9th CIRP Conference on Intelligent Computation in Manufacturing Engineering – CIRP ICME '14

Product Life Cycle Analytics –
Next Generation Data Analytics on Structured and Unstructured Data

Laura Kassner[a]*, Christoph Gröger[a], Bernhard Mitschang[a], Engelbert Westkämper[a]

[a]Graduate School advanced Manufacturing Engineering, University of Stuttgart, Nobelstr. 12, 70569 Stuttgart, Germany

* Corresponding author. Tel.: +49 711 68588402; fax: +49 711 68578402 E-mail address: laura.kassner@gsame.uni-stuttgart.de

Abstract

Existing analytics approaches on unstructured data around the product life cycle focus on isolated data sources from a single product life cycle phase, do not make use of structured data for holistic analytics and are typically cost-intensive and case-based, without a general framework. To address these issues, we present our Product Life Cycle Analytics (PLCA) approach for the holistic integration and analysis of unstructured and structured data from multiple data sources around the product life cycle. We survey structured and unstructured data sources around the product life cycle and discuss limitations of existing analytics. We develop a set of requirements for PLCA and present ApLAUDING, a reference architecture which meets all requirements, as well as an application scenario, and propose a strategy towards implementation.

© 2014 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license
(http://creativecommons.org/licenses/by-nc-nd/4.0/).
Selection and peer-review under responsibility of the International Scientific Committee of "9th CIRP ICME Conference"

Keywords: Analysis; Optimisation; Quality

1. Introduction

Large amounts of unstructured data, e.g., emails, failure reports and customer complaints, are abundant around the product life cycle and provide a huge potential for analytics-driven optimisation. Estimates of the proportion of unstructured data within enterprises range between 50% and 80% of all data [1]. In addition, external data relevant in particular to later phases of the product life cycle are mostly in unstructured form, for instance data from the social web such as blogs, tweets or forum posts.

Today, most unstructured data around the product life cycle lie untapped or are only accessed through manual analytics. Integration with structured data for holistic, automated analytics has only recently come into focus. Existing analytics approaches on unstructured data are fraught with three major insufficiencies limiting comprehensive business improvement: (1) They focus on isolated data sources from a single life cycle - for

example, data from a customer relationship management system are mined for frequent complaints without considering manufacturing failure reports related to the same product; (2) they do not make use of structured data for holistic analytics, e. g., to automatically correlate unstructured failure reports with structured performance data of a manufacturing execution system; and (3) implementations of data integration and analytics components are typically cost-intensive, manual and case-based, without a general framework. However, in an age of high competitive pressure, faced with megatrends like globalization, increased automation and changing demographics, enterprises need to develop analytics which recover untapped knowledge across the entire product life cycle in order to perform well in the global market.

To address these issues, we present our Product Life Cycle Analytics (PLCA) approach, a platform and reference architecture for the holistic integration and analysis of unstructured and structured data from multiple data sources around the product life cycle. It

bridges the gap which currently between full life cycle coverage and fully-fledged analytics.

The remainder of this paper is structured as follows: In ch. 2, we survey the different data types and sources which exist around the product life cycle. In ch. 3, we discuss existing analytics and integration approaches and their limitations. On this basis, we develop a catalogue of requirements for a product life cycle analytics architecture in ch. 4, outline ApPLAUDING as a reference architecture in ch. 5 and present first steps towards a concrete application in ch. 6.

2. Data Sources around the Product Life Cycle

Based on a literature survey [1–11] and a case study we conducted in the automotive industry (cf. 6.1), we compile an overview of the types of structured and unstructured data sources around the product life cycle.

We define structured data as stored in traditional databases, structured by rows and columns and mostly numeric in nature, and unstructured data as the content of documents such as text files, pdfs or image, audio and video files, in accordance with [1]. We put a strong focus on *textual* unstructured data. Fig. 1 shows the multitude of data sources which are created and accessed around the entire product life cycle, with a higher volume of structured data in the production and planning phases and a higher volume of unstructured data in design and usage phases (cf. [12]). Currently these data

Fig. 1. Structured and unstructured data sources around the product life cycle

sources exist in relative isolation, especially along the divide between structured and unstructured data. The only unstructured data sources regularly used for analytics today are social media content and customer feedback from the usage and maintenance phases.

3. Existing Approaches and their Limitations

We group existing approaches according to their focus on different aspects of the issue at hand: product life cycle management (3.1), integrating unstructured data with structured data (3.2) and applying analytics to unstructured data (3.3).

3.1. Product Life Cycle Management

The concept of closed-loop product life cycle management (PLM) [2,8] can be regarded as foundational for our work, since it addresses data integration around the entire product life cycle. [13] defines it as a strategic approach which attempts to guarantee access to product information at any point in the lifecycle as needed, to maintain information integrity and to manage business processes which create and distribute the information. Closed-loop PLM focuses on integrating structured data around the life cycle with the help of product-embedded identifiers. Neither unstructured data nor analytics are addressed.
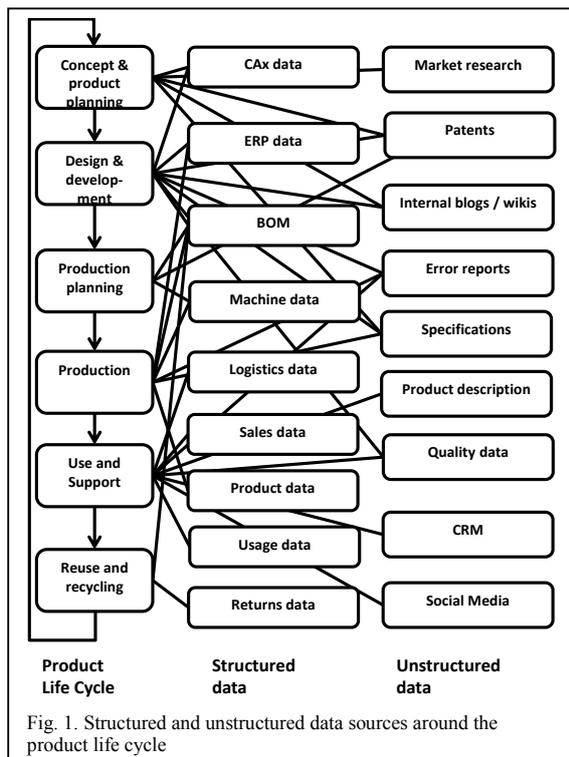
3.2. Integrating Unstructured Data

The Aletheia project (cf. e.g. [12,14]) provides an architecture that integrates structured and unstructured information from many different sources into federated repositories. Unstructured documents as well as facts extracted from text are stored in a dedicated "uncertain repository". The focus of Aletheia is on product data management; its functionalities are different forms of querying and exploration. There are no analytics beyond indexing, information extraction and ontology integration needed to provide searchability. Several industry use cases were implemented (e.g. [15]) but none with the focus on analytics.

The Advanced Manufacturing Analytics (AdMA) platform represents a reference architecture for data-driven manufacturing process optimization [16]. It comprises a holistic knowledge repository [17] integrating structured and unstructured data, but its analytics focus primarily on structured data.

3.3. Analytics for Unstructured Data

[18] propose a conceptual framework in which content collections of unstructured data are separated from structured data but extracted metadata on