A HYBRID APPROACH TO ASSESSING DATA QUALITY

(Research paper)

Philip Woodall
University of Cambridge
phil.woodall@eng.cam.ac.uk

Ajith Kumar Parlikad
University of Cambridge
ajith.parlikad@eng.cam.ac.uk

Abstract: Various techniques have been proposed to enable organizations to initiate procedures to assess and ultimately to improve the quality of their data. The utility of these assessment techniques (ATs) has been demonstrated in different organizational contexts. However, while some of the ATs are geared towards specific application areas and are often not suitable in different applications, others are more general and therefore do not always meet specific requirements. To address this problem we propose the Hybrid Approach to assessing data quality, which can generate usable ATs for specific requirements using the activities of existing ATs. A literature review and bottom-up analysis of the existing data quality (DQ) ATs was used to identify the different activities proposed by each AT. Based on example requirements from an asset management organization, the activities were combined using the Hybrid Approach in order to generate an AT which can be followed to assess an existing DQ problem. The Hybrid Approach demonstrates that it is possible to develop new ways of assessing DQ which leverage the best practices proposed by existing ATs by combining the activities dynamically.

Key Words: Data Quality, Information Quality, Assessing Information Quality, Information Quality Assessment, Data Quality Assessment, Hybrid Approach, Assessment Techniques.

1. INTRODUCTION

Assessing data quality (DQ) is an essential phase leading to DQ improvement. The aim of DQ assessment is to inspect data to determine the current level of DQ and the extent of any DQ deficiencies [5]. Many assessment techniques (ATs) have been proposed to support this endeavor, and these are typically part of a wider DQ methodology which also provides guidance on how to improve DQ (see for example, [3],[7],[14],[17],[20-22]). However, the focus of this paper is on only DQ assessment and the associated ATs. There are many methods which can be used as part of a DQ assessment such as interviewing, data modeling and gap analysis. The ATs support and guide the process of selection and combined usage of these methods to understand the current level of DQ.

Unfortunately, there are many different requirements related to DQ assessment because of domain and context differences associated with the data assessor. For example, a large financial institution with a high level of maturity with respect to data management and quality processes will have different needs than a small utility provider with a low level of maturity. Moreover, organizational information systems often contain different types of data (structured, semi-structured or unstructured) and therefore this imposes further constraints on selecting an AT because some ATs are focused towards a particular information system and a particular type of data [4]. With the gamut of possible requirements, data

assessors may be forced into selecting an existing AT which may not be wholly suitable for their given set of requirements.

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 activities suggested by existing ATs in order to meet all requirements of any person or organization needing to assess DQ. A literature review and bottom-up analysis of the existing ATs was used to identify the different activities proposed by each AT, and the Hybrid Approach comprises six steps which show how to combine these activities. We also suggest a way to document ATs so that data assessors can quickly determine whether a particular AT is suitable to assess DQ given their specific requirements.

The Hybrid Approach described in this paper shows how to develop ATs that cover only the DQ assessment part of a wider DQ methodology, which typically consists of performing a DQ assessment and then initiating a DQ improvement using the results of the assessment [4] (see Figure 1). Our experience working with organizations in the UK indicates that some organizations know what their DQ problems are, and, hence, require a method to assess and mitigate them, others need to determine what DQ problems are present within their data and which set of problems is a priority. The Hybrid Approach therefore starts with the general input of an 'initial motivation' which includes both of these example starting points while also catering for other initial motivations for conducting a DQ assessment.

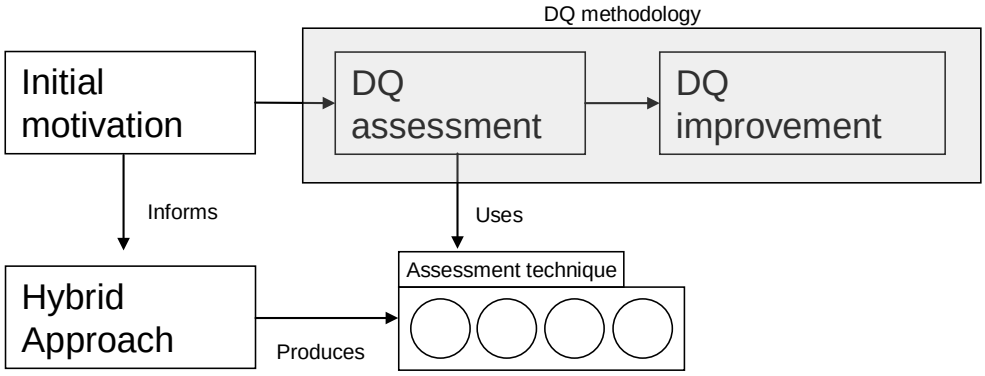

Figure 1: The Hybrid Approach integrated with DQ assessment and improvement

The terms data and information are used synonymously in this paper, and the rest of this paper is organized as follows: Section 2 describes the definitions used in this paper, a review of existing work on ATs and the breakdown of ATs into activities. Section 3 introduces the Hybrid Approach and describes how to use the approach to develop new ATs. Section 4 shows an example of the development of an AT using the Hybrid Approach based on the specific requirements from an aircraft maintenance, repair and overhaul (MRO) organization. Finally, Section 5 describes the limitations of the approach, and Section 6 presents conclusions and areas of future work arising from the Hybrid Approach.

2. DATA QUALITY ASSESSMENT TECHNIQUES

This work defines an AT to be *a technique for obtaining measurements of DQ and using these measurements 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) [5] 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, [17] 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 [4], 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. Clearly, these definitions 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.

DQ Assessment Technique name	Number of activities	Estimated time to implement (short, medium, long)	Type of data (structured, semi-structured, unstructured)			Information systems (monolithic, distributed, data warehouse, cooperative, web, paper-based)					
			s	**semi**	**un**	**M**	**D**	**DW**	**CO**	**W**	**P**
AIMQ (AIM Quality) [15]	*5*	*short*	*h*	*i*	*i*	*f*	*i*				
AMEQ-a (Activity-based Measuring and Evaluating of PIQ) [21]	*8*	*medium*	*h*	*i*		*f*					
CDQ-a (Comprehensive Data Quality Methodology) [3]	*8*	*medium*	*h*	*h*		*f*	*f*		*sf*		
COLDQ-a (Cost-effect Of Low Data Quality) [17]	*8*	*medium*	*h*	*i*		*f*					
DQA-a (Data Quality Assessment) [20]	*5*	*short*	*h*			*f*	*i*				
DWQ-a (Foundations of Data Warehouse Quality) [14]	*4*	*short*	*h*					*sf*			
IQM (Information Quality Measurement) [8]	*7*	*medium*	*h*	*h*						*sf*	
MMPRO (Methodology to draw up Data Quality Measurement Processes) [5]	*11*	*long*	*h*	*h*		*f*					
ORME-DQ (ORME Data Quality) [2]	*9*	*medium*	*h*	*i*		*f*					
PM (Prediction Markets) [19]	*3*	*short*	*h*	*i*	*h*	*i*					*f*
QAFD (Quality Assessment of Financial Data) [6]	*8*	*medium*	*h*			*f*					
TDQM-a (Total Data Quality Management) [22]	*8*	*medium*	*h*	*h*		*f*	*i*				
TQdM-a (Total Quality data Management) [7]	*12*	*long*	*h*	*i*		*f*	*f*				
UDQA (Utility-driven Quality Assessment) [9]	*6*	*short*	*h*			*f*					
sf = strongly focused *f* = focused *h* = can handle *i* = implicitly considered											

Table 1: A classification of DQ Assessment Techniques (extended from [4])

The ATs which were broken down into activities for the Hybrid Approach were extracted from the literature if they adhered to the above definition of an AT. Some ATs may not explicitly state how to determine the level of improvement required, so only the first part of the definition (*a technique for obtaining measurements of DQ*) was used to understand if a study contained an AT. The Scopus search engine, ACM digital library and proceedings of the International Conference on Information Quality were used to search for studies (papers/reports/books etc.) which contain ATs. The search continued for additional relevant studies by searching the references section of each study obtained. The following inclusion and exclusion criteria were used to select studies containing ATs:

Studies were selected if:

- the study contains an AT and describes what activities are involved.
- the study describes a DQ methodology and part of the methodology is an AT.

Studies were rejected if:

- the study does not describe an AT in sufficient detail
- the study only describes DQ improvement and not an AT.

In the case where multiple studies described the same AT, the study which described the AT in the most detail is cited. Furthermore, some studies described a complete DQ methodology and not just the assessment stage. In this case, only the activities related to the assessment stage were extracted from the studies—Total Data Quality Management (TDQM) is one example of a full methodology and therefore only the activities related to the DQ assessment were extracted, and this is referred to as 'TDQM-a' to indicate the DQ assessment part of the methodology. This convention is used for all of the ATs which are part of a full DQ methodology.

The resulting ATs are shown in the first column of Table 1, which lists the name (acronym) of the AT and the reference which proposed the AT. The additional data in this table is key for the selection of suitable ATs in the Hybrid Approach and includes the estimated time to implement the AT (based on the number of activities and time to conduct each individual activity), the type of data and the type of information systems to which the AT can be applied. The ATs have been developed with specific types of information systems and types of data as a focus. This is indicated in the studies describing the ATs explicitly or by the extent of the description provided for the type of data and systems. The user of an AT is clearly able to determine how to assess DQ for the type of data and system specified, but for other system types and types of data which are not mentioned it is not always clear. The extent of the description is indicated in Table 1 by the scale 'strongly focused', 'focused', 'can handle' and 'implicitly considered'. Cells which are left blank in this table indicate that the AT does not focus on the particular property. Parts of this table were developed previously (see [4]), and the extra data has been added using the same assumptions. For instance, implicitly considered is used when the AT does not explicitly mention the type of data or system, but the activities can be applied to it. Furthermore, the work in [4] does not include unstructured data in the table, however, it does mention that the AIM Quality (AIMQ) AT could be applied to unstructured data in his description. The 'unstructured' column is therefore listed in Table 1.

Selecting a particular AT or set of DQ projects to implement as part of a DQ methodology can be difficult for many organizations, and a model for determining the optimal project mix for improving data warehouse data quality is proposed in [1]. This model supports the selection of possible DQ projects, as determined by the data assessor, by considering constraints such as the value and cost of a project. However, the Hybrid Approach differs in that all contexts and not just data warehouses are considered, and the aim is to show how it is possible to combine and reuse the ideas from existing ATs.

Activities in data quality assessment techniques

The Hybrid Approach relies on viewing ATs as a series of activities which are performed to complete the assessment. In identifying the general activities (phases) of DQ methodologies, of which DQ assessment is a part, similar work used a top-down approach to extract the activities associated with DQ assessment. These activities include [4]:

- Data analysis, to understand the data and related architectural and management rules.
- DQ requirements analysis, to identify quality issues and set new quality targets.
- Identification of critical areas, to identify the most relevant databases and data flows.
- Process modeling, to produce a model of the processes producing or updating data.

- Measurement of quality, to select dimensions and associate metrics to these dimensions.

Activity	Definition of activity
Select DQ dimensions	The process of selecting or identifying DQ dimensions
Select DQ metrics	The process of selecting or developing DQ metrics.
Conduct analysis of results	The process of analyzing the values from the DQ measurement(s).
Formulate DQ goals	The process of determining the objectives of the entire DQ assessment process.
Identify reference 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 the 'real' value.
Select processes for the DQ assessment	The process of selecting business processes which will be focused on in the assessment.
Evaluate the DQ measurement process and identify potential improvements	The process of understanding if the DQ assessment was successful in meeting its aims (independent of the level of DQ) and identifying what improvements could be made to the DQ assessment process.
Perform subjective DQ measurement	The activity of measuring DQ by obtaining opinions of the current state of DQ.
Perform objective DQ measurement	The activity of performing DQ measurements on an actual data set.
Communicate and share the results	Communicate and share the results of the DQ assessment with relevant people.
Conduct a small-scale measurement	The process of conducting a small-scale measurement of DQ to test the measurement process before performing a full-scale measurement.
Perform a utility-driven DQ measurement	Perform a measurement to determine whether the presence of quality defects degrades utility of data, within a specific context of usage.
Define DQ requirements	The process of obtaining requirements which can be used to compare to the DQ measurement values (usually to determine the level of DQ improvement required).
Identify DQ problems	The process of determining the DQ problems.
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. This includes all enterprise objects.
Select organizational units for the DQ assessment	The process of selecting parts (units) of the organization which will be subject to the DQ assessment (includes any prioritization of the organizational units).
Group/organize data items	The process of grouping data items into categories (for example, grouping criteria could include the type of data, level of risk etc.).
Define IPs	The process of defining Information Products.
Select data items for the DQ assessment	The process of selecting the relevant data values, attributes, tables, information systems, paper files etc. which will be subject to the DQ assessment. This also includes the process of sampling the data to obtain the required data values.
Plan when to conduct the DQ assessment	The process of determining when to conduct the DQ assessment.
Identify DQ costs	The process of determining the economic losses caused by low DQ.
Perform organizational assessment	The process of performing an assessment of the organization in terms of its readiness to engage in DQ assessment (for example, a DQ maturity assessment).
Select a team for managing and executing assessment	The process of selecting people to coordinate, manage and execute the activities related to the DQ assessment.
Prioritize DQ dimensions	The process of determining which DQ dimensions are the most critical to focus the DQ assessment on.
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.

Table 2: Activities Associated with existing ATs

	DQ Assessment Techniques													
Activity	AIMQ	AMEQ-a	CDQ-a	COLDQ-a	DQA-a	DWQ-a	IQM	MMPRO	ORME-DQ	PM	QAFD	TDQM-a	TQdM-a	UDQA
Select DQ dimensions	1	4	5	4	1	2	1	4	5	1	3	3	5	1
Select DQ metrics	2	5	6	5	2	3	2	5	6	2	4	5		2
Conduct analysis of results	5	8			5		7	9			8			6
Formulate DQ goals		3											4	
Identify reference data													8	
Select processes for the DQ assessment			4	2					3				3	
Evaluate the DQ measurement process and identify potential improvements								11				8		
Perform subjective DQ measurement	3	7	7*		3		5*		7*	3	6	6*	9*	3*
Perform objective DQ measurement			8*	6	4	4	6*	8	8*		7	7*	10*	4*
Communicate and share the results.								10					11	
Conduct a small-scale measurement		6	2											
Perform a utility-driven DQ measurement														5
Define DQ requirements	4			7		1	4	3	9		5	2		
Identify DQ problems			1											
Model data creation and flow			3	1					1			4	6	
Select organizational units for the DQ assessment								2						
Group/organize data items											2		2	
Define IPs		2										1		
Select data items for the DQ assessment				3				6	4		1		1	
Plan when to conduct the DQ assessment								7						
Identify DQ costs				8					2				12	
Perform organizational assessment		1												
Select a team for managing and executing assessment								1						
Prioritize DQ dimensions							3							
Select a place where data is to be measured													7	

Table 3: The ordering of activities in the DQ assessment techniques (* = optional activity)

By contrast, our work applied a bottom-up approach to identify all the activities described by the existing ATs and these were collated to produce a list of distinct activities (see Table 2). Activities were extracted

from each study containing an AT by recording the AT stages/phases described by the study. The aim was to extract activities at a consistent level of granularity for all ATs, and once all activities had been extracted these were re-reviewed to ensure that the activities do not overlap and are not at inconsistent levels of granularity. This process also included combining similar activities described by different ATs to produce the final list of distinct activities shown in Table 2. The activities contained in each AT are shown in Table 3 and these are numbered to show the ordering of the activities within the AT. For example, the first activity (1) in the Utility-Driven Quality Assessment (UDQA) AT is 'Select DQ dimensions' and the final activity (6) is 'Conduct analysis of results'. Note that some ATs do not mention whether the measurement is required to be objective, subjective or both and in this case the number is shown with '*' to indicate that the activities are optional; in these cases the AT may therefore include subjective, objective or both types of measurement.

3. THE HYBRID APPROACH TO ASSESSING DATA QUALITY

The different components used by the Hybrid Approach to generate an AT for assessing DQ are shown in Figure 2. The development of an AT using the organization's requirements related to the DQ assessment and the requirements related to the initial motivation ensure that any resulting AT will meet the needs of the organization and DQ assessors.

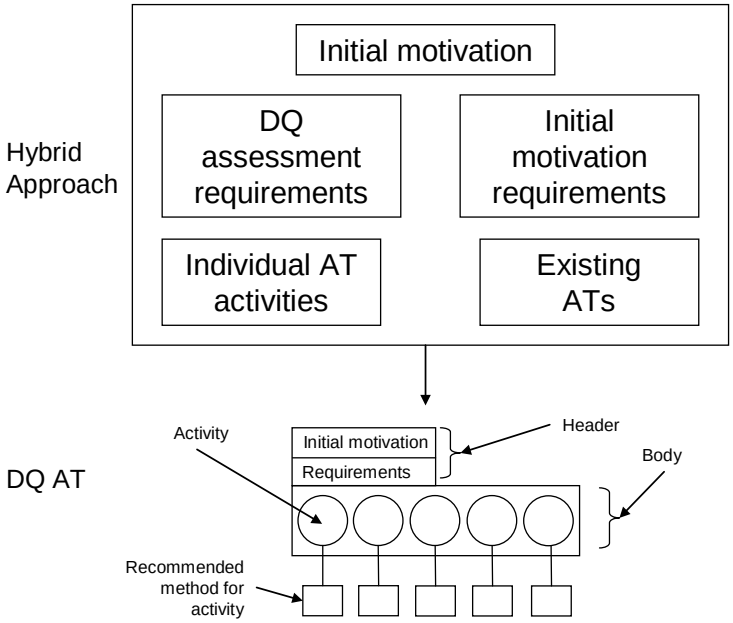

Figure 2: Components of the hybrid approach

The lower part of Figure 2 shows the structure of an AT produced by the Hybrid Approach. The header contains the initial motivation and the list of requirements. The activities which should be carried out as part of the assessment are shown in ovals in the main body of the AT. Finally, the recommended methods for carrying out an activity are attached below each activity. The methods describe the specific way of carrying out the activities. For example, a method for the 'define DQ requirements' activity could be to conduct an interview with key stakeholders.

We now describe the six steps of the Hybrid Approach which are required to develop an AT to assess the current level of DQ within an organization.

Step 1: State the initial motivation

The initial motivation drives the assessment process and is essential to inform DQ assessors of what the resulting AT should be used for. It is necessary therefore to determine the initial motivation and record this in the heading of the AT. Example initial motivations include:

- To assess a particular DQ problem which has been identified previously (example DQ problems include 'data is not accurate' and 'data is not easy to find').
- To determine and prioritize an organization's DQ problems.

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

For step 2, different companies will have different requirements which relate to the DQ assessment. This step requires the organization wanting to assess DQ to identify the requirements related to the DQ assessment. For example, an organization may require the assessment to be completed within a certain timescale and to understand the costs related to poor DQ. It may also be helpful to consider the initial motivation when identifying the requirements. For example, when the initial motivation is to assess a particular DQ problem, there may be requirements that relate to the problem, such as, the type of data, where it is stored, the sensitivity level, and these affect how the DQ problem can be measured. For instance, manual measurement of unstructured data contained on paper is a very different problem to automated measurement of the DQ of structured electronic data.

Step 3: Select AT activities which meet the requirements

The aim of this step is to select the relevant activities, from the list in Table 2, which meet the requirements that were identified in the previous step. This may not be possible for all the requirements and those which cannot be matched to activities should be used in step 4 in order to assist in selecting a base AT.

Step 4: Select a base AT

The aim of this step is to select an AT which closely matches the requirements (referred to as a base AT) by using both the set of 'required activities' from step 3 and the requirements which could not be matched to activities in step 3. The previous determination of the required activities simplifies the selection of a base AT because ATs can be selected if they contain the required activities and rejected if they do not. The breakdown of the ATs into activities (see Table 3) facilitates this selection. For the requirements which could not be matched to activities, additional knowledge of the ATs is required in order to determine which ATs meet the requirements. A structured way to conduct this step is to use the classifications of ATs such as the one presented in Table 1. These show whether the ATs are focused on specific types of data, types of systems and the estimated time to carry out the activities of the AT (see Table 1). Further work on the classification of ATs is needed to support this process.

Importantly, because an existing AT is used as a basis when developing a new AT, it ensures that the existing structures proposed by the existing ATs are captured, and, therefore, the best approaches to assessing DQ are used in the new AT. While it is possible to develop an AT by selecting individual activities, this could result in an AT which is not easy to implement and does not reuse the best combinations of activities. For example, the Data Quality Assessment (DQA-a) AT combines three activities for obtaining objective and subjective measurements of DQ and these measurements are compared to determine DQ deficiencies [20]; a new AT which uses only one of these activities is not capturing the intended approach correctly.

The result of this step is an AT containing all the required activities or, at least, an AT which can be extended to contain the required activities.

Step 5: Add and/or remove AT activities

It may be necessary to refine the AT from step 4 by adding or removing the necessary activities. For example, an AT from step 4 may contain activities which are not necessary for the assessment and can therefore be removed. Alternatively, an AT from step 4 may not contain all the required activities and so activities need to be added. Activities can be added from the existing set (as proposed by the existing ATs) or new activities can be proposed. As noted before, separating activities in this manner can be difficult, and, therefore, the Hybrid Approach recommends that clusters of related activities should be retained in order to both mitigate this problem and ensure that the best combinations of activities are reused.

Step 6: List recommended methods for the activities

Finally, for the activities in the AT, methods for how to carry out each activity should be suggested. Figure 2 shows how the recommended methods are attached to the activities in the AT. A method is a way of carrying out an activity and there are typically many different methods to choose from. Example methods for the 'perform subjective DQ measurement' activity include using interviews with stakeholders, and for the 'perform objective DQ measurement' activity, executing software which checks an information system for missing values could be used. Selecting the actual methods to use is the task of the DQ assessor using the AT, but it is important to provide recommended ('default') methods for guidance, which can be used without the user having to select their own. The studies that describe ATs usually propose certain methods for the activities in the AT, and if an AT has been selected as a base AT, then the methods proposed by this AT are usually good candidates to recommend. When using the Hybrid Approach, developers of ATs may also propose new methods for activities or suggest methods from other ATs which are suitable. In the final AT it is important to specify which AT the method has been cited from so that the DQ assessor can obtain more information about how to carry out the method.

Step 5 described how clusters of activities from existing ATs should be retained to ensure best practices are retained. Similarly, the methods from the ATs for these activities should also be clustered. For example, if two activities from an AT have been selected, then the methods proposed by the AT for these activities should also be recommended. At this stage, when all the steps have been completed, the AT is complete and should be documented as shown in Figure 2.

4. CASE EXAMPLE

In order to demonstrate how to use the Hybrid Approach to generate an AT, a scenario containing the requirements from an asset management (AM) company perspective is used. These requirements were obtained via interviews with professionals from a maintenance, repair and overhaul (MRO) organization in the UK (hereon referred to as 'Company A').

Asset intensive organizations typically own physical engineering assets such as trains, aircraft, underground water pipes, electricity pylons etc. and these organizations are responsible for the acquisition, deployment, maintenance and disposal of these assets. The capital invested in an organization's assets requires that maximum benefit is extracted from the assets throughout their lifecycle, which means that making sound decisions about managing the assets is critical [11],[18]. Basing decisions on poor quality data can potentially result in great economic losses [10]. Maintaining and providing good quality data is a difficult task, and many leading AM organizations are keen to identify areas where DQ can be improved. Assessing and improving the quality of AM data is a complex problem because it is not just stored in databases, paper-based systems are also used frequently.

Altogether, there are a variety of systems within which data is stored.

Scenario description

Company A is an MRO organization and a large part of their revenue and risk is associated with the management of their specialist aircraft tools, aircraft ground support equipment, and facilities (for example, a runway and hangars). Decision-makers in the AM unit of the organization have indicated that their biggest problem is that data is not easy to find. Typically, decision-makers need to determine the optimal timing to replace aircraft ground support equipment. The aim of the DQ assessment for Company A is therefore to assess the DQ problem of 'data is not easy to find' and to determine the level of DQ improvement required.

Only two people have been assigned the task of assessing DQ in the AM business unit of the organization, and these people must balance their existing AM duties with the DQ assessment. They are, therefore, only able to commit limited time to the DQ assessment and they are required to report their initial findings to their manager in approximately one month. Managers in the AM part of the organization have recognized that DQ is a problem and are keen to allocate more resources, but the only way to justify the cost of the extra resources is to quantify the costs of poor DQ to the senior managers.

The data in Company A is recorded in a variety of forms including structured, semi-structured and unstructured and it is stored in many different electronic and paper-based systems; typical systems include condition monitoring systems, paper-based inspection reports, numerous spreadsheets and databases and an Enterprise Asset Management (EAM) system. One of the main problems is that most of these systems need to provide information for operational decision-making as their primary purpose while also supporting higher levels of decision-making. Many AM organizations face this problem, for example, equipment maintenance and reliability data is typically needed firstly by reliability engineers for the determination of long-term maintenance strategies, and secondly by maintenance engineers and maintenance supervisors for addressing day-to-day maintenance issues [13]. These systems often do not support the first channel making it difficult to extract and integrate this information for higher-level decisions. Some of these problems are as a result of the lack of an overall data architecture and management strategy [16]. In Company A many of these systems are third-party systems which are difficult and/or expensive to 'open up' to extract information and share it with other applications. The end result is that the decision-makers are faced with having multiple disparate sources of information and they need to spend time to collate this information or find alternative ways of making the decision without the required data.

Using the Hybrid Approach

This section demonstrates how to develop a AT for the DQ problem 'data is not easy to find' using the scenario related to Company A presented in the previous section. This DQ problem has been selected as part of step 1 of the Hybrid Approach and the initial motivation is therefore 'to assess the problem of data is not easy to find'.

For step 2, the following requirements related to the DQ assessment have been identified:
The DQ assessment must:

1. Determine the costs caused by low DQ.
2. Obtain the DQ requirements (focus on the decision-maker).
3. Determine the level of DQ improvement required.
4. Be able to be carried out in a short time-scale
5. Be able to handle data in multiple forms including structured, semi-structured and unstructured forms.

6. Be able to handle data from multiple electronic computer systems and paper-based systems.
7. Identify the scenarios where data is required and how it is provided to the consumer.

Requirements five, six and seven relate to the DQ problem of not being able to find data easily and these were therefore indentified by also considering the requirements related to the initial motivation as well as the DQ assessment in general. Requirements five and six capture the type of data and where the data is stored. For requirement seven, an understanding of what data is required and how it should be provided to the consumer is critical because the consumer (AM decision-maker) is clearly not happy with the current situation and is not being provided with the required data.

For step 3, using the requirements related to the DQ assessment (the output of step 2), suitable activities were selected which map to these requirements (see Table 4); it was not possible to map requirements four, five and six to specific activities, so these requirements are not shown in Table 4.

Requirement number	Requirement	Suitable AT activity
1	Determine the costs caused by low DQ	Identify DQ costs
2	Obtain the DQ requirements (focus on the decision-maker).	Define DQ requirements
3	Determine the level of DQ improvement required.	Conduct analysis of results
7	Identify the scenarios where data is required and how it is provided to the consumer.	Model data creation and flow

Table 4: The selection of AT activities for each requirement

The selection of activities for requirements one and two is straightforward. However, for requirement three, in order to determine the level of DQ improvement required, it is necessary to obtain DQ measurements and compare them to DQ requirements, and hence, the 'conduct analysis of results' activity is used to make this comparison and determine the level of DQ improvement required. For requirement seven, modeling the data creation and flow should highlight the scenarios where data is provided to the consumer. Requirements four, five and six (see Table 5) are used in step 4 to help identify a base AT because it is not possible to identify specific activities for these requirements.

	Requirement (and requirement number)		
	The AT must take a short time to implement (4)	**The AT must handle multiple types of data (5)**	**The AT must handle data stored in multiple systems (6)**
ATs which meet the requirements	AIMQ, DQA-a, DWQ-a, PM, UDQA	AIMQ, PM	AIMQ, PM
ATs that do not meet the requirements	All other ATs (medium or long-term)	DQA-a, DWQ-a and UDQA (only structured)	-

Table 5: Selecting a base AT

Table 5 shows the selection of the base AT by identifying the ATs which meet each of the requirements. The classification of different properties of ATs (see Table 1) was used to assist this process. Proceeding from left to right in Table 5, the set of ATs which meet each requirement are listed. Note that not all ATs are considered for each requirement: Only the set appearing in the previous cell is considered. This is to

illustrate the process of filtering (reducing) the set of ATs when each requirement is applied.

Only five ATs are considered to take a short time to implement (AIMQ, DQA-a, DWQ-a, PM and UDQA) and these are listed under requirement four in the first column of Table 5. For the 'type of data' requirement it was not possible to find a suitable AT which can be applied to all types of data and so the ATs which could be applied to multiple types of data were selected. DQA-a, DWQ-a and UDQA can only be applied to structured data and so these ATs were rejected. Similarly with the information systems, no AT could be applied to multiple electronic systems and to a paper-based system, so this requirement was also relaxed to 'multiple systems'; no AT was rejected on this basis. AIMQ and PM were selected at this stage as candidates for being the base AT. Table 6 shows the number of required activities already contained in these ATs. AIMQ contains two out of the four required activities and PM does not contain any of the required activities. AIMQ is therefore favored over PM as a base AT.

AT	Identify DQ costs	Define DQ requirements	Conduct analysis of results	Model data creation and flow	Total number of matching activities
AIMQ		x	x		2
PM					0

Table 6: Activities contained in the ATs

Step 5 allows the base AT to be configured to the exact requirements by adding and removing the activities in the base AT. Using AIMQ as a basis, the 'model data creation and flow' and 'identify DQ costs' activities have been added before the selection of DQ dimensions activity so that suitable dimensions can be selected using the results of these activities. There are no unnecessary activities which need to be removed from this AT. Figure 3 shows the AIMQ AT with the two new activities added to the start. Unfortunately, the addition of these activities moves the AT closer to medium, in terms of time to conduct the assessment, and so in order to ensure that it can be completed within the timescale, methods which do not take too much time should be selected for each activity. The fact that the resulting AT contains seven activities indicates that assessing the DQ problem of data is not easy to find, with the existing activities from the ATs and with these requirements, is difficult to conduct within a short time period. However, it is possible to select methods for these activities which are not overly time-consuming. For the relevant activities, these suggested methods are shown with the final AT in Figure 3. This final AT also includes the updated requirements (which were modified because of the difficulty in selecting a base AT).

For the 'model data creation and flow' and 'identify DQ costs' activities the methods suggested by ORME-DQ can be used. This includes using matrices for the first method to represent the information flows where the goal is to provide a picture of the main uses of data, of providers, and of consumers of data flows [2], and using the hierarchy of costs for the second method to evaluate the economic costs of poor data. Note that both of the methods for these activities are taken from the same AT (ORME-DQ) following the idea that activities and methods should be clustered as much as possible. For the remaining five activities, the methods proposed by AIMQ are recommended which include using the PSP/IQ model of dimensions, using the instrument proposed by AIMQ which also contains measures for the dimensions in the PSP/IQ model, administering the questionnaire focusing on the decision-makers (see requirement 2), and using gap analysis. Again, the methods for these five activities have all been taken from the same AT rather than different ATs.

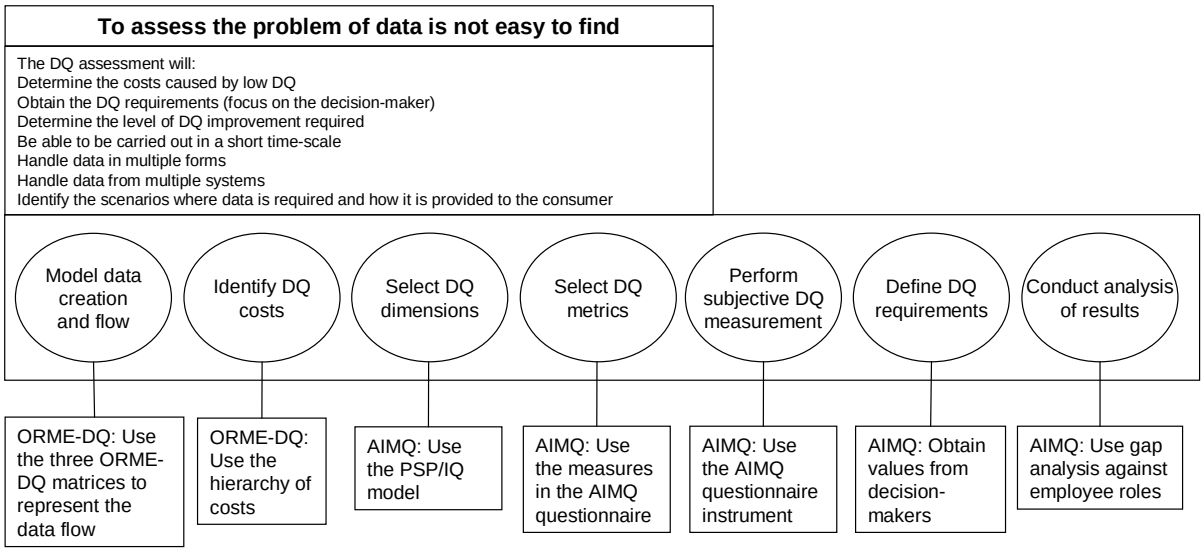

Figure 3: An AT for data is not easy to find

5. LIMITATIONS

The Hybrid Approach relies on being able to separate activities from the existing ATs and combine them with activities from other ATs—this is sometimes very difficult, however. For instance, Batini notes that some ATs are not extensible with respect to using dimensions and metrics from other ATs [4]. In these cases, while it would be very difficult to separate certain activities, in order to address this problem, the Hybrid Approach recommends retaining groups of activities which are closely linked. Moreover, it is often the case that the extensibility of existing ATs is not 'fixed' or 'open' in the Boolean sense, but can be measured on a continuous scale with some ATs being more open than others. Therefore, it may still be possible to use different metrics with an AT, but with increased difficulty to integrate the metrics for some ATs.

Another challenge is that the selection of a base AT requires a good knowledge of the different ATs and their idiosyncrasies. To support this selection process, the Hybrid Approach utilizes existing classifications of ATs (see Table 1) and the break-down of ATs into activities (see Table 3). This is to ensure that the selection can be structured and systematic, but also so that the required level of knowledge of the ATs is kept to a minimum. However, further classifications of ATs are required.

A newly developed AT alone is not a panacea to the problem of assessing DQ. To carry out the activities suggested by the AT requires the selection of methods for each activity. A good AT will guide the selection of methods by suggesting 'recommended methods', but in some cases it will be necessary for the data assessors to find and use other methods. In the latter case, the methods may have to be modified to ensure that the method for each activity integrate seamlessly.

6. CONCLUSIONS AND FUTURE WORK

Currently, no individual existing technique for assessing DQ is wholly suitable to assess DQ for all types of requirements due to the varying nature of requirements over time and organizational needs. The requirements may be different for every organization and even the same organization over time due to factors such as the change in the level of maturity of the organization. The proposed Hybrid Approach shows how to develop new DQ ATs by combining the activities from existing techniques in a way that meets differing requirements whilst still retaining the concepts and ideas incorporated in the existing DQ

ATs. The Hybrid Approach demonstrates that ATs need not be static where activities are predefined and do not change. It is possible to develop new ways of assessing DQ which leverage the best approaches proposed by existing ATs by combining the activities dynamically to generate new techniques for assessing DQ. The division of the existing ATs into a common set of activities enables future research to use the Hybrid Approach to develop ATs for any set of requirements and to document explicitly the context in which the AT can be used. This approach does not limit the development of new ATs and new activities, but rather it provides a way to capture the ideas so that they can be used in contexts which the existing or new ATs are not suitable. In fact, the application of the approach will illustrate where there are gaps in the current ATs and where new activities need to be developed. In the case of the AT developed in this work, it was difficult to find an AT which can handle both multiple types of data and multiple systems.

There are still a number of open questions regarding the Hybrid Approach. Whilst it is theoretically possible to combine activities from existing ATs, in practice this may be difficult. Future work on developing and reviewing ATs must establish the extent to which this is a problem. All proposed ATs should be evaluated with a standard review procedure which aims to ensure the feasibility to carry out the DQ assessment and whether the AT meets its requirements. The exact structure of the evaluation is an open question. Future work also needs to establish how to publish and share ATs that have been developed using the Hybrid Approach. Another open question is whether this is an interim approach which may lead to the development of a unified AT. For example, if many DQ problems can be attributed to a certain AT, could this AT be combined with other popular ATs to create a unified AT which would then be usable in all contexts? Moreover, using the Hybrid Approach could lead to a classification of DQ problems (based on how you need to assess the data related to the problems). A review of the previous research which attempts to categorize and identify types of DQ problems is given in [12]. Together, these could lead to an understanding of how techniques for assessing and improving one DQ problem could be reused to eliminate other similar DQ problems.

Future plans are to validate the Hybrid Approach by developing ATs for multiple organizations, based on their requirements, and then trialing each AT in order to demonstrate whether the approach is valuable. The next step after DQ assessment is DQ improvement, and, therefore, another future aim is to extend this work to include the improvement process so that both DQ assessment and DQ improvement can be tailored to requirements.

ACKNOWLEDGEMENTS

We would like to thank Alex Borek for helping to source papers and EPSRC for supporting this research.

REFERENCES

[1] Ballou, D.P., and Tayi, G.K., "Enhancing Data Quality in Data Warehouse Environments," *Communications of the ACM*, 42 (1), 1999, pp.73-78.

[2] Batini, C., Barone, D., Mastrella, M., Maurino, A., and Ruffini, C., "A Framework and a Methodology for Data Quality Assessment and Monitoring," *Proceedings of 12th International Conference on Information Quality*, 2007.

[3] Batini, C., Cabitza, F., Cappiello, C., and Francalanci, C., "A Comprehensive Data Quality Methodology for Web and Structured Data," *Proceedings of the 1st International Conference on Digital Information Management*, 2006, pp.448–456.

[4] Batini, C., Cappiello, C., Francalanci, C., and Maurino, A., "Methodologies for Data Quality Assessment and Improvement," *ACM Computing Surveys*, 41 (3), 2009, pp.1-52.

[5] Caballero, I., Verbo, E., Calero, C., and Piattini, M., "MMPRO: A Methodology Based on ISO/IEC 15939 to Draw Up Data Quality Measurement Processes," *Proceedings of the 13th International Conference on Information Quality*, 2008.

[6] De Amicis, F., and Batini, C., "A Methodology for Data Quality Assessment on Financial Data," *Studies in Communication Sciences*, 4 (2), 2004, pp.115–136.
[7] English, L., *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*, John Wiley & Sons, 1999.
[8] Eppler, M., and Muenzenmayer, P., "Measuring Information Quality in the Web Context: A Survey of State-of-the-art Instruments and an Application Methodology," *Proceedings of 7th International Conference on Information Quality*, 2002, pp.187–196.
[9] Even, A., and Shankaranarayan, G., "Understanding Impartial Versus Utility-Driven Quality Assessment in Large Datasets," *Proceedings of 12th International Conference on Information Quality*, 2007.
[10] Gao, J., Baškarada, S., and Koronios, A., "Agile Maturity Model Approach to Assessing and Enhancing the Quality of Asset Information in Engineering Asset Management Information Systems," *Proceedings of the 9th International Conference on Business Information Systems (BIS 2006)*, 2006, pp.486-500.
[11] Gao, J., Lin, S., and Koronios, A., "Data Quality in Engineering Asset Management Organizations - Current Picture in Australia," *Proceedings of the 11th International Conference on Information Quality*, 2006.
[12] Ge, M., and Helfert, M., "A Review of Information Quality Research," *Proceedings of the 12th International Conference on Information Quality*, 2007.
[13] Hodkiewicz, M., Kelly, P., Sikorska, J., and Gouws, L., "A Framework to Assess Data Quality for Reliability Variables," *Proceedings of the World Congress on Engineering Asset Management (WCEAM)*, 2006.
[14] Jeusfeld, M.A., Quix, C., and Jarke, M., *Design and Analysis of Quality Information for Data Warehouses*, *Proceedings of the 17th International Conference on the Entity Relationship Approach (ER'98)*, Springer, Lecture Notes in Computer Science, 1998, pp.349–362.
[15] Lee, Y.W., Strong, D.M., Kahn, B.K., and Wang, R.Y., "AIMQ: A Methodology for Information Quality Assessment," *Information & Management*, 40 (2), 2002, pp.133–146.
[16] Lin, S., Gao, J., Koronios, A., and Chanana, V., "Developing a Data Quality Framework for Asset Management in Engineering Organizations," *International Journal of Information Quality*, 1 (1), 2007, pp.100-126.
[17] Loshin, D., *Enterprise Knowledge Management: The Data Quality Approach*. Morgan Kaufmann Pub, 2001.
[18] Ouertani, M.Z., Parlikad, A.K., and McFarlane, D., "Asset Information Management: Research Challenges," *Proceedings of the 2nd International Conference on Research Challenges in Information Science (RCIS 2008)*, 2008, pp.361–370.
[19] Pierce, E., and Thomas, L., "Assessing Information Quality Using Prediction Markets," *Proceedings of 12th International Conference on Information Quality*, 2007.
[20] Pipino, L.L., Lee, Y.W., and Wang, R.Y., "Data Quality Assessment," *Communications of the ACM*, 45 (4), 2002, pp.211-218.
[21] Su, Y., and Jin, Z., "A Methodology for Information Quality Assessment in the Designing and Manufacturing Process of Mechanical Products," *Proceedings of the 9th International Conference on Information Quality*, 2004, pp.447-465.
[22] Wang, R.Y., "A Product Perspective on Total Data Quality Management," *Communications of the ACM*, 41 (2), 1998, pp.58-65.