After the approval of these 2 experts the database is send to the lead researcher for analysis.
14. The lead researcher filters the database for inconsistencies before he starts analysis.

WASHCost India

1. The system of data collection was designed in three rounds of pilot testing and revision of the research formats.
2. A consultant tested and reviewed the research formats that included the Qualitative Information System (QIS) research tool.
3. The system of data collection was validated by the most important stakeholders.

4. The name of the enumerator or field investigator who collected the information and the supervisor are stated on each questionnaire, format or survey.
5. During data collection in the field enumerators check each answer for logic and if unclear verify the answer.
6. During data collection in the field both the field investigator/enumerator and the field supervisor check all the collected research formats for inconsistencies and gaps. If consistencies or gaps are found the field investigator/enumerator tries to fill the gaps (by returning to the field) or verifies the information.
7. Both the field investigator/enumerator and the field supervisor sign the research format after checking the format for inconsistencies and gaps, stating that the information is correct.
8. The data is entered by experienced data entry staff.
9. The data base manager supervises data entry. The database manager has field experience and WASH expertise. He also visits the field teams regularly.
10. The database manager checks data entry. Out of a 100 records at least 5 records are checked at random. If more than 5 records are incorrect all data needs to be cross checked and re-entered.
11. The database manager cleans the database by filtering and screening for outliners with a mixture of cross tabulation and basic statistics. Errors are corrected and information is verified when needed.
12. The database is now *clean* and aggregated by the data manager.
13. The aggregated figures are checked by a 4 different experts for consistency. All inconsistencies are verified. After the approval of these 4 experts the database is send to the lead researcher for analysis.
14. The lead researcher filters the database for inconsistencies before he starts analysis.