Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Overview of international experiences with data standards and identifiers applicable for big data analysis[1]

Michal Piechocki,
Business Reporting-Advisory Group

[1] This paper was prepared for the meeting. The views expressed are those of the author and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

BRAG

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

Potential of use of big data analytical methods with standardised, regulatory, financial data sets

Author: Michal Piechocki (michal.piechocki@br-ag.eu)

Abstract

Financial regulators collect and process data, that both meets and contradicts volume, velocity and variety criteria commonly accepted for big data analysis. A number of regulatory data frameworks are described using international standards and identifiers, that may aid in improving efficiency of big data and machine learning algorithms, especially applied with granular data sets, which are increasingly requested by regulators. However, even application of big data methods requires understanding of the researched data and accuracy of information, for precise, unbiased identification of correlations and causations. This paper discusses how standards and identifiers, used across regulatory frameworks, may support application of big data analysis. The paper concludes with identification of further research fields, arising from combination of standardised, regulatory data pools with public data feeds, for discovery of new regulatory insights.

Keywords: data standards, SDMX, XBRL, ISO 20022, granular, transactional, big data, analysis, algorithms, LEI, UTI, UPI

JEL classification: C55, C8, E58, G28, 032, 033

Contents

Excerpts from central banks' leaders speeches

"Big data analytics enables better quantification and pricing of risks, and helps strengthen ex-ante risk resilience measures."

Ravi Menon, Managing Director Monetary Authority of Singapore

"I can see that we stand at the start of a period where central banks, like everyone else, will make use of "big data" and we should learn how to use them to maximize their benefits. While these potential benefits are large, the effort needed is equally significant. We need to invest in information technology infrastructure, but we also need to educate our statisticians how to deal with the new larger and more complicated data sets. (...) Instead of receiving readily usable processed information, we are beginning to demand from reporting agents huge amounts of granular information that is then processed in-house by our statisticians. There is a need to streamline the process of collecting data. In particular, we should exploit to the maximum synergies between the collection of supervisory and (traditionally) statistical data, by developing common definitions to the extent possible, or simple rules to transpose the ones into the others. (...) Central banks are leaving the small safe harbor of simple, aggregate data and are opening up to the brave new world of granular big data. In order not to get lost, we need new skills, more crew, that is statisticians, and stronger vessels, that is better and more versatile models"

Yannis Stournaras, Governor of the Bank of Greece

"Systemic risk, for example, is defined as the contribution of the distress of individual financial institutions (or a group of financial institutions) to overall stress in the financial system, with adverse repercussions on the real economy. The contribution of individual financial institutions to systemic risk is higher, the greater the risk of an individual institution, the larger an institution is (too big to fail), the more connected an institution is (too connected to fail), or the more financial institutions are exposed to common risk factors (too many to fail). This definition shows that systemic risk cannot be analyzed without making use of detailed and granular data on financial institutions. (...)Technological progress has contributed to improved access to micro data and to improved handling of large, granular datasets. (...)It will be possible to reap the full benefits of micro data in terms of efficiency and effectiveness of reporting only if there is close coordination between the scope of existing statistics and newly collected micro data. This may, in some instances, require the scope of existing statistics to be adjusted, and it requires detailed planning when designing new data requirements."

Prof Claudia Buch, Deputy President of the Deutsche Bundesbank

"We are also exploring how we - and others - could use the data the Bank collects more effectively. Big Data has the potential to help the Bank's policy committees identify trends in systemic risk and the economy."

Mark Carney, Governor of the Bank of England

BRAG

Overview of data standards and identifiers used in the financial industry

Introduction

The concept of application of analytical methods over voluminous, high-frequency and diversified data sets has settled well within regulatory environments. As highlighted by the MITSloan Management Review[1], over the past years a number of implementations indicated value of such analysis, for example for estimation of inflation, examination of housing and employment market conditions or studying the impact of high-frequency trading on stock markets by looking at equity transactions.[2] The most commonly used big data analysis methods including: association rule learning; classification tree analysis; genetic algorithms; machine learning; regression analysis; sentiment analysis; social network analysis and other, , coupled with granular data sets, promise, and in some cases already deliver, insightful results. In the process of applying big data methods over public and regulatory data sets researchers[3] and regulators[4] observed however a number of challenges related primarily to:

1. data governance, architecture and understanding;
2. data fragmentation in siloes;
3. performance of IT infrastructure;
4. statistical and analytical methods to limit false positives, data bias or noise accumulation and
5. knowledge and researches availability.

While some obstacles, such as 3 and 4, are gradually being overcome through technological advancements, other, like 1, 2 and 5 remain a key burden in realising the big data potential.

In this article, we will focus on understanding the conditions of and potential solutions to data governance, architecture, understanding and fragmentation hurdles, affecting efficiency of big data analysis.

1 https://sloanreview.mit.edu/case-study/better-data-brings-a-renewal-at-the-bank-of-england/

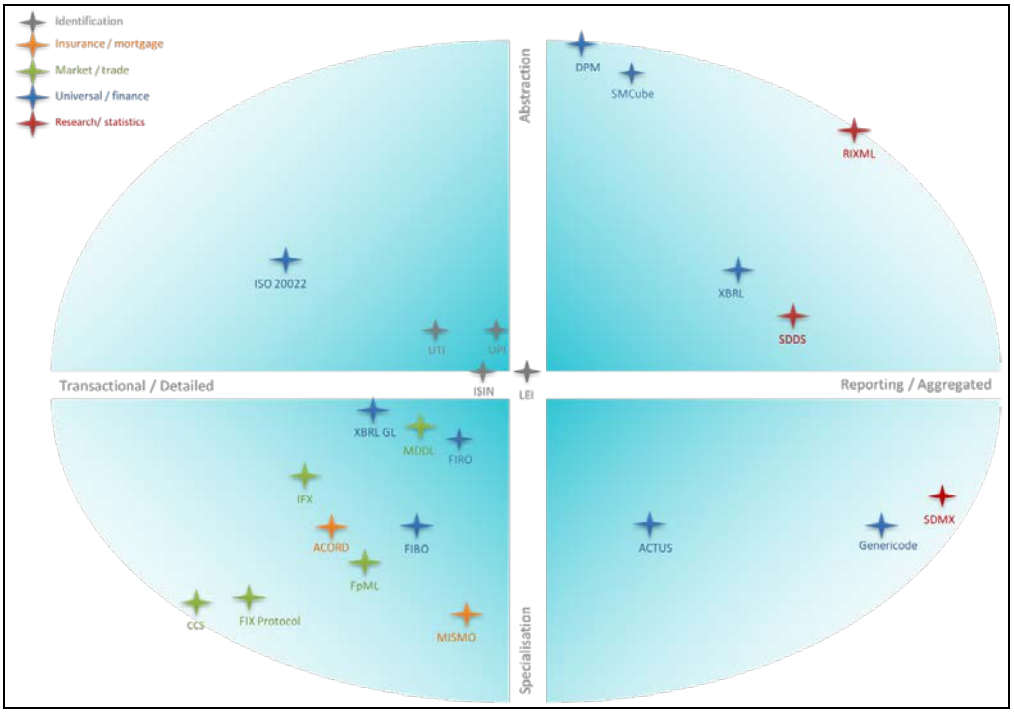
2 N. McLaren and R. Shanbhogue, "Using Internet Search Data As Economic Indicators," Bank of England Quarterly Bulletin 51, no. 2 (2011): 134-140; D. Pimlott and T. Bradshaw, "Bank of England Googles to Track Latest Trends," Financial Times, June 13, 2011; E. Benos and S. Sagade, "High-Frequency Trading Behavior and Its Impact On Market Quality: Evidence From the UK Trading Market," working paper no. 469, Bank of England, London, December 2012, www.bankofengland.co.uk; and E. Benos, A. Wetherilt, and F. Zikes, "The Structure and Dynamics of the UK Credit Default Swap Market," Financial Stability Paper no. 25, Bank of England, London, November 2013, www.bankofengland.co.uk.

3 Xu z, Shi Y., Exploring Big Data Analysis: Fundamental Scientific Problems https://link.springer.com/article/10.1007/s40745-015-0063-7

4 Nagel J. How the Banking Union has transformed banks' IT requirements http://www.bis.org/review/r141209a.pdf

Financial industry and regulatory data standards

Among the main tools utilised by financial regulators, in order to gain understanding of data they process, data standards and identifiers form an important subset, due to their role of enabling data collection, validation and organisation. Contrary to the popular expectation, there exists a large variety of financial data standards. Regulatory and industry experts forming the Frankfurt Group and its Technical Workshop[5] have analysed most common standards applied within the banking and insurance industries, and classified them according to the granular-aggregated axis and generic-specialised axis as presented on the Standards Map[6] below.



Picture 1: Financial data Standards Map

While classification of data standards and initiatives may be subject to experts' perceptions, the standards map demonstrates the heterogeneity of standardisation efforts, often competing across a variety of financial instruments, counterparties or other fields of interest. In summary, the map provides a classification of:

- 2 data description methodologies (DPM, SMCube) applicable to both granular and aggregated data sets;
- 4 granular data identifiers (ISIN, LEI, UTI, UPI);

5 The Frankfurt Group Technical Workshop (FGTW) on Data Standards Interoperability is a discussion forum, organised under auspices of the European Central Bank, gathering regulatory standards experts and conveying quarterly workshops on the topics of data standards, identifiers, methodologies and technologies. The author serves as a chairman of the FGTW.

6 The Standards Map was first published in the internal document of the FGTW: Piechocki M., McKenna K, Dill J. Note on Technical Vision of Standards Interoperability, 2014-06-23

- 16 data standards:
    - 11 granular standards (FixProtocol, FIBO, FIRO, CCS, FPML, MDDL, ISO20022, ACORD, IFX, MISMO, XBRL GL)
    - 5 aggregated standards ((XBRL, SDMX, RIXML, SDDS, Genericode)
- 1 data initiative (ACTUS) describing extremely granular-level data.

The table presents a brief explanation of each component positioned on the map:

Table 1: List of data standards used by financial regulators

Abbreviation	Full name	Purpose
ACORD	ACORD Data Standards and Framework	Data standards for life and annuity property and casualty and for Global Reinsurance & Large Commercial. Claims and settlements messages.
ACTUS	ACTUS Financial Research Foundation	Data and algorithmic standard aiming to break down the diversity in financial instruments into a manageable number of cash flow patterns
CCS	Clearing and connectivity standard	Clearing of OTS transactions
DPM	Data Point Model	Multidimensional data modelling
FIBO	Financial Industry Business Ontology	Define financial industry terms, definitions and synonyms using RDF/OWL and UML
FIRO	Financial Industry Regulatory Ontology	Ontology for description of financial services regulatory domain
FIXProtocol	FIX Protocol	Protocol for international real-time exchange of information related to the securities transactions and markets
FPML	Financial Product Markup Language	Business information exchange standard for electronic dealing and processing of financial derivatives instruments
Genericode	Generic Code	Generic code list representation
IFX	Interactive Financial eXchange	Interoperability of systems seeking to exchange financial information internally and externally
ISIN	International Securities Identification Number	Unique international identification of securities
ISO 20022	Universal financial industry message scheme	Universal financial industry message scheme
LEI	Legal Entity Identifier	Standard for identification of business entities
MDDL	Market Data Definition Language	Standard to describe financial instruments, corporate events and market related indicators
MISMO	Mortgage Industry Standards Maintenance Organization	Data standards that cover the entire mortgage life cycle
RIXML	Research Information Exchange Markup Language	Language for description of investment research documents and other research

SDDS	Special Data Dissemination Standard	Standard for dissemination of statistical information
SDMX	Statistical Data Metadata Exchange	Statistical time series
SMCube	Single Multidimensional Metadata Model	Model used to define the structure of a group of datasets that have been compiled following different modelling methodologies (e.g. SDMX, DPM/XBRL).
UPI	Universal Product Identifier	Unique identification of the OTC derivatives data elements
UTI	Universal Transaction Identifier	Unique identification of individual OTC derivatives transactions required by authorities to be reported to trade repositories
XBRL	Extensible Business Reporting Language	Electronic business reporting
XBRL GL	Extensible Business Reporting Language Global Ledger	Open standard for transactional reporting

It is noteworthy to mention that a number of other initiatives is under way, such as schema.org[7] approach to describe details of financial instruments and transactions, through advanced blend of ontological descriptions with elements of existing standards and identifiers.

Furthermore, it is necessary to mention that the above classification should not, by any means, be understood as canonical. Rather, it represents an early approach to use the key purpose or origin of the specific data standard for initial categorisation. Nevertheless, the authors of the Standards Map recognise that real-world application of various standards crosses boundaries indicated by original intents. For instance, SDMX and XBRL are used to collect highly-granular data in a number of regulatory projects, as will be discussed further in this article. Similarly, granular data standards are increasingly coupled with aggregation mechanisms, in order to reflect aggregated indicators, cubes or groups of data.

Three data standards have been identified as key for the financial industry, and most commonly applied across multitude of regulations: ISO 20022, SDMX and XBRL. It is important to note, that FIX Protocol has also been widely adopted, however, due to strong cooperation between FIXProtocol and ISO 20022, the latter was taken into account. The diagram presents a brief summary of the standards.

7 http://schema.org

Picture 2: Three most popular financial data standards

The ISO 20022 standard is a comprehensive, XML-based standardisation approach that includes a methodology, process and repository to be used by financial standards initiatives. As of date of publication of this paper the ISO 20022 describes processes, data repositories and messages for five domains: payments, securities, trade services, cards and FX.

The SDMX is an initiative and an XML-based standard led by major global regulatory and statistical bodies such as the IMF, the World Bank, the ECB or Eurostat. It describes time-series of data captured through variables according to a defined information model and exchanged through one of technical syntaxes supported by the standard.

The XBRL is an open, XML-based standard for exchange of multidimensional business information described in dictionaries called taxonomies, jointly with mathematical and logical business rules and exchanged as XBRL instance documents.

Application of data standards across regulatory data pools

Based on the European Union example it is possible to analyse which data pools, commonly collected and processed by financial regulators such as central banks, utilise data standards.

We have analysed 15 European Union regulations, initiatives or projects that include data standardisation within banking, insurance or capital market segments. The list of regulations follows:

1. Capital Requirements Directive IV / Capital Requirements Regulation (CRD / CRR)
2. Money Market Statistical Reporting (MMSR)
3. AnaCredit (AnaCredit)
4. Balance Sheet Items – Monetary Interest Rates (BSI-MIR)
5. Securities Holding Statistics (SHS)
6. European Markets Infrastructure Regulation (EMIR)
7. Markets in Financial Instruments Directive II / Markets in Financial Instruments Regulation (MiFID / MiFIR)
8. Securities Financing Transactions (SFT)
9. Undertakings for Collective Investment in Transferable Securities (UCITS)

10. Alternative Investment Funds Markets Directive (AIFMD)
11. Solvency II (Solvency II)
12. Target 2 Securities (T2S)
13. Single European Payments Area (SEPA)
14. Anti Money Laundering Directive IV (AMLD IV)
15. European Single Electronic Format (ESEF)

Each regulation was assessed for factors potentially applicable to big data analysis that is: volume, variety and velocity. For each regulation, a key data standard was identified.

	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

Picture 3: Financial regulations and data standards (EU)

Despite relative subjectivity introduced in quantifiers, it is possible to observe, that data sets already collected and processed by financial regulators, while independently not meeting criteria for big data analysis, treated jointly constitute a voluminous, diversified and high-frequency data pool, that may benefit from application of big data algorithms.

Furthermore, the table demonstrates that financial regulators are slowly, yet steadily, harmonising their data requirements and applying common standards, rather than developing custom approaches. Importantly, most regulatory initiatives rely heavily on data dictionaries, and regulators, such as the ECB, are implementing standardised data dictionaries both, within the organisation (in the ECB: Statistical Data Dictionary based on SMCube), and for communication with supervised parties (for banks: Banking Integrated Reporting Dictionary).

Potential of data standards for big data analysis

As demonstrated on the example of regulations from the European Union, a typical central bank may process standardised data sets that, jointly, may be subject of big

data analysis. For example, a central bank in Asia or Latin America may collect and process similar to CRD/CRR, Basel Acord-driven data sets, such as information on own funds, market, operational and credit risk, leverage, liquidity, large exposures etc. In addition, many central banks already collect detailed granular data on loans or securities, such as AnaCredit or SHS. Collection of granular data related to payments is, by nature, a common experience among central banks. Increasing number of these data sets are collected and processed using standards such as ISO 20022, XBRL or SDMX.

Standardisation of regulatory data brings about at least two advantages for application of big data algorithms:

1. potentially removes the burdens indicated in introduction section related to data governance, architecture and understanding and data fragmentation in siloes;
2. standardised dictionaries, schemas and identifiers provide for valuable inputs for big data algorithms such as: keywords, keys, links and relations.

The picture below presents potential inputs from common regulatory data standards, for a variety of 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.

Picture 4: Inputs for big data algorithms

Since many big data algorithms rely on character-based analysis, structured dictionaries, classifications, ontologies and categorisations provide significant input for training of networks and machine learning algorithms in analysis of regulatory data pools. Particularly methodologies such as the Data Point Model, SMCube or SDMX-IM (SDMX Information Model) may provide for important inputs for big data analytical approaches.

The potential expands, if financial regulators consider combination of regulatory data sets, with public and commercial data pools, through a variety of mash-up techniques. The picture below presents a sample of potential application cases, where regulatory, standardised data, mashed-up with public sets, may provide new insights.

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

Picture 5: Cases for potential application of big data analysis

The aggregated financial information collected in Solvency II tables, together with detailed technical risk provisions information, and detailed assets identification, combined with Internet-of-Things (IoT) sensors, delivering automated information from cars, households or inhabitants, may provide for better identification of insurance patterns, claims for technical risk provisions and actuarial assessments.

Information defined under proposed AMLD IV, combined with information from flight engines for suspicious travels and information from social media on excessive purchases, may support identification of suspects of money laundering.

Mashing-up of transactional data from trade repositories and securities information databases such as SHS, with family and social relations from social media, may aid in identification of potential insider trading schemes.

Similarly, family and social relations information mashed-up with loans data from CRD IV and AnaCredit, may provide for better identification of related borrowers of loans or identify relations between issuers and borrowers.

Last but not least, datasets like BSI-MIR, coupled with surveys and sentiment analysis from social media like Twitter or Facebook, may increase accuracy of inflation measurements.

Conclusions

Big data analysis promises valuable insights and discovery of new correlations and causations across voluminous, diversified and high-frequency data sets. While data sets typically collected by financial regulators like central banks may, individually, not meet criteria commonly accepted for big data analysis, the combination of regulatory data, coupled with public or commercial data, should open a new field of regulatory, supervisory and statistical financial analysis.

In order to realise the benefits of application of big data algorithms regulators should understand and standardise data sets they collect and process according to variety of regulations. Use of global, standardised identifiers should reduce potential bias and enable comparison of analytical results across industries, geographies or instruments.

Numerous standardisation efforts are under way and most advanced financial regulators have embarked on creation of data dictionaries, in order to introduce common understanding of data across organisation. These dictionaries are gradually being extended into the industry, to streamline data sourcing and mapping and therefore increase data quality and accuracy.

Application of big data algorithms over data already collected by regulators, combined with openly available sets, such as social feeds available through API (Application Programming Interface), provide central banks with unprecedented opportunity, to examine practical potential and value of big data analysis.

As highlighted by the speakers cited at the beginning of this paper, analysis of granular data is the new reality for central banks, however its efficient implementation requires understanding and mitigation of challenges identified by researchers and regulators. We, hopefully, demonstrated, that financial data standards, dictionaries and identifiers are not only a commonly applied foundation for building this efficiency, but they may also bring about unexpected value, through discovery and identification of new financial trends and phenomena.

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Overview of international experiences with data standards and identifiers applicable for big data analysis[1]

Michal Piechocki,
Business Reporting-Advisory Group

[1] This presentation was prepared for the meeting. The views expressed are those of the author and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

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

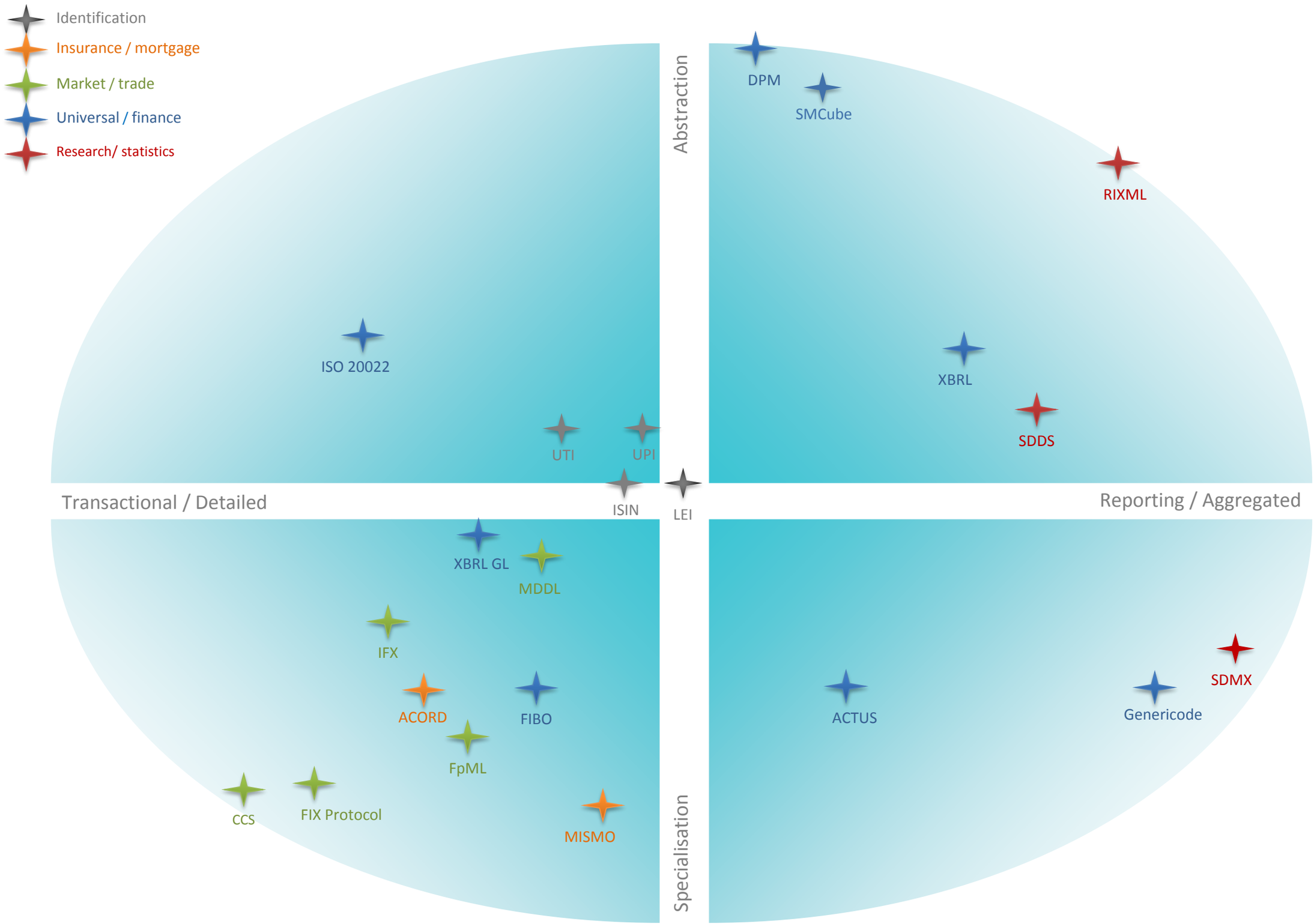
Bali, March 2017

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