Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 33 (2015) 35 – 40

www.elsevier.com/locate/procedia

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.

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

Selection and peer-review under responsibility of the International Scientific Committee of "9th CIRP ICME Conference"
doi:10.1016/j.procir.2015.06.008

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 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.

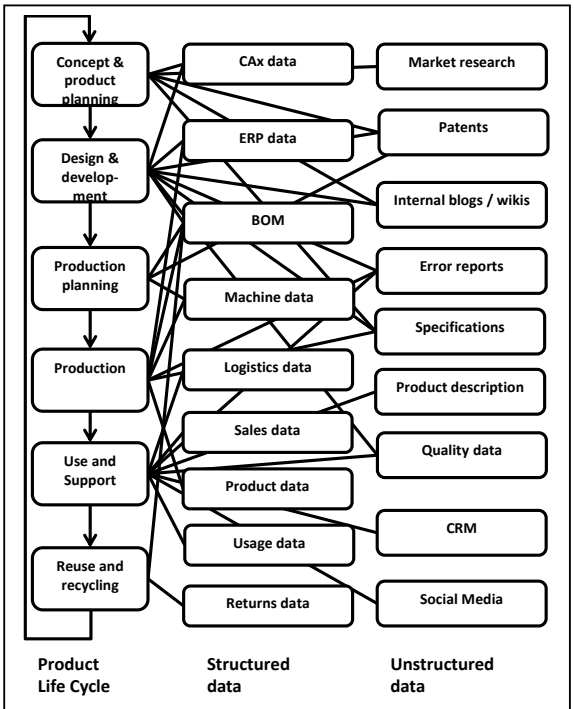

Fig. 1. Structured and unstructured data sources around the product life cycle

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

unstructured content are integrated into the structured data warehouse. The implementation and concept put forth by [9] introduce text analysis components for extracting information from unstructured sources into the ETL flow of a data warehouse and enrich the warehouse schemas with additional dimensions for the extracted information. The existing analytics infrastructure can then be used without modification on the extracted data. While these approaches to unstructured analytics go beyond data integration for searchability, they are still lacking in one crucial respect: If information extraction from unstructured data is to be conducted as a step in the ETL process, the type of information which will later be required for analytics must be known beforehand; therefore the dimensions of analytics that are possible will be severely restricted.

Case-based implementations of analytics for unstructured data exist: [9] extract keywords from automotive accident reports for quality early warning and root cause analysis. [19] and [20] mine social media sources and workshop reports for early warning and seasonal predictions in the automotive domain. While they present tailored and elaborate analyses, these implementations use single or homogeneous data sources from a single life cycle phase.

4. Requirements for Product Life Cycle Analytics

In 4.1, we present general core requirements that a product life cycle analytics architecture must meet and evaluate existing approaches with respect to these requirements. In 4.2, we sum up the specific challenges in unstructured data analytics.

4.1. Core Requirements

Based on the state of the field in the intersection of unstructured data integration and business analytics presented in ch. 3, we develop the following requirements R_i to be met by a holistic, integrated business analytics application of the next generation:

- Encompasses data from **all life cycle phases** (R1)
- **Integrates** structured and unstructured data (R2)
- Adheres to a general analytics and integration **framework** (R3)
- Includes **fully-fledged analytics** on structured and unstructured data – i.e. sophisticated, complex analytics which are applied to information from structured and unstructured data sources – and is capable of **deriving novel insights with added value** (R4)
- Can be **flexibly** expanded to include more data sources and more analytics tools (R5)

Table 1 shows to what extent the analytics approaches on unstructured data which we reviewed in ch. 3 [9,13,14,17,19,20] meet these requirements. As we can see, the case-based implementations [9,19,20] provide higher analytics capabilities while lacking full life cycle coverage, a framework and flexibility for expansion; whereas the approaches featuring a general framework and – sometimes – full life cycle coverage [13,14,17] do not provide analytics.

Table 1. Requirements met by approaches discussed

Approach	R1	R2	R3	R4	R5
[14]	√	√	√	Querying only	(√)
[13]	√	Struct only	√	--	(√)
[17]	Production only	√	√	Struct only	√
[9]	Accident DB only	√ (ETL)	(√)	(√)	--
[20]	Social media only	Unstruct only	--	(√)	--
[19]	Repair orders only	Unstruct only	--	(√)	--

4.2. Challenges in Unstructured Data Analytics

Textual data around the life cycle are characterized by domain-specific vocabulary and structure, however natural language processing (NLP) tools are trained on standardized language [9]. Federated data integration often relies on ontologies, which are costly to develop and maintain. In addition, semantic relatedness for the purpose of analytics is highly context-sensitive, so links between structured and unstructured data may need to be re-computed for new scenarios. A way of weighting the certainty of information has to be included in analytics algorithms. These challenges must be addressed at a state-of-the-art level in order to be able to provide fully-fledged and value-adding analytics on unstructured data (R4).

5. ApPLAUDING: An Architecture for Product Life cycle Analytics with Unstructured Data INteGration

In this chapter we present ApPLAUDING, a comprehensive conceptual architecture (R3) for product life cycle analytics consisting of an integration layer, an analytics layer and a presentation layer and addressing all five core requirements from ch. 4.1.

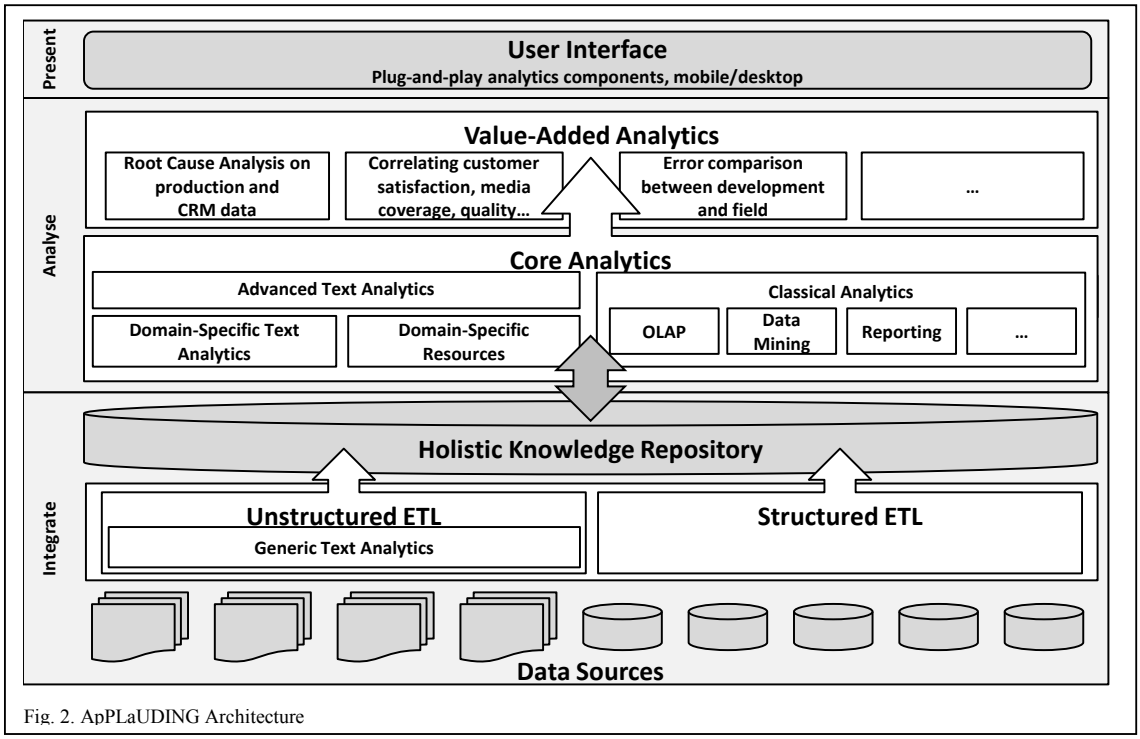

Fig. 2. ApPLaUDING Architecture

5.1. Integration Layer

The integration layer is subdivided into three components: data sources around the product life cycle (R1) filled with structured and unstructured data, an ETL layer, and a holistic knowledge repository along the lines of [16,17]. **Data sources** can be of various types, for example, relational data bases, content management systems, file systems, etc. The **ETL component** contains mechanisms for handling structured and unstructured data (R2): Structured data are loaded into the knowledge repository using traditional ETL techniques. Unstructured data are pre-treated with *generic text analytics* and thus enriched with structuring information (linguistic annotation), then the entire content of the unstructured documents is stored in the knowledge repository. This type of **unstructured ETL** distinguishes ApPLAUDING from other approaches to unstructured data integration [9,18] which only extract part of the information from unstructured data into the integrated repository – a core weakness which keeps them, but not ApPLAUDING, from fulfilling R5 for flexibility. The integration of generic text analytics into the ETL step also frees subsequent analytics on data from unstructured sources from costly preprocessing tasks.

The **holistic knowledge repository** is based on the concept presented in [17] and combines a structured data warehouse with unstructured data. Results from advanced and value-added analytics are also written into the knowledge repository – simpler examples are facts extracted from the textual data along with uncertainty weights (similar to [14]) and associations between structured and unstructured data concerning the same topic (as proposed in [17]).

Generic Text Analytics (Examples)

Part-of-Speech-Tagging: Associate each word with a tag indicating its word class (verb, noun, adverb, adjective…)
Named-Entity Recognition: Mark up persons, organizations, locations, and other types of entities

5.2. Analytics Layer

The analytics layer is divided into core analytics and value-added analytics services. The **core analytics** component provides analytics tools for both structured and unstructured data (R4): There are classical analytics tools for use on structured data – for example, OLAP, querying and reporting. These can also be applied to structured data originally extracted from unstructured data. For unstructured data, there exist two text analytics toolboxes: The first one contains domain-specific text analytics tools – e.g., the recognition of entities or expressions from a particular domain such as automobile parts or positive and negative sentiment expressions. These tools make use of domain-specific NLP resources such as taxonomies, wordlists / dictionaries and

schemas, which are also supplied as a separate part of the core analytics component. The second text analytics toolbox contains advanced analytics drawing on both the domain-specific resources and on analytics techniques from the domain-specific toolbox. Results of these analyses take the shape of further annotations or extracted information nuggets, which are also stored in the knowledge repository.

Advanced Text Analytics (Examples)

Topic Detection: Identify the topic of a sentence, paragraph or entire text
Classification and Clustering Algorithms: Classify texts based on (domain-specific or generic) linguistic features

By separating domain-specific resources and analytics as well as advanced analytics from the basic, generic text analytics in the ETL layer, we achieve a higher modularity and the ability to conduct analyses in a wider range of contexts than prior approaches to unstructured and structured analytics, which contributes to fulfillment of R5 and R2.

The **value-added analytics** component contains complex and custom analytics services which provide fully-fledged analytics capabilities (R4). They are crafted on a case-by-case basis, but their composition from modular components makes it easy to re-use them for different scenarios (R5, R2). An example of a complex value-added analytics case is the root cause analysis on data from different sources and life cycle phases, e.g. production data in combination with CRM data. Another example will be presented in 6.1.

5.3. Presentation Layer

The modularity and flexibility of ApPLAUDING are also reflected in the user interface, where analytics components can be combined into complex services via intuitive plug-and-play. The design of the presentation layer is subject to future work.

6. Towards an Application

Based on a case study of unstructured data formats in cooperation with an automotive OEM, we deduce a scenario to illustrate the application potential of PLCA in 6.1 and define an implementation strategy to realize ApPLAUDING in 6.2.

6.1. Application Scenario

Unstructured data in the form of textual quality reports are particularly predominant in quality management, especially in the development and support/maintenance phases of the product life cycle. Structured quality data may be associated with these textual reports or may be present in isolated databases, for example machine error codes from production or vehicle diagnosis data from maintenance. Textual and structured data from customer relationship management systems also contain quality information.

Going one step beyond previous analytics scenarios using automotive quality data [9,19–21], we can use data from several different sources to gain more sophisticated insights into quality analysis. A fully-fledged *error analysis* on unstructured quality reports and structured vehicle data from *development* and from the field *(aftersales)* can answer the following questions:

- Which errors or error types occur in the field that had already been observed in development and testing, and supposedly remedied?
- Which errors or error types occur in the field for which existing testing methods could be easily adapted?
- What are the quantitative and qualitative distinctions between errors from those two different life cycle phases?

For example, irregularities in the durability of the paint coat may have been observed in a car's prototype and remedied through a change of coating method. If similar irregularities appear again in after-sales vehicles and are reported through CRM or maintenance, discovering the connection between these two errors would be extremely valuable, since it indicates that the changes applied during development were not effective.

This scenario showcases all the features of the type of complex analytics use case for which we develop ApPLAUDING: Structured and unstructured data from different data sources and different life cycle phases (R1) must be integrated (R2). To compare errors from the two life cycle phases, generic, domain-specific and advanced text analytics must be applied to the quality reports, combined with classification algorithms on integrated structured and unstructured data (R4). Error quantities and re-occurrences can then be determined in a value-added analytics application. Parts of this application can be re-used for a qualitative comparison of error schemas between life cycle phases (R5).

6.2. Implementation Strategy

Based on the above application scenario, we are currently working on a prototypical implementation of ApPLAUDING focusing on the integration layer and the analytics layer.

Integration layer. The *data sources* for structured data are typical relational databases. Some unstructured

data in the shape of free-text fields can also be found here. Other unstructured data, e.g. customer complaints, may be stored in content management systems (CMS). The implementation of the *holistic knowledge repository* is based on the prototype presented in [17]. For transferring structured data into the knowledge repository, an out-of-the-box open source ETL solution (e.g. [22]) is used. For unstructured data, *generic text analytics* are implemented as a UIMA pipeline [23]. Several NLP toolkits are under investigation for quality and performance, e.g. [24,25] and [26].

Analytics layer. The *core analytics* is a combination of out-of-the-box classical analytics (e.g. [27]) and customized analytics tools. Domain-specific resources for text analytics will be compiled drawing from the experiences of [9,19–21]. For example, automotive parts and errors taxonomies can be derived from existing resources such as error code descriptions and enriched with synonyms using generic ontologies such as WordNet [28] and GermaNet [29]. Domain-specific basic and advanced text analytics tools will be developed on top of standard solutions with the help of these resources. For our application scenario, one *value-added analytics* application is a cross-life cycle error classifier / matcher, employing the aforementioned domain-specific error and part taxonomies.

7. Concluding Remarks

We have shown that there exists a gap in the analytics landscape with respect to full integration of structured and unstructured data around the product life cycle. We have evaluated existing analytics approaches against a newly developed set of requirements and found a dichotomy between case-based implementations with more advanced analytics and general frameworks with limited or no analytics. We have developed a reference architecture which meets all requirements for filling the analytics gap, and presented first steps towards application scenarios and an implementation for enhancing process effectiveness and efficency.

References

[1] Russom P. BI Search and Text Analytics. TDWI Best Pract Rep 2007.

[2] Ameri F, Dutta D. Product Lifecycle Management : Closing the Knowledge Loops. Supply Chain Manag An Int J 2005;2:577–90.

[3] Choudhary AKK, Tiwari MKK, Harding JA. Data Mining in Manufacturing: A Review Based on the Kind of Knowledge. J Intell Manuf 2004;20:501–21.

[4] Blumberg R, Atre S. The problem with unstructured data. DM Rev 2003:42–5.

[5] Dayal U, Castellanos M, Simitsis A, Wilkinson K. Data integration flows for business intelligence. Proc 12th Int Conf Extending Database Technol Adv Database Technol - EDBT '09 2009:1.

[6] Fiorentini X, Gambino I, Liang V, Rachuri S. An ontology for assembly representation 2007.

[7] Harding J a., Shahbaz M, Kusiak A. Data Mining in Manufacturing: A Review. J Manuf Sci Eng 2006;128:969.

[8] Jun H-B, Kiritsis D, Xirouchakis P. Research issues on closed-loop PLM. Comput Ind 2007;58:855–68.

[9] Lang A, Ortiz MMM, Abraham S. Enhancing Business Intelligence with unstructured data. Subs.emis.de 2009:469–85.

[10] Menon R, Tong LH, Sathiyakeerthi S. Analyzing textual databases using data mining to enable fast product development processes. Reliab Eng Syst Saf 2005;88:171–80.

[11] Sheth A. From Semantic Search & Integration to Analytics 2004:1–10.

[12] Konzag A, Dau F, Ko A, Ag BMW, Informationstechnik LI, Cottbus BTU, et al. Bereitstellung unstrukturierter Daten und Verknüpfung mit strukturierten Daten in der Konzeptentwicklung von Automobilen. 21st Symp Des X, … 2010:1–12.

[13] Kiritsis D. Closed-loop PLM for intelligent products in the era of the Internet of things. Comput Des 2011;43:479–501.

[14] Wauer M, Schuster D, Meinecke J. Aletheia: an architecture for semantic federation of product information from structured and unstructured sources. … Inf Integr Web- … 2010:8–10.

[15] Kunz S, Brecht F, Fabian B, Aleksy M, Wauer M. Aletheia--Improving Industrial Service Lifecycle Management by Semantic Data Federations. 2010 24th IEEE Int Conf Adv Inf Netw Appl 2010:1308–14.

[16] Gröger C, Niedermann F, Mitschang B. Data mining-driven manufacturing process optimization. Proc World … 2012.

[17] Gröger C, Schwarz H, Mitschang B. The Manufacturing Knowledge Repository. Consolidating Knowledge to Enable Holistic Process Knowledge Management in Manufacturing. Proc. 16th Int. Conf. Enterp. IInformation Syst. (ICEIS), 27-30 April. 2014, Lisbon, Port., 2014.

[18] Baars H, Kemper H-G. Management Support with Structured and Unstructured Data—An Integrated Business Intelligence Framework. Inf Syst Manag 2008;25:132–48.

[19] Schierle M. Language Engineering for Information Extraction. 2011.

[20] Bank M. AIM-A Social Media Monitoring System for Quality Engineering. 2013.

[21] Hänig C. Unsupervised Natural Language Processing for Knowledge Extraction from Domain-specific Textual Resources. 2012.

[22] Casters M, Bouman R, Van Dongen J. Pentaho Kettle solutions: building open source ETL solutions with Pentaho Data Integration. John Wiley & Sons; 2010.

[23] Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. Nat Lang Eng 2004;10:327–48.

[24] Toutanova K, Manning CD. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. Proc. Jt. SIGDAT Conf. Empir. Methods Nat. Lang. Process. Very Large Corpora, 2000.

[25] Finkel JR, Grenager T, Manning C. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proc. 43nd Annu. Meet. Assoc. Comput. Linguist. (ACL 2005), 2005, p. 363–70.

[26] Baldridge J, Morton T. OpenNLP 2004.

[27] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Software: An Update. SIGKDD Explor 2009;11:10–8.

[28] Fellbaum C. WordNet. Blackwell Publishing Ltd; 1999.

[29] Hamp B, Feldweg H. Germanet - a lexical-semantic net for german. Proc. ACL Work. Autom. Inf. Extr. Build. Lex. Semant. Resour. NLP Appl., 1997.