Quality, Trust, and Utility of Scientific Data on the Web: Towards a Joint Model

Matthew Gamble
School of Computer Science
University of Manchester
m.gamble@cs.manchester.ac.uk

Carole Goble
School of Computer Science
University of Manchester
carole.goble@manchester.ac.uk

ABSTRACT

In science, quality is paramount. As scientists increasingly look to the Web to share and discover scientific data, there is a growing need to support the scientist in assessing the quality of that data. However, quality is an ambiguous and overloaded term. In order to support the scientific user in discovering useful data we have systematically examined the nature of "quality" by exploiting three, prevalent properties of scientific data sets: (1) that data quality is commonly defined objectively; (2) the provenance and lineage in its production has a well understood role; and (3) "fitness-for-use" is a definition of utility rather than quality or trust, where the quality and trust-worthiness of the data and the entities that produced that data inform its utility. Our study is presented in two stages. First we review existing information quality dimensions and detail an assessment-oriented classification. We introduce definitions for quality, trust and utility in terms of the entities required in their assessment; *producer, provider, consumer, process, artifact and quality standard.* Next we detail a novel and experimental approach to assessment by modelling the causal relationships between quality, trust, and utility dimensions through the construction of decision networks informed by provenance graphs. To ground and motivate our discussion throughout we draw on the European Bioinformatics Institute's Gene Ontology Annotations database. We present an initial demonstration of our approach with an example for ranking results from the Gene Ontology Annotation database using an emerging objective quality measure, the Gene Ontology Annotation Quality score.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Data Sharing*; H.1.2 [**Information Systems**]: Models and Principles—*Human Factors*

General Terms

Human Factors

WebSci '11, June 14-17, 2011, Koblenz, Germany.

Keywords

Science, Data Quality, Trust, Data Sharing, Decision Networks

1. INTRODUCTION

Mechanisms such as peer review, curation and moderation have long been ingrained in the scholarly process in an effort to ensure quality and increase trustworthiness. However, these traditional mechanisms of quality control are being disrupted. The Web is facilitating a transformation in scientific practice, impacting dissemination of scholarly literature[26] and increasingly the way we share and consume scientific data[6][5]. Funding policy and initiatives such as the Open Knowledge Foundation[1], Science Commons[3], BioSharing[19] and the Royal Society's recently initiated SAPE working group[13] are driving a move towards a more open approach to data sharing. In parallel, data sharing platforms both large scale[7, 40, 9] and 'long tail'[16, 2] are emerging in an effort to enable scientists to share their data. The result of this open and Web-scale approach to data sharing is a data landscape with a wide spectrum of moderation and control, where quality is quickly becoming an issue[20]. The European Bioinformatics Institute's Gene Ontology annotation project for example provides a critical resource for functional genomics[15], currently providing annotations data for well over 100,000 species. With such high volumes of data from varying sources it can be difficult for users to assess the quality of that data[14]. The immediacy and scale of the Web coupled with the exponential increase in the volume of scientific data[25] has affected our ability to effectively judge the quality and trustworthiness of scholarly artifacts.

Before the scientific community can take full advantage of this Web of data it is necessary that mechanisms are introduced to evaluate the quality of scientific data at a Web scale. To achieve this we require a clear understanding of quality, its relationship to trust and how it ultimately informs utility.

2. WHAT IS DATA QUALITY?

The issue of data or information quality is one that is prevalent not just in the scientific domain, but any domain where critical decisions rely on high quality data. Historically driven from a business and organization management perspective[44] (with a focus on monolithic information systems[10]), research into methodologies and techniques to detect, assess and improve data quality issues has been conducted in fields such as statistics, business management and

computer & information science. Despite the diverse scope of study, common to almost all information quality research is the reduction of information quality to a multi-dimensional concept. Information quality then becomes an aggregation of the values of multiple information quality dimensions[37] such as accuracy, timeliness, reputation etc. The definition and subsequent classification of quality dimensions is central to many of the established information quality methodologies. Whilst there are many quality dimensions that are specific to a particular domain, there is also significant agreement on a core set[10] (see table 1).

To guide the modelling of information quality and quality dimensions, there are a number of definitions of quality from which the literature draws inspiration. By far the most frequently adopted is that of Juran, who describes data quality simply as "fitness for use"[28]. This definition typically leads to the assertion that the quality of data cannot be assessed independent of the data consumer[44] and that information quality is entirely subjective to the user and the context in which the data is to be used[12]. Alternatively the ISO 90001 standard defines quality as "the degree to which a set of inherent characteristics fulfils requirements". This alternative highlights what we see as a subtle but important variation, where requirements can be defined objectively and exist independent of an individual.

Information quality methodologies can be broadly viewed in two classes, organisational methodologies such as the Total Data Quality Management (TDQM) methodology[43] or Web-based such as Naumann's Subject-Process-Object methodology[37]. For organisational methodologies such as TDQM the focus is typically on intra-organisational data quality issues where the goal of the assessment is to inform quality-control procedures for large-scale monolithic information systems. In contrast, in the context of web-based information systems the goal is to support the data consumer in assessing the quality of data and ultimately decide if the data are of sufficient quality to meet their needs. In further contrast to organisational methodologies, data producer and consumer are typically unknown to each other. The introduction of uncertainty dictates that *trust* plays a central role in information quality assessment[21]. Prior work in web-based assessment therefore often adopts a primary concern of either information quality[12][31] or trustworthiness[21][29]. TDQM details an early, influential approach to modelling the data consumers view of information quality and partitions data quality dimensions into four classes: (1)*intrinsic data quality* - accuracy, objectivity, believability, reputation; (2)*contextual data quality* - relevancy, value-added, timeliness, completeness, amount of data; (3)*representational data quality* - interpretability, ease of understanding, concise representation, consistent representation; and (4)*accessibility data quality* - accessibility, access security. This classification, termed in [37] as a *semantic-oriented* classification, clearly supports a subjective view of information quality and as a result there is no clear separation of objective and subjective dimensions. The Subject-Process-Object methodology instead presents an *assessment-oriented* classification where quality dimensions are classified by its source ; subject - the user, process - the query process, or object - the data. Though still supporting a subjective view of information quality, this assessment-oriented approach goes some way to separating objective and subjective dimensions. We choose this assessment-oriented approach to guide our subsequent classification for this reason.

2.1 What is Data Quality in Science?

A scientist's decision as to whether to use data is informed by whether it is: (1) good when compared against norms and standards; (2) likely to be good given its provenance; and (3) a good fit to current needs. Rather than a single concern, these three interrelated concerns - *quality, trust and utility* - form the basis of a quality assessment for web-based scientific data. *Quality* has been the subject of systematic attempts at objective measurement. Communities and moderators define, promote, and adhere to standard data formats, vocabularies and minimum information models[19]. These act as a 'social technology'[17], negotiated by the community in an effort to support the asynchronous nature of science on the web[47]. Similarly, the quality of scientific data can be measured for accuracy and agreement against standards and benchmarks. The Gene Ontology Annotation Quality Score (GAQ)[14] is an example of such an emerging objective measure in the life sciences, defining a numeric measure of quality against which data descriptions (known as annotations) in the Gene Ontology Annotations database can be measured.

$$GAQ(a) = ECR_a \times depth_a$$

The GAQ score for an individual annotation is defined as the product of its depth in the ontology *depth* (accuracy) and the evidence code rank (ECR) of the annotation. Evidence codes indicate the process by which the annotation data was produced, for example by manual experimentation e.g. Inferred from Genetic Interaction (IGI) or computational analysis e.g. Inferred from Sequence (ISA). Similarly in the field of earth sciences, the widely adopted Spatial Data Transfer Standard (SDTS)[4], dictates a reporting requirement for information quality metadata such as positional accuracy, to aide in the effective exchange of spatial data.

These measures exemplify a scientific view of information quality, where there can be some context-dependant objective agreement of quality requirements between communities or for specific data sets.

Trust and reputation play a crucial role on the scholarly process, embodied in mechanisms developed over centuries. Attribution and citation machinery has been carefully developed to clearly define provenance. Citation counts, impact measures and emerging alt-metrics all serve as mechanic predictions of quality. Peer-review and past experiences give community-wide and personalised indicators of trust. Scientists are likely to select data from a source known to them [22] or widely regarded as trustworthy, even if objective measures of accuracy reveal this trust to be misguided [A. Williams pers. comm.].

Utility is a "fitness-for-use" interpretation of data. A measure requiring the scientific user to formally describe their information quality needs as *quality knowledge* has been explored in [35]. This consumer-centric approach, though motivated by objective quality measures in proteomics, presents the decision that data is "good enough" for current needs as a subjective one.

Our assessment-oriented approach has two primary components: (1) A classification of Quality Dimensions (each indicative of Quality, Trust or Utility) and the information entities that they are an assessment function of; (2) A method

Quality Dimension	Indicator Of	Function Of	IQ Study
Completeness	Quality	Artifact/Standard	[27, 43, 32, 42, 37, 11, 41, 18, 34]
Accuracy	Quality	Artifact/Process/Standard	[43, 8, 18, 42, 27, 37, 11, 29, 21]
Timeliness	Utility	Artifact/Consumer	[43, 32, 18, 27, 37, 11, 41, 34]
Consistency	Quality	Artifact/Process/Standard	[11, 41, 32, 18, 42, 43, 37, 21]
Accessibility/Availability	Utility	Artifact/Consumer	[43, 32, 18, 27, 11, 41, 37]
Reputation	Trust	Arti/Prod/Prov/Proc/Cons	[43, 32, 37, 11, 41, 21, 27]
Objectivity	Trust	Arti/Prod/Provi/Cons	[43, 32, 8, 37, 41, 29, 21]
Conciseness	Utility	Artifact/Consumer	[32, 18, 27, 43, 41, 37]
Relevance	Utility	Artifact/Consumer	[43, 32, 37, 41, 29, 11]
Understandability	Utility	Artifact/Consumer	[32, 37, 41, 43, 18, 11]
Believability	Trust	Arti/Prod/Prov/Cons	[43, 32, 37, 41, 21]
Interpretability	Utility	Artifact/Consumer	[43, 32, 27, 37, 41]
Currency	Quality	Artifact/Standard	[8, 18, 42, 11, 21]
Security	Trust	Arti/Prod/Prov/Proc/Cons	[43, 32, 37, 41]
Amount of Data	Utility	Artifact/Consumer	[43, 37, 41, 21]
Correctness	Quality	Artifact/Standard	[27, 34, 11, 29]
Value-Added	Utility	Artifact/Consumer	[43, 37, 41]
Stability/Volatility	Quality	Artifact/Process/Standard	[29, 34, 11]
Applicability/Appropriateness	Utility	Artifact/Consumer	[18, 32]
Authority	Trust	Producer/Provider	[21, 8]
Freedom from Errors	Quality	Artifact/Standard	[32, 41]
Recommendation	Trust	Arti/Prod/Prov/Proc	[21, 29]
Trustworthiness	Trust	Producer/Provider/Consumer	[29, 42]
Usefulness	Utility	Artifact/Consumer	[27, 34]
Cost	Utility	Artifact/Consumer	[11, 37]
Usability	Utility	Artifact/Consumer	[41, 32]

Table 1: Analysis of information quality dimensions. [38] was consulted as a guide to synonymous concepts in information quality.

of combining each Quality Dimension into one assessment measure that can be used in selection, ranking, and comparison decisions.

3. CLASSIFICATION OF DIMENSIONS

Table 1 details a review and analysis of twelve information quality studies[27, 43, 32, 42, 37, 11, 41, 18, 34, 29, 21, 8] from which we elicit a set of 26 information quality dimensions where there is some agreement. These quality dimensions are rarely functions of the data alone but instead a function of multiple entities. The Open Provenance Model[36] identifies three high level entities involved in the production of data: *agents* e.g. producer or provider, *processes* and *artifacts* i.e. data.

For example, scientific data exchange in the web is typically *asynchronous*[33] where a data provider agent lies between the producer agent and consumer agent. The producer deposits data with the provider, which may oversee its curation or modification, and the consumer agent accesses the data some time later. In total we identify 6 unique entities as potential subjects of assessment: producer, provider, consumer, process, artifact (the data itself) and a quality reference standard. For example, accuracy (an indicator of quality) may refer to the assessment of the artifact or an assessment of the process that produced the artifact, both requiring a reference standard to measure against. Authority (an indicator of trust) can be attributed to the provider or producer agents without a reference standard against which to measure. We can now specifically define our three classifications of quality dimensions in terms of their entities:

Quality Dimensions - a function of the *artifact* or *process* assessed against a *quality standard* independent of the consumer to provide a specific, objective measure of quality e.g. accuracy.

Trust Dimensions - a function of the *artifact*, *producer*, *provider* or *process* (along with perhaps the *consumer*) that can be assessed independent of a *standard* to provide a general prediction of quality e.g. reputation

Utility Dimensions - a function of the *artifact* and *consumer* to assess whether data are fit for purpose and meet the users subjective needs e.g. relevance.

To illustrate our classification consider an annotation from the GO Annotations database. Table 2 aligns an individual annotation with our entities and three example assessment dimensions accuracy, reputation, and timelines. The accuracy dimension is captured in the GAQ processes as the depth of the annotation, requiring the artifact and standard. We consider the set of producers of the annotation to be uniquely identified by the publication identifier (PMID) available in the reference column in for the annotation entry. A simple proxy for reputation is then a citation count of that publication. Timeliness as a measure of utility is a comparison of the date the annotation was updated with the consumer's requirement.

Entities	Example
Artifact	Annotation
Process	Evidence code e.g. ISA
Producer	Study that produced annotation e.g. PMID
Standard	Gene Ontology Annotation Quality Score
Provider	http://www.ebi.ac.uk/GOA/

Dimensions	Class	*Function of*
Accuracy	Quality	Annotation/GAQ
Reputation	Trust	PMID
Timeliness	Utility	Annotation/Consumer

Table 2: Annotation example.

From these classifications we observe a clear spectrum of objectivity. Quality is an entirely objective assessment against a reference standard. Trust may be objective based upon intrinsic qualities in entities such as popularity or subjective based upon the data consumers prior experience. This dual notion of trust correlates well with the established notion of global and local trust in prior work in computational trust[23]. An important consideration is the management of reputation information. An objective trust measure though intrinsic to the entity may require a trusted third party to provide information such as citation count or global reputation. Finally utility is perhaps intuitively entirely subjective and informed by the consumers contextual requirements, elicited at the time of the query. Ultimately these classifications serve to structure future assessment through a clear understanding of the entities required.

4. COMBINING ASSESSMENTS.

Having established our assessment-driven classification of quality, trust and utility dimensions we require a suitable mechanism to combine individual dimensions into an overall assessment. The quality dimensions of data are functions of the interaction of the multiple entities that produced it. An attempt to represent a dependency model of data lineage has been made by the web and eScience community through the development of models and vocabularies of provenance [36][46]. The Open Provenance Model serves to establish the lineage of data and the causal relationships between data, processes and agents. We similarly wish to capture the causal relationships between the quality, trust and utility scores of the entities involved in producing the data in question. Figure 1 shows the provenance graph for the annotation example in Table 2. Next we need a mechanism to use these dependency graphs to combine assessments into one numeric measure for decision making. For this we use decision networks, a well established formal approach to support decision making under uncertainty[30], ultimately providing a subjective score for utility.

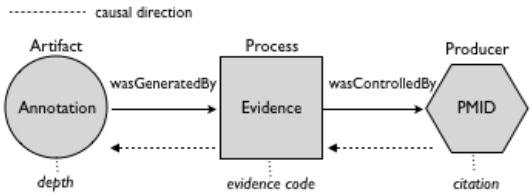

Figure 1: Provenance graph for annotation in the GO annotations database.

4.1 Decision Networks.

A decision network allows the modelling of causal relationships between interacting sets of variables, built upon probabilistic (or causal) networks. Networks are constructed as directed acyclic graphs where variables are represented by nodes in the graph and directed links between the nodes indicate causal relationships. Figure 2 illustrates a simple example. Using standard terminology there are three classes of variable; nature, decision, and utility.

Nature variables represent a set of mutually exclusive states along with a conditional probability distribution detailing the probability of each state occurring. Where a nature variable has parent nodes, these probabilities are stated as conditional on the parents possible states. Nature variables X and C in the example illustrate the conditional probability statement,

$$P(C_i \mid X_j) = y$$

that is *"given the observation of j in X the probability of i in C is y"*. The directed link between the two indicates cause (X) and effect (C) relationship.

A decision variable indicates a point in time of the modelled domain where a decision can be made. The domain of the decision node is an exhaustive set of mutually exclusive decisions. Finally, utility variables represent *utility functions*. A utility function serves to represent the subjective utility of any parent decision nodes, providing a utility score for each outcome.

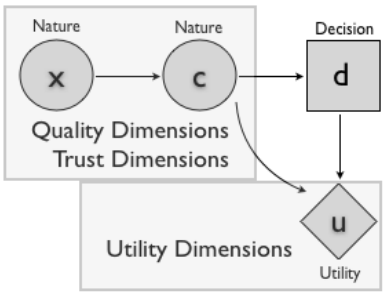

Figure 2: An example decision network.

4.2 Our Application of Decision Networks

In our application of decision networks the construction the network is driven by the entities in the provenance graph, and the dimensions available for their assessment. Each entity will have a corresponding nature variable, representing either the quality of trustworthiness of that entity. Quality and trust dimensions such as reputation will then inform these nature variables. Utility dimensionsx inform utility variables. In the network we will typically model one decision - whether to accept or reject the data - and one utility variable e.g. the utility of accepting a high quality artifact vs. the utility of accepting a low quality artifact.

4.3 Constructing the Decision Network

To illustrate our application of decision networks we return again to our example of the gene ontology annotation in table 2. Like [21] we demonstrate our assessment through the task of ranking, in our case by ranking a result set of gene annotations according to utility. Our task is therefore to transform a provenance graph for an artifact into a decision network suitable for scoring and ranking. Figure 1 provides a simple provenance graph for our artifact - a single annotation - highlighting the causal relationships between the entities. To construct the decision network we must (1) identify variables and capture the causal relationships between our assessment dimensions; (2) populate the conditional probability distribution for each node; and (3) construct a utility function. To guide the construction of the decision network we apply established *idioms*[39] in causal network modelling, Figure 3 shows the final network.

For our provenance graph the GAQ score coupled with the reputation dimension in table 2 provides us with a dimension for each entity, a quality dimension for the artifact, a quality dimension for the process and a trust dimension for the producer. We therefore introduce three variables, the quality of the artifact *qArtifact*, quality of the process *qProcess* and trustworthiness of the producer *tProducer*. We also introduce a variable for each element of metadata on which the assessments are based - *depth*, *evidence code*, *citation*. These variables are considered *background variables* - information that is available prior to the event and has a causal impact on the problem variables[30]. Citation count for the PMID is sourced from the publicly available citations data on PubMed Central[1]. Importantly, a PMID is not available for all annotations so we must ensure the network is flexible to missing information. *qArtifact* is the artifact about which

[1]http://www.ncbi.nlm.nih.gov/pmc/

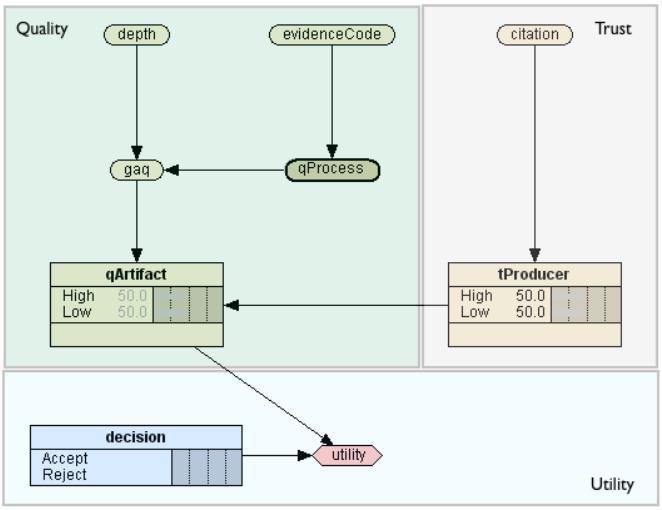

Figure 3: Decision network highlighting separation of quality, trust and utility assessment.

we are making a decision and therefore informs our utility function.

The quality of the process is captured in the GAQ scoring process as the evidence code rank (ECR). We model the *qProcess* as a *deterministic node* representing the evidence code rank given the *evidence code*.

The GAQ score is then a function of the *depth* and *qProcess* variables. We introduce the *gaq* variable as a *definition idiom* - combining two variables in a deterministic manner.

Given the causal relationships in our provenance graph we might have expected to model the network such that: *tProducer* has a causal effect on *qProcess* which in turn has an effect on *qArtifact*. However, *qProcess* is defined as a deterministic value which prevents information flow in decision networks. As a result any extra evidence about *tProducer* would not update qArtifact. To continue to capture the causal relationship we therefore introduce a causal link between *tProducer* and *qArtifact*. We model the domain of both *tProducer* and *qArtifact* as having two states $[high, low]$. The probability of each state is conditional on the parent nodes states. For *tProducer* this is *citation* and for *qArtifact* this is *gaq* and *tProducer*.

4.3.1 Discretization of Continuous Values

The score for *gaq* and *citation* are not discrete states for which we can assign probabilities but instead continuous variables. It is possible for nodes in a decision network to model continuous variables, but introduces the restriction that any child must also be continuous. For the purposes of this initial study we have chosen to discretize the continuous variables for *gaq* and *citation*. To discretize the variable *gaq* we examine the normal distribution function for the score in our result set (see Figure 4). We then partition *gaq* into a series of intervals between the upper and lower limits of the distribution. For *citation* our distretization is more intuitively informed as a simple partitioning based on the maximum citation count (8) observed in our result set (see table 3). Our network provides us with two condition probability distributions to populate $P(tProducer \mid citation)$ and

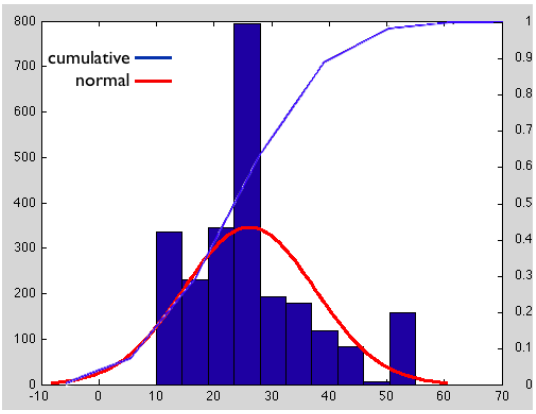

Figure 4: Normal and cumulative distribution of GAQ score for example query set.

$P(qArtifact \mid tProducer, gaq)$.

We continue the intuitive modelling of *citation* populating $P(tProducer \mid citation)$ as follows:

tProducer		
citation	High	Low
unknown	0.5	.05
0	0.5	0.5
1 - 2	0.6	0.4
3 - 4	0.7	0.3
5 - 6	0.8	0.2
>= 7	0.9	0.1

Table 3: *tProducer* probability distribution

Where there is no information or no citations, *tProducer* has an equal probability of being high/low i.e. the reputation is unknown. As the citation count increases so to does our belief in a high reputation and in turn trustworthiness.

qArtifact			
tProducer	gaq	High	Low
...			
High	30 - 35	0.62	0.38
High	35 - 39	0.77	0.23
High	40 - 44	0.88	0.12
High	45 - 49	0.94	0.06
High	55 - 59	0.99	0.01
...			
Low	30 - 34	0.31	0.69
Low	35 - 39	0.39	0.61
Low	40 - 44	0.44	0.56
Low	45 - 49	0.47	0.53
Low	50 - 54	0.49	0.51
Low	55 - 59	0.50	0.50

Table 4: *qArtifact* probability distribution (extract)

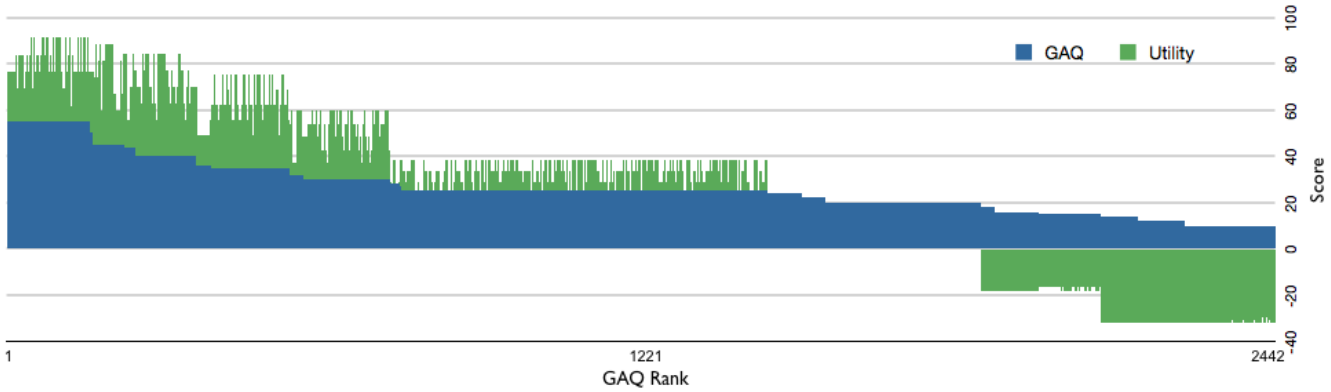

Figure 5: Comparison of annotation score and ranking for GAQ with utility score from the decision network.

To populate the conditional probability table for *qArtifact* (table 4) we are more systematic. We require a value which reflects the increased belief in the quality of the artifact as the gaq score increases. A convenient representation of this is the cumulative distribution for the *gaq* (see Figure 4).
To model

$$P(qArtifact = high \mid gaq, tProducer = high)$$

we populate entries with the cumulative distribution of the *gaq* interval. For,

$$P(qArtifact = low \mid gaq, tProducer = high)$$

the value is simply the inverse probability. Lastly to populate both

$$P(qArtifact = high \mid gaq, tProducer = low)$$
$$P(qArtifact = low \mid gaq, tProducer = low)$$

we wish to reflect the impact $tProducer = low$ has on our belief about the quality of the artifact. Here we have demonstrated with a weighting factor of 0.5.
For both *gaq* and *citation*, by sampling from the query set we achieve a discretization and probability values local to just this query set. Whilst this is sufficient for our ranking task to achieve utility scores comparable across queries we would need to choose a static discretization and values.

4.4 Utility Function and Ranking.

In our network the utility node has two parents, the decision node with the domain $[accept, reject]$ and the quality nature variable with the domain $[high, low]$. Our utility function is then a scoring of the potential outcomes, for example:

$$U(Accept \mid Quality = High) = 100$$
$$U(Accept \mid Quality = Low) = -30$$
$$U(Reject \mid Quality = High) = -50$$
$$U(Reject \mid Quality = Low) = 50$$

From the above scoring we achieve a relative utility such that $Accept \mid High > Reject \mid Low > Accept \mid Low > Reject \mid High$. An inherent advantage of the application of a utility function is the ability to capture subjective utility through alternative scorings of outcomes. A more risk averse user for example might instead prefer the relative utility of $Reject \mid Low$ to be higher than $Accept \mid High$. From our utility function we obtain two expected utility scores, one for *accept* and one for *reject*, where scores are weighted by the probabilities of the outcomes $Quality = High$ and $Quality = Low$.
As an initial illustration of the decision network we provide an example of our utility assessment for an entry in the GO annotations database (Q07812:GO:0006919) demonstrating that for a given annotation with a high GAQ and citation count we obtain a high utility score:

Background:
PMID = PMID:11912183,
$citation = 7$,
$depth = 11$,
$evidencecode$ = IDA,

Assessment:

$$qProcess = 5$$
$$gaq = 55$$
$$P(tProducer = high \mid citation) = 0.92$$
$$P(qArtifact = high \mid gaq = 55, tProducer) = 0.96$$
$$Utility(Reject) = -46.55$$
$$Utility(Accept) = 95.52$$

To further demonstrate the use of the decision network we compare the ranking of a set of annotations by utility with a ranking of the same set of annotations using the GAQ score. We expect the rankings to display a similar overall ranking. However where the GAQ score tends to cluster around values we expect the utility score to provide variation where citation data is available, increasing expected utility. Figure 5 shows the ranking of an annotations results set for the a query (*all annotations from manual experimentation for apoptosis - GO:0006915*) containing 2442 annotations. Ordered by GAQ score the graph also shows the corresponding utility score for each annotation at that GAQ rank. A Spearman's Rank Correlation co-efficient $rho = 0.91$ for the two rankings of this result set indicates that the utility ranking retains a significant overall correlation with the GAQ ranking. However we observe clear local variations where citation count as a measure of reputation impacts the utility

score translating into an increased ranking when compared to GAQ.

In summary these initial results show that we are able to combine individual assessment dimensions and achieve an expected utility suitable for scoring and ranking data. We do not yet know if the utility scores themselves are suitable for making decisions about individual annotations. Further work is required to understand how to both elicit and confirm the appropriate utility scorings. However, we believe the general application of decision networks to be a promising approach.

5. DISCUSSION AND FUTURE WORK

This work has been motivated by a desire to support the scientific user in assessing the quality of data on the Web. Adopting an assessment-oriented approach we have; (1) highlighted a clear separation between the concerns of quality, trust and utility for the scientific user; (2) through a comprehensive review of the literature introduced a classification of quality, trust and utility dimensions, and defined them in terms of the entities required for their assessment; and (3) detailed a mechanism for combining dimensions that respects the causal relationships between the entities that produced the data.

A scientist's decision whether to use data is informed by many different dimensions. Our classification of these dimensions provides a well defined structure for their assessment, supporting the objective view of data quality in science.

We have shown that the application of decision networks presents a promising and intuitive approach to combining individual assessments. The modular and flexible nature of decision networks also complements the multivariate nature of quality, trust and utility assessment. Through the ranking of Gene Ontology annotations we have demonstrate a potential application of utility assessment of scientific data. This work is a starting point from which we wish to further explore the application of decision networks to quality, trust and utility assessment. Where we have achieved an intuitive and manual construction of networks in this study we look to define a formal, automated approach.

The separation of quality, trust and utility also serves to structure future work. Objective quality assessment requires mechanisms for explicating, publishing and discovering quality standards on the Web. To support the subjective utility assessment requires a means of eliciting a utility function from the consumer. Finally, in trust assessment we wish to explore the application of dimensions of trust and reputation currently present in scholarly process.

Like [21] and [24] we have highlighted the key role that related entities and provenance play in quality and trust assessment. We look to ground future work in the Semantic Web technologies where we may take advantage of established provenance vocabularies as well as the available scientific data and linked structure of the Web of Linked Data. Our next step is to expand and apply our approach into online chemistry with a collaboration with the ChemSpider[45] team. The chemistry data sharing landscape is rich and complicated containing a large number of publicly available data sets with varying levels of quality, and varying levels of trust from the community. Interestingly a previous informal survey identified a lack of correlation between quality of the data sets, and the trustworthiness perceived by the community. This highlights a need to apply a systematic approach to quality assessment by exposing objective measures of quality and better support the users in assessing the quality of data.

6. REFERENCES

[1] Open knowledge foundation. `http://okfn.org/`.

[2] OpenWetWare. `http://openwetware.org/wiki/Main_Page`.

[3] Science commons. `http://sciencecommons.org`.

[4] Spatial Data Transfer Standard (SDTS). `http://mcmcweb.er.usgs.gov/sdts/standard.html`.

[5] Big Data Special. *Nature*, 455(7209), Sept. 2008.

[6] Special Issue: Dealing with Data. *Science*, 331(6018):639–806, Feb. 2011.

[7] Data observation network for earth (dataone). `http://www.dataone.org`, accessed 23/3/2010.

[8] J. E. Alexander and M. A. Tate. *Web Wisdom; How to Evaluate and Create Information Quality on the Web*. May 1999.

[9] A. Bairoch, R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L.-S. L. Yeh. The Universal Protein Resource (UniProt). *Nucleic acids research*, 33(Database issue):D154–9, Jan. 2005.

[10] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino. Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, 41(3):1–52, July 2009.

[11] C. Batini and M. Scannapieco. *Data Quality: Concepts, Methodologies and Techniques*. Data-Centric Systems and Applications. Springer Berlin Heidelberg, 2006.

[12] C. Bizer. *Quality-Driven Information Filtering in the Context of Web-Based Information Systems*. PhD thesis, Freie Universität Berlin, 2007.

[13] G. Boulton, M. Rawlins, P. Vallance, and M. Walport. Science as a public enterprise: the case for open data. *Lancet*, 377(9778):1633–5, May 2011.

[14] T. J. Buza, F. M. McCarthy, N. Wang, S. M. Bridges, and S. C. Burgess. Gene Ontology annotation quality analysis in model eukaryotes. *Nucleic acids research*, 36(2):e12, Feb. 2008.

[15] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, and R. Apweiler. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic acids research*, 32(Database issue):D262–6, Jan. 2004.

[16] D. De Roure, C. Goble, and R. Stevens. The design and realisation of the Experimentmy Virtual Research Environment for social sharing of workflows. *Future Generation Computer Systems*, 25(5):561–567, May 2009.

[17] P. N. Edwards. Beyond the ivory tower. "A vast machine": standards as social technology. *Science (New York, N.Y.)*, 304(5672):827–8, May 2004.

[18] M. J. Eppler and P. Muenzenmayer. Measuring Information Quality in the Web Context. In *Proceedings of the 7th International Conference on*

Information Quality (ICIQ-02), pages 187–196, 2002.

[19] D. Field, S.-A. Sansone, A. Collis, T. Booth, P. Dukes, S. K. Gregurick, K. Kennedy, P. Kolar, E. Kolker, M. Maxon, S. Millard, A.-M. Mugabushaka, N. Perrin, J. E. Remacle, K. Remington, P. Rocca-Serra, C. F. Taylor, M. Thorley, B. Tiwari, and J. Wilbanks. Megascience. 'Omics data sharing. *Science (New York, N.Y.)*, 326(5950):234–6, Oct. 2009.

[20] D. Fourches, E. Muratov, and A. Tropsha. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *Journal of chemical information and modeling*, 50(7):1189–204, July 2010.

[21] Y. Gil and D. Artz. Towards content trust of web resources. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(4):227–239, 2007.

[22] R. Giordano. The Scientist: Secretive, Selfish or Reticent? A Social Network Analysis. In *E-Social Science conference*, Ann Arbor, MI, 2007.

[23] J. Golbeck. *Compuing and Applying Trust in Web Based Social Networks*. Phd, University of Maryland, 2005.

[24] O. Hartig and J. Zhao. Using Web Data Provenance for Quality Assessment. In *Proceedings of the First International Workshop on the role of Semantic Web in Provenance Management at ISWC2009*, Washington D.C., 2009.

[25] T. Hey, S. Tansley, and K. Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, Redmond, Washington, 2009.

[26] D. Hull, S. R. Pettifer, and D. B. Kell. Defrosting the digital library: Bibliographic tools for the next generation web. *PLoS Comput Biol*, 4(10):e1000204, 10 2008.

[27] M. A. Jeusfeld, C. Quix, and M. Jarke. Design and Analysis of Quality Information for Data Warehouses. In *Proceedings of the 17th International Conference on Conceptual Modeling*, number 1, 1998.

[28] J. M. Juran. *Quality Control Handbook*. Mcgraw-Hill (Tx), 3rd edition, 1974.

[29] K. Kelton, K. R. Fleischmann, and W. A. Wallace. Trust in digital information. *Journal of the American Society for Information Science and Technology*, 59(3):363–374, Feb. 2008.

[30] U. B. Kjæ rulff and A. L. Madsen. *Bayesian Networks and Influence Diagrams*. Information Science and Statistics. Springer New York, New York, NY, 2008.

[31] S. A. Knight and J. Burn. Developing a framework for assessing information quality on the world wide web. *Informing Science*, 8, 2005.

[32] Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang. AIMQ: a methodology for information quality assessment. *Information & Management*, 40(2):133 – 146, 2002.

[33] P. Lord, A. Macdonald, R. Sinnot, D. Ecklund, M. Westhead, and A. Jones. Large-scale data sharing in the life sciences: Data standards, incentives, barriers and funding models. Technical report, 2005.

[34] A. Martinez and J. Hammer. Making quality count in biological data sources. In *Proceedings of the 2nd international workshop on Information quality in information systems - IQIS '05*, page 16, New York, New York, USA, 2005. ACM Press.

[35] P. Missier. *Modelling and Computing the Quality of Information in e-Science*. PhD thesis, The University of Manchester, 2008.

[36] L. Moreau, J. Freire, J. Futrelle, R. McGrath, J. Myers, and P. Paulson. *The Open Provenance Model : An Overview*, volume 5272 of *Lecture Notes in Computer Science*, pages 323–326. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[37] F. Naumann. *Quality-Driven Query Answering for Integrated Information Systems*, volume 2261 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2002.

[38] F. Naumann and C. Rolker. Assessment methods for information quality criteria. In B. D. Klein and D. F. Rossin, editors, *IQ*, pages 148–162. MIT, 2000.

[39] M. Neil, N. Fenton, and L. Nielson. Building large-scale Bayesian networks. *The Knowledge Engineering Review*, 15(3):257–284, Sept. 2000.

[40] H. Parkinson, U. Sarkans, M. Shojatalab, N. Abeygunawardena, S. Contrino, R. Coulson, A. Farne, G. Garcia Lara, E. Holloway, M. Kapushesky, P. Lilja, G. Mukherjee, A. Oezcimen, T. Rayner, P. Rocca-Serra, A. Sharma, S. Sansone, and A. Brazma. ArrayExpress – a public repository for microarray gene expression data at the EBI. *Nucl. Acids Res.*, 33(suppl_1):D553–555, Jan. 2005.

[41] L. L. Pipino, Y. W. Lee, and R. Y. Wang. Data quality assessment. *Communications of the ACM*, 45(4ve):211–218, Apr. 2002.

[42] M. Scannapieco, A. Virgillito, C. Marchetti, M. Mecella, and R. Baldoni. The architecture: a platform for exchanging and improving data quality in cooperative information systems1. *Information Systems*, 29(7):551–582, Oct. 2004.

[43] R. Wang. A product perspective on total data quality management. *Communications of the ACM*, 41(2):58–65, 1998.

[44] R. Y. Wang and D. M. Strong. Beyond accuracy : What data quality means to data consumers. *Journal of Management Information Systems*, 12(4):5–34, 1996.

[45] A. J. Williams, V. Tkachenko, S. Golotvin, R. Kidd, and G. McCann. ChemSpider - building a foundation for the semantic web by hosting a crowd sourced databasing platform for chemistry. *Journal of Cheminformatics*, 2(Suppl 1):O16+, 2010.

[46] J. Zhao. Guide to the Open Provenance Model Vocabulary. `http://open-biomed.sourceforge.net/opmv/opmv-guide.html`, 2010.

[47] A. S. Zimmerman. New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data. *Science, Technology & Human Values*, 33(5):631–652, Sept. 2008.