Applying Business Rules to Complex Patent Data Processing

Rex Bosma, European Patent Office, Data Flow Operations Manager, rbosma@epo.org

Jacob Feldman, OpenRules, Inc., Chief Technology Officer, jacobfeldman@openrules.com

October 2008

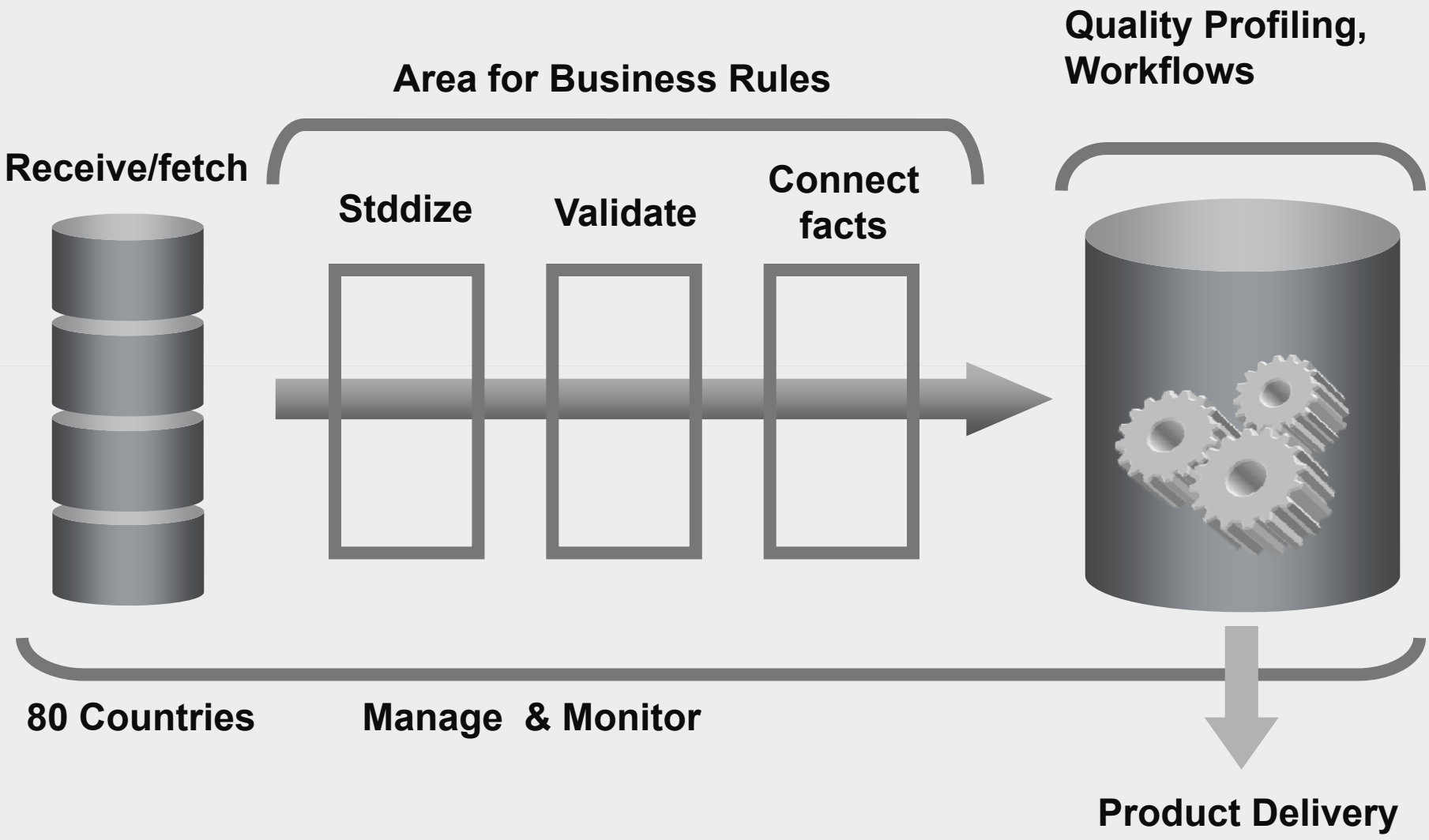
Outline

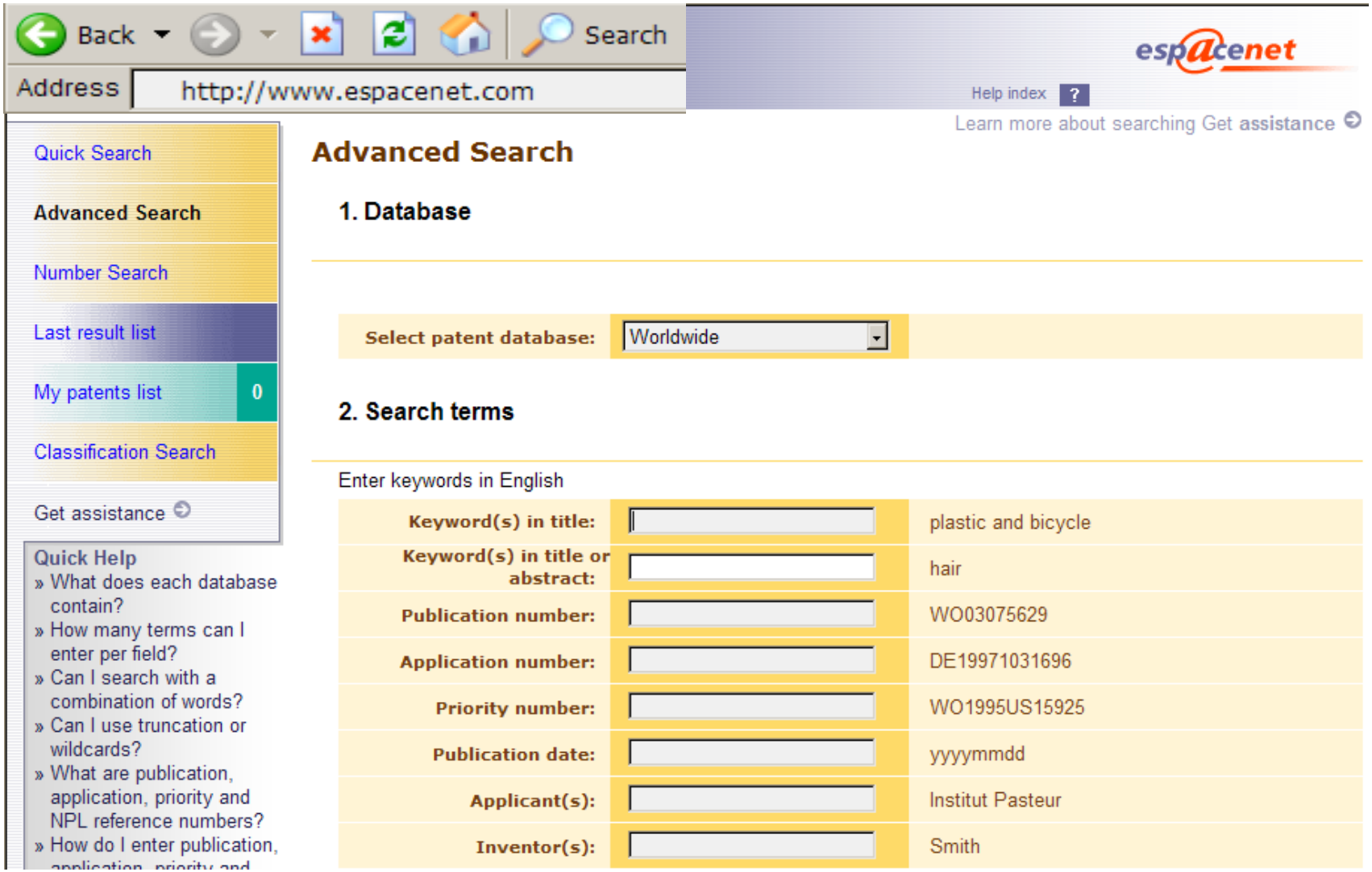
- Rules Based Patent Data Processing in EPO
- Moving Rules Management from IT to Business
- Handling Rules for 80 Countries
- Using Machine Learning for Rules Generation
- EPO Rules Repository: Test and Production Environments

The European Patent Office: Some Facts and Figures

- approximately 35 member states
- in 2006:
  - **200,008** patent applications
  - **170,000** patent searches
- numbers are still growing
- **6500** staff
- 60% patent examiners
- 1100 administrative support staff

The Business Domain of Patent Data Processing




Back
Search
Address
http://www.espacenet.com
espacenet
Help index
Learn more about searching Get assistance
Quick Search
Advanced Search
Number Search
Last result list
My patents list 0
Classification Search
Get assistance
Quick Help
» What does each database contain?
» How many terms can I enter per field?
» Can I search with a combination of words?
» Can I use truncation or wildcards?
» What are publication, application, priority and NPL reference numbers?
» How do I enter publication,
Advanced Search
1. Database
Select patent database: Worldwide
2. Search terms
Enter keywords in English
Keyword(s) in title: plastic and bicycle
Keyword(s) in title or abstract: hair
Publication number: WO03075629
Application number: DE19971031696
Priority number: WO1995US15925
Publication date: yyyymmdd
Applicant(s): Institut Pasteur
Inventor(s): Smith

Cellular phone equipped with a function of measuring gaseous components

Publication number: EP1046910
Publication date: 2000-10-25
Inventor: ITO YUSUKE (JP); KOIZUMI KUNIYOSHI (JP); KENMOCHI HIROKI (JP)
Applicant: TANITA SEISAKUSHO KK (JP)
Classification:
- international: ***G01N33/497; H04M1/02; H04M1/21;*** *H04M1/725;* **G01N33/483; H04M1/02; H04M1/21;** H04M1/72; (IPC1-7): G01N33/497
- European: G01N33/497; G01N33/497A; H04M1/02A; H04M1/21
Application number: EP20000108669 20000420
Priority number(s): JP19990115600 19990423; JP19990149405 19990528

Also published as:
- US6858182 (B1)
- EP1046910 (A3)
- EP1046910 (B1)
- DE60021016T (T2)
- CN1136750C (C)

patent family

Cited documents:
- JP5028324U
- US4823803
- EP0885591
- WO9312604
- JP5322816

more >>

citations

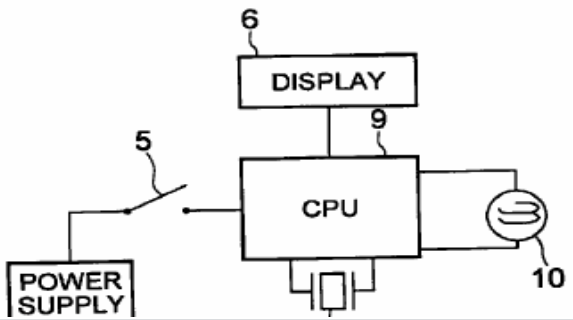
View INPADOC patent family
View list of citing documents
View document in the European Register

Report a data error here

Abstract of **EP1046910**

Disclosed is an improved exhalation gaseous component gauge including, in a palm-sized casing having exhalation taking-in and taking-out slots made on its front and rear sides respectively, at least an electric power supply, a CPU connected to the power supply via an associated switch, a semiconductor gas sensor connected to the CPU and a display device fixed to the front side of the casing, and connected to the CPU. The semiconductor gas sensor is so placed as to permit the breath out of one's mouth to flow over the semiconductor gas sensor on the way from the exhalation taking-in slot to the

FIG.1



Standardization of Patent Number Formats

HANDBOOK ON INDUSTRIAL PROPERTY INFORMATION AND DOCUMENTATION

STANDARD ST.6

RECOMMENDATION FOR THE NUMBERING OF PUBLISHED PATENT DOCUMENTS

Revision adopted by the SCIT Standards and Documentation Working Group at its second session on December 6, 2002

Examples of presentation of publication numbers according to this recommendation:

2001-12345	2001/12345		
2001/1234567	2001/1,234,567	2001/1.234.567	2001/1 234 567
1234567890	1,234,567,890	1.234.567.890	1 234 567 890

10/2003/123456 for a patent for invention
20/2003/123456 for a utility model publication
30/2003/123456 for a design patent, etc.

Business Rules

Patent Standards and Business Rules

- Patent Laws and the number system
  - patent laws differ between countries
  - subject to changes in time

- World Intellectual Property Organisation WIPO proposes standard ST.6 for Patent Number Formats
  - even if adopted, there is a back file of old format

- Not all National Patent Offices have implemented proposal

- Differences in the ***references*** to Patent Numbers by
  - Examiners world wide
  - Patent agents
  - National Patent Offices

Samples of Reality: Facing Bad Data Quality

DE data

10 2004 012 935.5
102005028814.6
43 44 042.8
PCT/DE00/03936
299 09 715 U

US data

148,630
10/003,847
09/239,652
96 756807
60171249
08127,853
&11055832
^09982553

JP data

2002-109652
10-278397
4-347239
HEI.10-292853
106503/2002(PAT.)
P. HEI. 10-108768
>3-56489 +0 U

Business Rules + Noise

Patent Processing Flow ("Good Old Way")

Patent Publication
Data from >**80 countries**

1.Different file formats:
- **XML**
- **many others**

2.Different number standards

PC environment:

Pre-processing
-Rexx and Perl

Mainframe:
- **COBOL** programs
  - 120 country programs
  - standardisation modules

Business Rules Are Everywhere

rejects

manual corr.

Patent Information Resource DocDB DB2

Before BR: “Good Old Way” Summary

- Business specifies → developers implement
- Cobol monolithic data flow process
  - all logic in code spread through modules
- Accumulated solutions to different problems
  - developers lost overview
  - for business: difficult to change standards and formats
- IT change lifecycle days - weeks
- Documentation not always synchronous or available

Looking for a New Rules-Based Approach

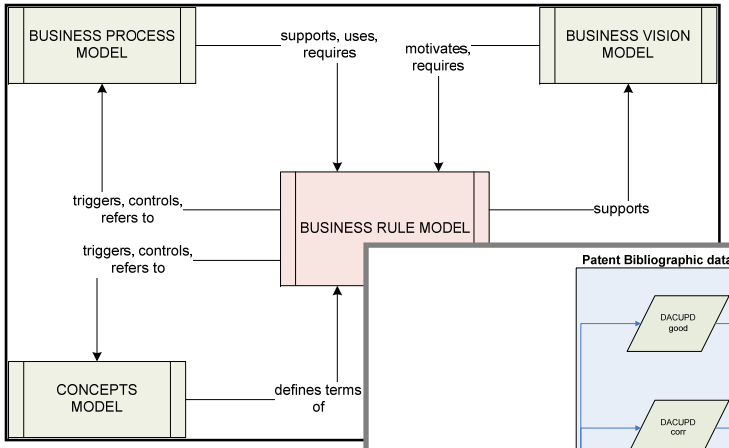
- **Solutions promised by:**
  - Business Rules Approach
  - Business Process Management
  - ETL (Extract, Transform, Load) Products

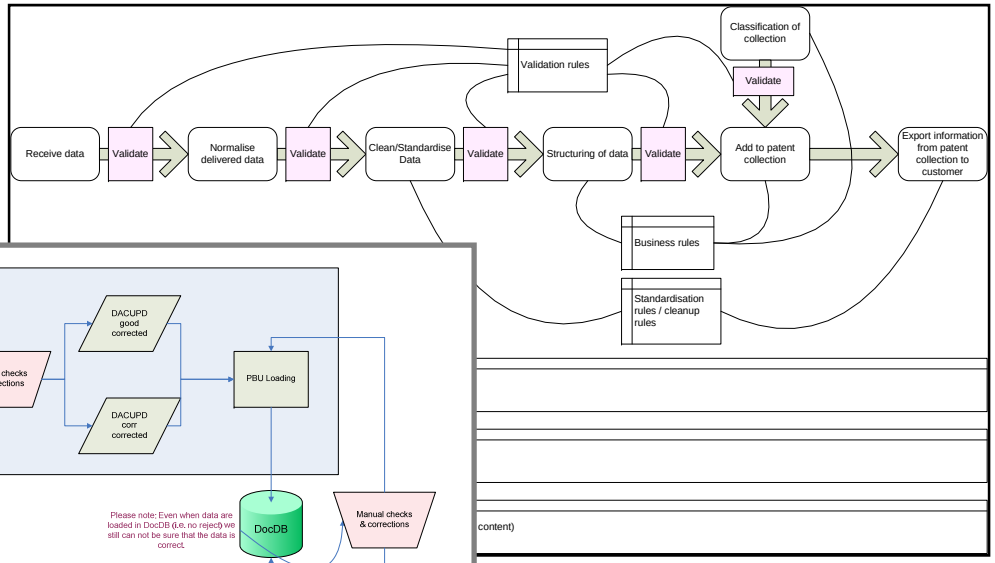
- **Finding the Architecture that fits**:
  - Wide range of choices:
    - Service Oriented Architecture
    - OSGi
    - Other

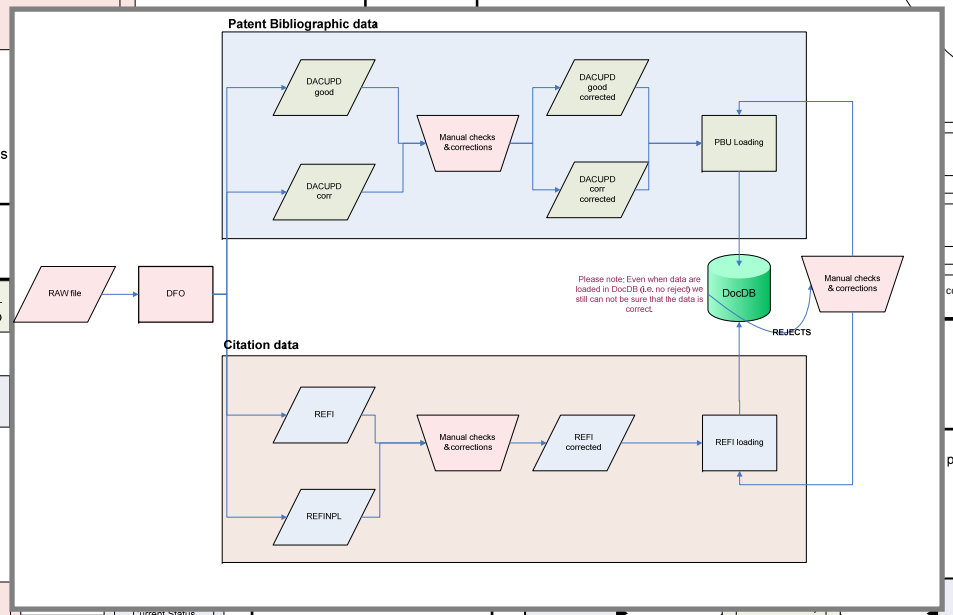
It takes time to understand:

- Where Business Rules fit
- What Business Rules in the context *really* are
- How Business Rules relate to ETL and Data Quality:
  creating value from raw data
- Who's going to do it
  - developers
  - business users / analysts

Rules and Processes Models Created, but...

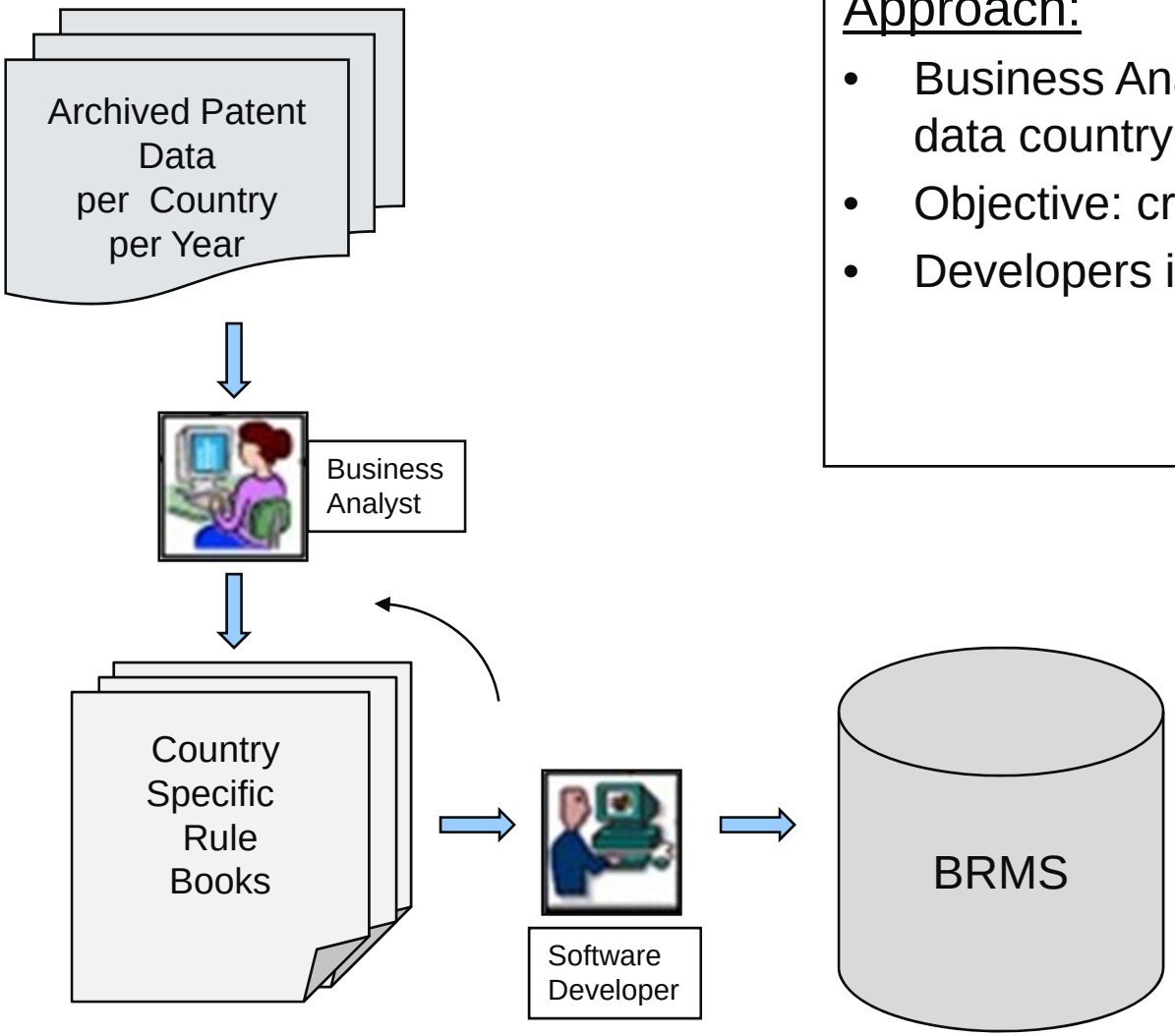






Following the BR methodology,
Paper Rules and Process Models were created but…

Business Rules Approach I



Approach:

- Business Analyst analyzes archive patent data country-by-country year-by-year
- Objective: create Country Rule Books
- Developers implement the rules

Results:

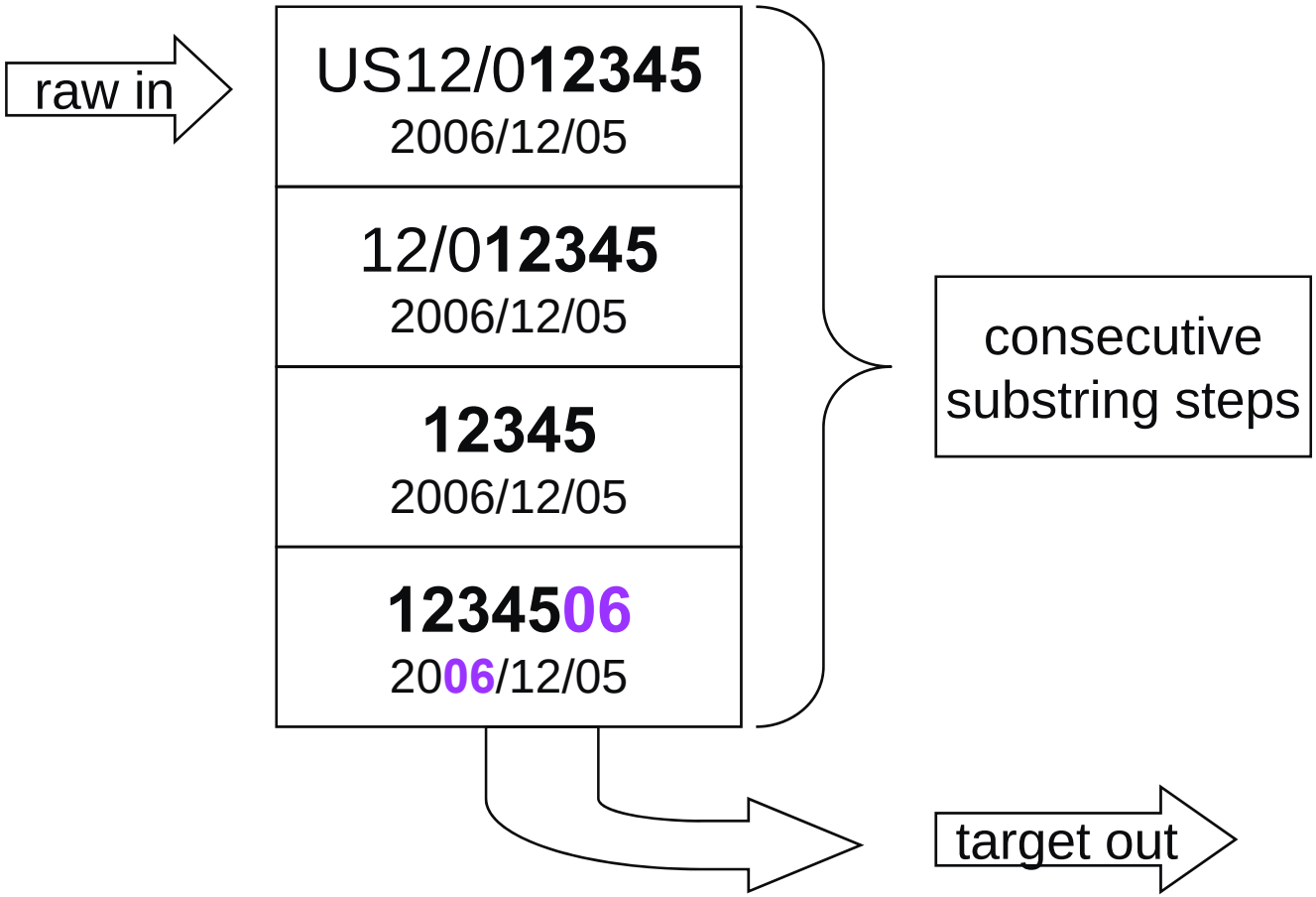
- Externalized business logic
- IT change advantages

BUT:

- Rule Books take too long
- It's still “good old way”: business defines -> developer implements

Approach I: A Fundamental Detail

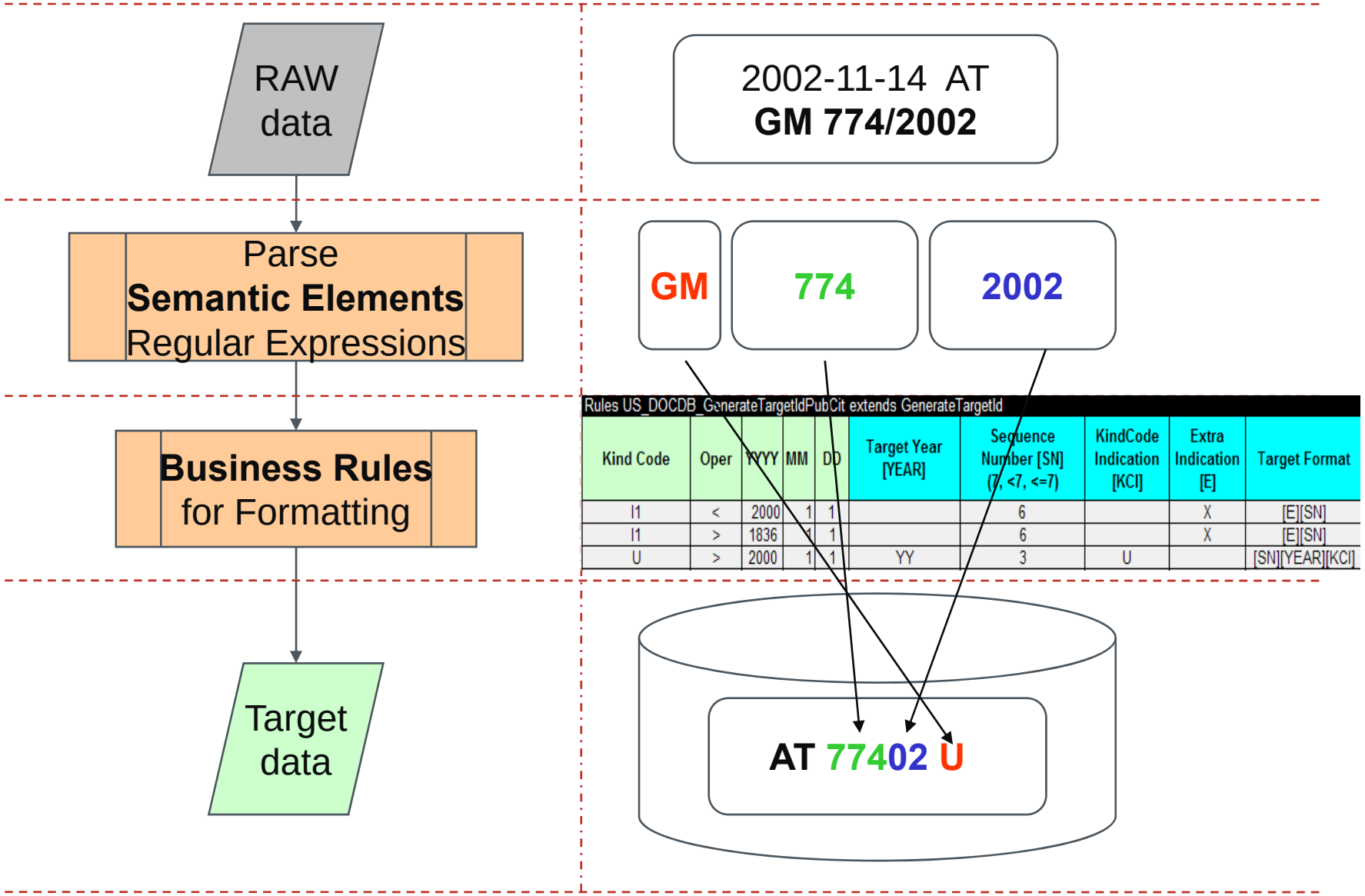
- In fact still based on coding
  - the same **string manipulations** from Cobol
    were transferred to little macros



Switching to Open Rules with Excel

- We learned:
  - Not all products that claim to be BR are really BR
  - BR approach requires a real Rules Engine
  - ... *and* a real BRMS(!)
- Evaluation of Open Source Rules Engines
  - Orientation to Business Analysts
  - Powerful Decision Tables
- Why Open Rules:
  - Fits the tabular logic of our Data Flows
  - Good documentation and support: low learning curves
  - Jump start consulting
  - Business Analysts have no learning curve for Excel tables and can work independently on
    - rules
    - test cases
  - Minimized developer input for creation of rules
  - Fits Architecture
  - Easily adapted to domain specifics

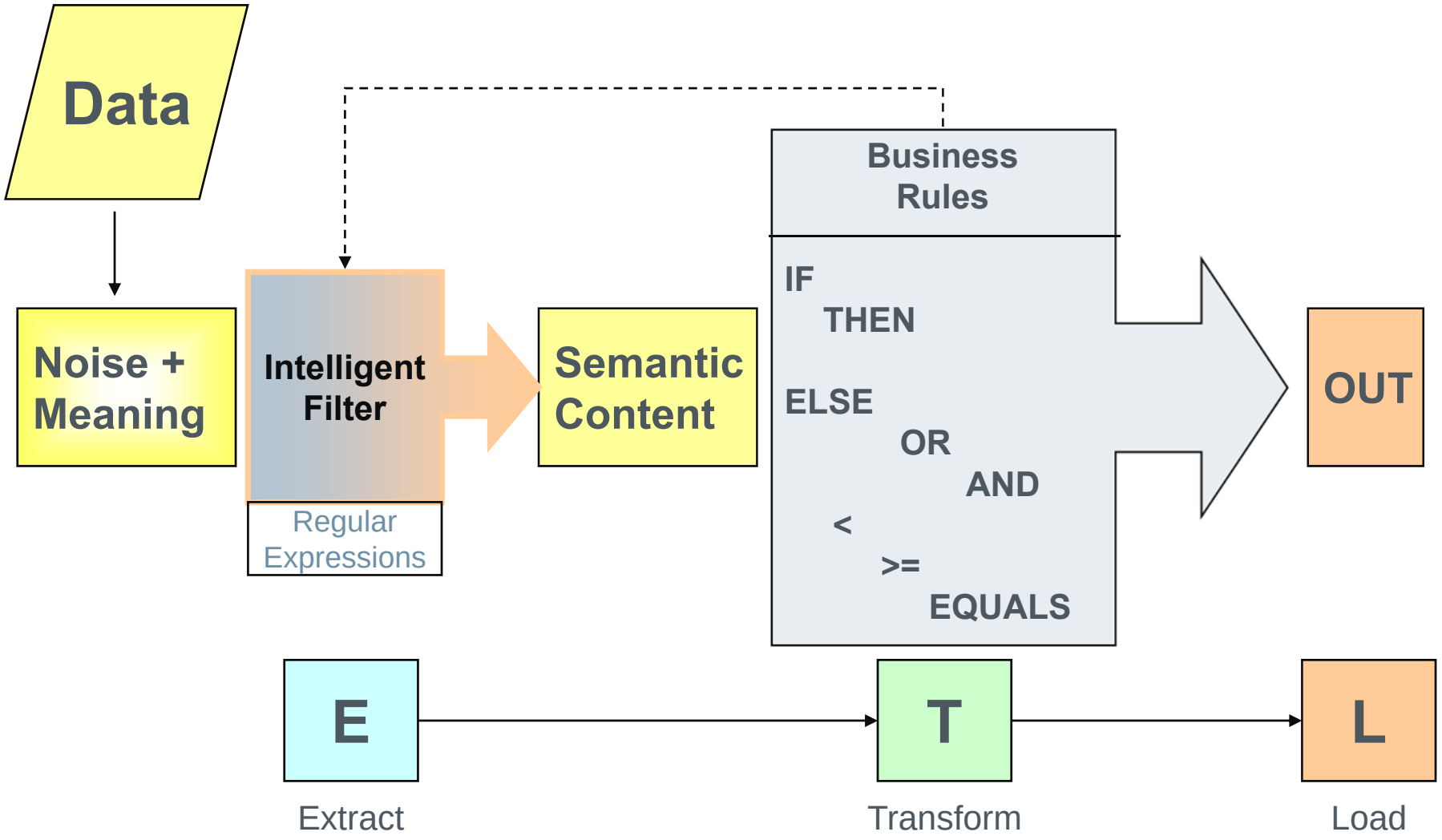
Apply Rules to Patent Data: Example



Rules US_DOCDB_GenerateTargetIdPubCit extends GenerateTargetId

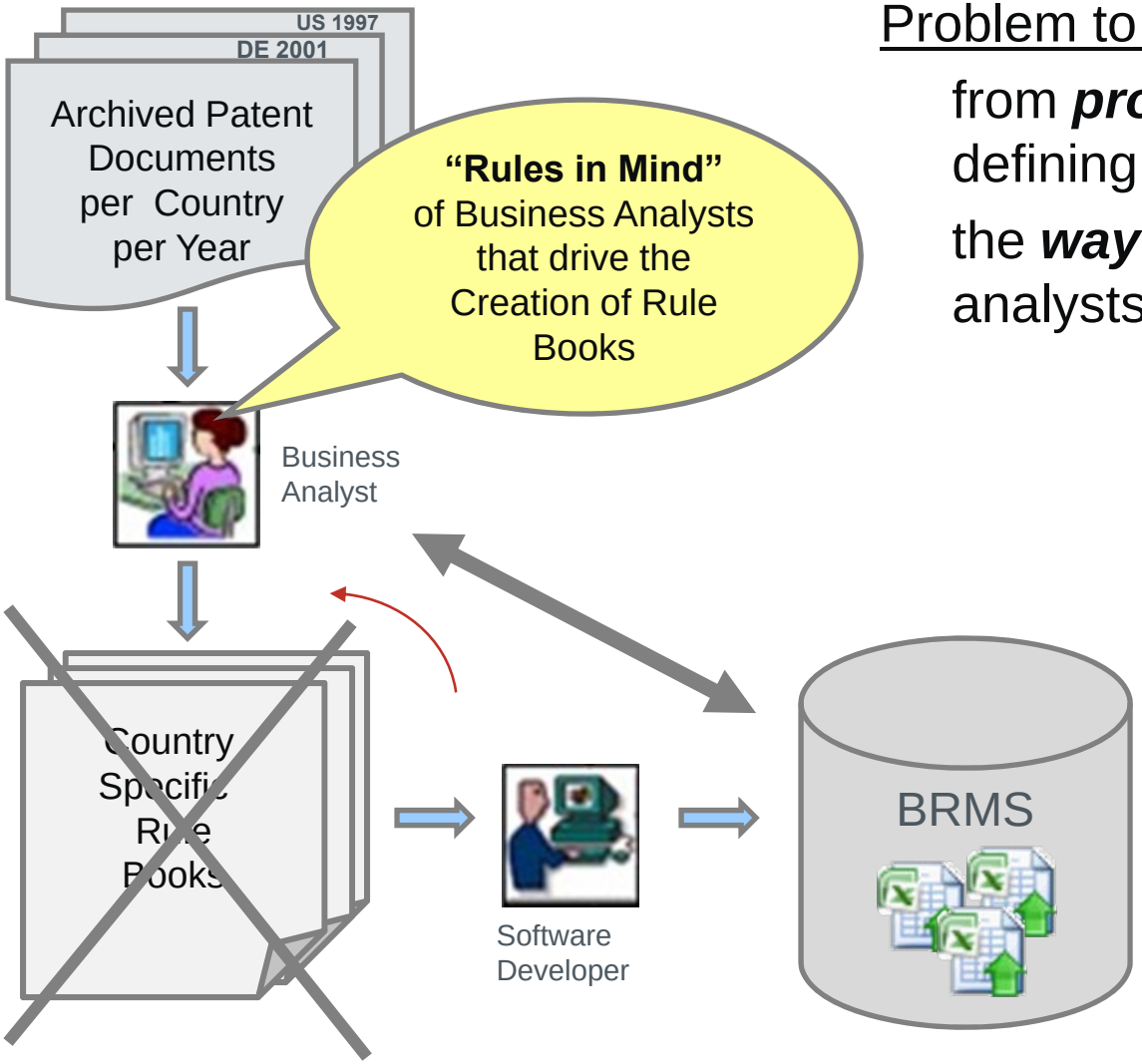
Kind Code	Oper	YYYY	MM	DD	Target Year [YEAR]	Sequence Number [SN] (7, <7, <=7)	KindCode Indication [KCI]	Extra Indication [E]	Target Format
I1	<	2000	1	1		6		X	[E][SN]
I1	>	1836	1	1		6		X	[E][SN]
U	>	2000	1	1	YY	3	U		[SN][YEAR][KCI]

Data Flow Platform and Business Rules



Powerful Data Quality Platform

The *Real* Business Rules Approach



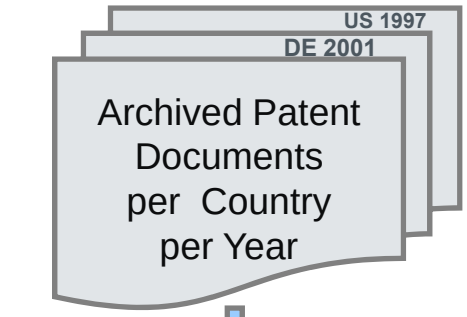
Problem to be solved:

from ***products of thought*** to defining a structure that defines the ***way of thinking*** of business analysts.

New Results:

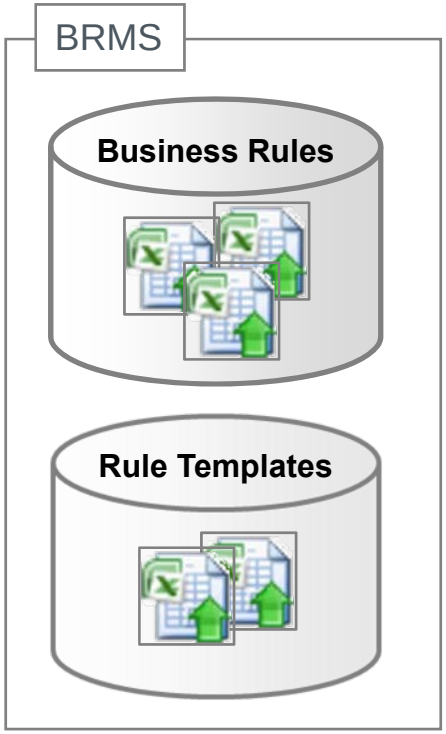
- Business Analysts create and execute Business Rules and Test Cases
- Use Friendly Excel Formats
- Developers help to create Rule Templates but never write Business Rules themselves

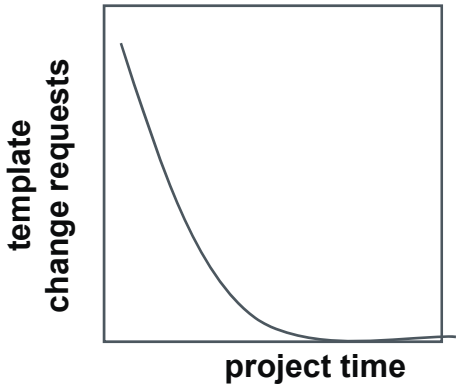
Our Present BR Implementation Scheme



Business Analyst

Software Developer





What Developers Do

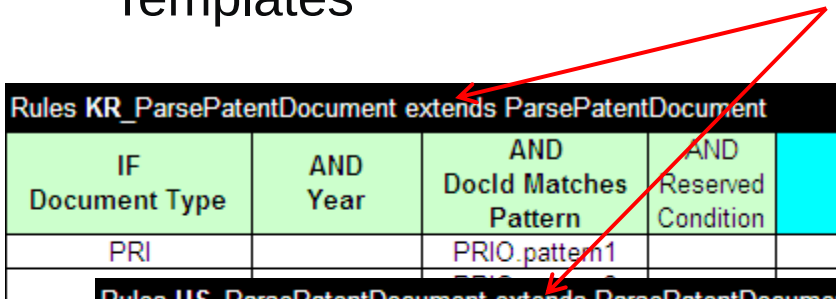
- Create a Business Object Model (XML, Java) to support EPO specific terms and facts
- Create Rule Templates using OpenRules Excel tables to support major types of country-independent Business Rules

Rules String ParsePatentDocument(StandardizationEngine eng)						
C1	C2	C3	C4	A1		
eng.isDocumentType(doctype)	value.compare(eng.doc().date.year)	eng.checkPattern(pattern)		eng.applyPattern(pattern, message, status); return status;		
String doctype	CompareToInt value	String pattern		String pattern	String status	String message
IF Document Type	**AND Year**	**AND DocId Matches Pattern**	**AND Reserved Condition**	**THEN Apply Pattern**	**AND Set Status**	**AND Produce Error**

Business Object Model Used with Java Snippets

What Business Analysts Do

- Create Business Rules for different countries
- Concrete Country Rules Extend the same generic Rule Templates



Rules KR_ParsePatentDocument extends ParsePatentDocument

IF Document Type	AND Year	AND DocId Matches Pattern	AND Reserved Condition	THEN Apply Pattern	AND Set Status	AND Produce Error
PRI		PRIO.pattern1		PRIO.pattern1		

Rules US_ParsePatentDocument extends ParsePatentDocument

IF Document Type	AND Year	AND DocId Matches Pattern	AND Reserved Condition	THEN Apply Pattern	AND Set Status	AND Produce Error
PUB		PUB.pattern1		PUB.pattern1		
PUB		PUB.pattern2		PUB.pattern2		
PUB		pattern3		pattern3		
PUB		pattern4		pattern4		
CIT		pattern5		pattern5		
CIT		pattern1		pattern1	SUCCESS	
CIT		pattern2		pattern2		
CIT		pattern3		pattern3		
CIT		pattern4		pattern4		
APP		pattern6		pattern6		
PRI		pattern7		pattern7		
				NONE	ERRORS	ERROR: No patterns found

What Business Analysts Do (2)

- Created Rules Test Cases for different countries for different countries using simple Excel Tables with Expected Results
- Execute Tests and Maintain Business Rules

Data PatentApplication applications

ID	Country Code	Year	Month	Day	Kind Code	Document Number	Expected Target ID
APP1	US	2002	12	24	A	11123457	

Data PatentPublication publications

ID	Country Code	Year	Month	Day	Kind Code	Document Number	Expected Target ID
PUB1	US	2006	5	30	A	D0513662	
PUB2	US						
PUB3	KR						
PUB4	KR						
PUB5	KR						

Data PatentCitation citations1

id	countryCode	date.year	date.month	date.day	kindCode	rawId
ID	Country Code	Year	Month	Day	Kind Code	Raw Document Number
Cit1	US	2006	5	30		D0513662
Cit2	US	2004	11	12	A	D0513670

Data PatentPriority priorities1

id	countryCode	date.year	date.month	date.day	kindCode	legalStatus	rawId	expectedTargetId
ID	Country Code	Year	Month	Day	Kind Code	LegalStatus	Document Number	Expected Target ID
PR1	KR	2004	01	17	A		PCT2004123456	*20040123456*
PR2	KR	1993	10	16	A	Cont A	KR 1993/25632.52	*932563252*
PR3	KR	2005	01	07	A	Cont	A-2004-12345	*200512345*
PR4	KR	2004	10	16	A		P.10-2004-2619567	*20042619567*
PR5	KR	1988	01	04	F		10-1987-576.2	*885762*
PR6	KR	1992	08	19	F		10 1992 7748844	*927748844*
PR7	KR	2005	02	28	F	Cont	10/2005/16604	*20050016604*

Using Country Specific Patterns

- EPO applies country-specific rules for converting different patent numbers to different target formats
- All rules are created based on the same templates
- Rule Name defines
  - Country (the first prefix)
  - Target Standard (the second prefix)

Rules US_DOCDB_GenerateTargetIdAppPri extends GenerateTargetId

Kind Code	Oper	YYYY	MM	DD	Target Year [YEAR]	Sequence Number [SN] (7, <7, <=7)	KindCode Indication [KCI]	Extra Indication [E]	Target Format
A	<	2001	0	0		6			[YEAR][SN]
A	>=	2001	0	0	YY	7			[YEAR][SN]
B	>=	1994	0	0	CCYY	6			[YEAR][SN]
F	<	1994	0	0	YY	7			[YEAR][SN]
F	>=	1994	0	0		6		DES	[E][SN]
									NONE

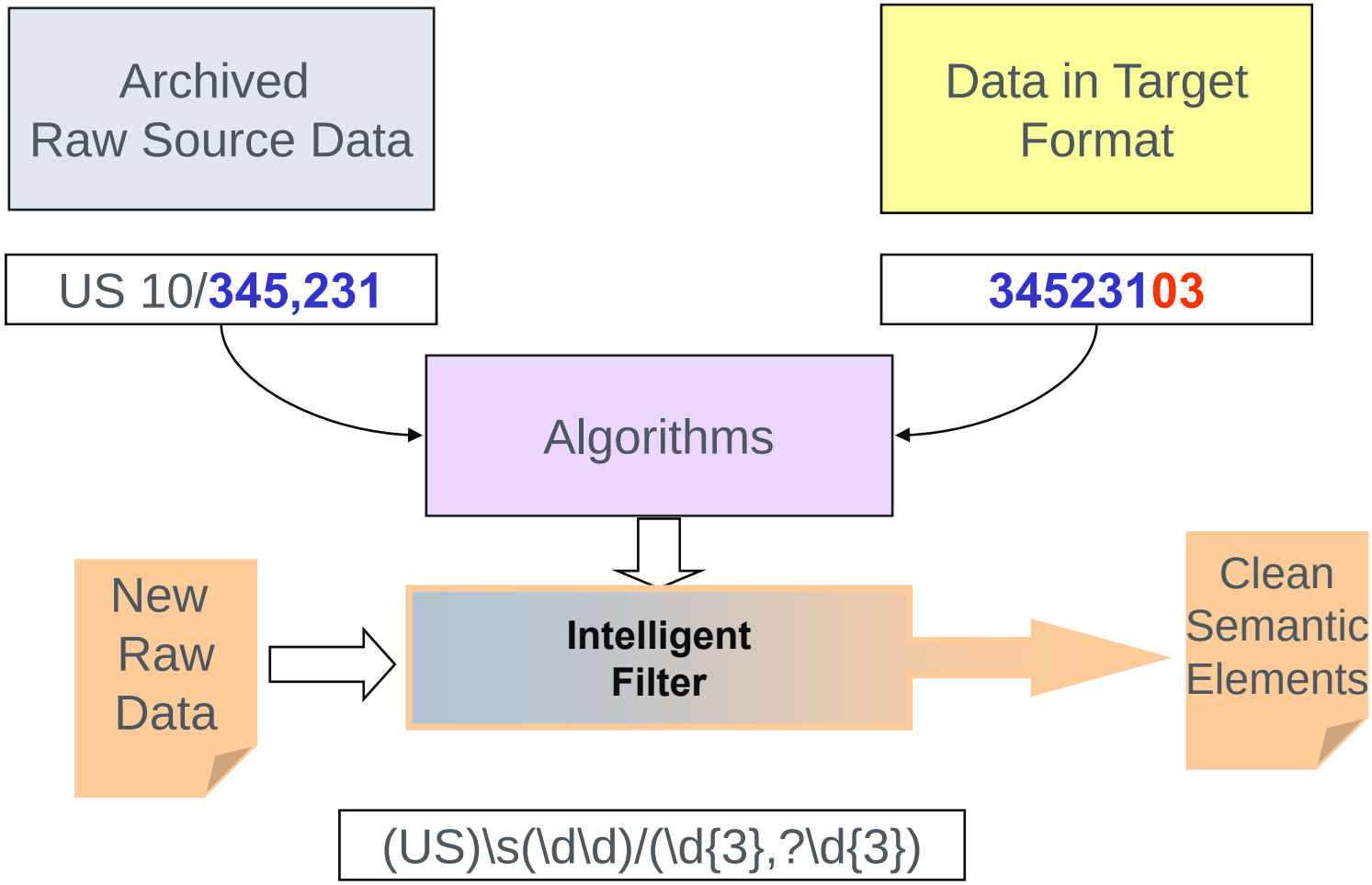
Generating and Using Regular Expressions

To parse raw numbers we apply different Regular Expressions with data capturing groups

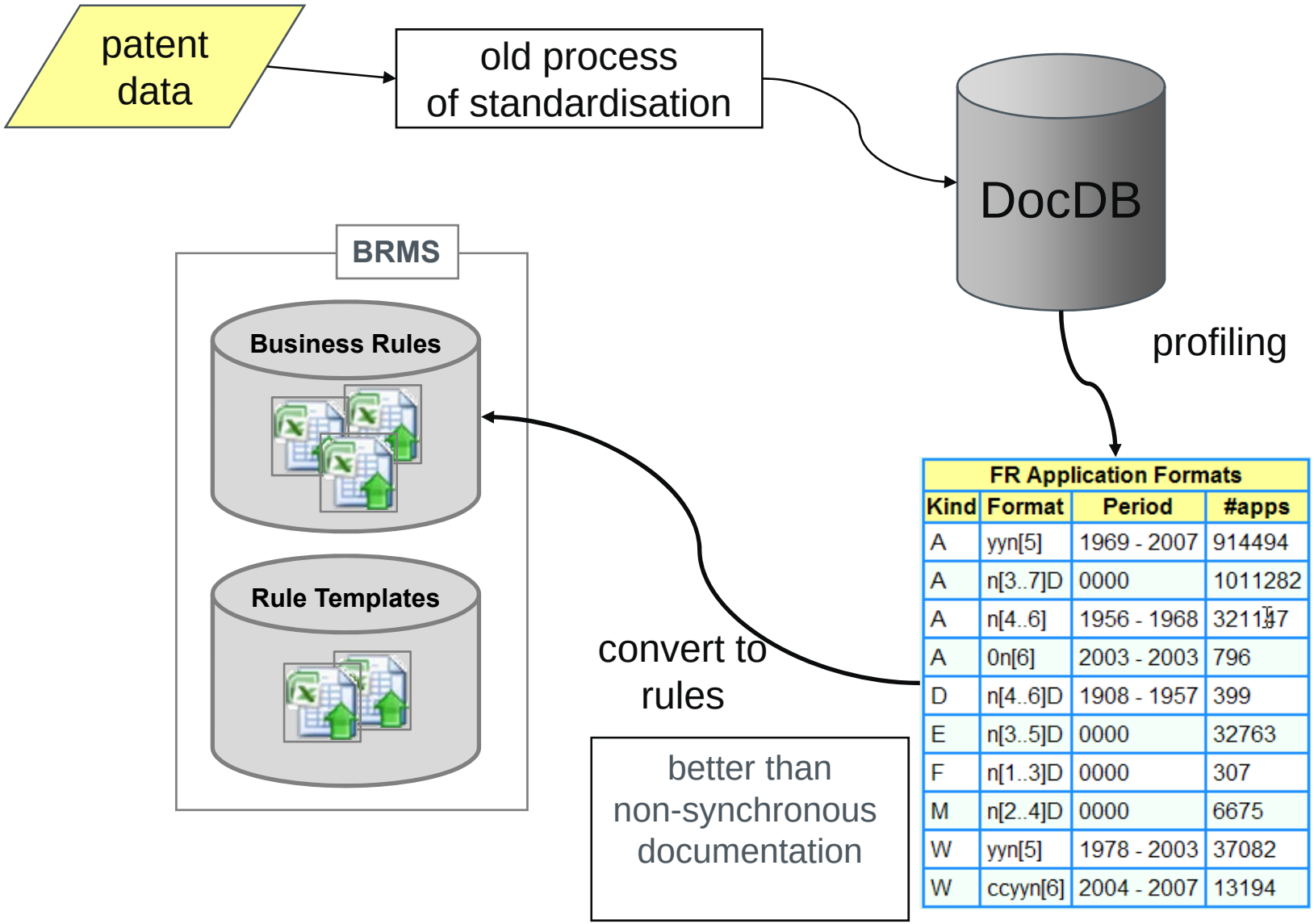
Rules KR_CreatePatterns extends CreatePatterns					
Pattern Name	**Regular Expression**	**Year Group**	**Sequence Number Group**	**Kind Code Indication Group**	**Protection Type Group**
PRIO.pattern1	([A-Za-z]{0,2})\W?(\d\d)?\W?(19|20\d{2})\W(\d+)	3	4	-1	1
PRIO.pattern2	[A-Za-z]{0,2}\W?(\d\d)?\W?(19|20\d{2})\W(\d+)\W?(\D?|\d)	2	3	-1	-1
PRIO.pattern3	[A-Za-z]{0,2}\W(19\d\d|20\d\d)\W(\d+\W?\d{0,2})	1	2	-1	-1
PRI.pattern4	[A-Za-z]{0,2}\W?(\d\d)?\W?(19|20\d{2})\W(\d+)\W?(\D?|\d)	2	3	-1	-1
PRI.pattern5	(PCT)(\d{4})(\d{6})	2	3	-1	1
CIT.pattern1	(\d+)\W(19\d\d|20\d\d)	2	1	-1	-1
NONE	x	-1	-1	-1	-1

- Initially the Regular Expressions were created manually
- Then EPO applied different Machine Learning (ML) techniques that generate regular expressions after analyzing multi-year patent data for different countries
- Such integration of ML and BR technologies shows the real benefits

Example Using Data Mining for Rules Generation 1



Using Data Mining for Rules Generation 2



FR Application Formats			
Kind	Format	Period	#apps
A	yyn[5]	1969 - 2007	914494
A	n[3..7]D	0000	1011282
A	n[4..6]	1956 - 1968	321147
A	0n[6]	2003 - 2003	796
D	n[4..6]D	1908 - 1957	399
E	n[3..5]D	0000	32763
F	n[1..3]D	0000	307
M	n[2..4]D	0000	6675
W	yyn[5]	1978 - 2003	37082
W	ccyyn[6]	2004 - 2007	13194

Target DB Profiles into Rules

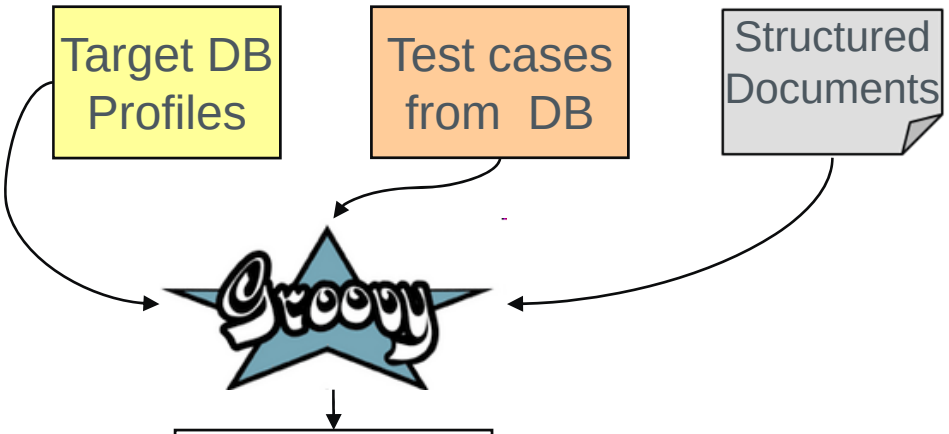
DataBase Profiling result

kind	format	from - until	#occur
M	n[2..4]D	0000	6675
W	yyn[5]	1978 - 2003	37082
W	ccyyn[6]	2004 - 2007	13194

Rules US_DOCDB_GenerateTargetIdPubCit extends GenerateTargetId

Kind Code	Oper	YYYY	MM	DD	Target Year [YEAR]	Sequence Number [SN] (7, <7, <=7)	KindCode Indication [KCI]	Target Format
I1	<	2000	1	1		6		[E][SN]
I1	>	1836	1	1		6		[E][SN]
U	>	2000	1	1	YY	3	GM	[SN][YEAR][KCI]
W	<	2004	1	1	YY	5		[YEAR][SN]
W	>=	2004	1	1	CCYY	6		[YEAR][SN]

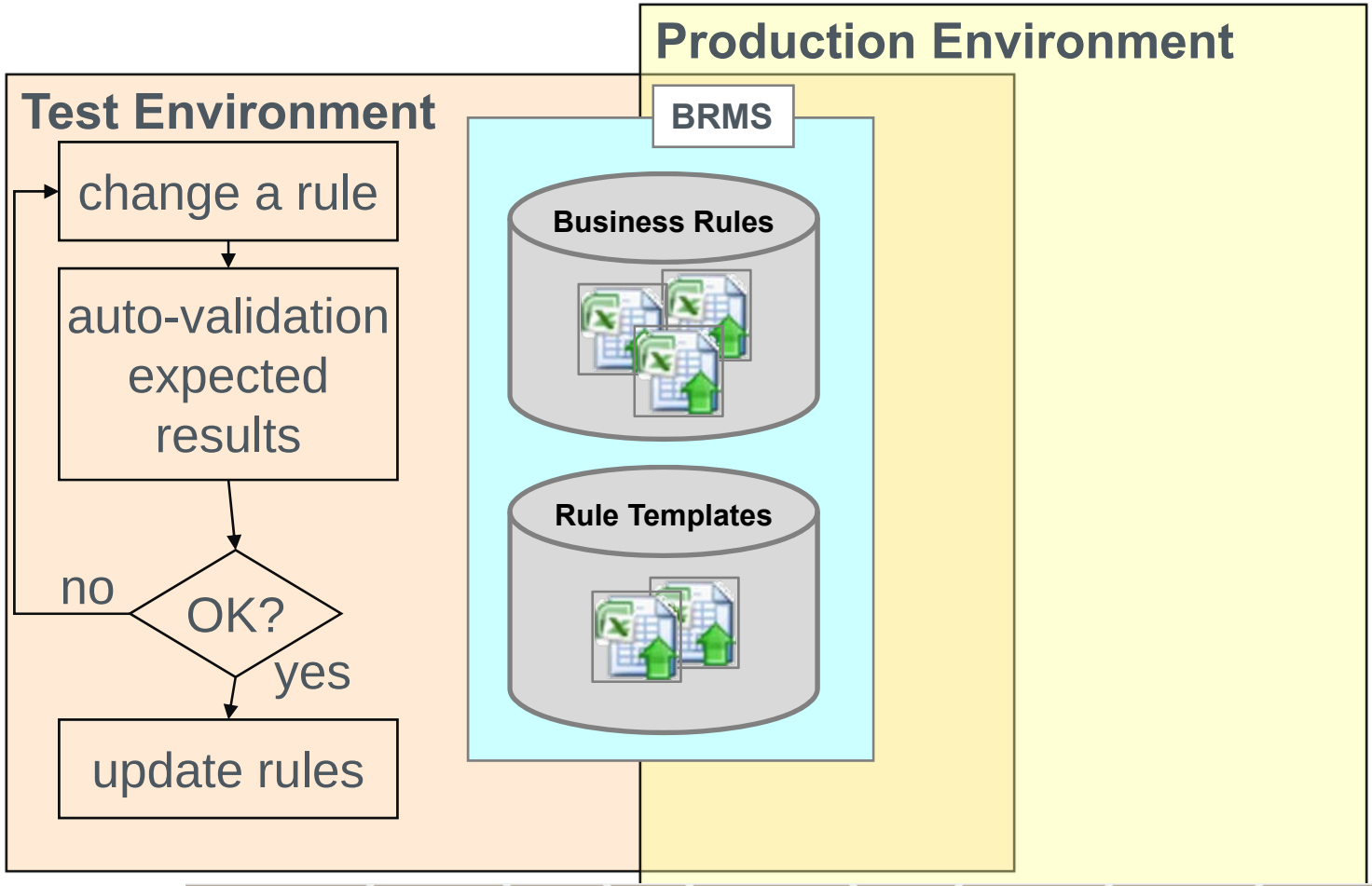
Preparing OpenRules Excel Data Automatically



massive test data + additional info

Data PatentPriority priorities1								
countryCode	date.year	date.month	date.day	kindCode	legalStatus	rawId	expectedTargetId	
Country Code	Year	Month	Day	Kind Code	Legal Status	Document Number	Expected Target ID	%
US	2000	11	22	A		721516	72151600	11.55
US	2004	09	01	P		60606281	60628104	7.49
US	2003	02	28	P		60/451,064	45106403	6.05
US	2001	11	21	W		PCT/US01/44186	0144186	5.88
US	2002	02	05	P		2002 355405	35540502	5.32
US	1990	04	27	A		515.858	51585890	5.23
US	2001	01	05	A		09754593	75459301	4.60
US	2000	12	01	P		250621 P	25062100	3.30
US	1998	11	17	A		09/193,491	19349198	3.05
US	2004	09	15	A		10941594	94159404	2.81
US	1996	01	29	A		08/593.265	59326596	2.78
US	2003	02	26	A		10/375.839	37583903	2.60
US	1995	01	09	W		US9500305	9500305	2.30

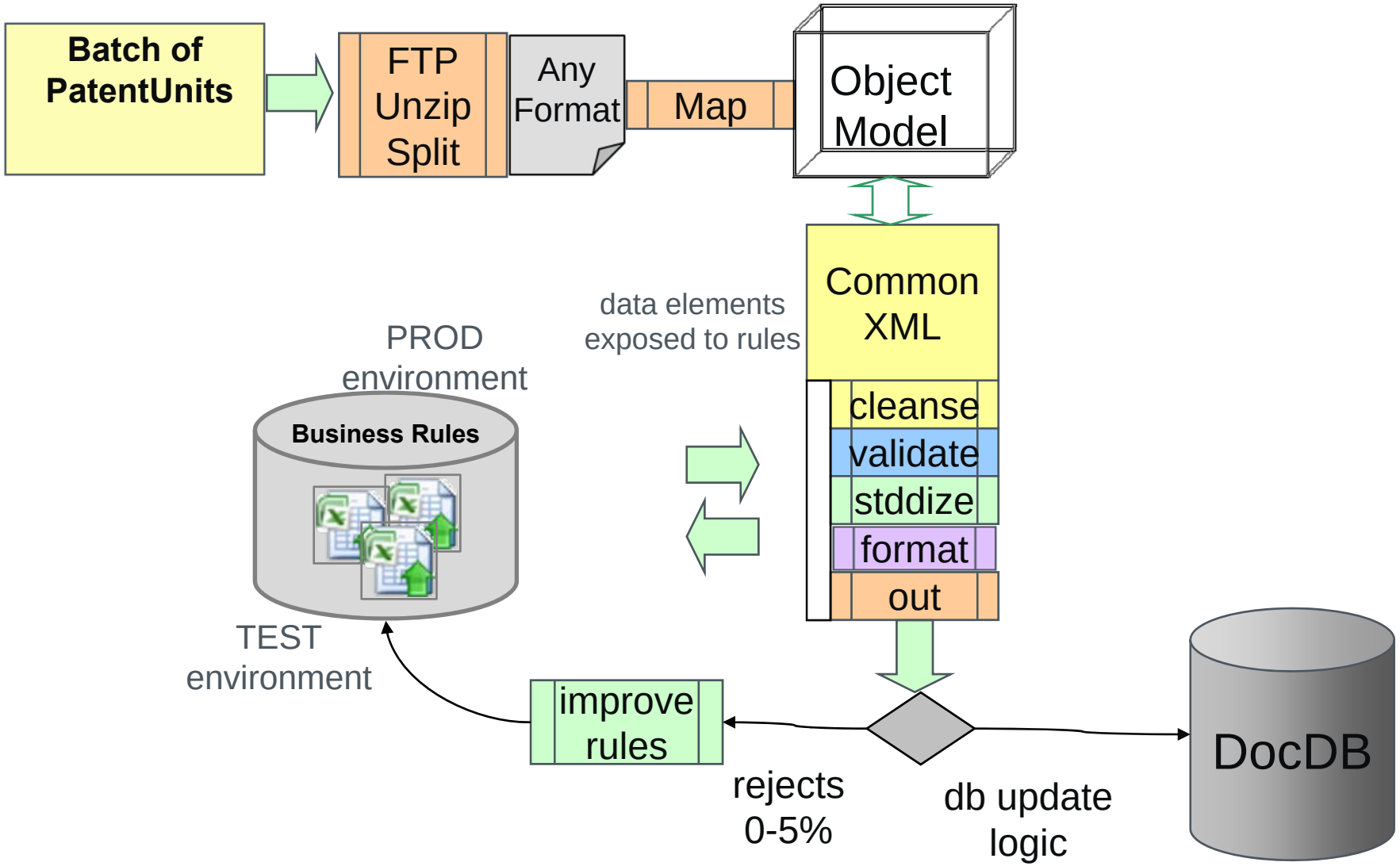
A Rules Repository for Test and Production



regression test report

Document Date (dd/mm/yyyy)	Country Code	Kind Code	LMI	Protection Type	Self Claim	Raw ID	Expected ID	Target ID	Rule Engine Comments
30/5/2006	US	S1	-	DESIGN	-	D0521711	D521711	D521711	SUCCESS
24/12/2002	US	F	-	DESIGN	Yes	29173103	17310302	17310302	SUCCESS

...through the Business Rules



Challenges

- **New Roles and Responsibilities in Rule Management**
- **Management of a Rule Repository for >100 data flows**
- **Change in culture and mindset for users *and* developers**
- **Changing Business Rules dynamically in production: Surrounding the flexibility with security measures**

Summary

The new Business Rules based Data Flow Platform:

- Improved Data Quality with BR:
  - Improved data cleansing
  - Consistent data validations
  - Central service for data formats and standards
  - More self-adapting

- Tabular logic of OpenRules fits Patent Data processing logic
- Data Mining + OpenRules great support for Business Analysts
- Patent Number Format changes can be readily introduced
- Additional rule types can be added
- New country bibliographic data easy to add

END

Questions?