A Framework for the Evaluation of Business Models and its Empirical Validation

Jean-Paul Van Belle
Department of Information Systems, University of Cape Town, South Africa
jvbelle@commerce.uct.ac.za

Abstract: This article describes a proposal for a framework to evaluate and compare enterprise models. It suggests three major categories for grouping the model evaluation criteria: syntactic, semantic and pragmatic analysis. The paper draws on a wide literature to present a large selection of criteria and to operationalise their measurement by means of several possible metrics. As an empirical validation test, a selection of metrics for eight of the criteria has been calculated for fifteen large enterprise models. Their interpretation supports the usefulness and validity of the overall framework. Various attempts at deriving a composite overall quality score are discussed, but there is less confidence in the validity of this approach.

Keywords: Enterprise modelling, evaluation framework, system analysis metrics.

1. Introduction

The trend towards model-driven development and the growing automation of much of systems programming, increases our reliance on high-quality models. Furthermore, the move from functional "silo" applications to enterprise-wide systems has widened the scope of business domain models to integrated enterprise-wide models. This move is amplified by the need to build enterprise-wide data warehouses to satisfy the managers' desire for integrated information. Modelling, and enterprise modelling in particular, has become serious business.

But, despite the abundance of evaluation and comparison frameworks for development and modelling *processes* (*conf.* methodology engineering), there appears to be a relative dearth of guidance available on how to evaluate the actual *output* of the modelling activity: how does one evaluate a completed enterprise model?

This article proposes a comprehensive, systematic yet interdisciplinary framework for the evaluation of domain models. It also addresses the concern of empirical testing of the proposed framework. This is done by selecting a representative sample of criteria within the framework – those that could most easily be automated – and testing them against fifteen medium-sized enterprise models. The intent is not to present a final and conclusive set of criteria for the evaluation of models, but rather to illustrate the wide range of possible metrics as well as the obstacles which must be given consideration when comparing models.

2. Research objective

The purpose of this paper is the empirical validation of a comprehensive framework for the analysis and evaluation of enterprise models. No satisfactory framework for enterprise model evaluation was found in the literature, although a number of candidate frameworks for the evaluation of other "intellectual products" exist. These are used as inputs to guide the building of an integrated and comprehensive framework. Since the development of a theoretical framework does not contribute towards science unless it is accompanied by a verifiable and methodological testing and evaluation of the framework itself, an empirical validation of the framework is also presented. For this, a "test bed" was constructed, consisting of fifteen publicly available enterprise models from different reference disciplines.

The framework includes *intrinsic* qualities (absolute measures that can be computed for one specific model) and *comparative* qualities (relative measures that compare models). Some of these entail a ranking or judgement (better, worse) whereas other measures merely differentiate (e.g. model A is more like model B, whereas models C, D and E form a separate family). Because of practical and methodological reasons, the emphasis was on *static* models. However, a parallel and completely independent research effort developed a similar framework to cater for dynamic models (Taylor 2003). Also, although the framework focuses on enterprise models, it is quite possible to use the framework for the evaluation of models from *other domains* such as embedded systems, or specific functional areas within the enterprise.

Finally, the framework was developed from an *interdisciplinary* perspective. Since enterprise models themselves originate from such diverse sources as systems theory, computer science, ERP, accounting, linguistics and systems engineering, the proposed framework sought to incorporate contributions from those areas and

Reference this paper as:
Van Belle J-P (2006) "A Framework for the Evaluation of Business Models and its Empirical Validation" *The Electronic Journal Information Systems Evaluation* Volume 9 Issue 1, pp 31-44, available online at www.ejise.com

others, such as construction architecture, complexity theory and aesthetics.

3. Prior research

A rich body of literature exists on evaluation criteria for models. These stem from a variety of reference disciplines: methodology engineering, systems engineering, ontology research, modelling, etc. Criteria can be arranged in a flat (i.e. unstructured sequential) list, in a hierarchical tree structure or in the form of a framework based on some theoretical structuring concept or theory.

Many authors have suggested lists of criteria to evaluate models (e.g. Benyon 1990:66, Korson & McGregor 1992, Claxton & McDougall 2000, Halpin 2001). Additional criteria are those that define high-quality data, such as those proposed by Orli (1996) and Courtot (2000). Ontology researchers have also proposed their own criteria for enterprise models (e.g. Fox 1998, van Harmelen 1999, Noy 2001). Williams (1996) lists 45 requirements for enterprise reference architectures.

A few authors go beyond unstructured lists and organise their evaluation criteria into frameworks, usually in the context of evaluating the quality of modelling approaches and methodologies (e.g. Khaddaj 2004, Brazier 1998). Structures which organise the different criteria can also be matrix presentations of quality factors; these are often used in a software engineering context: McCall's quality factors, the very similar Böhm model and Gillies' hierarchical quality model (Böhm 1976, Gillies 1997).

Most frameworks suffer from the *grounding problem*: they lack an underlying theoretical or philosophical basis for the framework dimensions. Although it is acknowledged that the frameworks may still be valuable and valid – as long as they are based on the principles of soundness and completeness – some authors (e.g. Frank 1999) stress the value and importance of a strong theoretical grounding for an evaluation framework. The framework proposed below roots itself firmly in the discipline of semiotics and draw on a fundamental distinction which has proved valuable in many different contexts.

4. The model evaluation framework

The proposed framework will be developed in the following sequence. Firstly, the two major dimensions, or the conceptual structuring principles, will be explained and motivated. Next, the framework will be "populated" with evaluation criteria. Finally, some sample metrics will be presented for the various criteria.

4.1 Framework dimensions or structuring principles

The first and major dimension is grounded in linguistics, information and communication theory and semiotics. The key distinction used in the framework is the fact that all models – or indeed any informational object – have a syntactic, semantic and pragmatic aspect (see Table 1).

- *Syntax* refers to the type of constructs and the legal ways of combining them i.e. the physical or logical tokens which represent the information..
- *Semantics* refers to the meaning: the sense of the information obtained by interpreting the token or signifier.
- *Pragmatics* refers to the context i.e. considerations, issues and background information which influences or moderates the interpretation of the information.

Table 1: Main classification dimension of proposed analysis framework.

Classification concept	Related terms and mappings
Syntax	Symbols, form, shape, structure
Semantics	Meaning, denotation, sense
Pragmatics	Background, situation, context

The *syntactic* analysis deals with the structural model relations, i.e. shape and form of the entities and their relationships and groupers. It treats all entity and relationship names, therefore, as mere "alphanumeric labels". The *semantic* analysis of models is concerned with the intrinsic *meaning* of the model i.e. the relationship with and mapping to the underlying domain reality that the model represents. The essence of semantic analysis is to unravel the *meaning* of the name (label, word, token) used for a specific model element (entity, relationship, grouper). Put another way, semantic analysis is concerned with the correspondence (mapping, projection, validity) between the model (abstract or intellectual construct) and the underlying domain (reality). Whereas syntactic analysis is fairly technical and easy to automate, semantic analysis involves the more tricky matters of meaning and interpretation and thus lends itself not quite as easily to objective and/or automated analysis. *Pragmatic* model analysis requires the consideration of information regarding the use, environment or context of the model i.e. information outside the model. The analysis techniques falling under this heading include the face validity, degree of use, authority of model author, availability, cost, flexibility, adaptability, model currency, maturity and support. Most analysis relies on the searching and ranking of certain specific information details, often involving a degree of subjective interpretation and an understanding of business issues.

Interestingly, Fabbrini (1998) used the same distinction between semantics, syntax and pragmatics in the context of assessing the quality of a software architecture, but he interpreted the terms quite differently and did not operationalise their approach in any practical way. Equally, Brinkkemper (1996:214) makes reference to the doctoral research of John Venable whereby, in his search for a theoretical basis for the CoCoA methodology, he distinguished his criteria for the evaluation of conceptual data models between those relating to semantic concepts, syntactic constructs and more generic requirements such as views and modularization. These are by no means the only explicit references of the three categories in the context of information systems analysis: the fairly well-known publications from both Benyon (1990) and Stamper (1987) made references to these distinctions, and most computer science research in formal (programming) languages uses the distinction between syntax and semantics extensively. In particular, Stamper proposes his *semiotic framework* to classify information on six different levels: three on the human information level: social world, pragmatics, semantics; and another three on the IT platform: "syntactics", empirics and physical world (1995). He also maintains that too much of the research focuses on the "syntactic" elements. Finally, a parallel but independent research effort concerning the quality of process models uses the same framework distinctions between syntactic, semantic and pragmatic issues [Taylor 2003].

A second organising principle, summarized in Table 2, is not as clear-cut and is presented here mainly as an ordering mechanism within each column of the framework. It will not be explored further and is left for future research purposes. It is proposed that, within each category above, measures range from "absolute" to "relative". For some criteria factors, an absolute value can be established ("more is better") whereas for others, no ideal value can be determined *ex ante*, since the desired target value depends on factors not directly related to the intrinsic nature of the model. It must be recognized that this distinction is not a categorical classification but rather a continuum: there is a fuzzy cross-over area in the middle where it is relatively difficult (or arbitrary) to decide whether a given quality measure is absolute or relative.

Table 2: Possible second dimension for proposed analysis framework.

Classification concept	Related terms and mappings
Absolute measures	Theoretical; "das Model an Sich"; the model as object; objective standards; intrinsic qualities; technical factors; "Conforms to specification"; computer science; academic.
Relative measures	Applied; "das Model für Uns"; the model as subject; subjective standards; extrinsic qualities; business factors; "Fit for purpose"; information systems; practitioner.

4.2 Populating the framework with detailed criteria

Table 3 lists the proposed model evaluation criteria within the framework structure. Criteria are grouped into conceptual clusters, which can be sub-divided if one wishes. Note that only simple or primitive evaluation criteria are listed here. Composite concepts consist of sub-criteria and can therefore often not be allocated to one specific framework cell. An example is *usability*, which includes all of the pragmatic and many semantic and syntactic criteria. Another example is model *dependability*, as discussed by Barbacci (1995).

The various criteria are drawn from different sources, which often ascribe different meanings to certain criteria and, conversely, different authors sometimes use different terms to describe a similar concept. To indicate this overlap in meaning, criteria were grouped into "clusters".

Table 3: Populated framework for model analysis.

	Syntactic	Semantic	Pragmatic
Absolute	Size	Genericity: universality & technical independence	Validity: authority & user acceptance
	Correctness; error-free; integrity; consistency	Completeness (domain coverage); conciseness; efficiency	Flexibility; expandibility; portability; adaptability
	Modularity; structure; hierarchy.	Expressiveness Similarity and overlap with other models	
Relative	Complexity; density	Perspicuity; comprehensibility; understandability; self-descriptiveness	Price; cost; availability Support
	Architectural style	Documentation	Purpose; goal; relevance; appropriateness

4.3 Proposed metrics for the evaluation criteria.

The empirical validation of the framework requires that the each of the evaluation criteria can be measured or calculated. In order to operationalise the framework, many measures were adopted from the rich literature whilst others were newly developed. Although many more were calculated than shown, only the ones that were found to display some validity are listed in Table 4 which also presents the final proposed version of the framework.

Table 4: Summary of validated framework metrics and measures.

	Criterion	Suggested metric / measure
SYNTACTIC	Size	CASE (concept) count and adjusted CASE count
	Correctness; error-free; integrity; consistency	Syntax error, consistency and standards level score
	Modularity	Number of groupers, group levels and diagrams
	Structure; hierarchy	Multiple inheritance; mean inheritance depth, reuse ratio.
	Complexity; density	Relative connectivity; average fan-out; plot of Fruchterman-Reingold (for similar-sized models); harmonic mean of fan-out; fan-out distribution (chart); fan-out model signature.
	Architectural style	Layout aesthetics
SEMANTIC	Genericity	% mapping to domain
	Coverage	Domain coverage score; core concept coverage
	Completeness	Ranking of absolute lexicon coverage
	Efficiency; conciseness	Relative lexicon coverage
	Expressiveness	Average expressiveness score
	Similarity & overlap with other models	Plot of similarity coefficients; most similar neighbours; similarity dendogram; most important concepts.
	Perspicuity; comprehensibility; understandability; readability	Normalised rank-adjusted weighted perspicuity count based on user lexica
	Documentation	Completeness, extensiveness, readability (Flesh Reading Ease score)
PRAGMATIC	Validity: authority & user acceptance	Academic author citations
	Flexibility; expandability; adaptability	Composite flexibility score
	Currency; maturity	Descriptive table & taxonomy
	Purpose; goal; relevance; appropriateness	Descriptive table
	Availability	Medium & status
	Cost	Purchase cost
	Support	Tool & vendor support, user base

Formatted: Bullets and Numbering

4.4 Sample of enterprise models

In order to demonstrate the feasibility of using the framework to evaluate real-world, industrial-strength systems, fifteen medium-sized to large enterprise data models were captured. Models had to have at least 100, preferably more than 200 entities and needed to be publicly available. They are grounded in a wide variety of reference disciplines, but the prototypical example of an enterprise data model is the model underlying Enterprise Resource Planning (ERP) systems.

The following gives a brief overview of the generic enterprise models which were selected, grouped according to their reference discipline (Van Belle, 2002a). The database with the models is available in XML format for research purposes on request.

Two ERP models underlying the leading integrated enterprise applications were captured: the *SAP* R/3 (Sheer, 1998) and *BAAN* IV. The latter was "re-engineered" from Perreault (98). Four generic data model libraries were captured from their respective published books: *Hay* (1996) and *Silverston* (1997; 2001); Marshall's *BOMA* (Marshall, 2000) and *Fowler*'s analysis patterns (Fowler, 1997). Two smallish academic enterprise reference models were found: *Purdue*'s Reference Model for CIM and *ARRI*'s Small Integrated Manufacturing Enterprise Model in IDEF0 notation (Williams, 1991) in DFD notation. As an example of a data warehousing model, *Inmon*'s set of high and mid-level data models was gleaned from his website. Two framework derived models were *AKMA*'s Generic DataFrame and IBM's San Francisco "*SanFran*" (predecessor of WebSphere). Finally, three enterprise ontologies were selected: the Enterprise Ontology developed by the *AIAI* in Edinburgh (Uschold, 1998), *TOVE* from EIL in Toronto and a subset of the *CYC* Upper Ontology was created containing all organisation and enterprise-related concepts.

Although the meta-model used for capturing the above models contains 7 meta-entities and numerous meta-attributes, the only three meta-entities used below are the meta-concepts of "*entity*" (sometimes called *object* or *concept*; their graph-analytic equivalent is a *node* or *vertex*), the "*grouper*" construct and "*relationship*" (equivalent to a *connection, link, edge* or *arc*). The latter can be subdivided into hierarchical structuring relationships of the "IS-A" type (reflecting specialisation/generalisation) and "proper" domain relationships reflecting some semantic or domain-inspired link between entities.

4.5 Syntactic analysis

The syntactic analysis deals with the purely structural aspects of the model. Its analysis techniques are mainly derived from software engineering and computer science. This includes a large variety of standard syntactic metrics relating to size, grouping, layering, inheritance structure, and network visualisation, as well as some less standard metrics such as interface aesthetics (i.e. the visual beauty or tastefulness of graphically drawn models). Perhaps the most obvious criterion is model size. Many *size* measures were calculated of which three are listed in table 5: the total number of entities or classes in the model, the CASE size or concept count (the number of entities + relationships + grouper elements) and an expanded concept count whereby additional meta-modelling elements such as attributes are also taken into account. Although the measures are very correlated coefficient, the expanded CASE size is preferred because it favours the more fully specified models above the shallower models. For instance, the badly specified *Inmon* model now drops its relative ranking, although there is still no accounting for model documentation and description. It is perhaps somewhat surprising to see the SAP and BAAN models rank behind the data models published by Hay and Silverston but both the CYC and TOVE ontologies are indeed very sizeable.

A second criterion is model *correctness.* Models were rated using a composite "correctness score" consisting of a score for the amount of errors, the degree of (in)consistency and the use of and adherence to notational standards. A typical correctness issue is two diagrams showing different cardinalities for a given relationship between the same entities .

Formatted: Bullets and Numbering

The "scoring system" used for both errors and consistency problems gave 3 for no problems or errors, 2 for minor, 1 for medium and 0 for major problems. In addition, a score between 0 up to 2 was allocated for adhering to stringent standards for naming, diagramming etc. The combined "correctness score" thus ranges from 0 to 8; a higher the score indicating a more correct model. Not surprisingly, the well-validated SAP model achieves the highest score with most of the other well-known models following immediately after. Some lesser-known individual research models obtain relatively low scores.

Table 5: Syntactic model analysis.

	Size			Correctness				Complexity					
Model	Nr of Entities	CASE Size	Expanded CASE Size	Accuracy	Consistency	Standards	Combined Score	Cyclomatic Complexity	Relative Connectivity	Average Fan-Out	De Marco's Data Bang	Average Data Bang	(Harmonic) Mean Fan-out
AIAI	94	270	510	2	3	2	7	30	1.82	3.32	220	4.99	1.81
AKMA	82	565	769	1	2	2	5	6	1.15	2.18	160	2.85	1.54
ARRI	128	430	790	2	2	1	5	79	2.09	3.31	592	4.97	1.81
BAAN	328	1086	1927	2	2	2	6	377	2.29	5.24	2018	8.7	2.23
BOMA	183	552	770	3	2	2	7	65	1.68	3.00	557	4.35	1.81
CYC	777	2623	4537	2	2	2	6	511	2.32	3.60	3507	5.49	1.94
Fowler	120	375	579	2	2	2	6	37	1.67	2.76	372	3.92	1.71
Hay	291	1292	3465	2	3	2	7	491	3.13	6.17	2470	10.5	2.42
Inmon	427	2429	2670	1	1	1	3	17	1.08	2.14	682	3.03	1.22
NHS	269	751	1460	0	3	2	5	48	1.7	2.54	622	3.6	1.52
Purdue	106	343	866	0	0	1	1	136	2.11	5.03	711	7.99	3.82
SAP	396	1218	1917	3	3	2	8	285	1.97	3.73	1851	5.64	2.30
SanFran	109	332	532	1	0	0	1	74	1.68	3.47	520	5.25	1.95
Silverston	267	1269	2235	2	3	2	7	114	1.51	3.08	950	4.55	1.76
TOVE	564	1937	2042	2	1	2	5	678	2.28	4.51	3876	7.19	2.33

Model *complexity* has many interpretations: Edmonds (1999) calculates 48 different syntactic complexity metrics. In an attempt to minimize the influence of "size", model complexity in this framework refers to the average "density" of the model network. Of the various *complexity* metrics which were calculated for the models; Table 5 lists only McGabe's (1976) cyclomatic complexity, relative connectivity (including the inheritance structure), average fan-out; and De Marco's data bang (Shepperd 1995). However, none of these appear to convey the subjective feeling of model network density well. However, *the frequency distribution of the entity fan-outs* for each model were found to yield a distinctive and characteristic signature of the underlying model complexity (see Van Belle, 2002b). The most descriptive statistic for this average "density" of the model network turned out to be the *(harmonic) mean of the fan-out distribution*.

Inmon's low complexity is typical of data warehousing conceptual models whereas the ontologies and financial models score relatively high. Since it is relatively easy for a small, compact model to achieve relatively high complexity (e.g. Purdue), the harmonic mean fan-out is best compared among similarly sized models e.g. SAP and Baan exhibit similar complexity as do the fairly comparable BOMA, Fowler and Silverston.

4.6 Semantic analysis

Semantic model analysis refers to the relationship of a model with the domain it is representing. This type of analysis is grounded in linguistics, ontology research and lexicography. Much of the analysis concentrates on similarity, correspondence and cluster analysis. It proved to be a challenge to eliminate subjectivity from the metrics thus preference was given, where feasible, to automatically calculated or computer-generated measures.

Perhaps the most straightforward criterion is the *expressiveness* of a model. This depends on the richness of the modelling language used and is determined by looking at the number of (different) modelling language constructs used by each model. An expressiveness score can be calculated as the weighted index of the number of meta-model attributes covered in a model. In our case, the metric used the following expressiveness qualities (unless otherwise specified, a weight of 1 was applied): degree of formality (weight of 3); diagrams; directed graph; use of generalisation; depth of inheritance tree; multiple inheritance; number of grouper levels (x2); entity definitions; entity examples; entity attributes; relationship names; relationship role names (x2); relationship cardinalities; relationship types; definitions for relationships/groupers; and the use of additional constructs such as constraints.

It is no surprise that the three ontologies in the sample (CYC, TOVE and AIAI) – which use semantically very rich languages - and the object-oriented models (BOMA, SanFran) all score very high. The exact composition of the expressiveness metric can be modified to suit the ultimate requirements of the model analysis.

Model *perspicuity* and *readability* refer to the extent to which the model can be understood or comprehended by the intended users or readers of the model and how self-describing the model is. The perspicuity analysis was based on matching all model element names against common domain vocabulary lists. Although several business lexicons were investigated, the use of Someya's (1999) well-validated corpus-based business language list annotated with word frequency statistics yielded the most valid results, especially if slightly more sophisticated wordlist preparation and matching algorithms are used. The GPC ("Gross Perspecuity Count") measures what percentages of model element labels exist in the business word list, whereas the NRAWPC Normalizes for the size of the model, is Rank-Adjusted for word use frequency and applies a Weighting to concatenated or multiple word labels.

Formatted: Bullets and Numbering

Table 6: Semantic and pragmatic model analysis.

	Expressiveness		Perspicuity		Completeness		Authority		Flexibility			
Model	**Raw Expressiveness**	**Weighted Expressiveness**	**GPC**	**NRAWPC**	**Completeness1**	**Completeness2**	**Google Page-Rank™ for model**	**PageRank™ for organisation URL**	**Digitally available?**	**Customizable/ reusable**	**Implementation independence**	**Overall flexibility score**
AIAI	10.5	**13.0**	85%	**68%**	80	**272**	7	**7**	Yes	Some	Low	**1.50**
AKMA	8.8	**10.0**	94%	**75%**	74	**233**	5	**5**	No	No	High	**1.00**

	Expressiveness		Perspicuity		Completeness		Authority		Flexibility			
ARRI	7.3	**8.5**	86%	**77%**	121	**346**	3	**5**	No	No	Med.	**0.50**
Baan	6.3	**8.0**	95%	**81%**	235	**636**	0	**7**	Yes	Some	Med.	**2.25**
BOMA	9.7	**12.0**	91%	**77%**	156	**452**	0	**4**	Yes	Some	High	**2.25**
CYC	9.5	**12.0**	89%	**74%**	590	**1143**	6	**7**	Yes	Some	Med.	**2.25**
Fowler	6.8	**8.5**	88%	**68%**	100	**336**	0	**7**	No	Yes	High	**2.00**
Hay	9.3	**12.0**	93%	**76%**	201	**574**	5	**5**	No	Yes	High	**2.00**
Inmon	6.5	**7.5**	91%	**76%**	356	**840**	4	**6**	No	Some	Med.	**1.00**
NHS	8.3	**9.0**	86%	**70%**	144	**398**	4	**6**	Yes	No	Med.	**1.50**
Purdue	6.5	**7.5**	93%	**79%**	116	**383**	4	**6**	No	Limited	Med.	**0.75**
SanFran	7.7	**9.5**	90%	**76%**	99	**310**	5	**9**	Yes	Yes	High	**3.00**
SAP	8.8	**10.5**	94%	**82%**	236	**632**	0	**8**	Yes	Some	Med.	**2.25**
Silverston	8.8	**11.5**	95%	**81%**	141	**461**	0	**5**	Yes	Yes	High	**3.00**
TOVE	9.5	**12.5**	77%	**60%**	226	**571**	4	**6**	Yes	Some	Med.	**2.00**

It is again encouraging to find the measure validated by the high ranking of the well-known ERP models as well as the published data models. Models with obscure or even obtuse language score very low. It is, of course, possible to adopt a more simplistic approach by using a more general and more easily computable readability index such as a Flesh Reading Ease or Flesch-Kincaid Grade Level score. However, the results of this were found to be far less valid (see Van Belle 2004).

A third criterion attempts to measure the semantic equivalent of the syntactic size metric: model *completeness.* It measures how much of the domain has been covered. It requires an accepted and complete description of the model domain against which each model can be mapped. Since this is hardly ever available, an approximation will normally be used. Here we used the same business lexicons as were used for the perspicuity analysis and calculated how many distinct concepts within the business lexicon are covered – the more words or concepts that were covered, the better. The analysis was then enhanced by using an intermediary translation process with synonyms as found in WordNet ("completeness2") to enable the mapping of meanings instead of word tokens. Not surprisingly, the larger models tend to be more complete although some interesting observations can be made. The ERP and data warehousing models cover most of the business domain. But the second-largest (syntactic size) model, *Hay*, drops dramatically in position when looking at completeness. This is a strong motivation in favour of complementing standard syntactic measures with semantically based metrics.

4.7 Pragmatic analysis

Pragmatic model analysis, as defined above, is concerned with criteria which cannot be assessed purely on the basis of the information contained within the model, but which require the consideration of information regarding the use, environment or context of the model i.e. information outside the model. Unfortunately, most pragmatic criteria involve a substantial degree of subjectivity. Also, by definition, many criteria depend on the actual purpose for which the evaluation is carried out. For this reason, the discussion is limited to the two pragmatic measures which are believed to be more objective and universally applicable. In a real-world model analysis, several other criteria are likely to be important and should be included on a as-needed-basis, e.g. cost, tool support etc.

Model *authority* refers to the acceptance of the model by practitioners in the field. Depending on the publishing medium of the model, metrics that are reasonably easy to collect are the relative sales ranking for book-based models (e.g. by Amazon.com), and popularity of a web page by counting the number of external hyperlinks to the page for web-based models, e.g. using the Google PageRank™ system. Another pragmatic proxy used in the commercial world is *authoritative validity* i.e. the reputation of the authoring person, team or organisation. This can be measured by the number of author citations which measures and ranks the academic standing or authority of the lead author associated with the model relatively accurately, especially within the same reference disciplines. Table 6 lists the PageRank metrics for both the webpage listing the model as well as the home page of the author (organisation). The highest authority is accorded to large commercial organisations such as IBM (SanFran) and the ERP systems, as well as the better known research organisations (ontologies).

Model *flexibility* is concerned with how well models can be changed or *adapted* to different situations. A composite flexibility measure was calculated which awarded a score for three aspects of flexibility: the model's availability in digital format, its customisability or reusability and its implementation independence. Models designed specifically as templates for model

Formatted: Bullets and Numbering

development such as Silverston and SanFran scored maximally. The ontologies and ERP models also scored high, whereas the more obscure academic models (ARRI, Purdue) tended to score well below average.

4.8 Composite model evaluation score

It is a natural question to investigate whether an overall score can be calculated, combining the values of the various component metrics to capture, perhaps, the essence of overall model quality. In an attempt to achieve this, those metrics which appeared to be the most valid or representative measure for each criterion in this research were selected from tables 5 and 6. These were expanded CASE size, correctness, harmonic mean of fan-out for complexity, weighted expressiveness, NRAWPC for perspicuity, completeness2, Google PageRank™ for website authority and the flexibility composite score as explained above. In each of these eight measures, each model was given a ranking from 1 to 15 against the other models based on its scores, giving rise to a vector of 8 ranking scores for each model.

This vector of rankings for a given model was processed in four different ways:

- The *average* ranking could be calculated.
- A model's *median* ranking was calculated
- The number of times that a model was ranked in the "bottom *half*" of the models was deducted from the number of times that it featured in the "top half" of the models.
- The number of times that a model was ranked in the top *quartile* minus the number of times it found itself in the bottom quartile.

Formatted: Bullets and Numbering

Note that this procedure ignores relative weighting of criteria as well as the absolute differences in the underlying scores. However, this section is merely intended as a first-cut evaluation of the overall feasibility of calculating a composite score. Table 6 details the results. The section with "calculated scores" gives the raw values of the above four calculations. Since these scores are not normalised, they provide little information. Therefore, the last four columns convert the raw scores into a relative position by ranking each model from 1 to 15.

Table 7: Composite model ranking.

	Calculated (or Raw) Scores				Overall Ranking based on …			
Model	**Average Rank**	**Median Rank**	**#[R<8] - #[R>8]**	**#[R<5] - #[R>11]**	**Average Rank**	**Median Rank**	**#[R<8] - #[R>8]**	**#[R<5] - #[R>11]**
AIAI	8.3	9	-1	0	**9**	**9**	**10**	**8**
AKMA	11.3	11.5	-7	-4	**15**	**14**	**15**	**15**
ARRI	10.0	10.5	-5	-1	**14**	**13**	**13**	**9**
BAAN	5.1	4	6	3	**4**	**3**	**2**	**5**
BOMA	6.9	6.5	2	2	**7**	**6**	**7**	**6**
CYC	4.4	3	6	5	**2**	**1**	**2**	**2**
Fowler	9.6	11.5	-2	-3	**12**	**14**	**11**	**14**
Hay	4.9	4	6	4	**3**	**3**	**2**	**3**
Inmon	9.1	9.5	0	-2	**10**	**10**	**8**	**12**
NHS	9.9	9.5	-5	-2	**13**	**10**	**13**	**12**
Purdue	9.1	9.5	-2	-1	**10**	**10**	**11**	**9**
SAP	3.6	3.5	8	6	**1**	**2**	**1**	**1**
SanFran	8.1	8	0	-1	**8**	**8**	**8**	**9**
Silverston	5.5	5	4	4	**5**	**5**	**5**	**3**
TOVE	6.8	6.5	4	1	**6**	**6**	**5**	**7**

The first observation is that the rankings are relatively robust, regardless of which procedure is used. In fact, it was found that even including additonal measures (i.e. more than 8 metrics) does not necessarily affect the relative ranking of a model significantly (Van Belle, 2004).

A more detailed evaluation can be made by looking at the overall rankings of the models, *regardless of reference discipline*. The *top scoring* models are: Scheer's SAP reference model, CYC's enterprise sub-ontologies, Hay's data model and Baan's reference model. From a subjective evaluation of these models, based on their study throughout this research, a strong case can be presented that they indeed represent the best overall models in the database, not only in their respective disciplines but also from an overall quality perspective. It is a particularly interesting vindication (and validation) of the framework that these models originate from three fairly different reference disciplines, yet the framework allows an interdisciplinary comparison in spite their fundamentally different modelling

approaches and philosophies. They are followed by a set of "close seconds", namely Silverston's data models, the TOVE ontology and Marshall's BOMA patterns.

Similarly, the models ranked at the *bottom of the scale* are indeed the "worst" models from a qualitative perspective. AKMA is a fairly shallow model. ARRI is a tiny model, focussing on a manufacturing environment. NHS is a vertical model which is very specific to the health care industry. Although large, Inmon appears to be an inconsistent and extremely shallow model. Again, it is interesting that the framework manages to identify problematic models, regardless of the underlying reference discipline.

An alternative way of analyzing the rankings is by comparing the relative positions of models *within the same reference discipline*. Between the two ERP models, SAP fairly consistently beats or matches Baan across various criteria. It must be stressed that the Baan model here is not the original conceptual model but a model re-engineered from a description of its relational implementation. This guarantees a quality handicap so the results should not be read as reflecting on the actual ERP packages. Nevertheless, there is support for the contention that the SAP model has a better theoretical foundation as well as represents significantly more analysis effort.

Comparing the data models, it must be admitted that the quality difference between Silverston and Hay is a tough call. However, of the more "pattern-like" models, BOMA is definitely cleaner than SanFran, bearing in mind that the author experienced significant methodological problems in capturing SanFran. The lagging position of Fowler, however, is very debatable and it may well have been penalized because of its highly conceptual nature.

The comparative scores for the ontology-based models correspond perfectly with the amount of ontology engineering and analysis effort invested in each of the models. CYC represents by far the most effort, followed by TOVE and then by AIAI. However, although smaller, AIAI is a more homogenous and conceptually higher level model, which is perhaps not fully reflected in the score. Within the CIM models, it must be recognized that although ARRI is a much cleaner, more correct and rounded model than Purdue, the latter is a better representation of the enterprise domain and represents significantly more modelling effort than ARRI.

This section illustrates that that the interpretation or even overall value of a composite index is limited. This reinforces an early comment from the field of system engineering metrics: "[C]alculating and understanding the value of a single overall metric for […] quality may be more trouble than it is worth. The major problem is that many of the individual characteristics of quality are in conflict; added efficiency is often purchased at the price of portability, accuracy, understandability, and maintainability." (Böhm 1978:ix)

5. Conclusion

The overall research objective was to present and empirically validate a framework for the comparative evaluation of enterprise models. This research suggests that the proposed framework is a highly productive and valid approach for evaluating enterprise models. As shown in Addendum 1, almost all of the evaluation criteria which have been suggested in the literature can easily be accommodated by the framework.

The most useful dimension within the framework is the separation of syntactic, semantic and pragmatic criteria or factors. Each of these sets of criteria has a very distinct tone, reflecting a certain paradigmatic approach. Syntactic analysis has a strong pure science and engineering flavour, drawing heavily from computer science metrics, systems engineering graph-theoretical and even architectural concepts. Semantic analysis relies mainly on lexicography and computational linguistics, as well as more conceptual information sciences such as meta-analysis or information frameworks. Finally, pragmatic analysis, focused on practical business or commerce issues such as support, pricing, purpose, organizational impact etc. The framework thus brings together the basic constituent reference disciplines of information systems.

The framework was validated empirically using a sample consisting of fifteen relatively large enterprise models. Eight criteria were selected from the framework for which a number of possible metrics, including some newly proposed ones, were calculated.

Overall, it was quite easy to generate fairly valid metrics for the framework criteria. Interestingly, it was possible to calculate fairly objective metrics for semantic and pragmatic analysis, although care has to be taken with the interpretation of the results.

A combined overall model ranking, intended to represent some type of composite quality index,

appears to have some but fairly limited face validity.

Apart from the value of the framework to *classify* existing criteria, the framework should also be seen as an ideal way for the *creative design or generation* of new criteria or measures. Indeed, many metrics suggested in this research were inspired by thinking about model evaluation along the dimensions identified by the framework.

There is still considerable scope for future research and development of the framework. The development of more refined or alternative metrics, especially for the pragmatic analysis, would be useful. A more theoretically valid approach to combining individual criteria into composite metrics, such as model quality or usability, is also still an open research question, although this may well prove to be an illusive goal. More emphasis could also be placed on structural relationships (such as inheritance) and grouper constructs. Finally, the framework should be empirically validated within other modelling domains.

Formatted: Bullets and Numbering

References

Avison, D.E. & Fitzgerald, G. (1995) Information Systems Development: Methodologies, Techniques and Tools. McGraw Hill: London.

Barbacci, M.R.; Klein, M. H.; Longstaff, T. A. et al. (1995) Quality Attributes. Technical Report 95-TR-021, Software Engineering Institute.

Benyon, D. (1990) Information and Data Modelling. Blackwell: Oxford, UK.

Böhm, B. et al. (1978). Characteristics of Software Quality. Elsevier North-Holland: New York.

Brazier, F.M.T. & Wijngaards, N.J.E. (1998) "A Purpose Driven Method for the Comparison of Modelling Frameworks".In Proceedings of the Eleventh Workshop on Knowledge Acquisition, Modeling and Management, Banff, Canada, 18-23 Apr 1998.

Brinkkemper, S.; Lyytinen, K. & Welke, R.J. (Eds.) (1996) "Method Engineering: Principles of Method Construction and Tool Support" In Proceedings of the IFIP TC8, WG8.118.2 Working Conference on Method Engineering, Atlanta, Georgia, 26-28 Aug 1996.

Chapman & Hall, London.Courtot, T. (2000). What to Look for in Packaged Data Models/Databases. In Proceedings of the Meta Data Conference, Arlington, Virginia, 19-23 Mar 2000.

Claxton, J.C. and McDougall, P.A. (2000). Measuring the Quality of Models. The Data Administration Newsletter, 2000:14. [Online] http://www.tdan.com/i014ht03.htm.

Edmonds, B. (1999) Syntactic Measures of Complexity. Doctoral Thesis, University of Manchester.

Fabbrini, F.; Fusani, M. & Gnesi, S. (1998) "Quality Evaluation based on Architecture Analysis" In Proceedings of the International Workshop on the Role of Software Architecture in Testing and Analysis (ROSATEA'98), Marsala, Italy, 30-Jun to 3-July 1998.

Fowler, M. (1997) Analysis Patterns. Addison-Wesley: Reading (MA).

Fox M.S. & Gruninger M. (1998) "Enterprise Modelling" The AI Magazine, Fall 1998: 109-121.

Frank, U. (1999) "MEMO: Visual Languages For Enterprise Modelling" Arbeitsberichte des Instituts für Wirtschaftsinformatik (Universität Koblenz-Landau) Nr 18.

Gillies, A. (1997). Software Quality: Theory and Management. Thomson: London.

Halpin, T. & Bloesch, A. (1999) "Data Modeling in UML and ORM: A Comparison." The Journal of Database Management, 10 (4): 4-13.

Hay, D.C. (1996) Data Model Patterns. Dorset House: New York.

Khaddaj, S. & Horgan G. (2004) "The Evaluation of Software Quality Factors in Very Large Information Systems" Electronic Journal of Information Systems Evaluation, 7 (2): 43-48.

Korson, T. & McGregor, J.D. (1992) "Technical Criteria for the Specification and Evaluation of Object-Oriented Libraries" Software Engineering Journal, 7 (3): 85-04.

Marshall, C. (2000) Enterprise Modelling with UML. Designing Successful Software through Business Analysis. Addison-Wesley, Reading (MA).

McGabe, T.J. (1976) "A Software Complexity Measure" IEEE Trans. Software Engineering, 2 (Dec 1976): 308-320.

Ngo, D.; Chek L., Teo, L. et al. (2000) A Mathematical Theory of Interface Aesthetics. Unpublished working paper. [Online] http://www.mi.sanu.ac.yu/vismath/ngo/.

Noy, N.F. & McGuinness, D.L. (2001) Ontology Development 101: A Guide to Creating Your First Ontology. SMI technical report SMI-2001-0880.

Orli, R.; Blake, L.; Santos, F. & Ippilito, A. (1996) Address Data Quality and Geocoding Standards. Unpublished report. [Online] http://www.kismeta.com/Address.html.

Oxford. (1979) The Oxford Paperback Dictionary. Oxford University Press: Oxford.

Perreault, Y. & Vlasic, T. (1998) Implementing Baan IV. Que: Indianapolis, Indiana.

Scheer, A.-W. (1998) Business Process Engineering. Reference Models for Industrial Enterprises. Springer-Verlag: Berlin (2nd ed).

Shepperd, M. (1995) Foundations of Software Measurement. Prentice-Hall: London.

Silverston, L., Inmon W.H. & Graziano, K. (2001) The Data Model Resource Book. A Library of Universal Data Models For All Enterprises. J. Wiley: New York (2nd ed).

Someya, Y. (1999). A Corpus-based Study of Lexical and Grammatical Features of Written Business English. Masters Dissertation, Dept of Language and Information Sciences, University of Tokyo.
Stamper, R. (1997) "Semantics. Critical Issues" in Information Systems Research, Boland, R.J. & Hirschheim, R.A. (Eds.), J. Wiley: Chichester, 43-78.
Taylor, C. & Sedera, W. (2003) "Defining the Quality of Business Process Reference Models" in Proceedings of the 14th Australasian Conference on Information Systems (ACIS), Perth, 1-3 Dec 2003.
Uschold, M., King, M., Moralee, S. & Zorgios, Y..(1998) The Enterprise Ontology. The Knowledge Engineering Review, Vol. 13. Special Issue on Putting Ontologies to Use.
Valente, A. & Breuker, J. (1996). Towards Principled Core Ontologies. In Proceedings of the Tenth Workshop on Knowledge Acquisition for Knowledge-Based Systems, Banff, Canada, Nov 1996.
Van Belle, J.P. (2002a) A Survey of Generic Enterprise Models. Proceedings of the 32nd Annual SACLA Conference, Port Elizabeth (South Africa), June 2002.
Van Belle, J.P. (2002b) Towards a Syntactic Signature for Domain Models: Proposed Descriptive Metrics for Visualizing the Entity Fan-out Frequency Distribution. Proceedings of the SAICSIT Conference, Port-Elizabeth (South Africa), Sep 2002.
Van Belle, J.P. (2004) "A Framework to Evaluate the Quality of Information System Models." Ingénierie des Systèmes d'Information, Special Issue on IS Quality, 9 (5) (in publication).
Van Harmelen, F. & Fensel, D. (1999) "Practical Knowledge Representation for the Web" in Proceedings of the IJCAI-99 Workshop on Intelligent Information Integration, Stockholm, Sweden, 31 Jul 1999.
Williams, T.J. (1996) "The Needs of the Field of Integration" in Architectures for Enterprise Integration. Bernus, P.; Nemes, L. & Williams, T. (ed). Chapman & Hall, London: 21-31.
Williams, T.J. (ed). (1991) A Reference Model For Computer Integrated Manufacturing (CIM). A Description from the Viewpoint of Industrial Automation. CIM Reference Model Committee, International Purdue Workshop on Industrial Computer Systems, the Instrument Society of America: Research Triangle Park (North Carolina).

Addendum: Evaluation Criteria Mapped Against the Framework

The following table gives an overview of some of the model evaluation criteria found in the literature and how they are mapped into the proposed model evaluation framework categories (left three columns).

Model characteristic	Benyon 1990	Fox 1993	Fox 1998	Valente 1996	Courtot 2000	Chen 1998	Orli 1996	Claxton 2000	Witt 1994	Crockett 1991	Powel 1996	Halpin 2001	Frank 1999	Korson 1992	V.Harmelen 1999	Brazier 1998	Williams 1996	Noy 2001	Swartout 1996	McGall 1997	Boehm 1997	Gillies 1997	Syntactic	Semantic	Pragmatic
Abstract basic concepts				■					■			■		■		■									
Adheres to standards						■					■						■	■		■			■		
Aesthetics/appeal																						■	■		
Alternatives/comparison						■																	■	■	■
Architecture					■				■														■		
Availability					■																				■
Basis for communication	■							■		■							■				■			■	
Competent/problem oriented		■														■			■						
Computer manipulation	■	■	■							■															■
Conciseness																				■	■		■		
Consistency (model)					■	■	■	■		■				■			■			■	■		■		
Consistency (representation)	■						■			■	■			■						■	■		■		
Controlled vocabulary																		■						■	
Coverage/Domain (extent)					■						■			■	■	■								■	
Docs different levels														■										■	
Docs indexed alpha/keyw/struct														■				■						■	
Docs organised & structured														■										■	
Documentation											■			■		■				■				■	
Economics/costs																	■					■		■	■
Effect on business/integration					■						■			■					■					■	■
Efficiency (representation)		■	■					■						■			■			■	■	■	■	■	■
Efficiency/minimality (constructs)			■	■								■	■				■			■	■		■		
Executability																■	■							■	■
Expressiveness										■		■					■		■					■	
Extensible/Customizable/Modifiable		■			■						■			■			■	■	■	■	■	■		■	■