PROBLEM JAKOŚCI DANYCH W HURTOWNIACH

Błaszczyk Katarzyna
Knosala Ryszard

Wprowadzenie

Wdrażanie systemu hurtowni danych jest przedsięwzięciem długofalowym o strategicznym znaczeniu dla przedsiębiorstwa. Praktyka współczesnego zarządzania wskazuje na istnienie wielu problemów związanych z tym procesem. Jak wynika z badań projektu wrażania hurtowni danych, przyczyną takiego stanu jest szereg elementów począwszy od błędnego zrozumienia potrzeb przedsiębiorstwa, zły dobór narzędzi i technologii, małą wydajność, po brak procedur uwzględnienia zmian i zagrożeń systemu [Adel99, Wojt05]. Analitycy zidentyfikowali także inny problem, który ma ogromny wpływ na jakość finalnego produktu - problem jakości gromadzonych i analizowanych w systemie danych. Czynnik ten, zarówno jak poprzednie, ujawnia ryzyko na etapie definicji potrzeb użytkowników, planowania oraz kontroli zmian, jednak jest często bagatelizowany. Jak sugeruje artykuł [Wang95] ilość błędów w bazach danych sięga 10 % wszystkich gromadzonych informacji. Przedsiębiorstwa rzadko kontrolują i prowadzą analizy jakości danych, wobec tego nawet nie zdają sobie sprawy z wagi problemu oraz z finansowych i społecznych strat z tym związanych. Przykładowo amerykańska firma pocztowa U.S. Postal Service (USPS) odnotowała w 2001 roku stratę 1,8 bilionów dolarów z powodu źle zaadresowanej

korespondencji – błędnych lub przestarzałych danych [Smal06]. Inne doniesienia mówią o tym, że firmy produkcyjne tracą ponad 25% obrotu z powodu niedbałości tych praktyk - odsetek ten wzrasta do 40% w przypadku firm usługowych [Jark03].

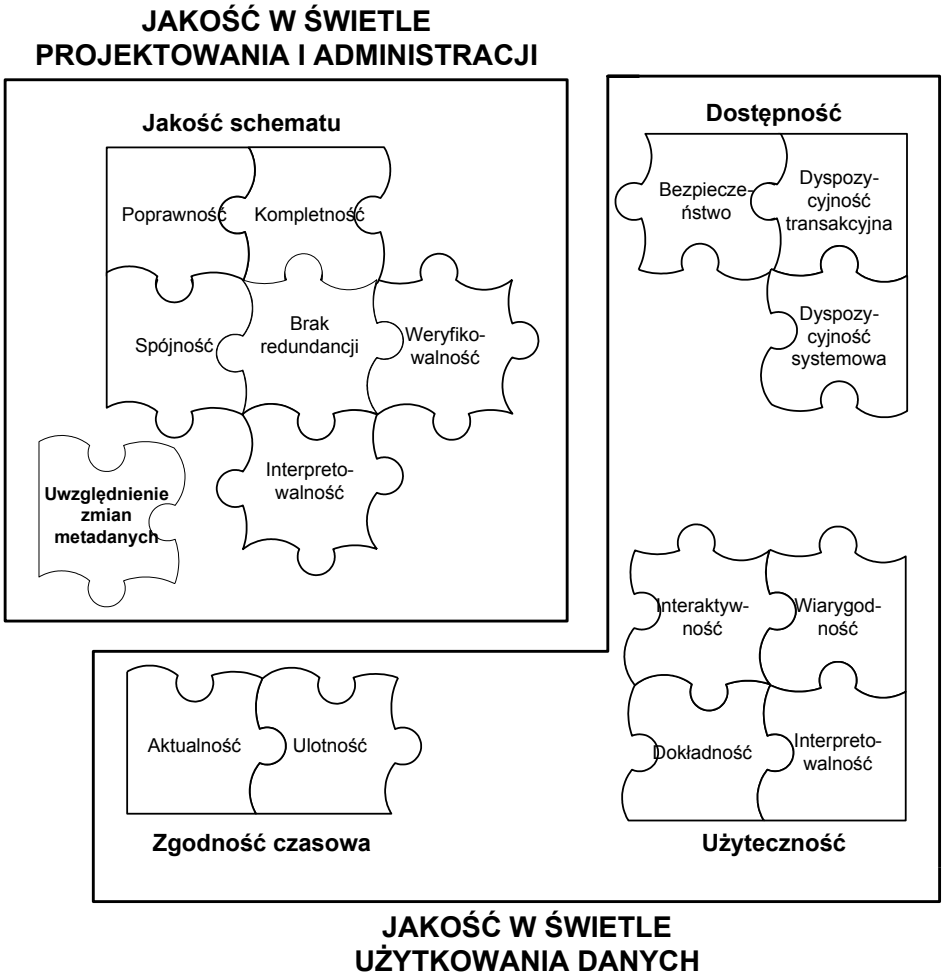
Nieprawidłowe informacje powodują, iż tysiące reklam i korespondencji nie trafia do potencjalnych klientów. Zduplikowane rekordy, brakujące pola, niezrozumiałe dane czy błędnie wpisane informacje przyczyniają się do tego, iż decyzje podejmowane na podstawie zawartości hurtowni danych są często kwestionowane i muszą być weryfikowane przez praktykę gospodarczą.

W ostatnim raporcie Gartner Group przepowiada, że około 50 % projektów hurtowni danych w 2007 roku skończą się niepowodzeniem właśnie z powodu braku staranności w tworzeniu standardów jakości danych [Beal05].

Definicja jakości danych w hurtowni

Poszukując przyczyn oraz możliwości projektowania i budowy hurtowni danych spełniającej kryteria jakości, należałoby najpierw sprecyzować to pojęcie. Według [Stec05] „jakość danych należy rozumieć jako kwalifikację poprawności danych, ale także ich przydatności". Jakość wobec tego powoduje, iż dane dla użytkownika zaczynają nabierać pewniej wartości utylitarnej. Jak można więc zauważyć, pojęcie to jest subiektywne, zależne od punktu widzenia zainteresowanych grup. Dlatego też definiując cechy jakości danych nie należy zapomnieć o ich przeznaczeniu.

Poniżej zostaną określone wymiary jakości danych dla trzech kluczowych użytkowników systemu: programisty, administratora i końcowego użytkownika - decydenta. Każdy z tych osób odgrywa inną rolę w procesie projektowania i użytkowania systemu, wobec tego ich wymagania odnośnie kryteriów jakości będą się nieco różniły między sobą.



Rys 1. Model jakości danych z uwzględnieniem grup użytkowników

Źródło: opracowanie własne na podstawie [Jark03]

Jakość w świetle projektowania i administracji

Jakość w świetle projektowania i administracji jest związana z rolą projektanta, programisty i administratora hurtowni danych.

Projektant systemu, jak i programista są zainteresowani mierzeniem jakości schematów hurtowni danych (nowo tworzonych lub już istniejących) oraz jakości metadanych. Etap tworzenia systemu ze względu na konieczność zaprojektowania i wdrożenia elementów procesu ETL jest stosunkowo trudny, jednakże ma jednak największy wpływ na jakość końcową danych systemu. Administrator hurtowi danych jest natomiast zainteresowany raportowaniem błędów, dostępem do metadanych oraz sprawdzaniem ich aktualności. Wysoka jakość schematów danych i metadanych umieszczonych w tej strukturze jest zatem celem działania tego użytkownika systemu.

Na jakość w świetle projektowania i administracji składają się dwa główne wymiary: jakość schematu i uwzględnienie zmian metadanych [Jark03].

Jakość schematu określa zdolność modelu (struktury) do przechowywania informacji pozwalającej na dokładne opisanie istniejącej sytuacji rzeczywistej. Składa się ona z szeregu wymiarów odzwierciedlających jej pełną funkcjonalność [Wojt05, Jark03]:

- **poprawność** (validation) – stanowi właściwe zrozumienie obiektów świata rzeczywistego, właściwe zrozumienie wymagań użytkowników, poprawne schematy źródeł oraz uzasadnione obliczenia i agregacje w systemie,

- **kompletność** (completeness) – stopień kompleksowego ujęcia niezbędnej wiedzy w strukturach hurtowni danych oraz systemach źródłowych, uzupełnienie brakujących rekordów i pól,
- **spójność** (consistency) – oznacza ujednolicenie i integrację przechowywanych w systemie informacji - zgodność czasową oraz zgodność danych na poziomie atomowym,
- **brak redundancji** (reduction of recurrence) – stopień usunięcia pokrywających się informacji gromadzonych w systemie,
- **weryfikowalność** (traceability) – oznacza iż, wszelkie wymagania użytkowników końcowych, projektantów i administratorów powinny być możliwe do wyśledzenia w schemacie hurtowni danych,
- **interpretowalność** (interpretability) – stanowi właściwe opisanie składników struktury, co pozytywnie wpływa na łatwość administrowania i używania systemu.

Kolejnym elementem oddziaływującym na jakość projektowania i administracji hurtownią danych jest **uwzględnienie zmian metadanych**. Jak wiadomo, zarówno źródła danych jak i wymagania w stosunku do systemu ulegają ewolucji w czasie. Niezbędne jest wobec tego odzwierciedlenie tych zmian w strukturze metadanych systemu [Jark03].

Jakość w świetle użytkowania danych

Celem tworzenia hurtowni danych jest umożliwienie złożonej analizy danych, w taki sposób, aby można było najefektywniej wydobyć, interesujące z punktu widzenia użytkownika, informacje. Toteż podstawą funkcją systemu jest wykorzystywanie danych w procesie przetwarzania zapy-

tań oraz ukazywanie wyników. Użytkownik końcowy natomiast wymaga od systemu, aby informacje jakie uzyskał były jak najlepszej jakości tj. użyteczne i zgodne czasowo a wyniki analiz, tworzone za pomocą narzędzi OLAP, były generowane szybko, w postaci czytelnej i zrozumiałej.

Jednym z głównych wymiarów jakości w świetle użytkowania danych jest **dostępność** (accessibility). Oznacza ona możliwość dostępu do danych za pomocą kierowanych do bazy zapytań. Następujące kryteria pozwolą dokładniej opisać ten składnik jakości [Jark03]:

- **bezpieczeństwo** (security) – ograniczenie dostępu do informacji (autoryzacja) z uwagi na rolę i wymagania użytkowników,
- **dyspozycyjność transakcyjna** (transactional availaibility) – procent czasu w jakim zgromadzone dane w systemie są dostępne dla użytkownika,
- **dyspozycyjność systemu** (system availability) – procent czasu w jakim system hurtowni danych jest dostępny dla użytkownika.

Innym, ważnym dla decydenta, wymiarem jakości jest **użyteczność** (usefulness) - stopień pożyteczności i funkcjonalności uzyskiwanych informacji. Kryteriami związanymi z tym wymiarem są:

- **interaktywność** (interactivity) – możliwość nawiązania łatwej komunikacji między systemem a użytkownikiem, uzyskiwanie dodatkowych informacji np. o czasie odpowiedzi na zapytania,
- **interpretowalność** (interpretability) – stopień zrozumienia informacji będących wynikami oraz możliwość ich analizy pod kątem odzwierciedlenia w rzeczywistości,

- **wiarygodność** (credibility) – przekonanie o prawidłowości otrzymanych wyników,
- **dokładność** (accurancy) – zgodność wartości przechowywanych z wartościami rzeczywistymi.

Ostatnim wymiarem jakości w świetle użytkowania danych jest ich **zgodność czasowa** (timelines). Wymiar czasu jest podstawowym elementem w hurtowni danych. Dzięki niemu jest możliwe otrzymanie historii podmiotu oraz analiz z uwzględnieniem różnych przedziałów czasowych. Jednakże, aby otrzymywane dane były wiarygodne, niezbędne staje się restrykcyjne przestrzeganie **aktualności** (currency) gromadzonych w systemie informacji. Za aktualność w tym wypadku określa się różnicę pomiędzy datą zmiany stanu rzeczywistego a datą odnotowaną w źródle danych, bądź hurtowni danych. Wobec tego dane w hurtowni nie są może najświeższe, ale aktualne na dany moment czasu. W stosunku do specyficznych przeznaczeń systemu (bank, giełda papierów wartościowych) wyznacznikiem jakości może także stać się aktualność zrozumiana jako stopień świeżości danych - różnica pomiędzy czasem zmiany stanu rzeczywistego a czasem odnotowania tej zmiany w systemie. Wymiar **ulotność** (volatility) natomiast oznacza okres, w jakim dane pozostają aktualne w świecie rzeczywistym (rocznie dezaktualizuje się ok. 36 % danych) [Jark03].

Omówiony powyżej model jakości danych w hurtowni jest zamieszczony na rysunku 1. Jak można zauważyć, każdy z wymiarów jakości stanowi pojedynczą cząstkę (puzzel) całości modelu. Dopiero zapewnienie przez system, w sposób spójny i konsekwentny, jakości wszystkich wymiarów wpływa na końcową jakość danych produktu.

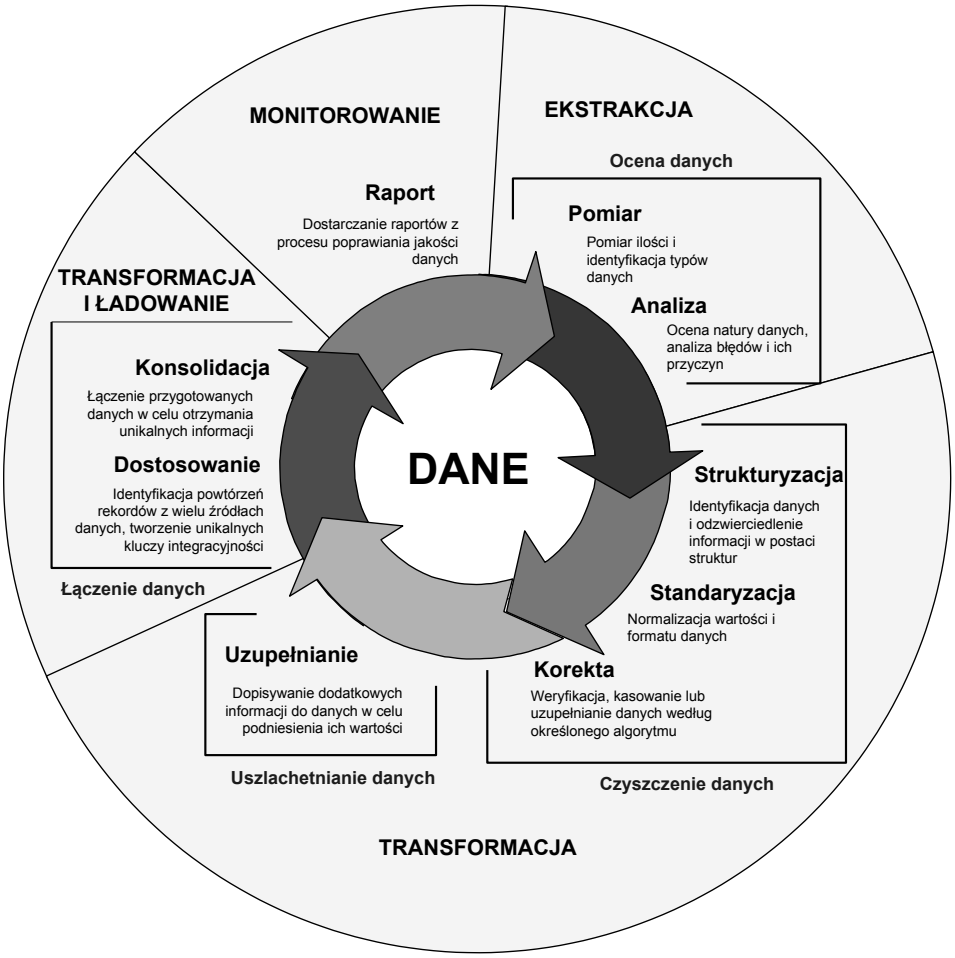
Rozwiązania w dziedzinie jakości danych w hurtowniach

Analizując sposoby rozwiązania problemu jakości danych w hurtowni, należy wziąć pod uwagę poszczególne etapy tworzenia i użytkowania systemu. Niektóre z nich mają bowiem wyjątkowy wpływ na efekt końcowy produktu. Jak już zostało wspomniane, takimi etapami są (por. [Jark03]):

- budowa schematu logicznego hurtowni danych,
- ekstrakcja, transformacja i ładowanie danych do hurtowni.

Właściwe zaprojektowanie schematu logicznego hurtowni danych decyduje o poprawności, kompletności oraz znacznej spójności i integracji danych pochodzących z wielu źródeł. Jeśli podczas opracowania struktury modelu nie uwzględniono wszystkich wymaganych zasad integralności, dane gromadzone będą niepełne, błędne, bądź też nadmiarowe. Projektant, w celu uskutecznienia swojej pracy, może wykorzystać narzędzia dotyczące modelowania danych, projektowania baz danych, integracji schematów, zarządzania metadanymi, inżynierii odwrotnej danych oraz narzędzia CASE [Jark03].

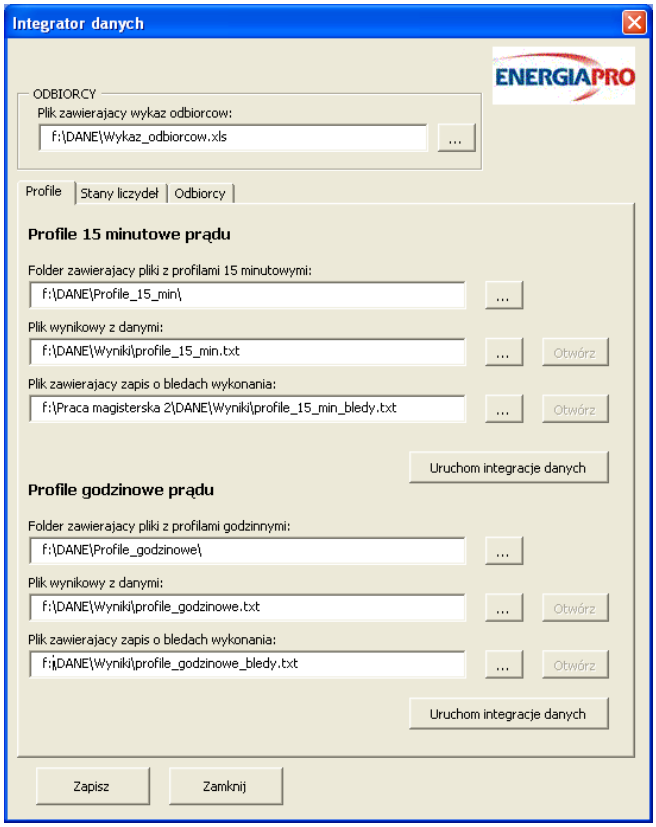
Ponadto na jakość danych w hurtowni wpływa w głównej mierze jakość danych pochodzących ze źródeł. Błędne informacje w bazach transakcyjnych mogą spowodować brak poprawności danych w systemie, a co za tym idzie niewiarygodne analizy. Dlatego też, tak ważnym elementem w procesie tworzenia i użytkowania systemu jest mechanizm ładowania i aktualizacji danych (ETL). Na tym etapie należy zaprojektować i wdrążyć szereg procesów usprawniających poprawianie i utrzymanie jakości danych (rys. 2).



Rys 2. Etapy poprawiania jakości danych w mechanizmach ETL
Źródło: opracowanie własne na podstawie [Busi06, Data06]

Dla dokładnej analizy (czyszczenia) danych źródłowych, w przypadku małej liczby informacji, często wystarcza znajomość filtrowania danych (np. język SQL) i ich przetwarzania (np. przy pomocy programów analitycznych, statystycznych, czy nawet arkuszy kalkulacyjnych) [Stec05]. Przykładowo rysunek 3 przedstawia prostą aplikację napisaną dla Energii Pro przy pomocy języka VBA (makro programu Excel), który umożliwia

oczyszczenie, zebranie i integracje danych dotyczących stanów licznikowych oraz zapisanie ich w postaci wynikowych plików tekstowych. Pliki te mogą być już bezpośrednio załadowane do bazy hurtowni danych.



Rys 3. Program VBA oczyszczający i integrujący dane stanów licznikowych

Źródło: opracowanie własne

W przypadku wielkich przedsiębiorstw wymagane będzie zastosowanie specjalnych narzędzi. Na tym etapie rozwiązaniem problemu może być **zapora jakości danych** (data quality firewall) [Beal05]. Zapora jakości danych działa podobnie jak zapora sieciowa (firewall): potrafi filtrować pakiety przesyłanych przez nią danych a także w ramach możliwości

potrafi je czyścić. Narzędzie to jest umiejscowione pomiędzy źródłami danych i strukturą hurtowni współpracując z procesami ekstrakcji, transformacji i ładowania. Problemem pojawiającym się tutaj jest sprzężenie zwrotne. Jak już zostało podkreślone, często problem jakości danych wynika bezpośrednio ze błędnych danych gromadzonych w bazach transakcyjnych. Wobec tego, jednym z rozwiązań problemu byłoby, po zidentyfikowaniu błędów przez zaporę, poinformowanie baz transakcyjnych i poprawienie informacji u samych źródeł. Jednakże administratorzy baz danych są często nie zainteresowani tymi praktykami, z uwagi na brak potrzeby na precyzyjne dane i tworzenie historii [Stec05].

Obecnie dostępne na rynku narzędzia pozwalające na ocenę, czyszczenie i uzupełnianie danych to m.in. QualityStage (Ascential Software), Trillium Software i First Logic. Te programy pomogą w wykryciu i automatycznej naprawie wielu głównych formatów danych, np. słownictwa, typowych imion i adresów (używając wielojęzykowego słownika imion i adresów), formatu daty, kodów pocztowych, adresów mailowych, telefonów. Pozwalają na usunięcie powtarzających się danych, błędów oraz niedokładności [Tril06]. Dla pewnego rodzaju „brudnych danych" automatyczne sprawdzanie jest dobrym rozwiązaniem, ale pełne naprawianie danych jest niemożliwe. Dla przykładu, aby automatycznie sprawdzić wiek osoby można ustawić, że wiek musi zawierać się w przedziale od 18 do 65. Gwarantuje to tylko naprawianie nietypowych wartości. Jednakże, jeżeli ktoś pomyli się i wpisze do systemu 45 (zamiast 54 lat), tego program już nie jest w stanie wychwycić.

Podsumowanie

Każdego roku, przedsiębiorstwa wydają biliony dolarów wdrażając systemy oparte na hurtowniach danych (CRM, ERP). Raporty donoszą, iż duży odsetek projektów kończy się fiaskiem albo przekracza swoje budżety o 65 - 75% właśnie z powodu nie wystarczającej jakości gromadzonych danych. W obecnym wieku informacji – problem ten będzie miał coraz większe znaczenie. Im szybciej przedsiębiorstwa włączą procesy służące zapewnieniu dobrej jakości informacji do operacyjnych i strategicznych działań biznesowych firmy, tym będą miały większą możliwość konkurencyjnego działania poprzez zwiększenie swojej efektywności. Należy jednakże podkreślić, iż jakość danych powinna być postrzegana całościowo, jako zespół wszystkich wymiarów jakości informacji i metadanych. Natomiast mechanizmy kontroli jakości winny obejmować kompleksowe działania, zarówno w sferze projektowania jak i pierwszego zasilania i każdorazowej aktualizacji hurtowni danych. Nie należy także zapominać o jakości danych gromadzonych w bazach transakcyjnych i innych źródłach informacji.

Literatura

[Adel99] Adelman S., Moss L.: Indicators for Success: Data Warehouse Critical Success Factors and Measuring Results, Part 2, DM Direct Newsletter, 1999.

[Wojt05] Wojtachnik R.: Problemy we wdrażaniu hurtowni danych, Gazeta IT nr 9 (39), 19 październik 2005.

[Smal06] Smalltree H.: Postal service delivers data quality, SearchDataManagement.com, 25 styczeń 2006.

[Beal05] Beal B.: Half of data warehouse projects to fail, SearchCRM.com, 09 marzec 2005.

[Wang95] Wang R. Y., Reddy M. P., Kon H. B. : Towards quality data: An attribute-based approach, Decision Support Systems, 13 (3/4), 1995.

[Jark03] Jarke M., Lenzerini M., Vassiliou Y.,Vassiliadis P.:Hurtownie danych. Podstawy organizacji i funkcjonowania, WSiP, Warszawa, 2003.

[Stec05] Stecyk A.: Jakość i integralność informacji w hurtowniach danych, Gazeta IT nr 9 (39), 19 październik 2005.

[Busi06] Business Objects: http://www.firstlogic.com/dataquality/default.asp.

[Data06] DataFlux, a SAS Company: http://www.dataflux.com/Data-Management/index.asp.

[Tril06] Harte – Hanks Trillium Software: http://www.trilliumsoftware.com/site/content/index.asp.

Mgr inż. Katarzyna Błaszczyk
Prof. dr hab. iż. Ryszard Knosala
Instytut Inżynierii Produkcji
Politechnika Opolska
45-370 Opole, ul. Ozimska 75
tel.: (0-77) 423-40-35
e-mail: blaszczyk@po.opole.pl
knosala@po.opole.pl