Overview of international experiences with data standards and identifiers applicable for big data analysis

Michal Piechocki
Chairman | Frankfurt Group Technical Workshop
Director | XBRL International Board of Directors
CEO | BR-AG

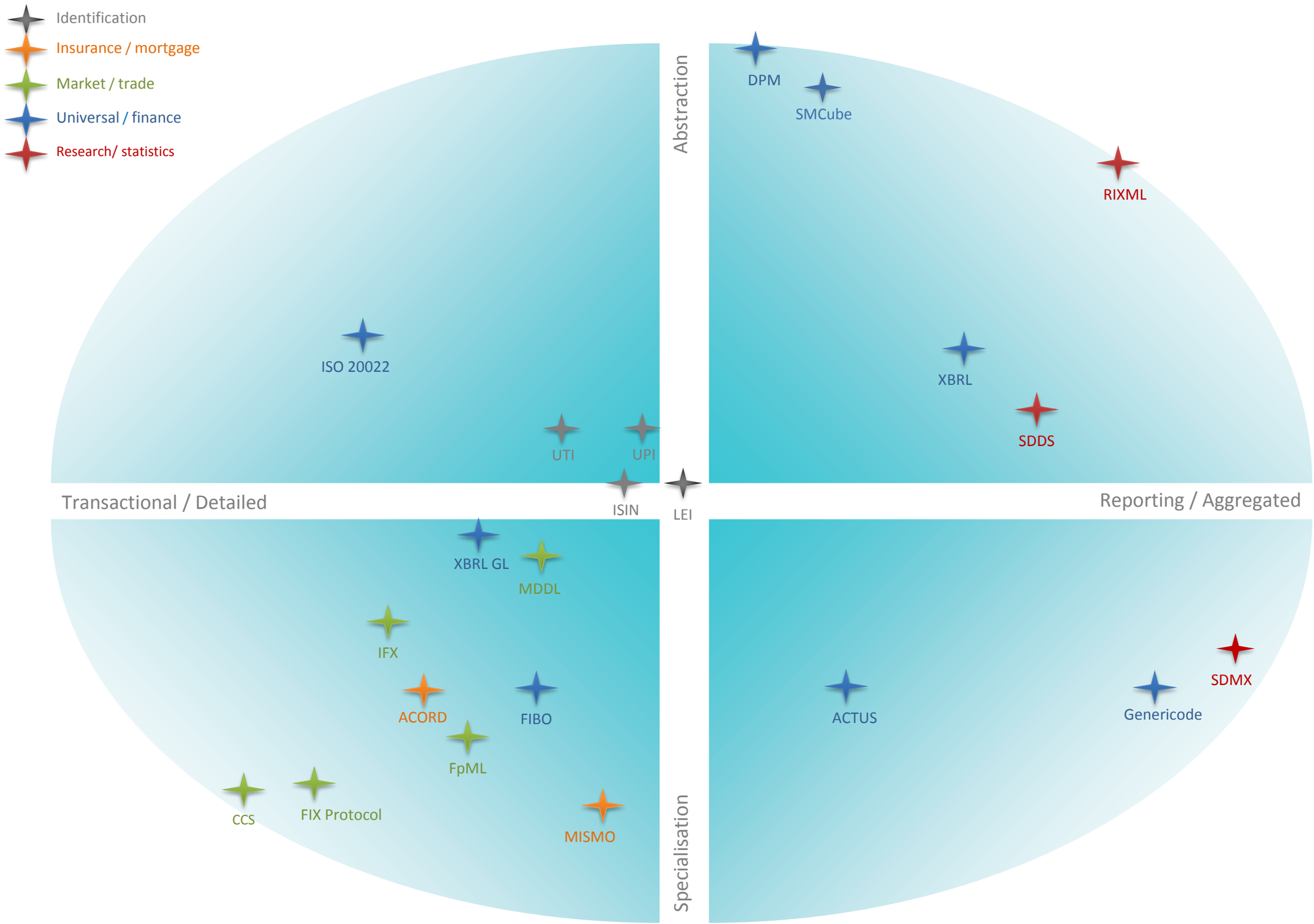
Bali, March 2017

Agenda

- Overview of data standards and identifiers used in the financial industry
- Analysis of data frameworks applicable to financial institutions
- Verification of big data requirements
- Forward-thinking considerations

Central banks, financial supervisors and financial institutions operate at least several data standards and a few identifiers

Data standards and identifiers: map



BRAG

Key data standards in the financial sector include SDMX, XBRL/DPM and ISO20022 and are applicable across multitude of regulations

Key data standards: highlights

SDMX / SDMX-IM	ISO 20022	XBRL / DPM
Statistical	Transactional & business	Supervisory & business
Flows, categories, sets, code lists, concepts, keys, group keys, dimensions, attributes, measures, representations, topics VTL, registries	Dictionary, business process, business domains, business concepts, message concepts Transportation, e-Repository	Dictionary, domains, domain members, hierarchies, dimensions, concepts, facts, linkbases, links Versioning, Rendering, Formula, InlineXBRL, OIM, registries

Based on the European example a typical financial regulator uses a large number of data pools stemming from variety of regulations

Financial data frameworks: overview

1. Capital Requirements Directive IV / Capital Requirements Regulation
2. Money Market Statistical Reporting
3. AnaCredit
4. Balance Sheet Items – Monetary Interest Rates
5. Securities Holding Statistics
6. European Markets Infrastructure Regulation
7. Markets in Financial Instruments Directive II / Markets in Financial Instruments Regulation
8. Securities Financing Transactions
9. Undertakings for Collective Investment in Transferable Securities
10. Alternative Investment Funds Markets Directive
11. Solvency II
12. Target 2 Securities
13. Single European Payments Area
14. Anti Money Laundering Directive IV
15. European Single Electronic Format

Individually none of these data pools falls into the category of big data analysis, but together they may constitute a data lake applicable for big data algorithms

Financial data frameworks: mix

	STANDARD	VOLUME	VARIETY	VELOCITY
CRD IV / CRR	DPM / XBRL	MIXED	MIXED	INFREQUENT
MMSR	ISO20022	GRANULAR	STRUCTURED	FREQUENT
AnaCredit	N/A	GRANULAR	STRUCTURED	INFREQUENT
BSI-MIR	SDMX	AGGREGATED	STRUCTURED	INFREQUENT
SHS	SDMX	GRANULAR	STRUCTURED	INFREQUENT
EMIR	ISO20022	GRANULAR	STRUCTURED	FREQUENT
MiFID II/MiFIR	ISO20022	GRANULAR	STRUCTURED	FREQUENT
SFT	ISO20022	GRANULAR	STRUCTURED	FREQUENT
UCITS	CUSTOM	AGGREGATED	MIXED	INFREQUENT
AIFMD	CUSTOM	MIXED	MIXED	INFREQUENT
Solvency II	DPM/XBRL	MIXED	MIXED	INFREQUENT
T2S	ISO20022	GRANULAR	STRUCTURED	FREQUENT
SEPA	ISO20022	GRANULAR	STRUCTURED	FREQUENT
AMLD IV	UNKNOWN	MIXED	MIXED	FREQUENT
ESEF	inlineXBRL	AGGREGATED	MIXED	INFREQUENT

Importantly data standards, identifiers and dictionaries provide for valuable inputs for big data algorithms: keywords, keys, links and relations

Inputs for big data algorithms

Inputs	Algorithms	Function
• **SMCube Dictionaries** • **Data Point Model Dictionaries** • **SDMX Schemas and Information Model** • **ISO20022 Business Concepts Dictionary** • **XBRL Taxonomies** • **Legal Entity Identifier** • **Universal Transaction Identifier** • **Universal Product Identifier** • **ISIN** • **Ontologies** • **…**	Levenshtein distance	Metric of minimum number of single-character edits required to change one character sequence into another.
	Damerau–Levenshtein	Variation of Levenshtein measuring number of required edits and character transpositions.
	Needleman–Wunsch	Dynamic programming Algorithm based on DNA sequence matching, adopted to character sequences.
	Bitap algorithm with modifications by Wu and Manber	Discrete test whether text contains sequence approximately equal to given pattern. Approximate equality is measured with Levenshtein of given maximum distance.
	n-gram	Statistical analysis of sequence of speech or text (syllables, letters, words …) trying to predict next element of a sequence based only on value of previous element.
	BK-tree	Configuration of character sequences similarity organized in trees based on particular metric (usually Levenshtein)
	Soundex	Phonetic algorithm for indexing words by English pronunciation. Allows words to be matched eliminating differences in spelling.

If we consider these pools jointly with variety of identifiers and potential of mash-up with other data sets the big data algorithms become even more useful

Potential applications

Case	Data frameworks	Data to mash-up
Better identify insurance patterns and claims for technical risk provisions and actuarial assessments	Solvency II	IoT (sensors) / automated information from cars / households / health
Identify suspects of AML	AMLD IV	Information from flight engines for suspicious travels / information from social media on excessive purchases
Identify potential insider trading schemes	MIFIR / EMIR / ESEF / SHS	Family and social relations from social media
Identify related borrowers of loans or relations between issuer and borrower	CRD IV [LE] / AnaCredit	Social, business and family relations from social media
Increase inflation measurement accuracy	BSI-MIR	Surveys, sentiment analysis from social media

THANK YOU

Michal Piechocki

e: michal.piechocki@br-ag.eu
m: +48505558628

Acknowledgments: Michal Skopowski