2012 EL Project Title and Number: Engineering Data Quality Measurement

Program Title: Systems Integration for Manufacturing and Construction Applications

Principal Investigator: Edward Barkmeyer, 734

Project Staff:

Name
Ed Barkmeyer
Don Libes
Simon Frechette
Albert Jones
Fabian Neuhaus
Antoine Gerardin
Severin Tixier

Date Prepared: September 2, 2011

Summary: The development and production of complex products demands the cooperation of engineering and operations activities across a network of companies, and the successful exchange of information among them. Standards assist in the mechanics of information transfer, but they don't guarantee consistency, completeness and timeliness of the information – fitness for purpose. Data quality problems of these kinds cause delays and "do-overs" that create significant cost in time and money in 50% of American industry[1] and impede efficient product development and manufacture. There is no current technology for measuring the fitness for purpose of data sets. The EDQM project is developing the measurement science for a new kind of data quality measurement, using emerging knowledge engineering techniques to trace data provenance, to analyze information content, and to measure consistency with fitness rules and background engineering knowledge. The EDQM technology will enable early diagnosis and repair of data quality problems, and the elimination of resulting delays, rework, and other related cost, thus removing a key barrier to the overall productivity of manufacturing networks.

[1] Data Quality Pro Survey: April, 2009

Description:

Objective: To improve engineering data quality by prototyping a new measurement science for the consistency, completeness and timeliness of information. By the end of FY2013, the project will have demonstrated use of a prototype methodology and metrics for measuring fitness for purpose of data sets in a real manufacturing environment, and have preliminary estimates of its value.

What is the new technical idea? Development and production of a complex product involves the capabilities of a network of independent companies, each with its own software systems providing its relevant knowledge to other participants and acquiring information they possess. The incoming information arrives from multiple sources in different formats with overlapping content. In this environment, erroneous data and misinterpretation of shared data have become a source of significant delays and cost.

Standards have been developed to minimize the number of forms, and current tools test for conformance of data sets to the standards. But, in addition to outright errors in content, there are significant variations in the usage and interpretation of standard data elements. This often results in 'valid data' that is incomplete for the receiver's purposes, or interpreted differently by the sender and the receiver. Incompatible interpretation of design information, for example, delayed the Airbus 380 by 18 months[2]. Similar problems in engineering, in production specifications, in materials shipment data, significantly reduce the productivity of production networks.

The new EDQM technology is based on two ideas. First, messages contain a set of statements that convey part of the sender's knowledge about things of interest to the receiver. This knowledge can be integrated by the receiver with other knowledge the receiver possesses and other knowledge provided to the receiver from other partners. Data quality is about the consistency of the receiver's derived model of the world with the actual state of the world. Consistency is crucial because that model will be used in making engineering and operations decisions.

Second, knowledge engineering technologies can support automated reasoning to make assessments about the collections of such statements that appear in messages. The nature of the assessment is not just the abstract validity of data values, but, rather the completeness and consistency of the model of the world that they convey. Such technologies have recently become commercially viable. The main technical idea for the EDQM project is to enhance and adapt this technology to create a measurement science methodology for analyzing the fitness for purpose of the body of information provided by messages exchanged among engineering and operations partners in the production network.

The project will produce a collection of tools that will be able to measure (1) conformance of the data sets to standards and usage rules, (2) completeness and timeliness of the information with respect to usage purposes, (3) consistency of the information provided by multiple sources, and, (4) consistency with background knowledge of the engineering domain. This will enable early

[2] http://www.businessweek.com/globalbiz/content/oct2006/gb20061005_846432.htm

diagnosis of serious quality problems and reconciliation of conflicting information, before they create significant losses in time and money.

What is the research plan? As noted, the EDQM will develop a methodology and a collection of associated tools to analyze fitness for purpose of messages exchanged among production network partners. The technical approach involves three major tasks.

- adaptation of automated reasoning tools to support integrated data provenance information and specialized capabilities for time and measurements
- dynamic conversion of messages and database information as needed into standard forms that can be used by an automated reasoning tool
- support for specification of knowledge and fit-for-purpose evaluation criteria in a form that is readable by engineering and operations personnel

This approach can be used to test the validity of an individual data item – a single sentence – against any set of criteria. More importantly, it can be used to test the consistency and completeness of a composite set of information from multiple sources, as a body, with respect to expectations of the recipient, and background knowledge possessed by the recipient.

The initial version of the EDQM toolkit will be completed by within the first quarter of FY2012 and then demonstrated to industry. The next steps are to mature the EDQM toolkit and to identify specific industrial partners and associated industry scenarios for evaluation of the toolkit for full scale industrial data quality assessment. At least two such partners/scenarios, probably in supply chains and materials management, will be identified by mid-2012, and the corresponding industrial use experiments undertaken in FY2012 and in FY2013.

At the end of FY2013 we will have evaluated a new measurement science methodology supporting a new dimension in measuring engineering data quality and supporting its improvement. The technology can be tailored and expanded to further engineering and operations needs. We expect this to lead to development of commercial tools supporting the technology, thereby enabling the removal of a common barrier to the productivity of manufacturing networks.

Major Accomplishments:

Recent Results:
Outputs:

- Demonstration of the alpha version of a tool (“PrIKL”) that validates data quality queries against manufacturing knowledge represented in a formal logic language that is an extension of ISO 24707 Common Logic (Jan 2011).
- Demonstration of an alpha version of a tool (“Recon”) that translates a manufacturing knowledge specified in a restricted form of English into ISO Common Logic (June 2011).
- W3C member submission “Validating Semantic Web Data with OWL Integrity Constraints” (co-authored by Evan Wallace)
- Paper on how to fix the semantics of modules in ISO/IEC 24707 (co-authored by Fabian Neuhaus), accepted for publication in journal “Applied Ontology”.

- Submission of OMG standard Date/Time reference ontology (Sep 2011)

Outcomes:
- Published work on provenance capture in an extension of Common Logic is used in commercial software tool (Highfleet Inc.)

Standards and Codes:
The EDQM project relies on representations of engineering knowledge in logic-based knowledge representation languages, in particular ISO/IEC 24707 Common Logic and W3C Web Ontology Language (OWL). One challenge is that these languages need to be evolved and extended to meet the requirements for data quality measurement; NIST staff has already worked in this area (see outputs above). Another challenge is that there is no standard about the integration of different Common Logic and/or OWL artefacts. To address these challenges:
- Wallace supports the development of extensions to OWL; in particular integrity constraints.
- Neuhaus participates in ISO TC37 "Ontology Integration and Interoperability“ standard (FY2012-2013)
- Neuhaus participates in the development of an addendum to ISO/IEC 24707.