

Towards a Quality Model for Effective Data Selection in Collaboratories

Yogesh L. Simmhan, Beth Plale, Dennis Gannon

Computer Science Department, Indiana University, Bloomington IN 47405

{ysimmhan, plale, gannon}@cs.indiana.edu

Abstract

Data-driven scientific applications utilize workflow frameworks to execute complex dataflows, resulting in derived data products of unknown quality. We discuss our on-going research on a quality model that provides users with an integrated estimate of the data quality that is tuned to their application needs and is available as a numerical quality score that enables uniform comparison of datasets, providing a way for the community to trust derived data.

1. Introduction

The increasing ability to sense the world around us due to pervasiveness of wireless networking and sensor technologies has resulted in a proliferation of *data-driven applications* in scientific computing. These applications can be composed of hundreds of inter-connected tasks with ingest and processing routines that handle gigabytes to terabytes of data in a single workflow run [1]. The data products derived from workflow execution are stored in data repositories and shared with collaborators across the virtual organization [2] or in community-wide holdings. This puts scientists in the predicament of having multiple sources from which to select input data for their computational investigations, posing interesting challenges in selecting of the “right” data.

Specifically, in this paper we address the question of: *how does a scientist select the best quality dataset(s) for their application from multiple that qualify?* For example, in meteorology forecasting, model generated data is often used as input to subsequent model runs. Model data with similar uses but varying qualities can be generated by altering the physics or the input data. From our experience in the LEAD [1] collaboration and compute environment, a large number of scientists may be willing to publish model-generated data as a community resource. Current approaches to determine the quality of the data rely on knowledge about an individual’s expertise, or the reputation of an organization. These “soft” qualities do not scale to hundreds to thousands of users.

This challenge drives the need for the quantification of quality metrics that help a user determine “goodness-of-fit” between data and the user’s need. Metrics translate a user’s subjective perception of data quality into a tangible numerical *quality score* that enables uniform comparison of qualities across datasets, promotes creation of better quality data by providing quality feedback to its creators, helps data custodians decide on archive and caching policies based on data quality, and automates searching of the “best” input datasets for workflows at runtime.

Overall, the ability to rate a data’s quality establishes the user’s trust in the data they use.

This paper presents on-going research that addresses the above problems through a quality model for data products in collaboratory environments. It makes the following specific contributions:

1. We identify the four key factors that go into determining the quality of datasets, namely, *intrinsic metadata*, *data provenance*, *quality of service*, and *community perception* of the data.
2. We define a *quality model* and introduce the concepts of *quality profiles* and *quality constraints* that enable users to specify their quality needs in a portable and reusable form.
3. We propose a unique system of *quality metrics* for evaluating the quality factors to translate the subjective user quality perception into a usable *quality score*.

The remainder of the paper is organized as follows: we introduce terms used in this paper in section 2, describe our quality metrics in section 3 and show their role in the quality model in section 4, present our prototype plans in section 5, provide related work in section 6, and explore future work in section 7.

2. Definitions

The quality model we present uses an abstract *metadata model* that is independent of the metadata representation, but we assume it is structured and its schema is known in advance. A metadata document instance comprises of a set of typed name–value(s) attributes. Those attributes upon which users base their expectation of the data are termed *quality factors* [3]. These are abstract high-level indicators understood by the user and map to one or more attributes in the metadata. A *quality metric* is a function that operates over the numerical value of a set of quality factors to produce a quality score based on user constraints. The four quality metrics we define are *intrinsic metadata*, *provenance*, *quality of service*, and *community perception*.

The *quality score* is an interval measurement, based on a standard -7 to $+7$ integer scale, that is an assessment of the degree to which a quality factor holds for the data. It permits counting, ordering, and differences of scores, and also ratios between scores if they are offset appropriately.

Users specify their quality needs using *quality constraints* over metadata attributes, and a collection of these constraints forms the *quality profile* for the user or their application. A quality constraint defines ways to measure and scale a metadata attribute that will result in its quality score, and the importance assigned to that attribute. Quality profiles can import quality constraints

```

switch (AuthorCName) {
  case IN {'Bugs', 'Daffy'} : return qualityScore 7;
  case == 'Elmer' : return qualityScore -3;
  default : return qualityScore 0;
} && return weight 0.3;
switch (provenanceProcess) {
  case == 'WRFSim' : return qualityScore provenanceScore();
  default : return qualityScore -7;
} && return weight 0.1;
switch (transferTimeSecs) {
  case < 10 : return qualityScore 7;
  case > 60 : return qualityScore -7;
  default : return qualityScore 0;
} && return weight 0.4;
return qualityScore expertUserScore() && weight 0.2;

```

Figure 1. Sample user-defined quality constraints on Intrinsic Metadata (Author), Provenance (Process), QoS (timeliness) and Community perception (Expert's score).

from a parent quality profile to better manage and reuse the profiles, and scale its evaluation. This allows users to have a global quality profile, define application-specific quality profiles that inherit from it, and share their profiles. The quality profile encapsulates the query and ranking information needed to perform the data discovery.

3. Proposed Quality Metrics

We propose four classes of metrics for evaluating the quality of scientific data, namely, *intrinsic metadata*, *data provenance*, *quality of service (QoS)*, and *community perception*, with a quality score evaluated for each and combined to form an overall quality score. These classes are similar to content/cognitive, economic, and social information filtering techniques [4], with provenance introduced as a novel filter. Each of these is a function over a (potentially overlapping) subset of attributes of the metadata model. These attributes can come from different sources and need not be physically present in a single metadata document. A sample set of user-defined quality constraints given in **Error! Reference source not found.** describes measurement and aggregation rules specified on the metadata attributes. These constraints are evaluated over the metadata of each dataset to determine its quality score.

Intrinsic Metadata. The intrinsic metadata quality score makes use of the intrinsic metadata of the data, which are the inherent properties that can be discovered by having physical access to the data [5]. Some examples include author, published date, data source, domain keywords, and quality control flags. Since these attributes describe heterogeneous characteristics of the dataset, their semantic meaning is apparent only to the user. Users specify quality constraints over these metadata attributes to indirectly measure their quality score and provide weights for aggregation. This metric is similar to conventional search services that match metadata

properties through boolean queries and, in addition, we include weights to make the results more meaningful.

Provenance. Searching over provenance [6, 7], the derivation history of the data, is less intuitive and has not been investigated in detail. Provenance is an important quality metric since the derivation process has significant implications on the data's quality, and errors introduced by faulty data tend to inflate as they propagate to data derived from them. This is especially so in workflows, which execute several processes, generating intermediate and final data products whose quality depends on preceding workflow steps. The process that generated the data, its configuration parameters, and the input data form the provenance metadata, and the metric is a function over them. The input data is concisely represented in the metadata model by its quality score, which captures its different quality facets. The intrinsic metadata attributes of the deriving process and non-data parameters are other attributes used by the provenance metric.

We address the non-intuitive nature of specifying constraints on provenance attributes by using machine-learning techniques to aggregate them into a quality score. The *decision tree* inductive machine-learning technique [8] is well suited to classifying discrete and continuous valued attributes into quality scores. The tree, associated with the deriving process, is constructed by supplying it sample provenance metadata of datasets generated using the same process and having known quality scores. We use just the intrinsic quality scores for the sample datasets to circumvent the bootstrapping problem. Once constructed, the decision tree can automatically determine the provenance quality score (the *provenanceScore*() function in Figure 1) for datasets derived from that process using the provenance attributes.

Quality of Service. The quality of service metric for a data product present in a data repository measures the ability to access the data product and transfer it to a remote location for an application's use at a certain resource cost [9, 10]. The accessibility of a dataset depends on the *availability* of the repository, its *reliability*, the *timeliness* of the data transfer (dependent on the *dataset size* and transfer *throughput*), and *access restrictions* on the data product, and is bound by the *cost* affordable for that resource. While quality of service has been used by resource brokers for resource scheduling and replica selection [11], they are usually not exposed to the user, restricting the user's ability to trade-off the quality of service with other quality factors of the dataset during data selection.

Since all attributes relevant to quality of service are numerical, in a naive case, this metric uses the product of reliability and availability with the weighted sum of transfer time and resource cost, given the prerequisite that the data can be "purchased" is met. Other complex quality constraints based on a function over these attributes can be provided by the user's quality constraints. We envisage

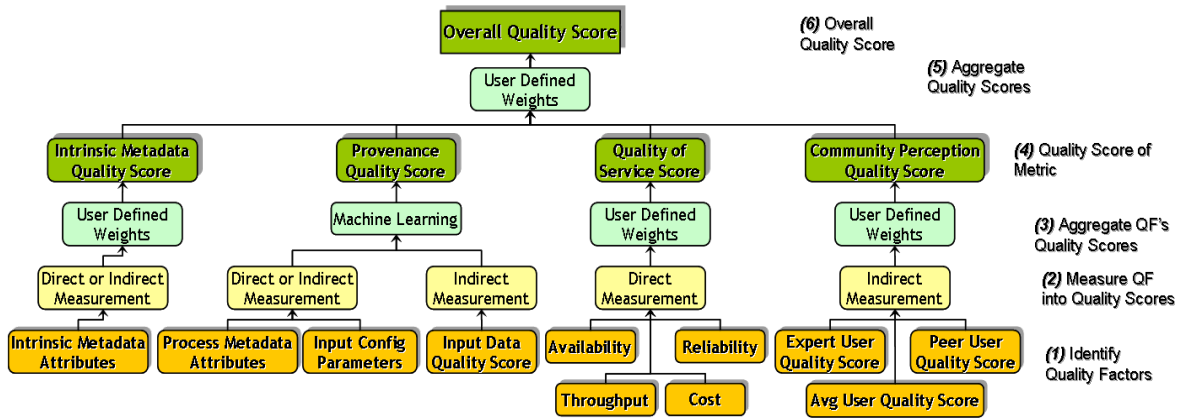


Figure 2. Quality model to derive overall quality score from quality factors, showing measurement and aggregation techniques used by each quality factor

leveraging existing resource brokers that can provide a quality of service rating for the data products.

Community Perception. For the purposes of defining the community quality constraint for a user, community members can broadly be classified into *experts*, *peers*, and “everyone” based on their domain knowledge. While experts have superior knowledge, peers fall under the same skill level as the user. Each of these three groups may have rated a particular dataset through their quality profiles and these scores form the quality factor for the dataset. Users can weigh these pre-computed scores (e.g *expertUserScore()* in Figure 1), possibly based on the skill level of the respective groups, to get an aggregate. In addition, users can use the *data usage frequency* of that group to weigh their quality scores, or consider the usage frequency as a separate quality factor. This provides a simple, yet useful, capability to share data ratings with the community that is especially relevant to beginner members, such as students.

4. Quality Model

The quality model is a data model that relates the quality factors with ways to measure and aggregate them into a quality score using metrics. Figure 2 illustrates the quality model that derives the overall quality score for a data product using the various quality factors and their metrics.

Before they can be used within a quality metric, quality factors need to be quantified by measuring them either *directly* (if numerical value is provided), or *indirectly* (by translating from discreet attribute values to numbers), and then scaling them to the quality score scale. For example, the availability of the data is a quality factor that is directly measured as the percentage of time that the data is accessible online, while the existence of certain keywords in a metadata attribute is an indirect measure that needs to be quantified using user directions. While directly measured values may have linear scaling or step-wise scaling based on value ranges, indirect measurement requires further user involvement to translate semantic, higher-level terms to quality score equivalents. This user input can come in the form of

explicit ratings given to data products, by studying *user behavior*, or through *user-defined rules*. In the former, users rate individual datasets on the quality score scale but the score applies to all attributes of the data, not just a single quality factor. This has more user overhead than a study of their behavior that collects data usage patterns and converts them to quality scores. However, neither are effective for a large data space and, instead, we focus on users defining quality rules that apply to entire classes of data. Rules defined over metadata attributes map their values or a range of values to a certain quality score. For example, as in figure 1, a user may specify that data created by members ‘Bugs’ and ‘Daffy’ are trustworthy and give them a rating of +7, while those from other members are given a rating of 0. In addition to matching ranges, users can also specify aggregations over array values and simple mathematical functions such as offsetting and scaling of the values to get the quality score. Figure 1 has pseudo-code for sample rules on the different metadata attributes, that each returns a quality score and a weight.

Quality metrics use different techniques to combine the quality scores from individual quality factors into a quality score for that metric. The naïve solution averages all the quality scores, and an improvement to it allows *user-assigned weights* for each quality factor (Figure 1). This allows users to decide how important each attribute in the metadata is for their application and expresses the metric as a linear combination of quality factor scores. When the aggregation is not as obvious, *machine learning* techniques can be used to derive a linear or non-linear function over the attribute values by training it against known quality score for data products, as discussed for the provenance metric. The *overall quality score* is a weighted aggregation of quality scores from each metric and forms the “bottom-line” quality score that is most inclusive of the user’s quality constraints.

5. Implementation

We are presently prototyping a data quality framework for the Linked Environments for Atmospheric Discovery (LEAD) [1] meteorological research and education

project. As part of the framework, we are building a data quality broker that accepts users' quality constraints, illustrated in Figure 1, and evaluates the data quality score for raw and derived data products generated from weather simulation workflows. The broker is intended to work with external metadata providers, such as resource brokers providing quality of service information, metadata repositories that hold intrinsic metadata [1], and provenance services [12] that supply provenance and data usage information. Some of the potential hurdles we foresee are in dynamically aggregating user-specific quality scores for hundreds of thousands of data products, continuous gathering of data availability and throughput statistics, and in sampling active data and workflows when building the decision trees. We are hoping our experience with designing the quality model and implementing it in a Grid environment will address some of the data quality issues and expose further research avenues.

6. Related Work

Data repositories in Grids have *metadata catalogs* [9, 10, 13] that can be searched for datasets. Several workflow systems use them for runtime data and resource selection using users' search terms. These catalogs usually restrict queries to the intrinsic metadata of the data product without the ability to holistically search on provenance and quality of service metadata too. They are also constrained to sorting the results on attribute fields, lacking sophisticated ranking means.

DaQuinCIS [14] is a data quality architecture that uses feedback on published data to evaluate the credibility of the publisher, similar to our community perception metric. Data search results are ranked on the trust in the data provider, accuracy dimensions, and some QoS attributes, but no quality metrics or scores exist. They do not consider provenance as a quality factor either.

Data quality management techniques for business information systems [15] have proposed quality dimensions to evaluate information quality, but no automated systems exist to collect and rate the information dynamically. *Statistical quality control* techniques that determine the accuracy and errors in data are commonly used for raw instrument data but are less amenable to derived data. When present, they can be used as quality factors for the intrinsic metadata metric.

7. Ongoing Work

Estimating the data quality of derived data scientific products is exploratory research that poses unique challenges in collaborative workflow environments. The publishers of data products may wish to advertise the quality of their data through quality scores, and ways to resolve expected and perceived data qualities need study. Users familiar with keyword searches need intuitive ways of defining quality constraints that satisfy their

requirements, with the potential for automating this process. Finally, there remains the open question of how best to evaluate the effectiveness of the quality metrics and accuracy of the quality scores in meeting user's data needs.

8. References

- [1] B. Plale, D. Gannon, D. Reed, S. Graves, K. Droegemeier, B. Wilhelmson, and M. Ramamurthy, "Towards Dynamically Adaptive Weather Analysis and Forecasting in LEAD," in *ICCS workshop on Dynamic Data Driven Applications*, 2005.
- [2] I. Foster, C. Kesselman, and S. Tuecke, "The Anatomy of the Grid: Enabling Scalable Virtual Organizations," in *International Journal of Supercomputer Applications*, vol. 15, 2001.
- [3] N. F. Schneidewind, "Software Metrics Model For Quality Control," in *IEEE METRICS*, 1997, pp. 127-136.
- [4] T. W. Malone, K. R. Grant, and F. A. Turbak, "The information lens: an intelligent system for information sharing in organizations," in *CHI*, 1986, pp. 1--8.
- [5] S. Weibel, J. Godby, E. Miller, and R. Daniel, "OCLC/NCSA Metadata Workshop Report," in *The OCLC/NCSA Metadata Workshop I*, 1995.
- [6] Y. Simmhan, B. Plale, and D. Gannon, "A survey of data provenance in e-science," in *SIGMOD Record*, vol. 34, 2005, pp. 31-36.
- [7] R. Bose and J. Frew, "Lineage retrieval for scientific data processing: a survey," in *ACM Comput. Surv.*, vol. 37. New York, NY, USA, 2005, pp. 1--28.
- [8] T. M. Mitchell, "Machine Learning," 1997.
- [9] I. T. Foster, J.-S. Vöckler, M. Wilde, and Y. Zhao, "The Virtual Data Grid: A New Model and Architecture for Data-Intensive Collaboration," in *CIDR*, 2003.
- [10] R. W. Moore, A. Jagatheesan, A. Rajasekar, M. Wan, and W. Schroeder, "Data Grid Management Systems," in *IEEE/NASA Conference on Mass Storage Systems and Technologies (MSST)*, 2004.
- [11] S. Vazhkudai, J. M. Schopf, and I. T. Foster, "Predicting the Performance of Wide Area Data Transfers," in *IPDPS*, 2002.
- [12] Y. L. Simmhan, B. Plale, D. Gannon, and S. Marru, "A Framework for Collecting Provenance in Data-Centric Scientific Workflows," in *Submitted to Intl WWW Conference*, 2006.
- [13] E. Deelman, J. Blythe, Y. Gil, C. Kesselman, S. Koranda, A. Lazzarini, G. Mehta, M. A. Papa, and K. Vahi, "Pegasus and the Pulsar Search: From Metadata to Execution on the Grid," in *Applications Grid Workshop, PPAM*, 2003.
- [14] M. Scannapieco, A. Virgillito, C. Marchetti, M. Mecella, and R. Baldoni, "The DaQuinCIS architecture: a platform for exchanging and improving data quality in cooperative information systems," in *Information Systems*, vol. 29. Oxford, UK, UK, 2004, pp. 551--582.

[15] R. Y. Wang, M. Ziad, and Y. W. Lee, "Data Quality," 2002.