UNIVERSITY OF
CAMBRIDGE
Cambridge Service Alliance

Data Quality Assessment: The Hybrid Approach

Philip Woodall, Alexander Borek, and Ajith Kumar Parlikad

This is a working paper

Why this paper might be of interest to Alliance Partners:

Various techniques have been proposed to enable organisations to assess the current quality level of their data. Unfortunately, organisations have many different requirements related to data quality (DQ) assessment because of domain and context differences. Due to the gamut of possible requirements, organisations may be forced to select an assessment technique which may not be wholly suitable for their requirements.

The approach presented in this paper can help your organisation to develop an assessment technique that is suitable for your requirements whilst leveraging the best practices proposed by existing assessment techniques.

August 2013

Find out more about the Cambridge Service Alliance:
Linkedin Group: Cambridge Service Alliance
www.cambridgeservicealliance.org

UNIVERSITY OF CAMBRIDGE
Cambridge Service Alliance

Data Quality Assessment: The Hybrid Approach

Philip Woodall[a], Alexander Borek[a], and Ajith Kumar Parlikad[a]
[a]University of Cambridge, Department of Engineering, Institute for Manufacturing, 17 Charles Babbage Road, Cambridge, CB3 0FS, UK.

Various techniques have been proposed to enable organisations to assess the current quality level of their data. Unfortunately, organisations have many different requirements related to data quality (DQ) assessment because of domain and context differences. Due to the gamut of possible requirements, organisations may be forced to select an assessment technique which may not be wholly suitable for their requirements. Therefore, we propose and evaluate the Hybrid Approach to assessing DQ which demonstrates that it is possible to develop new techniques for assessing DQ, suitable for any set of requirements, while leveraging the best practices proposed by existing ATs.

1. Introduction

The quality of an organisation's data is paramount to its success, and poor data quality (DQ) can have disastrous and even life-threatening consequences. The explosion of the Challenger space shuttle and the mistaken shooting down of an Iranian civilian aircraft are two high-profile examples where DQ was a causing factor [1]. While these may be extreme examples, many organisations rely on having good quality data to make decisions for their every-day operations. Furthermore, even organisations with advanced data management practices, that implement continuous improvement methodologies, find that employees have a greater need for high quality data [2].

To be assured that data is "fit for use"—where "fit for use" data is of required quality [3]—the first step is the DQ assessment (see Figure 1), the aim of which is to inspect data to determine the current level of DQ and the extent of any DQ deficiencies [4]. Many assessment techniques (ATs) have been proposed to support this endeavour, and these are typically part of a wider DQ methodology which also provide guidance on how to improve DQ (see for example, [5–9]). The focus of this paper is on DQ assessment and the associated ATs rather than the DQ methodology or DQ improvement. There are many methods which can be used as part of a DQ assessment such as interviewing, data modelling and gap analysis. The ATs support and guide the process of selection and combined usage of these methods to enable organisations to understand their current level of DQ.

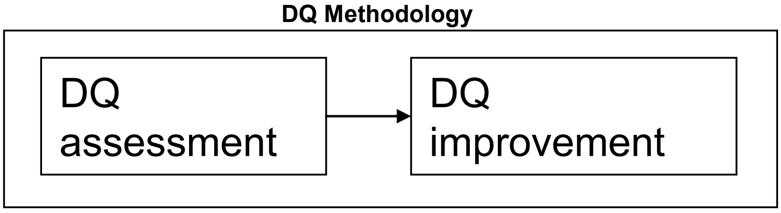

Figure 1: Relationship of DQ assessment and improvement

Unfortunately, there are many different requirements related to DQ assessment because of domain and context differences associated with different organisations. For example, a large financial institution with advanced data management practices will have different needs than a small utility provider with one or two information systems. With the gamut of possible requirements, organisations may be forced into selecting an existing AT which may not be wholly suitable for their given set of requirements; this could lead to the unnecessary execution of DQ related activities or the omission of essential activities as part of an assessment.

Prior to this research, a series of informal interviews were conducted to understand various organisational requirements related to existing and planned DQ assessments within the organisations. From these interviews, one maintenance, repair and overhaul (MRO) organisation indicated that they had three requirements: to determine the actual costs caused by low DQ, to model the way data is created and how it flows, and to gather existing data models. No current AT can meet these requirements because no single AT advises how to conduct all of these activities.

The Hybrid Approach is proposed to address this problem and the aim of this approach is to show how new ATs can be developed by combining the existing activities in order to meet all requirements of any organisation needing to assess DQ. The Hybrid Approach therefore avoids the problem of having to complete unnecessary activities, and also provides the ability to take activities from one AT and integrate them with activities from other assessment techniques to produce a fully customised AT.

For the Hybrid Approach, the existing ATs have been divided into their constituent activities. These activities were then analysed to understand the order in which they should be placed in and whether any activity is dependent on another activity. Finally, a four step process is described that shows how to develop a new AT based on the existing activities and their ordering and dependency constraints.

The terms data and information are used synonymously in this paper, and the rest of this paper is organised as follows: section 2 describes DQ assessment techniques and section 3 presents the methodology used to develop the Hybrid Approach. The selection of ATs, the extraction of activities from these ATs, and details of how the activities should be ordered and what activities are necessary for new ATs is described in section 4. Section 5 lists the steps required to develop a new AT by combining the activities, and section 6 presents the results of applying the approach within London Underground. Section 7 discusses the evaluation of the Hybrid Approach and, finally, section 8 presents the conclusions.

2. Data Quality Assessment Techniques

Data quality ATs form a core part of the Hybrid Approach and this work defines an AT to be a series of activities that are used to complete a DQ assessment. A DQ assessment is defined as a process for obtaining measurements of DQ to determine the current state of DQ. The current state can then be used to determine the level of DQ improvement required. In general, DQ measurements are obtained by determining values for different metrics; for example, counting the number of missing entries in a database. To determine the level of DQ improvement required, measurements can be compared to reference values, such as DQ requirements, which could state how many missing entries can be tolerated for the data to be ‘fit for purpose’. This definition of DQ assessment follows the unified terminology of the Data Quality Measurement Information Model (DQMIM) [4] where the idea of assessment is to make a judgment about DQ measurements (to determine the level of DQ improvement required). This is a common definition, although the exact terminology is not always used in a uniform way; for instance, [10] defines assessment as the “means to identify and document those areas with greatest need of improvement as well as provide a baseline against which further improvements can be measured”. In a review and classification of the ATs [11], measurement is defined as the process of obtaining values for DQ dimensions and assessment is when these values are compared to reference values to enable a diagnosis of quality. A multitude of DQ dimensions exist and they help to categorise DQ problems, examples include accuracy, completeness, consistency, timeliness [12]. Clearly, these definitions of DQ assessment capture the idea of measurements being just values and assessment being the application of judgment to these values to determine the level of DQ improvement required.

2.1. Configuring Assessment Techniques

DQ projects can be selected by considering constraints such as the value and cost of a project, and a method for this has been proposed previously [13]. A DQ assessment could be one of these projects and therefore the Hybrid Approach, which describes how to configure the assessment, operates from a less general perspective. The research on selecting DQ projects could be used at a more strategic level to complement the Hybrid Approach by showing what set of projects (or DQ assessments) need to be performed to address DQ on a larger scale throughout an organisation. Some of the existing ATs also suggest that the assessment should be configured, such as by planning what activities need to be done [6,7]. This planning only refers to the activities within a single AT and does not consider what other activities, from external ATs, need to be included. Furthermore, the existing research only acknowledges the fact that an AT should be configured rather than providing specific guidance on how to do it successfully.

3. Methodology

The Hybrid Approach was developed in six main stages, starting with the task of obtaining the existing ATs from the literature, then extracting the activities from these ATs, checking the extraction results and resolving any problems, and, finally, refining the approach based on feedback from a review (see the numbered boxes (tasks) in Figure 2).

For each of the six tasks, different methods were used (see dashed lines in Figure 2) to ensure that the Hybrid Approach satisfied the following quality goals from an existing evaluation framework [14]:

- Validity - All statements made by the approach are correct and relevant to the problem, for the set of statements that are worth testing as valid.

- Completeness - The approach contains all the statements about the domain that are relevant, for the set of statements that are worth trying to find.

- Comprehension - The approach is adequately understood by its target audience.

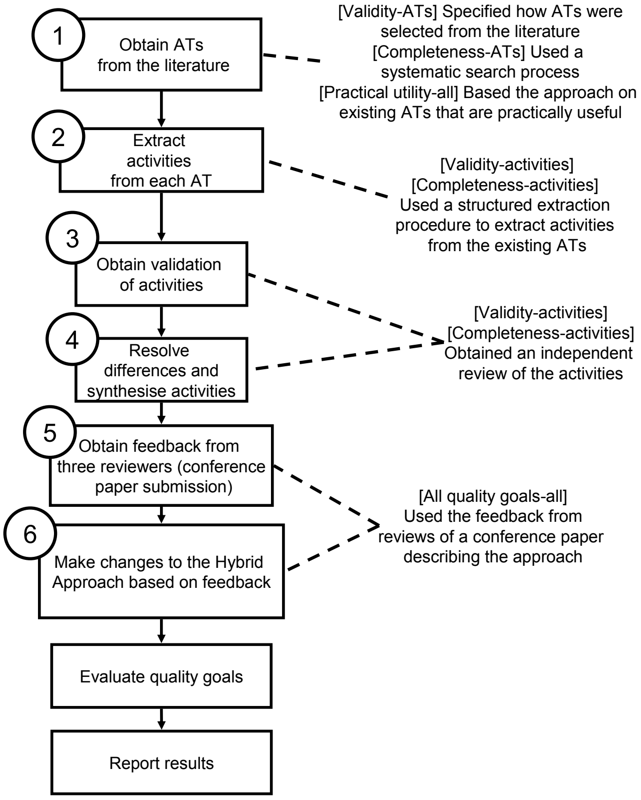
1
Obtain ATs from the literature
[Validity-ATs] Specified how ATs were selected from the literature
[Completeness-ATs] Used a systematic search process
[Practical utility-all] Based the approach on existing ATs that are practically useful
2
Extract activities from each AT
[Validity-activities]
[Completeness-activities]
Used a structured extraction procedure to extract activities from the existing ATs
3
Obtain validation of activities
4
Resolve differences and synthesise activities
[Validity-activities]
[Completeness-activities]
Obtained an independent review of the activities
5
Obtain feedback from three reviewers (conference paper submission)
6
Make changes to the Hybrid Approach based on feedback
[All quality goals-all]
Used the feedback from reviews of a conference paper describing the approach
Evaluate quality goals
Report results

Figure 2: The Hybrid Approach Development Process

- Understandability - As far as feasibly possible the approach is presented in an understandable format.

- Test coverage - The approach has been adequately tested in terms of feasible test coverage. Feasible test coverage means that there may be other relevant tests, but it is not worthwhile identifying and performing them.

- Practical utility - The utility of the approach is the extent to which it improves some aspect of performance for the target audience or provides non-trivial insights into the phenomenon being studied.

- Future resilience - The above quality goals remain stable or improve as the approach is used.

To obtain a more detailed evaluation of the most important components of the Hybrid Approach, the following two components were given special attention with regards to meeting the quality goals:

- The set of existing ATs used for the approach [ATs]
- The set of extracted activities from the ATs used in the approach [activities]

The reference to the above components and each of the quality goals is made within the square brackets in Figure 2; in cases where the entire approach is evaluated, this is referred to as ‘all’ in Figure 2.

The evaluation framework is well suited to evaluate the Hybrid Approach because it is useful for approaches that do not produce an output value that can be compared to an actual value. It is also useful in cases where the use of the approach and non-use of the approach are difficult to trial with all contextual factors remaining consistent. This is the case with the Hybrid Approach where it is very difficult to stabilise contextual factors when conducting multiple DQ assessments and no one value indicates the success of the approach over another.

An important aspect of the evaluation framework is that it specifies that methods introduced to achieve these goals should be separated from the means to assess the goals. Therefore, the framework was used both during the development of the Hybrid Approach (in an attempt to help achieve the goals—see Figure 2) and also to evaluate the approach after its development (to assess the extent to which the goals have been satisfied—see Section 7).

The original evaluation framework contains the 'syntactic correctness' goal (meaning that all statements in the model adhere to the syntax of the language used for the model). This goal was not measured for this work because the Hybrid Approach is not expressed using a formal language. Furthermore, the last quality goal (future resilience) was not specified in the original framework and has been added by this work because it was noticed at the end of the development of the Hybrid Approach that this goal needs to be considered. For this reason, this goal has only been evaluated (see section 7.4), but was not considered during the development.

The following subsections discuss the specific methods used for each task in Figure 2.

3.1. Selection of ATs

The development of the Hybrid Approach started with the task of obtaining the existing ATs from the literature (see Figure 2, task 1). The search was done systematically to minimise the chances of missing an AT (completeness quality goal), and the following filtering criteria were used to ensure that the ATs are valid and practically useful:

Studies were selected if:

1. the study contains an AT (according to the definition in Section 2) and describes what activities are involved
2. the study describes a DQ methodology and part of the methodology is an AT
3. the study contains an AT that has been subject to a rigorous review (as required by papers in high ranking journals or ATs described in peer reviewed books)
4. The study contains an AT that has been subject to an actual implementation and successful trial of the approach

Studies were rejected if:

5. the study does not describe an AT and the activities in sufficient detail to enable a DQ assessor to clearly and easily implement the activities
6. the study only describes DQ improvement and not an AT

The Scopus search engine, ACM digital library, Google books and proceedings of the International Conference on Information Quality were used to search systematically for studies (papers/reports/books etc.) which contain ATs. Special attention was given to ensure that the digital libraries covered the relevant journals (such as Information and Management and Communications of the ACM etc.). Moreover, the search continued for additional relevant studies by searching the references section of each study obtained.

3.2. Extraction of activities

Task 2 involved extracting the activities from each AT, and this was done using a structured extraction procedure to ensure that no activities would be missed and that each activity is valid (see Figure 2, task 2). Activities were extracted by reviewing each source describing an AT and recording the activities it describes in an 'extraction table'. Table 1 shows a sample part of an extraction table for two activities in the TQdM-a AT. The table records the name, description, general comments related to the activity, and the link to the study (where the activity is described in the original source).

AT reference	TQdM-a [6]		
Comments	none		
Activity	Definition of activity	Comments	link to study
…			
Model data creation and flow	The process of understanding and creating a model of the way data is created, updated, deleted and is transferred from one source to another.	none	Page 160, Step 3: "determine all business process and applications, and all who create or update a group of data"
Select a place where data is to be measured	Select the place where data is to be measured based on the objectives for measurement. This includes determining when and where to measure the data or specifying who will give subjective opinions.	none	Page 164, Step 4: "Select a place where data is to be measured"
…			

Table 1: A Sample of the Data Extraction Table

In addition to the structured extraction process, the activities were subject to an independent review process shown in Figure 3. This provided a secondary check that the activities are correct and relevant and to detect any activities that were missed from the extraction (see Figure 2, tasks 3 and 4).

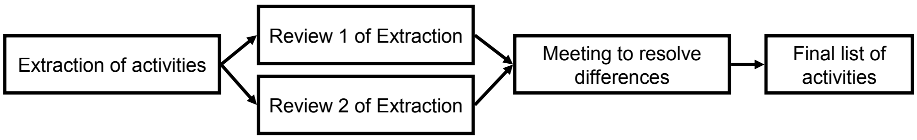

Figure 3: Process for Validating the Activities

3.3. Determining the Ordering of Activities

A conference paper (see [15]) describing the approach was peer reviewed by three reviewers as part of a conference submission procedure for task 5. The aim was to use the review process as a means to check all the quality goals for the entire approach; although it was not possible to ensure that the reviewers gave specific attention to all quality goals, useful feedback was obtained. In addition to the conference review process, general comments from attending the conference contributed to the quality goals. In particular, the need to provide guidance on how the activities should be ordered in a new AT was suggested as a means to improve the approach, and this was developed in task 6 (see Figure 2). The extraction procedure was extended to record, for each activity in each AT, the dependencies and any ordering constraints described by the AT. Furthermore, for the final evaluation of the Hybrid Approach, a series of DQ assessments were carried out as part of a trial in a UK-based organisation that manufactures car parts, and these assessments also provided information about the ordering of activities.

4. Results of the Selection of ATs and Extraction of Activities

The complete list ATs, as found by the systematic search process, are shown in Table 2, which lists the name (acronym) of the AT, its full name, and the study which proposed the AT. For the CDQM-a AT, two studies described the same AT and both of these were selected and used to extract the activities. Some studies described a complete DQ methodology (which includes DQ improvement) and not just the assessment stage. In these cases, "-a" is added to the end of the AT name to indicate that the DQ assessment is part of the full methodology; for example, TDQM is the full methodology and 'TDQM-a' is the DQ assessment part.

DQ Assessment Techniquename	Full name	Study
AIMQ	AIM Quality	[16]
CDQM-a	Complete Data Quality Methodology	[5] and [17]
COLDQ-a	Cost-effect Of Low Data Quality	[10]
DQA	Data Quality Assessment	[18]
EDQP-a	Executing Data Quality Projects	[7]
SODQA-a	Subjective-Objective Data Quality Assessment	[8]
TDQM-a	Total Data Quality Management	[9]
TQdM-a	Total Quality data Management	[6]

Table 2: DQ Assessment Techniques that meet the Selection Criteria

4.1. DQ Assessment Activities

The activities, extracted from the selected ATs, are shown in Table 3. The first column contains the name of the activity and an abbreviation of the activity in parentheses, a description of the activity is given in the second column, and the ATs that contain the activity are listed in the final column.

Activity + (abbreviation)	Definition of activity	Source AT(s)
Communicate and share the results (Com.)	Communicate and share the results or current progress of the DQ assessment with relevant people.	TQdM-a EDQP-a COLDQ-a
Conduct analysis of results (Analysis)	The process of analyzing the values from the DQ measurement(s).	All

Define DQ requirements (Reqs.)	The process of defining what level of DQ is required (for example, setting minimum thresholds that DQ must meet). DQ requirements can be compared to the measurement values to determine the level of DQ improvement required.	AIMQ TDQM-a COLDQ-a EDQP-a CDQM-a TQdM-a
Expose the DQ assessment project to senior management (Expose)	Expose and establish senior management support for the DQ assessment project.	COLDQ-a EDQP-a
Group/organize data items (Group)	The process of grouping data items into categories (for example, grouping criteria could include the type of data, level of risk etc.).	TQdM-a EDQP-a
Identify and prioritise the organisational problems (Probs.)	Based on what is known at the start of the assessment, list the specific problems focussing on problems that relate to DQ.	TQdM-a EDQP-a CDQM-a COLDQ-a
Identify DQ costs (Costs)	The process of determining the business impact and/or economic losses caused by low DQ (note that business impacts may not only be financial).	TQdM-a COLDQ-a CDQM-a EDQP-a
Identify DQ dimensions (Dims.)	The process of identifying dimensions or using an existing model of DQ dimensions e.g. PSP/IQ DQ.	AIMQ SODQA-a TQdM-a TDQM-a COLDQ-a EDQP-a CDQM-a (DQA)
Identify DQ metrics (Metrics)	The process of identifying, developing or using an existing set of DQ metrics.	All
Identify reference data (Ref. data)	The process of determining comparison data which can be used as input to the selected metrics. For example, one metric for measuring accuracy requires the stored value to be compared to the 'real' reference value; this process attempts to determine and document the 'real' value.	TQdM-a (AIMQ) (SODQA-a) (TDQM-a) (COLDQ-a) (EDQP-a) (DQA) (CDQM-a)
Model data creation and flow (Model)	The process of understanding and creating a model of the way data is created, updated, deleted and is transferred from one source to another.	TQdM-a TDQM-a COLDQ-a EDQP-a CDQM-a

Perform objective/subjective DQ measurement (Measure)	The process of obtaining DQ measurements from an actual data set or by obtaining (subjective) opinions of the current state of DQ.	All
Select a place where data is to be measured (Place)	Select the place where data is to be measured based on the objectives for measurement. This includes determining when and where to measure the data or specifying who will give subjective opinions.	TQdM-a EDQP-a DQA (AIMQ) (SODQA-a) (TDQM-a) (COLDQ-a) (CDQM-a)
Select data items (Data items)	The process of selecting the relevant data values, attributes, tables, information systems, paper files etc. which will be subject to the DQ assessment. This can also include the process of sampling the data to obtain the required data values.	TQdM-a TDQM-a EDQP-a DQA (AIMQ) (SODQA-a) (COLDQ-a) (CDQM-a)
Select processes (Process)	The process of selecting business processes that will be focused on in the assessment.	TQdM-a COLDQ-a CDQM-a
Gather general meta data (Meta)	The process of gathering relevant meta data such as data models.	DQA
Perform data profiling (Profile)	The process of examining the data and collecting statistics and information about that data such as distribution of values.	DQA
Validate the DQ metrics (Val. Metrics)	The process of checking that the DQ metrics and the implementation of DQ metrics are correct.	TDQM-a DQA

Table 3: Activities Associated with Existing ATs

As a result of the review process of the activities (see Figure 3), a total of 16 disagreements were found and resolved by either: removing an activity that was mistakenly added; adding an activity that was missed; or refining the name, description, and/or link to the study. Two of the key changes included: (1) changing the description for “Identify DQ dimensions” to account for the case where an existing model of dimensions is already available and can be used without having to select specific individual dimensions and (2) adding the “Analysis of results”, “Identify and prioritise the organisational problems”, and “Identify DQ costs” activities to TQdM-a, which were originally missed out erroneously.

Another observation concerning the activities is that one activity extracted from EDQP-a described planning the DQ assessment project (where the required activities from the EDQP-a AT are selected). This activity was removed from the extraction because if the Hybrid Approach is used, it becomes the planning activity and therefore describes how to select the relevant activities from not just EDQP-a, but from all the ATs in Table 1.

4.1.1. Problems with the Extraction Process

One problem with extraction procedure used to extract the activities from the ATs was identified: some sources that describe an AT deliberately present the set of activities at a high level of abstraction and therefore miss out some of the more "obvious" activities that must be performed. The extraction procedure alone therefore gives a false picture of the full set of activities in each AT. To address this problem, a further check of the activities was carried out by actively checking if any other activities should be attributed to the ATs that do not describe all activities. As a result of this process, the Dims., Data items, Place and Ref. data activities were found to be part of all ATs but are not described in some. As an example, for the Data items activity, Figure 2 in the paper describing the SODQA-a AT shows "Dataset in use" [8], which implies that the data set has been selected. However, the paper does not describe any activity related to selecting the dataset and hence this was not extracted as part of the original extraction procedure. The activities that have been attributed to an AT are shown with the AT in parentheses in Table 2.

4.2. Classification of Activities

The activities in Table 2 were classified as either 'recommended' or 'optional' based on whether they are included in every existing AT or not. Recommended activities are those that are recommended to be considered for inclusion in every new AT. According to this classification, the following are recommended activities:

1. Select data items
2. Select a place where data is to be measured
3. Identify reference data
4. Identify DQ dimensions
5. Identify DQ metrics
6. Perform measurement
7. Conduct analysis of the results

The measurement process (6) obtains values for the dimensions (4) and metrics (5) for a given set of data items (1) and is the heart of an AT. Furthermore, the values from the measurement are meaningless until some level of interpretation/analysis is applied (7). The place to measure the data (2) needs to be identified in order to know where to apply the metrics in the measurement process, and the reference data activity (3) is conditional depending on what is being measured. For example, reference data may be needed as input to a metric that measures accuracy—in this case, the "real" value (reference value) is needed as well as the recorded value. If the metrics do not require reference data, then it is not necessary to identify reference data.

4.3. Ordering and Dependencies between Activities

It is necessary to know what order the activities should be placed in so that newly developed ATs are usable and do not contain clearly un-implementable links between activities. In addition to ordering, it is also necessary to know, if an activity is added to an AT, what other activities also need to be added (because the first activity cannot be completed in isolation); this type of relationship is referred to as an activity dependency. Ordering indicates whether given any two activities that are present in an AT, which activity must be completed before the other. Dependencies indicate, given the inclusion of one activity in an AT, what other activities must be added to the AT. The dependency gives some indication of ordering because, in many cases, the reason that one activity needs another activity to be included in the AT is because the additional activity needs to be completed before the other activity. There are cases, however, where given an activity in an AT, no other activity needs to be included in the AT, but it is still necessary to determine the ordering between the activities; both concepts therefore need to be considered separately.

Figure 4 shows the resulting ordering and dependencies between all activities. In this figure, a black dot with an arrow signifies the starting position, the grey boxes indicate the recommended activities, the boxes with grey diagonal lines represent activities that can be placed in various positions subject to constraints (referred to as variable activities), arrows indicate necessary ordering between activities, and dashed arrows indicate dependencies between activities—for example, if the Group activity is included so must the Data items activity. In any part of the AT in Figure 4 where there are multiple paths, these can be done in parallel until the arrows converge on an activity. In which case, all previous activities must be completed before proceeding.

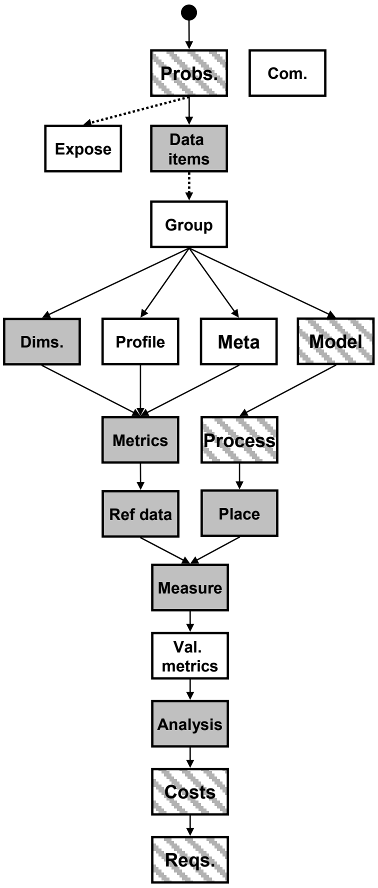

Figure 4: A Generic AT

The Com. activity has no links to other activities and can be done at any position in the AT to keep people informed of progress. In fact, EDQP-a suggests that the Com. activity can be done at various points in the AT to maintain a high level of communication about the assessment to external stakeholders. The Com. activity is therefore a type of variable activity with no constraints on where it may be placed.

Two of the trial assessments (conducted in a UK car parts manufacturer organisation) provided useful findings for this research. In particular, the Data items activity can be preceded by the Process activity and the Dims. activity can be preceded by the Probs. activity. Appendix A gives explanations for the ordering of these and all the other activities (excluding the variable activities, which are described in the next section).

4.3.1. Variable Activities

Five of the activities were found to have different possible positions in the AT depending on what is required from the combination of activities, and these are referred to as variable activities. The placement options are shown in tables 4 to 8—that is, the list of activities that can be placed before and after the main activity and a description is given for why the variable activities can be placed in the specified order. In the Tables, italics are used to show dependent activities. For example, Table 6 shows the three activities that Costs is dependent on, and at least one of these activities must precede the Costs activity. Note that the other activities (including the recommended activities) should always be used in the order shown in Figure 4.

Main activity	Probs.	
Activity	Position	Comments
Process	Before	The people that are associated with the selected processes can help with identifying the DQ problems.
none	Before	It is useful to start the assessment with an initial understanding of current organisational DQ problems.
Measure	After	It does not make sense to identify suspected DQ problems after a full measurement process has indicated the actual DQ problems.
Expose	After	One way to gain management support is to show how each identified DQ problem affects the organisation. (Probs. must be placed before Expose.)
Dims., Data items or Group	After	The dimensions, data items (or group of data items) can be selected for the assessment that are relevant to the DQ problems identified.
Costs*, Reqs.*, Model*	After	*see Probs. in the Costs, Reqs. and Model tables.

Table 4: Placement Options for the Probs. Activity

Main activity	Reqs.	
Activity	Position	Comments
Probs.,	Before	The requirements should be relevant to the organisational DQ problems.
Data items or Group, Process	Before	The requirements should be specified for the selected data items (or group of data items) and processes where, for example, the people associated with the processes can help specify the requirements.
Measure	Before	DQ requirements can be specified for each DQ problem identified from the measurement.
Costs	Before	An understanding of costs can be used to set realistic DQ requirements that are feasible to achieve.
Model*	After	* see Reqs. in the Model table

Table 5: Placement Options for the Reqs. Activity

Main activity	Costs	
Activity	Position	Comments
Probs.,	Before	Organisational problems have an associated financial loss and therefore should be identified before identifying DQ costs. (Either Probs., Measure or Analysis is necessary for Costs, i.e. must be placed before)
Measure or Analysis,	Before	The financial impact of the actual DQ problems found in the measurement process (or subsequent analysis) can be determined. (Either Probs., Measure or Analysis is necessary for Costs, i.e. must be placed before)
Model	Before	The model can be used to identify particular transactions/processes that are affected by the DQ problems and the financial impact of the failed transaction/process.
Data items	After	It is possible to drive the selection of data items by first identifying financial losses caused by DQ and then selecting the data items that are relevant to these losses.
Reqs.*	After	*see Costs in the Reqs. table
Expose	After	Demonstrating the financial impact of the DQ problems is a good way to gain management support.

Table 6: Placement Options for the Costs Activity

Main activity	Model	
Activity	Position	Comments
Data items or Group	Before	The model can focus on the selected data items (or group of data items) rather than all possible data items in the organisation to reduce the scope of the modelling.
Reqs.	Before	If the intent is to add the DQ requirements to the model, then the requirements need to have been gathered before modelling.
none	Before	An assessment project that is not overly time-constrained can conduct the modelling activity first and use it to identify likely areas containing DQ problems. Modelling first, without any information to reduce the scope, can be very time-consuming.
Probs.	Before	The modelling task can be limited to the areas that relate to the DQ problems previously identified. Modelling is therefore more focussed and faster to complete compared to attempting to model all information flows.
Measure	After	The model is used to identify areas that need to be subject to DQ measurements and therefore needs to be conducted before the measurement process.
Process*	After	*see Model in the Process Table
Place	After	The model can be used to identify relevant places in the flow of data that can be used to obtain the measurements.

Table 7: Placement Options for the Model Activity

Main activity	Process

Activity	Position	Comments
Data items or Group	Before	Processes that use and require the previously identified data items (or group of data items) can be selected.
Model	Before	The model can be used to provide information on where data is likely to be poor quality and the processes related to these areas can be selected.
Data items or Group	After	Data items can be selected that are used in the selected processes.

Table 8: Placement Options for the Process Activity

Figure 4 shows an AT containing all the activities and represents one configuration of the variable activities that satisfies all the constraints. The Probs. activity can be used first because eight activities are recommended to be done after this activity (see Table 4) and can benefit from an initial identification of DQ problems. The modelling of data creation and flow (Model activity) is done only for the set of data items identified by the Data items and Group activities, which, in turn, are relevant to the identified DQ problems (see Table 7). The Process activity uses the model to select the processes that will be the focus of the assessment (see Table 8); note that the processes use and are relevant to the data items from the Data items and Group activities because the model has been developed for these data items. The Place activity uses the model to identify relevant places in the flow of data that can be used to obtain the measurements (see Table 7). At the end of the AT, the costs due to poor DQ are determined for the actual DQ problems identified by the measurements and the analysis of the measurements. Note that the Costs activity must precede at least one of the Probs. Measure or Analysis activities (see Table 6). In addition, the model is also useful with regard to helping to identify the costs of any failed transactions/processes (see Table 6). Finally, DQ requirements can be specified and the understanding of costs can be used to set realistic DQ requirements that are feasible to achieve (see Table 5).

5. Steps to Develop a DQ Assessment Technique

A simple four step procedure is proposed that shows how to use the results from the previous section to develop a new AT that is suitable for specific organisational requirements.

5.1. Step 1: Determine the aim of the assessment

The aim drives the assessment process and is essential to inform DQ assessors of what the resulting AT should be used for. The aim will vary depending on what the organisation intends the assessment to achieve. Example aims include:

- To measure a particular DQ problem which has been identified previously,
- To determine and prioritize an organisation’s DQ problems and obtain measurements for each problem.

Continuing the case of the MRO organisation described in the introduction, the organisation intends to perform a DQ assessment in the asset management (AM) part of the organisation (which is responsible for managing the equipment throughout the organisation), and the aim of the assessment is: to identify what financial effect DQ is having on the AM part of the organisation and to identify why people do not want to use the data in the main AM information system.

5.2. Step 2: Identify the company requirements related to the DQ assessment

Different companies will have different requirements which relate to the DQ assessment. This step requires the organisation wanting to assess DQ to identify these requirements related to the DQ assessment. To ensure the relevance of the requirements, it is useful to check that each requirement follows from the aim (step 1) and therefore contributes to achieving this aim.

The MRO organisation indicated that they have four requirements:

1. determine the actual costs caused by low DQ,
2. obtain an initial estimate of costs to justify the resources for the assessment,
3. model the way data is created and how it flows,
4. gather existing data models.

Managers in the organisation recognised that poor DQ was one possible source of financial loss and therefore they wanted to identify if and how much poor DQ was costing the business. Furthermore, having never completed a DQ assessment before, the managers want an initial estimation of costs early in the assessment in order justify the resources being put into the assessment process (they also want a more detailed estimation of costs later). The reason for the third requirement is that the organisation has lost track of where data originates and ends up and the documentation of this is out of date. The existing documentation of the data models is also out of date or missing entirely, hence the need for the fourth requirement.

5.3. Step 3: Select AT activities which meet organisational requirements

The aim of this step is to select the relevant activities, from the list in Table 2, which meet the organisational requirements related to the DQ assessment. Some requirements may not be applicable to activities, and, in this case, the remaining requirements should be considered when configuring the activities in the next step.

Requirement	Suitable AT activity
Determine the actual costs caused by low DQ	Identify DQ costs
Obtain an initial estimate of costs to justify the resources for the assessment	Expose the DQ assessment project to senior management and Identify DQ costs
Model the way data is created and how it flows	Model data creation and flow
Gather existing data models	Gather general meta data

Table 9: Mapping of Requirements for the MRO organisation

The requirements from the MRO organisation are shown in the ‘requirements’ column of Table 9 and the mapping of these requirements to activities is shown in the ‘suitable AT activity’ column. The first, third and fourth requirements map easily to activities, whereas the second requirement maps to both the Expose and Costs activities and also specifies that costs need to be determined early in the assessment program. This last part of the second requirement, which does not map to an activity, is used to configure the positions of the activities in the AT in the following step.

5.4. Step 4: Configure the activities in the AT

The aim of this step is to arrange the activities (including the required activities from step 3) into a sensible order and include any activity dependencies. The ‘recommended’ activities should be strongly considered to be included and only removed if there is a sound reason for not needing to perform them. To help with this step it is useful to start with the generic AT in Figure 4 and remove the activities that are not needed and then move any remaining variable activities to the desired positions. Figure 5 shows the final AT suitable for the MRO organisation's requirements; activities surrounded by a thicker border are the required activities from Table 9.

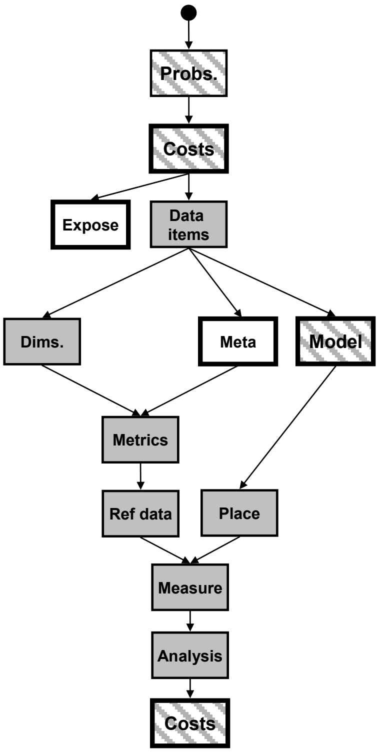

Figure 5: A New AT for the MRO Organisation

In Figure 5, the Costs activitiy has been added twice. The first instance of this activity will be used to obtain an initial estimate of the financial impact of poor DQ and has therefore been placed second in the AT preceded only by the Probs. activity, which must be included before Costs (see Table 6). The second instance of Costs (at the end of the AT) will be used to obtain the actual cost of the DQ problems and is preceded by the Measure and Analysis activities so that the costs can be determined for the identified DQ problems. The Expose activity follows the Probs. and Costs activity because it is necessary to identify initial DQ problems before conducting Expose. Furthermore, the Costs activity can also be a useful input to the Expose activity, and this ordering also satisfies the second requirement of the MRO organisation: the management want to see some initial estimate of the costs of poor DQ. Finally, the position of the other activities has been retained from the AT shown in Figure 4 because there is no need to reorder these. Figure 5 therefore presents an AT that is a perfect fit for the MRO organisation's requirements and shows how the activities from multiple ATs can be combined in a 'best practice' order.

6. Applying the Hybrid Approach within London Underground

In order to evaluate whether the Hybrid Approach is of any practical utility, the approach was trialled within London Underground Limited (LUL). A new AT was developed and carried out to assess the current state of DQ in the signalling, control and information asset group of LUL. The data used by this group includes maintenance data (such as what infrastructure/equipment etc. has and needs to be replaced) about the train lines operated by LUL and is, therefore, critical for the safety of the passengers.

The manager of this data is keen to ensure that it is of the highest quality and wants to detect any possible DQ problems. The aim of the assessment was, therefore, to determine what the existing DQ problems are in the signalling, control and information asset group of LUL and obtain measures for these problems.

Requirement	Suitable AT activity
Determine what information systems and specific data to focus on (within in the signalling, control and information asset group)	Select data items
Group the data into categories according to use.	Group/organize data items
Inform the relevant people of the results after the data has been measured	Communicate and share the results

Table 10: Requirements for LUL

There are many information systems in use in this group of LUL and therefore one requirement for the assessment was to select the data/systems that would be the focus of the assessment (see the first requirement in Table 10). Furthermore, this data is used by many people for different purposes; thus, the second requirement was to group the data into categories of use so that the final results would be meaningful to the different user groups. Finally, the last requirement was to disseminate the results to the various groups of people who use the data after the data has been measured for levels of quality. The managers of each group could then decide whether any necessary future action is required to improve the data.

The three matching activities for these requirements are Select data items, Group/organize data items, and Communicate and share the results (shown in Table 10). The last requirement also specifies that communicating the results should occur at the end of the assessment when the final results have been obtained. This requirement helps in Step 4 of the Hybrid Approach to indicate that the Com. activity needs to be placed at the end of the AT. The resulting AT, which matches the requirements of LUL, is shown in Figure 6. Note that none of the variable activities were included in the AT because they are not needed to meet the requirements of LUL.

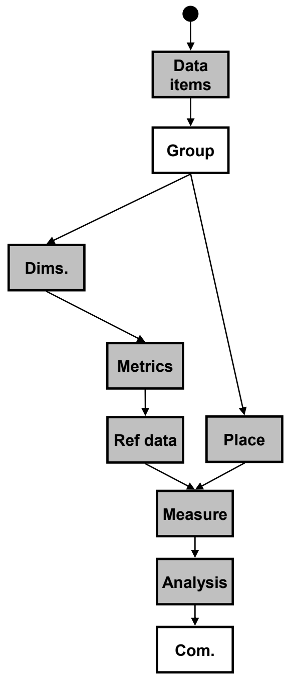

Figure 6: AT for London Underground

The assessment started with the selection of data items because the manager at LUL wanted to identify if there are any problems with the data in the signalling, control and information asset group. As mentioned before, many systems are used in this group and so this activity firstly identified one of the larger information systems and secondly selected a set of data items from within this system. The data items all contained condition-related data about one of the underground train lines. As part of the grouping activity, the data items were identified as being used by either maintenance or finance or both. Over 50 thousand rows (instances of physical assets/equipment) were extracted from the information system into a spreadsheet (the spreadsheet having been identified as the place from which to measure the data). This provided a snapshot of the data from which to check the quality without any chance of affecting the data in the live system; this is a very simple version of a staging area [18]. From the data, two dimensions were identified as being necessary to check: completeness and conformance to business rules—see [6]. These dimensions were then used in the Metrics activity to guide the construction of the metrics. No reference data was needed for the metrics (both completeness and conformance to business rules could be measured without needing additional data) and so the Ref. data activity was not carried out. Many of the metrics for business rule conformance were of the following form: "number of violations of the business rule" / "total number values inspected" which gave an indication of the proportion of errors. The metrics for completeness were similar except that they checked for missing values rather than "number of violations of the business rule".

The metrics were coded in software and executed on the spreadsheet as part of the measurement activity, and the results were analysed and documented in a final report that was sent to LUL. The analysis included grouping the results according to the users of the data items (identified in the grouping activity). Finally, this report was distributed to the relevant people in LUL as part of the communicate the results activity.

Clearly stating the aim of the AT (see Step 1 of the Hybrid Approach) was essential as it was used in the documentation of the final assessment report. This informed the readers of the report exactly what the assessment had focussed on and what it aimed to achieve. The report is confidential to LUL and therefore no details of the results are presented.

6.1. Changes to the Hybrid Approach after the Trial

Although the main aim of the trial with LUL was to evaluate the practical utility of the approach, one additional finding about the Ref. data activity was included in the Hybrid Approach. Originally, the only information available about the ordering of the Ref. data activity came from the TQdM-a AT: the reference data must be associated with the particular data items or group of data items (see Appendix A) and should therefore come after the Data items and Group activities. However, in addition, the trial demonstrated that the metrics indicate whether reference data is needed. For instance, a metric that measures the difference between a real-world value and a stored value (the accuracy dimension) needs reference data (the real-world values) as input. Whether or not reference data is needed is therefore known after defining the metrics. This finding was therefore included in the arrangement of activities shown in Figure 4 and added to the descriptions in Appendix A.

7. Discussion of the Quality Goals

The following subsections present a discussion of the extent to which the quality goals set out in the introduction have been met by the Hybrid Approach.

7.1. The Practical Utility of the Approach

The result of the assessment has provided LUL with a current state assessment of the level of DQ and an understanding of what DQ problems exist in one of their main maintenance systems. This meets the main aim of the assessment, as specified by LUL. With the Hybrid Approach, this was achievable without having to spend time conducting many of the activities that are specified by existing ATs and were not needed to achieve the aim of the assessment. In this respect alone, the Hybrid Approach is therefore a useful mechanism for organisations to ensure that DQ assessments are focussed only on their aims and that cost, time and resources are saved by not conducting unnecessary activities.

In addition to the practical utility goal, the development and the implementation of the AT with LUL covers two more of the quality goals by demonstrating that the activities within the LUL AT are valid (correct and relevant to the problem) and the actual trial of the approach in this real scenario forms part of the test coverage goal.

7.2. The Validity and Completeness of the List of ATs and Activities

An additional literature search was carried out by an independent researcher to determine the extent to which the list of ATs used in the Hybrid Approach is valid (by checking whether the existing ATs would be selected again) and comprehensive. Furthermore, promising discoveries of new papers/books etc. found by the authors that related to DQ assessment were also checked for ATs. Although these searches were not systematic and tended to be sporadic, they did provide the opportunity to find a new AT because they were conducted over a longer period of time.

The results of both of these searches found two new studies which could possibly contain ATs: [19,20]. Both of these were published after the main literature review and therefore could not have been found initially, confirming that the initial review was feasibly complete.

One of the new studies [20] did contain a new AT and this was used to confirm the validity and completeness of the existing set of activities. To do this, an extraction procedure was performed that checked for existing and new activities in the new AT. The following activities were confirmed by the extraction: Process, Data items, Costs, Meta, Model, Dims, Metrics, Measure, Analyse and Com. The Ref. data activity could only be inferred to be present because it was not explicitly described; the only reference to this was regarding the accuracy dimension where the study describes, in order to check for accuracy, a "comparison to a system of record" is needed [20]. No new activities were found confirming the completeness of the activities.

During the extraction, notes were made regarding the ordering of activities, and this study states that the Model activity is useful as an input to the Place activity because the documentation of the information flows can be used to identify the best place to inspect the data—this confirms the original finding (see the Place activity in Appendix A).

7.3. The Understandability and Comprehension of the Hybrid Approach

In addition to the London Underground trial of the Hybrid Approach, a series of smaller DQ assessments were carried out in a UK-based organisation that manufactures car parts. These were carried out by an independent data assessor (a student at the Institute for Manufacturing in Cambridge) to evaluate the understandability and comprehension quality goals. The aim was to determine whether, using the existing documentation of the approach (including the previous conference paper, list of steps and activities, existing studies describing the ATs, etc.) the assessor could follow the entire approach and carry out an assessment without assistance from the developers of the approach. The assessor had limited experience of DQ and therefore prior reading about DQ assessment, and DQ research in general, was done before attempting to understand and carry out the Hybrid Approach. During the assessments, which were developed and conducted on site at the organisation, the first author (PW) maintained telephone contact with the assessor to monitor the progress and help with any problems encountered.

The results indicate that the assessor had no problems understanding the approach and was able to produce and conduct a number of useful DQ assessments in different parts of the organisation. In fact, these assessments were able to provide feedback that could be incorporated in the Hybrid Approach with regard to the ordering of activities; these were described in section 4.3 and are labelled as "trial" in the table in Appendix A.

7.4. Future Resilience of the Approach

The Hybrid Approach incorporates ATs developed from 1998 to 2008 and all of these ATs advocate using very similar approaches to assess DQ. No one AT differs drastically from the rest in its approach. The assessment process is therefore fundamentally very similar between all of these ATs despite the ten year time span. Furthermore, the attempt to validate the validity and completeness of the activities and ATs (see section 7.2) confirmed most of the existing activities and did not find any new activities. Currently, therefore there is no work indicating that these approaches are in any way outdated, and since the Hybrid Approach is based so heavily on these, this provides an assurance that it will also be resilient into the future. However, new activities are likely to be presented in the future, and for the Hybrid Approach to remain current it must demonstrate that it can accommodate these along with any new evidence related to the inputs and outputs of the activities. For the latter, new evidence concerning the ordering has already been incorporated from the various trials of the Hybrid Approach. The Hybrid Approach is also open to the inclusion of new activities by simply including them in the current list of activities. The only stipulation is that, for the activity to be useful, the ordering and dependencies should be specified in the same way as for the existing activities.

8. Conclusion

Data is a critical asset in today’s organisations, and problems with the quality of this data can have catastrophic and even life-threatening consequences—especially in the case of the London Underground, where data about the maintenance of the underground train lines is used to make decisions about when to maintain the various assets and equipment that transport people continuously around London.

The first step towards high quality data for any organisation is the DQ assessment. This can provide an indication of the current level of DQ in the organisation and is the basis for initiating actions to improve DQ to desirable levels. Currently, no individual existing AT is wholly suitable to assess DQ for all types of requirements due to the varying nature of organisational requirements. The requirements may be different for every organisation and even the same organisation over time due to factors such as changes in the information systems used. The proposed Hybrid Approach shows how to develop new ATs by combining the activities from existing techniques in a way that meets differing requirements whilst still retaining the best practice concepts and ideas incorporated in the existing ATs. It also shows what activities can be omitted and carried out in parallel, even when activities have been combined from different existing ATs and no AT describes all the activities and how they should be combined. For the DQ assessment, this affords savings in costs, time and resources which organisations are constantly striving to contain.

Acknowledgement

We would like to thank EPSRC (grant EP/G038171/1) for supporting this research. We also thank London Underground for their assistance with the DQ assessment.

References

[1] C. Fisher, B. Kingma, Criticality of Data Quality as Exemplified in Two Disasters, Information & Management. 39 (2001) 109-116.

[2] J.M. Pearson, C.S. McCahon, R.T. Hightower, Total Quality Management. Are information systems managers ready?, Information & Management. 29 (1995) 251-263.

[3] R.Y. Wang, D.M. Strong, Beyond Accuracy: What Data Quality Means to Data Consumers, Journal of Management Information Systems. 12 (1996) 5-34.

[4] I. Caballero, E. Verbo, C. Calero, M. Piattini, MMPRO: A Methodology Based on ISO/IEC 15939 to Draw Up Data Quality Measurement Processes, in: The 13th International Conference on Information Quality, 2008.

[5] C. Batini, M. Scannapieco, Data Quality: Concepts, Methodologies and Techniques, 1st ed., Springer, 2006.

[6] L. English, Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits, John Wiley & Sons, 1999.

[7] D. McGilvray, Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information, Morgan Kaufmann, 2008.

[8] L.L. Pipino, Y.W. Lee, R.Y. Wang, Data Quality Assessment, Communications of the ACM. 45 (2002) 211-218.

[9] R.Y. Wang, A Product Perspective on Total Data Quality Management, Communications of the ACM. 41 (1998) 58-65.

[10] D. Loshin, Enterprise Knowledge Management: The Data Quality Approach, Morgan Kaufmann Pub, 2001.

[11] C. Batini, C. Cappiello, C. Francalanci, A. Maurino, Methodologies for Data Quality Assessment and Improvement, ACM Computing Surveys. 41 (2009) 1-52.

[12] D. Ballou, H. Pazer, Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems, Management Science. 31 (1985) 150–162.

[13] D. Ballou, G. Tayi, Enhancing Data Quality in Data Warehouse Environments, Communications of the ACM. 42 (1999) 73-78.

[14] B. Kitchenham, S. Linkman, S. Linkman, Experiences of Using an Evaluation Framework, Information and Software Technology. 47 (2005) 761-774.

[15] P. Woodall, A. Parlikad, A Hybrid Approach to Assessing Data Quality, in: Proceedings of the 2010 International Conference on Information Quality, 2010.

[16] Y.W. Lee, D.M. Strong, B.K. Kahn, R.Y. Wang, AIMQ: A Methodology for Information Quality Assessment, Information & Management. 40 (2002) 133–146.

[17] C. Batini, F. Cabitza, C. Cappiello, C. Francalanci, A Comprehensive Data Quality Methodology for Web and Structured Data, Int. J. Innov. Comput. Appl. 1 (2008) 205–218.

[18] A. Maydanchik, Data Quality Assessment, Technics Publications LLC, 2007.

[19] L. English, Information Quality Applied: Best Practices for Improving Business Information, Processes and Systems, John Wiley & Sons, 2009.

[20] D. Loshin, The Practitioner's Guide to Data Quality Improvement, Morgan Kaufmann, 2011.

9. Appendix A: Explanations for the order of activities

Main Activity	Activities which also need to be placed in the AT before the Main Activity	Reason
Com.	-	This can be done at any point in the AT to keep people informed of progress
Analysis	Measure	It is only possible to analyse results after the results have been obtained from performing the measurements
	Measure and Reqs.	The DQ requirements can be compared to the values from the measurement in the analysis stage.
Group	Data items	To determine related data items and groupings, the initial set of data items being used for the assessment is needed
Dims.	Probs. (trial)	The dimensions can be selected which are relevant to the DQ problems identified.
	Data items, or Group	The dimensions can be selected that are relevant to the selected set of data items (or group of data items).
Metrics	Dims.	The metrics are developed from each selected dimension because the dimensions define what needs to be measured.
	Profile	Results from data profiling can indicate the types, ranges and distribution of data values, which can help with developing metrics that need to inspect these values
	Meta	Meta data can be used to help develop the metrics
Ref. data	Data items or Group	The reference data must be associated with the particular data items or group of data items that need to be checked
	Metrics	The metrics show whether reference data is needed as input to the metrics. (This was a finding from the LUL study)
Measure	Metrics	The measurement is carried out for each identified metric.
	Ref. data	The reference data may be needed as input to metrics that are being measured.
	Process	The measurement could focus on obtaining measurements for data items in the selected process(es).
Data items	Costs	The financial impact of DQ for different sets of data needs to be known before selecting data items if the assessment focuses on reducing financial losses caused by DQ problems
	none	This can be the first activity
	Probs.	Data items that are relevant to the organisational problems can be selected.
	Process (trial)	Data items can be selected that are used in the selected processes.
Process	Model	The model can be used to provide information on where data is likely to be poor quality and the processes related to these areas can be selected

	Group or Data items	Processes that use and require the previously identified data items should be selected. If the data items have been grouped, then the Group activity provides the input rather than the Data items activity.
Meta	Data items or Group	Knowing which data items (or group of data items) are being assessed will help reduce the scope of gathering meta data (only meta data about the relevant data items needs to be gathered).
Profile	Data items or Group	Profiling is done for a specific set of data items (or group of data items)
Val. metrics	Measure	To validate the metrics e.g. check for false positives and false negatives etc. it is necessary to have first performed a measurement and review the results of the application of the metrics.
Expose	Probs.	The best way to gain management support is to show how each identified problem affects the organisation.
	Costs	Demonstrating the financial impact of the DQ problems is a good way to gain management support.
Place	Data items or Group	It is necessary to know what data items (or group of data items) need to be measured before determining where they will be measured.
	Model (confirmed in AT used for evaluation)	The model can be used to identify relevant places in the flow of data that can be used to obtain the measurements.