Document and Corpus Quality Challenges for Knowledge Management in Engineering Enterprises

DIRK AHLERS, NTNU – Norwegian University of Science and Technology
JOHN KROGSTIE, NTNU – Norwegian University of Science and Technology

Categories and Subject Descriptors: H.2.m [**Information Systems**]: Miscellaneous

General Terms: Design, Documentation

Additional Key Words and Phrases: Document Quality, Corpus Quality, Linkability

ACM Reference Format:
Dirk Ahlers and John Krogstie. 2015. Document and Corpus Quality Challenges for Knowledge Management in Engineering Enterprises. *ACM J. Data Inform. Quality* 6, 4, Article 14 (October 2015), 2 pages.
DOI:http://dx.doi.org/10.1145/10.1145/2818379

1. INTRODUCTION

Enterprise data is an amalgam of mostly semi-structured and unstructured data and documents stored in heterogeneous systems. The available structure is often not readily apparent or modelled to be useful. Formats such as PDF, DWG, Excel, or Word offer a high grade of flexibility; the issue is rather that their freeform content does not divulge its structure and meaning. In the case of binary formats as used in CAD or simulation tools, even basic textual content may be missing. Structured metadata is only sometimes available.

When taking a step back, we see a challenging quality issue not based on individual documents, but on the whole corpus of enterprise documents. Our specific background is an engineering setting with large sets of documents from whole development and manufacturing lifecycles[1]. An individual document is of limited use; the organizational knowledge is distributed through many separate documents and entities. In some cases, it is easy to find it, but in other cases, heterogenous documents all over the organization make up the knowledge of, e.g., how to build a complex processing plant. For such complex retrieval tasks, different types of relations between documents and their entities have to be identified, such as same author, same or similar parts, part of same project, subpart or subproject, predecessor, precondition, clarifications, updates, vendor lists, financial or structural relations, similar tasks in previous projects, and many more [Ahlers and Mehrpoor 2014].

[1] The research leading to these results has received funding from the European Union Seventh Framework Programme ([FP7/2007-2013]) under grant agreement n^{o}284613 for the LinkedDesign project.

Author's addresses: D. Ahlers and J. Krogstie, Sem Sælandsvei 7-9, 7491 Trondheim, Norway; email : {dirk.ahlers|john.krogstie}@idi.ntnu.no

DOI:http://dx.doi.org/10.1145/10.1145/2818379

In such settings, document integration and inter-linking issues in the whole heterogeneous corpus pose the largest challenge in identifying these higher-order application-driven relations. We propose a new quality metric to handle this challenge.

2. RESEARCH DIRECTIONS

Document linkage can be understood as data integration [Martin et al. 2014] on an enterprise corpus level. Drawing connections between documents makes the corpus much more valuable as discussed above. Currently, documents can be overlayed with extracted entities, annotated to terms [Barczyński et al. 2010], and semantic relations based on entity occurrence [Peukert et al. 2015] can be used to generate typed links between documents. After deriving semantics from documents, a logical next step is deriving semantics from unstructured corpora. This is where the aforementioned task- and project-based higher-order relations can be derived.

To better understand this challenge, we propose a new measure, *linkability*, as a joint quality measure of a corpus and its documents to be semantically linkable to each other, based on specific use cases. Linkability can be understood as an extension of other quality indicators. Document quality is often modelled as an intrinsic measure. In the corpus case, it has to be extended with extrinsic features such as searchability, findability, retrievability [Azzopardi and Vinay 2008], with a focus towards connectivity and graph-based exploration and navigation along semantic relations. This is related to quality features [Krogstie 2013] of semantic quality as completeness, correctness, consistency, accuracy; and deontic quality, concerning the fitness for use cases.

For example, given a system of linking algorithms, knowledge bases, and ontologies, the missing factor is the quality of the corpus itself in this system. A corpus invariably is heterogeneous, both in terms of document formats and in terms of depth and granularity of content and annotations. For linkability, high quality would mean documents that have a rich accessible content that can be linked on entity and also higher-order levels to support complex retrieval tasks. For example, structured textual documents are easier to process that 3D data. However, if the corpus (or parts of it) are stored in a DMS or DBMS, metadata such as author names, titles, project names etc. can raise the quality of these documents and thus of the corpus.

The challenge requires to move from data to documents (which might contain extractable data or entities), and further to corpora, viewing them as a graph-based collection, and then to find and refine meaningful quality metrics for the corpus which will subsequently feed into the methods for generation of complex higher-order relationships. We aim to formally link this research to document and corpus quality issues and applications in the enterprise domain. This challenge occurs in any organisation with large amounts of heterogenous documents, not only the engineering context we used as a starting point, similar to the variability issue in big data. Solving it will enable better understanding and utilisations of organisations' hidden knowledge.

Thus, there is a strong need to define the quality of a corpus based on the quality of its constituent documents and the quality of their semantic connections.

REFERENCES

Dirk Ahlers and Mahsa Mehrpoor. 2014. Semantic Social Recommendations in Knowledge-Based Engineering. In *SP 2014: Workshop on Social Personalisation at Hypertext 2014*. CEUR-WS.

Leif Azzopardi and Vishwa Vinay. 2008. Retrievability: An Evaluation Measure for Higher Order Information Access Tasks. In *CIKM '08*. ACM, 561–570.

Wojciech M. Barczyński, Falk Brauer, Adrian Mocan, Marcus Schramm, and Jan Froemberg. 2010. BI-Style Relation Discovery among Entities in Text. In *ICDEW 2010*. IEEE.

John Krogstie. 2013. Evaluating Data Quality for Integration of Data Sources. In *The Practice of Enterprise Modeling*. Lecture Notes in Business Information Processing, Vol. 165. Springer, 39–53.

Nigel Martin, Alexandra Poulovassilis, and Jianing Wang. 2014. A Methodology and Architecture Embedding Quality Assessment in Data Integration. *J. Data and Information Quality* 4, 4, Article 17 (2014).

Eric Peukert, Christian Wartner, and Erhard Rahm. 2015. A Smart Link Infrastructure for Integrating and Analyzing Process Data. In *BTW 2015*.