

Golden Bullet

Intelligent Classifier

**User-friendly Semi-Automatic
Product Classification System**



1. Product Classification Problem
2. Workflow driven approach / blended State-of-the-Art
3. Technical bits / Achievements
 - Software architecture
 - Java XML Registries
 - User taxonomies
 - Packages and API access
4. Conclusions and Outlook
5. Online Demo



1. E-Catalogs contain thousands of cryptic product descriptions
 1. CAREPAQ BUREAU PROSIGNIA3YRS/SITE/J+1/TEL
 2. TRAINING ACT/ASEEXCEPT TRU64UNIX and OPENVMS
 3.
2. Businesses have to deal with thousands of e-catalogs
3. Classification standards have tens of thousands of product categories (*21192 in UNSPSC 8.04*)
4. **The result:** high manual classification effort is required



- Product classification for Warehouse Catalogues requires
 - tremendous labor expensive,
 - complicated, time-consuming and error-prone efforts
- many standards (e.g. UNSPSC, eCI@ss, ebXML, GPC, ...),
 - ~20.000 classes each!,
 - millions of products
- Current SOA: Outsourcing to low-salary countries or use of (counterproductive) low level quality software tools with 25% failure rates
- GoldenBullet (GB IC) offers the exclusive "semi-automatic" functionality to support the classification by manual intervention and to achieve by "learning" a classification level of 95% and speed up the process up to 60 times
- The inclusion of the GB IC into will be an innovative creation of added value and help to reduce outsourcing of labor.

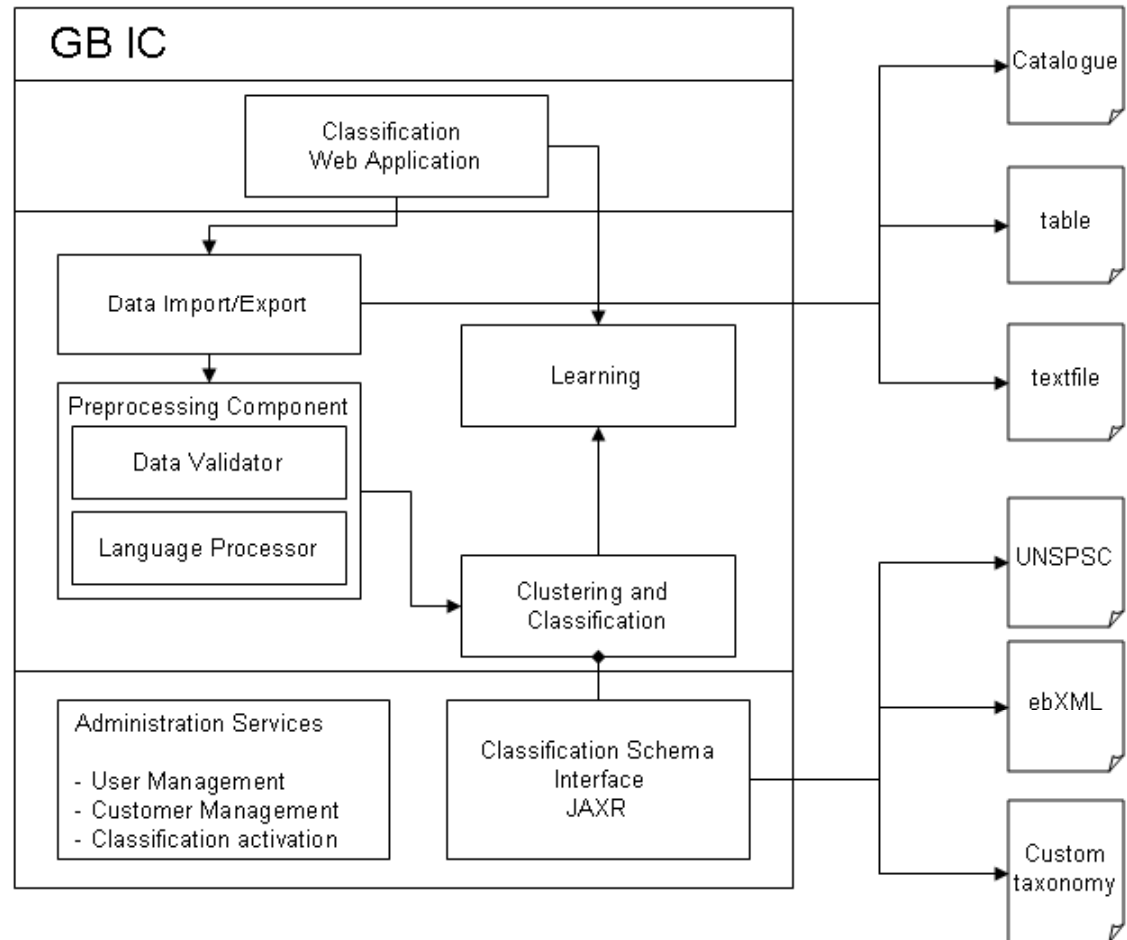


1. GoldenBullet **semi-automatically classifies** product descriptions into ***arbitrary classification standards*** by employing
 1. NLP techniques to preprocess descriptions
 2. Clustering methods to generate representative sub-sets of e-catalog
 3. Machine learning techniques to train the system and automatically generate ranked classification options
2. The user **approves** or **corrects** the proposed classification
3. GoldenBullet **constantly learns** from the user choices and **updates** the classification options
4. The classification rules can be widely exploited as (e.g. as ***Web Service***)



Mapping the workflow to functional modules:

- Standard Software Platform (Java EE)
- Separation of concerns
- Workflow support - implemented in the GUI



Enhanced usability:

- Software can be deployed in a **Java Enterprise Edition Application Server** (e.g. Tomcat, all major vendors)
- The Java EE **XML Registry** is instrumented for storing and accessing classification schema data
- Enables customer **catalogue** taxonomies to be **stored** and **exchanged** over a common format or through a standard API (JAXR).



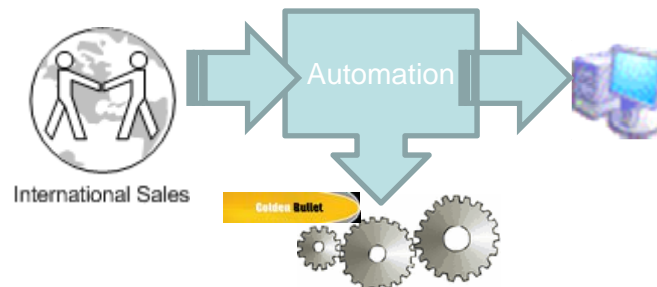
1. Customer specific or industry standard taxonomy (e.g. eClass) is registered – classification domains are defined



2. Manual classification of a representative sample – training of taxonomy specific rules



3. Automated classification using GB Web service – integration in existing eCatalogue processing workflows



- Contact established
- GB demo @SYSTEMS trade fair
- Today: plug – able functions, configuration at contract time
- Emerging: platform eFire
 - Easy integration
 - Ad hoc adaption of services
- Possible exploitation scenario as category search
- Possible improvement with multiple categories



Conclusion:

1. GoldenBullet is a semi-automatic product classification system that offers significant reduction of e-catalog classification effort
2. GoldenBullet transIT project results considerably improve (re-) usability and robustness of the system
3. Results facilitate maintenance and provision of service
4. Industry contacts provided:
 - value-able insights
 - Test catalogue from eHealth
 - test cases
 - Possible ways of exploitation



In future we aim at:

- industry line specific implementation for
 - eTourism & eSales
 - eHealth
 - Social Networks
- Service improvement:
 - Provision of a pre classification (e.g. LifeScience)
 - Refinement of user interfaces
- specialized Web Services with fixed training sets
 - Presumably customer specific
 - Still integrated in the overall code base



- Questions so far?

- <http://www.gbclass.com>



- Slides needed in case there is no internet connection or the
- Demo is not reachable
- No user is registered for demo



1. Project is ProIT funded in co-operation with transIT and CAST
2. Duration 1st September 2007 - 31st August 2008 – extended to Nov 30th 2008
3. Primary Objectives fulfilled:
 - Submission of a debugged, robust and marketable GB IC Prototype
 - Extended Usability and Robustness, Extended Reusability
4. Completed tasks & Status:
 - Worked out contract for handling IPR between stakeholders (UIBK, Excogito NL, BvW Global Pty; Including foundation regulations for marketing and selling)
 - 1st and intermediate report with deliverable of the technical specification accepted by CAST and transIT
 - Cooperation with industrial partner Excogito
5. Extended tasks:
 - Identification and preparations for further exploitation scenarios



1. Project is ProIT funded in co-operation with transIT and CAST
2. Duration 1st September 2007 - 31st August 2008
3. Objectives:
 - Submission of a debugged, robust and marketable GB IC Prototype
 - Extended Usability and Robustness
 - Extended Reusability
4. Completed tasks & Status:
 - Worked out contract for handling IPR between stakeholders (UIBK, Excogito NL, BvW Global Pty)
 - Including foundation regulations for marketing and selling
 - 1st report with deliverable of the technical specification accepted by CAST and transIT
 - Cooperation with industrial partner Excogito

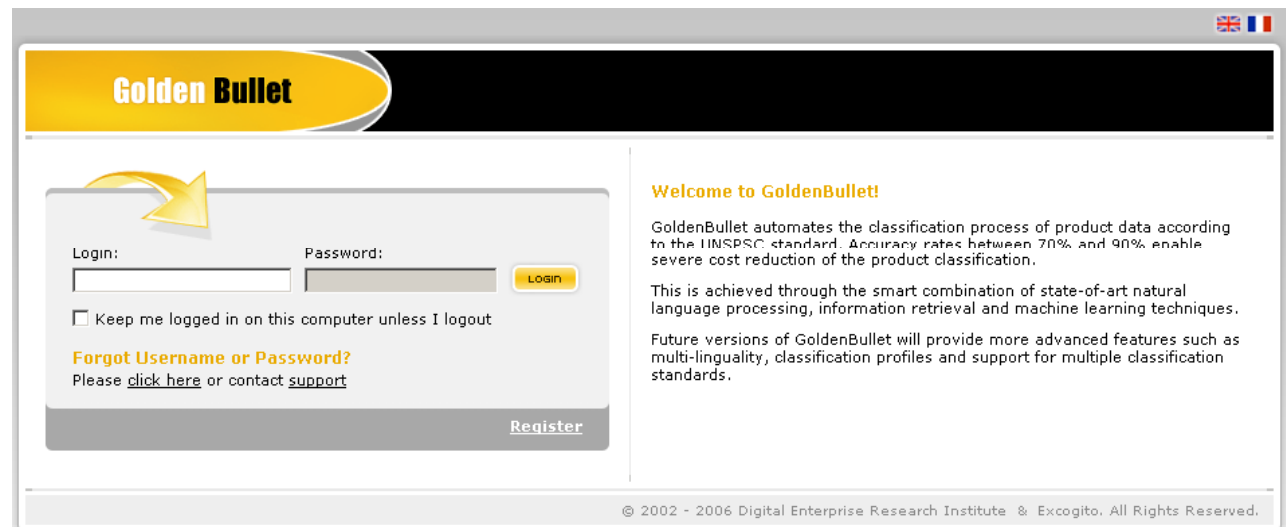


1. Wizards
 1. Data Import/Export
 2. Simple and Expert Training
 3. Classification
2. E-Catalog and UNSPSC Browsers



GoldenBullet IC has an integrated GUI style and continuous designed and brand-like Interface.

- Recognition
- Usability through commonly used symbols
- + improved backend allows to use user specific schemata



- Several file types
- + flexible handling of csv table styles
- + selection of desired / specific schema
- + selection of classification domain (see Training)

Import Data Wizard

Import Data

File type: Text (.txt)

Delimiter: \t Quoted:

File name:

Classification Options

Domain: Default

Scheme: Default

Export Data Wizard

Export Data

File Format: Tab-delimited Text (.txt)

File Name:

File Specifications:

- Available export formats: .TXT, .XLS (planned support)



- + Capable of unlimited catalogue sizes
(only limited by physical borders of the machines)
- + Sort function
- + Search in Item descriptions
- + Extended catalogue management (edit, save, etc.)

The screenshot shows the Golden Bullet E-Catalog Browser interface. At the top, there is a yellow header with the 'Golden Bullet' logo and a navigation bar with 'Settings | Help | Logout'. Below the header, there are tabs for 'Data', 'Expert Training', and 'Classify'. A 'New' dropdown menu and an 'IMPORT DATA' button are visible. A 'Hide selected columns' checkbox is present, along with 'ADD POSITION' and 'DELETE POSITION' buttons. The main content is a table with 10 rows and 6 columns: #, Company, Product ID, Short Description, Long Description, and UNSPSC Code. The table contains various technical specifications and descriptions. At the bottom of the table, there are 'ADD POSITION' and 'DELETE POSITION' buttons, and an 'EXPORT DATA' button. The footer of the interface includes the copyright notice: '© 2002 - 2006 Digital Enterprise Research Institute & Excogito. All Rights Reserved.'

#	Company	Product ID	Short Description	Long Description	UNSPSC Code
1	empty1	empty2	empty3	CAREPAQ 4YRS ONSITE9X5 SITE J+1 ARMADA M700 NS	n/a
2	empty1	empty2	empty3	IMPRIMANTE IJ 6001200X1200 DPI NS	n/a
3	empty1	empty2	empty3	17IN MV740 28MM 1024X768ONLY F/ PRESARIO 5/7000 IN	n/a
4	empty1	empty2	empty3	PRESARIO 7990 ATH-800W/ 15 IN FR	n/a
5	empty1	empty2	empty3	64MB MEMORY SDRAM PC100FOR PRESARIO NS	n/a
6	empty1	empty2	empty3	CAREPAQ BUREAU PROSIGNIA3YRS/SITE/J+1 NS	n/a
7	empty1	empty2	empty3	CAREPAQ YEAR 2000 HEALTHCHECK 1 ADDITIONAL SERVER FR	n/a
8	empty1	empty2	empty3	WARRANTY EXTENSION 3YRSEXCHANGE STANDARD SITEJ+1 NS	n/a
9	empty1	empty2	empty3	CAREPAQ BUREAU PROSIGNIA3YRS/SITE/J+1/TEL NS	n/a
10	empty1	empty2	empty3	TRAINING ACT/ASEEXCEPT TRU64UNIX and OPENVMS NS	n/a



Automatically created representative sub-catalog is provided to the user for semi-automatic classification

+ introduced classification domains to avoid overfitting

Golden Bullet
Settings | Help | Logout

Data Expert Training Classify

Hide selected columns

File Name: non-classified-data-10.txt
Total approved: 0 out of 9 (0.0%)

Pages: 1

#	Company	Product ID	Short Description	Long Description	UNSPSC Code	
<input type="checkbox"/> 1	empty1	empty2	empty3	CAREPAQ YEAR 2000 HEALTHCHECK 1 ADDITIONAL SERVER FR	[100] 81111812 - Computer hardware mainten	
<input type="checkbox"/> 2	empty1	empty2	empty3	CAREPAQ BUREAU PROSIGNIA3YRS/SITE/J+1 NS	[100] 81111812 - Computer hardware mainten	
<input type="checkbox"/> 3	empty1	empty2	empty3	IMPRIMANTE IJ 6001200X1200 DPI NS	[100] 81111812 - Computer hardware mainten	
<input type="checkbox"/> 4	empty1	empty2	empty3	17IN MV740 28MM 1024X768ONLY F/ PRESARIO 5/7000 IN	[100] 43172401 - Moniteurs	
<input type="checkbox"/> 5	empty1	empty2	empty3	WARRANTY EXTENSION 3YRSEXCHANGE STANDARD SITEJ+1 NS	[100] 81111812 - Computer hardware mainten	
<input type="checkbox"/> 6	empty1	empty2	empty3	TRAINING ACT/ASEEXCEPT TRU64UNIX and OPENVMS NS	[100] 81111812 - Computer hardware mainten	
<input type="checkbox"/> 7	empty1	empty2	empty3	CAREPAQ 4YRS ONSITE9X5 SITE J+1 ARMADA M700 NS	[100] 81111812 - Computer hardware mainten	
<input type="checkbox"/> 8	empty1	empty2	empty3	PRESARIO 7990 ATH-800W/ 15 IN FR	[100] 43171803 - Postes de travail ou bureau	
<input type="checkbox"/> 9	empty1	empty2	empty3	64MB MEMORY SDRAM PC100FOR PRESARIO NS	[100] 43171904 - Modules de memoire	

Pages: 1

RECLASSIFY
TRAIN & CLASSIFY

© 2002 - 2006 Digital Enterprise Research Institute & Excogito. All Rights Reserved.



Automatically created classification options are proposed to the user for approval

+ undo / redo functionality

+ rules applied are „unlearned again“ upon undo action

Data Expert Training Classify

Hide selected columns

File Name: non-classified-data-10.txt
Total approved: 0 out of 10 (0.0%)

Pages: 1

<input type="checkbox"/> #	<input type="checkbox"/> Company	<input type="checkbox"/> Product ID	<input type="checkbox"/> Short Description	<input type="checkbox"/> Long Description	<input type="checkbox"/> UNSPSC Code
<input type="checkbox"/> 1	empty1	empty2	empty3	CAREPAQ 4YRS ONSITE9X5 SITE J+1 ARMADA M700 NS	[100] 81111812 - Computer hardware mainten...
<input type="checkbox"/> 2	empty1	empty2	empty3	IMPRIMANTE IJ 6001200X1200 DPI NS	[100] 81111812 - Computer hardware maintenance or sup
<input type="checkbox"/> 3	empty1	empty2	empty3	17IN MV740 28MM 1024X768ONLY F/ PRESARIO 5/7000 IN	[65] 43171904 - Modules de memoire [52] 43171806 - Serveurs
<input type="checkbox"/> 4	empty1	empty2	empty3	PRESARIO 7990 ATH-800W/ 15 IN FR	[46] 43171801 - Agenda electronique [38] 43171803 - Postes de travail ou bureau
<input type="checkbox"/> 5	empty1	empty2	empty3	64MB MEMORY SDRAM PC100FOR PRESARIO NS	[34] 81111901 - Recherche d'informations dans une base [26] 43171902 - Unite de traitement centrale (CPU) proce
<input type="checkbox"/> 6	empty1	empty2	empty3	CAREPAQ BUREAU PROSIGNIA3YRS/SITE/J+1 NS	[25] 43172509 - Imprimantes a jet d'encre [24] 44101700 - Printer and photocopier and facsimile acc
<input type="checkbox"/> 7	empty1	empty2	empty3	CAREPAQ YEAR 2000 HEALTHCHECK 1 ADDITIONAL SERVER FR	[23] 43172401 - Moniteurs [100] 81111812 - Computer hardware mainten...
<input type="checkbox"/> 8	empty1	empty2	empty3	WARRANTY EXTENSION 3YRSEXCHANGE STANDARD SITEJ+1 NS	[100] 81111812 - Computer hardware mainten...
<input type="checkbox"/> 9	empty1	empty2	empty3	CAREPAQ BUREAU PROSIGNIA3YRS/SITE/J+1/TEL NS	[100] 81111812 - Computer hardware mainten...
<input type="checkbox"/> 10	empty1	empty2	empty3	TRAINING ACT/ASEEXCEPT TRU64UNIX and OPENVMS NS	[100] 81111812 - Computer hardware mainten...

Pages: 1

RECLASSIFY
CLASSIFY ALL



The Browser allows the user:

- to locate an appropriate category in the current schema
- manually assign it to a product description
- + search the current schema for the current item
- + search the internet for information on the current item

Change UNSPSC code

#	Company	Product ID	Short Description	Long Description	UNSPSC Code
6	empty1	empty2	empty3	CAREPAQ BUREAU PROSIGNIA3YRS/SITE/J+1 NS	81111812

Wrap long description

Segment

- 80000000 - Management and Bu
- 81000000 - Engineering and Res**
- 82000000 - Editorial and Design
- 83000000 - Public Utilities and P
- 84000000 - Financial and Insura
- 85000000 - Healthcare Services
- 86000000 - Education and Traini
- 90000000 - Travel and Food and
- 91000000 - Personal and Domes
- 92000000 - National Defense an
- 93000000 - Politics and Civic Aff

Family

- 100000 - Professional engineering
- 110000 - Computer services**
- 120000 - Economics
- 130000 - Statistics
- 140000 - Manufacturing technolo
- 150000 - Earth science services

Class

- 11500 - Software or hardware en
- 11600 - Computer programmers
- 11700 - Management information
- 11800 - System administrators**
- 11900 - Information retrieval sys
- 12000 - Data services
- 12100 - Internet services
- 12200 - Software maintenance ar

Commodity

- 807 - Stockage des donnees
- 808 - Analyse de systemes
- 809 - Installation de systemes
- 810 - Software coding
- 811 - Technical support or help de
- 812 - Computer hardware mainte**
- 813 - Computer software mainten
- 814 - Co location service
- 815 - Printer maintenance or supp
- 816 - Mainframe computer mainte
- 817 - Telecom equipment mainte

