A data quality framework applied to e-government metadata

A prerequsite to establish governance of interoperable e-services

Per Myrseth, Jørgen Stang and Vibeke Dalberg
DNV Sustainability and Innovation
Det Norske Veritas
Oslo, Norway
per.myrseth, jorgen.stang, vibeke.dalberg@dnv.com

Abstract—**Data quality is becoming increasingly important in information critical systems and the data, also coined information product (IP)[1] is frequently considered as an item with equal importance to the physical product. Several frameworks for both data quality processes and data quality methods have been defined and are routinely applied to transactional and master data systems. However, it has been suggested that the same principles could also be successfully implemented for metadata repositories[7]. This article outlines how data quality processes and a data quality framework both based on semiotics[2] and Wang[1] are being applied to monitor and improve the content in an e-government metadata repository. In the work described here, the term *data* is used to denote data instances whereas *metadata* denotes the reusable data definitions.**

Keywords - metadata, metadata quality, e-government, data quality, quality processes, ontology management

I. Introduction

As governments establish e-government services and open up to share and utilize the vast potential value of collected public data, both new and commonly accepted measures for data governance and data quality are needed[3][4][5][6]. To maximize the benefits we rely on sharing and reusing definitions and terms. Data quality and data governance issues need to be addressed to ensure a coherent operation on data of known and monitored quality. The private sector has faced and are facing similar challenges and several cases have been described for how to define corporate wide data governance regimes and frameworks to assuage data rework and waste, further, the consequences of not addressing these issues have been estimated both as damage to brands and as direct costs.

The potential cost savings of achieving reuse of terms and data across governmental bodies will both boost productivity and leverage novel uses of the distributed data sources. Open and shared (and trusted) governmental data can be used as reference for other applications as an integral part of a data validation regime.

The work presented here describes a quality framework for metadata and how it fits into the work processes of establishing and maintaining metadata in a national metadata repository. We believe the framework is useful for actors responsible for metadata within a single domain. Harmonization between domains could be done by a slightly different process, but with a similar quality framework. Our motivation for defining the framework is the repeating challenges learned in our work with data quality and inter-connected public services where metadata very frequently is used to measure data quality, to enable integration of online services and to assure compliance to regulations. We therefore need metrics defining and measuring the quality of the metadata.

The repository used in our case study is designed to be populated by metadata of three different levels of abstraction. These are: (i) term level, for describing the meaning of a term, (ii) structure level, which contain groups of terms frequently reused, and (iii) implementation level containing terms and structures implemented in actual individual e-services. The data quality of the metadata in the repository is monitored continuously both within and across domains using processes designed to support a data quality maturity model using both generic indicators and business rules to support data quality metrics for syntactic, semantic and pragmatic[2] evaluation of the data. Dashboard visualization is used to provide snapshots in time, trends and high level drilldown points. The process described helps those responsible for the metadata to define the level of metadata quality needed to meet certain business goals and data quality goals.

Grimstad et al [10] argues that top management needs to get involved in metadata governance and funding. Our metadata measurements, visualizations and trend analysis can help management understand the status and progress in metadata governance. On a general basis we believe our quality framework is independent of the actual case study and the metadata repository used as a test bed.

II. Background

The Norwegian government has initiated several activities to capture common terms and their meaning as used in private sector reporting obligations and across governmental functions such as handling of taxes, social benefits and national registers. The intention is both to improve the cooperation between departments as well to make the government - public interface more efficient and less prone to duplicated and inconsistent data. The public interface, digital or paper based, is used as a source for identifying terms and their meaning. Well defined terms and concepts are used as dynamic building blocks for domain specific taxonomies and even more formal ontologies. This, in turn, is intended to support a wide adoption of the terms, both within the domain but also across domains.

We use the word *term* to capture the definition and the relationships between the definitions.

The underlying data model of the repository content tends to be complex and this in turn advocates for best practices and modeling rules to ensure a consistent and compatible implementation. The entire process of preparing, capturing, building, implementing and governing the model definition and model implementation is manual, and hence there will be discrepancies in and a varying quality of the content entered in the metadata repository. Some of the quality issues can be addressed by measuring the compliance to the defined best practices and rules, however, metadata quality indicators are also required to alert users of potential errors or low quality by measuring statistical outliers, general completeness and integrity.

The ultimate goal of the metadata quality process is to improve the users trust in the metadata repository and to make sure the metadata is of known quality which can be traced over time to produce current status, goals and trends. The visualization of the metadata quality must be comprehensible and made available and appropriate to all levels of users and provide a good connection to underlying processes that supports both the reactive short time cleansing as well as proactive upstream process improvements. The current regime was used to notify users of deteriorating quality within a particular function thus preventing any implementations based on the flawed metadata.

III. E-GOVERNANCE METADATA

The metadata model defined for this particular e-governance collaboration project attempts to model each governmental function as a separate domain where common cross-functional terms (such as person) are identified and administered in a core area accessible to all domains. As shown in figure 1, each separate domain implements a three layered structure (term – structure - implementation) which will typically be developed in a top-down and/or bottom-up manner. Terms which are identified as reusable across the domain are defined as individual items in the *terms* layer (eg. person, name and address), terms which can be connected to form new useful aggregated entities are identified and linked in the *structure* layer, and domain specific implementations of either terms and/or structures are defined in the *implementation* layer. In the top-down modeling approach, domain experts will define general terms which should form a basis for the structure and implementation layers, whereas in the bottom-up approach the terms and structures will be derived from the actual implementations. The chosen approach will vary between domains and will to a certain degree be dependent on the expert resources available, however, it is desirable that metadata models are balanced in the sense that neither terms nor implementations are overrepresented. This criteria will also be used as a part of the metadata quality indicator measures as described in a later section.

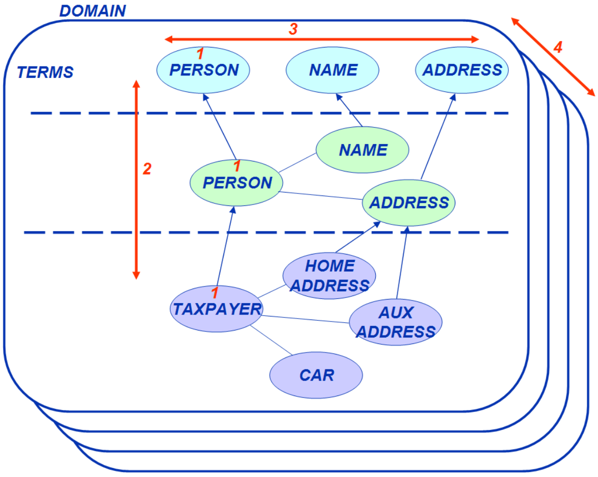

Figure 1. Metadata definitions and quality measures

IV. METADATA MANAGEMENT PROCESSES

Establishing and maintaining metadata models in a national metadata repository should be a lifecycle process, ensuring the natural evolution of the metadata models. Existing literature in metadata models engineering typically focuses on parts of the process, often with less focus on maintenance, evolution and adoption[14]. Therefore we propose adapting parts of the theory behind ontology evolution[13], even though a metadata model is not necessarily a formal ontology. Automatic and manual metadata quality tests should be performed at various stages in the lifecycle process; when evaluating existing metadata models for potential reuse, when considering harmonisation of metadata, during validation and verification, as an initiator to change and evolution, and as a reporting facility to management and budget processes in the enterprise.

V. METADATA QUALITY

The metadata quality can be measured along dimensions similar to those commonly used for data quality[1], such as *completeness*, *integrity*, *accuracy, consistency* and *timeliness*. At this stage in the work, we also partially adopt the semiotic data quality framework[2] where dimensions are categorized into *syntactic*, *semantic* and *pragmatic* data quality measures. On a general note on data quality, syntactic refers to validations with respect to a predefined schema and/or set of programmatic rules, semantic applies to conformance with the immaterial object or real world physical objects the data intends to represent, whereas pragmatic denotes the users perceived quality of the data. In the context of metadata quality, syntactic still applies to schema compliance, however, the real world mapping typically performed by semantics could be substituted with measures for how well the metadata represent the (sometimes abstract) entities they are meant to represent. Further work is required to map this out in more detail, the work described here has so far identified provenance as a pragmatic[2] quality metric and this is measured by traceability of documentation source. The provenance metric (traceability) is fundamental when establishing and rating trustworthiness of external sources. Also, trusted reference sources (trusted surrogates) are identified to measure metadata accuracy, i.e. conformance to other metadata repositories, legal definitions, international standards etc. The pragmatic category for metadata should capture the general suitability of the definitions, i.e. measure how well we are able to reuse terms, if the set of terms is perceived by the user to be complete enough and at the right granularity to handle a defined task, and to perform mappings across domains. In addition, two separate regimes are implemented; *metadata profiling* and *metadata quality measures,* also commonly termed indirect and direct data quality. The metadata profiling activity provides indicators of quality issues but does not offer an absolute threshold or definition for those values that represent an error. The missing values or anomalies detected by profiling provide candidates for the absolute data quality measures where strict definitions of errors are required. Typically the error definitions are formulated in a service level agreement (SLA).

Figure 1 shows the meta model definition, as described previously, as well as possible levels of metadata quality measures (numbered 1-4 in the figure, 5 and 6 are added here);

1) *Measures for individual terms, structures and implementations to determine missing or inaccurate values. E.g. all terms should have a source to verify authenticity and reliability, and so terms can be measured on completeness and tracability.*

2) *Measures across the* terms – structure – implementation *axis will be able to determine usage characteristics such as the imbalanced hierarchies mentioned previously. If a term is defined and never used in the structure or implementation layers, this represents network redundancy, similarly, if several implementations define similar items, this could indicate a potental term abundancy or term overlap. Examples include the structure entities* **Person, Name** *and* **Personname.** *A name matching algorithm could be applied to suggest that* **Personname** *is redundant and that the implementations should rather use* **Person** *and* **Name***.*

3) *Measures within each layer (terms, structures and implementations) can identify possible definition overlaps or other inconsistencies. By use of experience we will be able to use predefined graph patterns and anti-patterns to identify eventual problems or new recommended patterns. Figure 2 shows how edit distance[9][12] can be used to classify terms based on the spelling characteristics. Terms with edit distance close to 1 are substantially different whereas a value of 0 indicates term overlap and should be further investigated. The same method can be used to calculate similarities between terms and reference data to measure compliance with standard terminology.*

4) *The most challenging measures will span all domains and possibly use semantic matchmaking[8] to identify core model candidates. Also, generic measures from each domain can be displayed in a domain wide dashboard. This will potentially uncover gaps and highlight different practices and a relative level of maturity.*

5) *To support trending and measuring the impact of any alleviating activities, measures will also support the time dimension. Each metadata quality metric is considered as a key performance indicator (KPI) yielding a status, trend and target value.*

6) *If the metadata are represented or can be exported in a formal ontology language (e.g. OWL) semantic analyses at a more advanced level can be performed.*

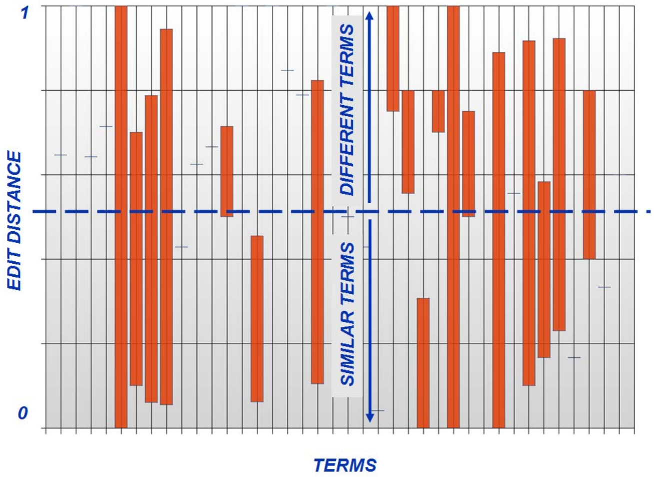

Figure 2. Edit distance applied to terms to identify overlaps

Currently, only measures for type 1 and 2 from the above list have been implemented. Table 1 indicates the type of metadata quality category that is supported for each metadata model layer. Syntactic measures include measuring the presence of required attributes and general schema compliance, semantic includes traceability of sources and pragmatic includes the usefulness of the defined terms for users involved in modeling the implementation layer. The syntactic measures can be performed automatically, the semantic measures both automatically and by sampling, and the pragmatic category is measured by frequency of use and by user feedback.

TABLE I. SEMIOTIC CLASSIFICATION OF DATA QAULITY MEASURES BY METADATA LAYER

	SYNTACTIC	*SEMANTIC*	*PRAGMATIC*
TERMS	✓	✓	✓
STRUCT.	✓		
IMP.	✓		

Pragmatic and semantic measures have not yet been identified for the structure or implementation layers, however, as the implementation matures the matrix is expected to include measures for all categories across all layers. For the syntactic category all three levels can be validated with respect to the repository schema and metadata modeling guidelines.

We suggest using a methodology to identify *semantic drift* [11] in our metadata repository based on usage of concept signatures (text vectors) that are constructed on the basis of how concepts (the definitions of terms in our repository) are used, linked and described. By comparing term/concept signatures for one term at different snapshots in time, we see how the semantics of a term evolves and how its relationships to other terms gradually reflect these changes.

VI. CASE STUDY

A number of syntactical measures have already been implemented in the *terms* layer and reporting devices such as charts and feedback loops are in place to handle metadata quality issues. In a current and real case completeness is measured continuously for a particular domain. In this case, completeness is defined as the existence of *definitions* and *documentation* for each defined term. This is required information as domain experts rely on the given definitions and documentations to assess terms for consistency and overlap. Normally, domains will show a gradual improvement for this measure as definitions and documentation are added as the domain matures, however, the completeness measure indicated that all definitions and documentation were missing. This was due to a trivial transformation error and was routinely fixed for the next load, however, the measurements enabled prompt notification of the relevant resources; the developers were notified so that they could identify the error and the domain modelers were made aware of the malfunction and the possibility of side effects.

As the metadata quality measures are developed further and are implemented as an integral part of the domain modeling process it is expected that the measures can be used both reactively to notify errors (as shown here), but also as a proactive tool to identify best practices and support both modeling teams and users. Also, consistency measures can aid the development of candidates for core definitions used across domains.

VII. SUMMARY

High quality metadata is a prerequisite for successful governance of e-services. We argue that metadata quality should be subjected to the same data quality regimes commonly implemented for master data and transactional data.

The article has illustrated how the quality for e-government metadata can be successfully monitored along several dimensions using the semiotic data quality framework for defining *syntactic*, *semantic* and *pragmatic* data quality metrics. Also, metadata quality processes have been suggested to support all *reactive*, *proactive* and *continuous* measurements. In particular, the syntactic metadata quality metrics have been implemented to measure metadata *terms*, *structures* and *implementations* whereas the semantic metrics have been applied to detect inconsistencies between metadata categories and domains as well as to measure provenance (source traceability).

The time dimension is added to the measurements to support trending and continuous improvement processes. Further work is required to evaluate and improve the current reporting and process support tools, however, even within the relatively short time the described framework has been active, an example is given indicating the potential usefulness of monitoring simple metrics to warn owners and users of metadata domains suffering acute metadata quality issues.

Further work is required to be carried out to develop and benchmark the tools and methods described here.

ACKNOWLEDGMENT

This research was carried out as part of the Semicolon project (project no. 183260/S10), funded by the Norwegian Research Council, Det Norske Veritas and other Semicolon participants.

REFERENCES

[1] Y.W. Lee, L.L. Pipino, J.D. Funk and R.Y. Wang, "Journey to Data Quality", The MIT Press, 2006

[2] R.J. Price and G. Shanks, "Empirical Refinement of a Semiotic Information Quality Framework", Proceedings of the 38th Hawaii International Conference on System Sciences, 2005

[3] T. Margaritopoulus, M. Margaritopoulus, I. Mavridis and A. Manitsaris, "A Conceptual Framework for Metadata Quality Assessment", Proceedings of the International Conference on Dublin Core and Metadata applications, 2008

[4] X. Ochoa and E. Duvall, "Quality Metrics for Learning Object Metadata", Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications, 2006

[5] N. Dushay and D.I. Hillman, "Analyzing Metadata for Effective Use and Re-use", Proceedings of the Dublin Core Metadata Conference, 2003

[6] J. Barton, S. Currier and J.M.N. Hey, "Building Quality Assurance into Metadata Creation", Proceedings of the Dublin Core Conference, 2003

[7] D. McGilvray, "Executing Data Quality Projects", Morgan Kaufman, 2008

[8] G. Shu, O.F. Rana, N.J. Avis and C. Dingfang, "An Ontology-based Semantic Matchmaking Approach", Advances in Engineering Software, vol 38, January 2007

[9] C.D. Manning, P. Raghavan and H. Schutze, "Introduction to Information Retreival", Cambridge University Press, 2009

[10] T. Grimstad and P. Myrseth, "Information Governance and Metadata Strategies as a Basis for Cross-sector e-Services", Accepted at ICEE 2011

[11] J. A. Gulla et. al., "Semantic Drift in Ontologies", WEBIST 2010 - International Conference on Web Information Systems, 2010

[12] J. Stang, T. Christensen, D. Skogan, A. Kvalheim and T.A. Ihrgens, "A Generic Data Quality Framework Applied to the Product Data for Naval Vessels", Presented at the International Conference on Information Quality (ICIQ), MIT, Boston, 2008

[13] Eds J. Davis, R. Studer, P. Warren, "Semantic Web Technologies. Trends and research in ontology-based systems", Wiley 2006.

[14] H. Suonuuti, "Guide to Terminology", 2nd ed. Nordterm 8. 2001, 2nd ed ISBN 952-9794-14-2, 2001