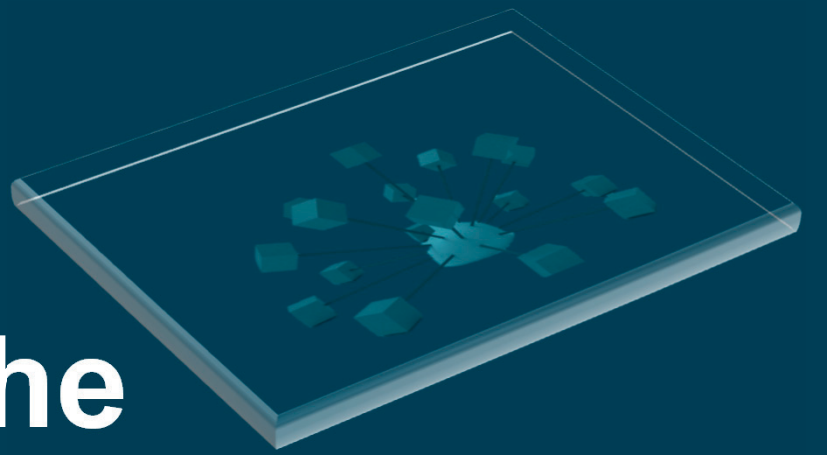
Data sharing experiences at the Bank of Italy



G-20 Workshop on Data Sharing
Frankfurt, 31 January – 1 February 2017

Data sharing experiences at the Bank of Italy

by Giovanni Giuseppe Ortolani

Banca d'Italia
Directorate General for Economics, Statistics, and Research
Statistical Data Collection and Processing Directorate

SUMMARY

1. BoI and data sharing: overview

2. How we deal with the granularity/confidentiality trade-off: two case studies

3. Conclusive remarks

1

BoI & data sharing

Overview

1. BoI & data sharing
The Bank of Italy

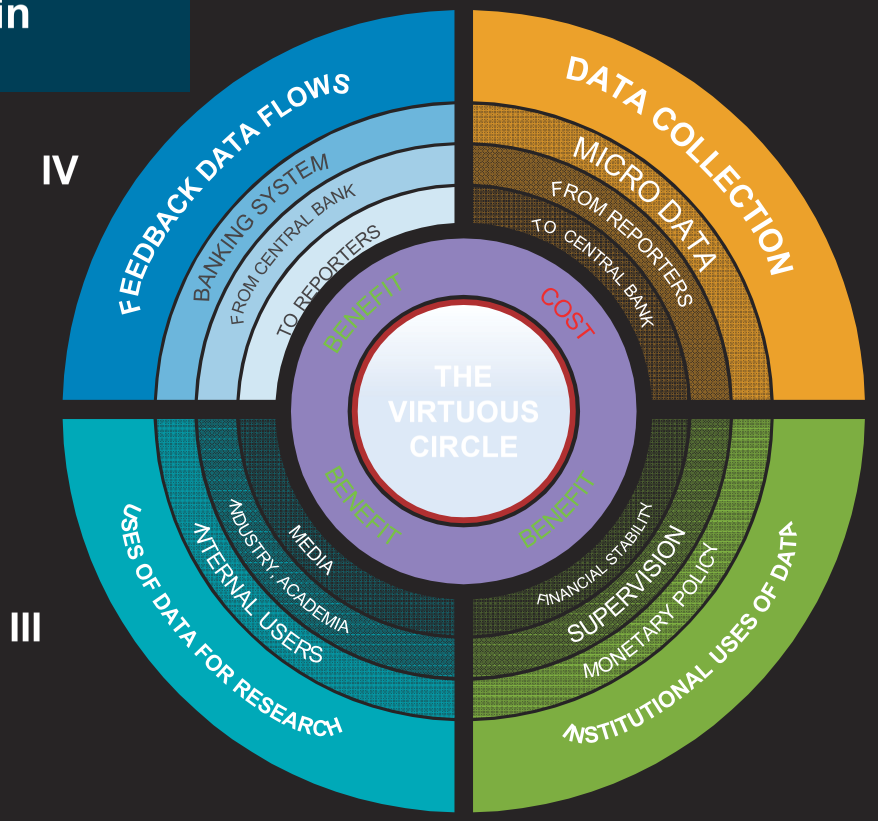
- A strong data producer and a heavy data user: from individual firm information for micro and macroprudential supervision to macroeconomic statistics (e.g. financial accounts, bop/iip).

- Attaching high priority to data sharing, compatibly with the legal framework, to improve coherence and comprehensiveness of the information provided and to reduce the reporting burden.

- Long tradition: sharing of granular information with external users (BoI's survey of households income and wealth) started in the late '80s and is constantly expanding.

1. BoI & data sharing

Enabling a positive feedback: quality in input = quality in output



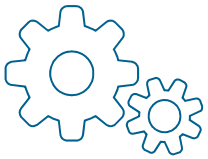
1. BoI & data sharing
BoI's highly integrated statistical system

INPUT

Coordinated data collection from reporting agents, taking into account all information needs.

THROUGHPUT & OUTPUT

Data processing and quality management follow a common approach: unique IT environment, single data dictionary, common data warehouse.



GOVERNANCE

High-level internal "Statistics Committee", including both data producers and users.

1. BoI & data sharing
How do we support data sharing ?

- This high level of integration of the statistical system is the primary strategy to support data sharing, as it ensures consistency of concepts, classifications and other standards of data representation and elaboration.

- In addition, the BoI devised

  - ad-hoc inter-institutional agreements

    A recent example: the agreement signed with central banks of England, Germany, France, and Portugal for exchanging experiences on statistical handling of granular data ("INEXDA").

  and

  - technical arrangements, leading to the development of targeted statistical products and tools (a selection of them are the subject of this presentation).

Case studies

How we deal with the granularity/confidentiality trade-off

2. Case studies

2. Case studies
1. Statistical return flows

Statistical return flows (SRF)

Statistical products elaborated by the BoI and specifically addressed to the reporting intermediaries (mostly banks). They are produced on the basis of the information received by the reporters and provided, free of charge.

Objective

Provide reporting entities information that they can incorporate into their internal information systems. This contributes to the qualitative and quantitative development of intermediaries' management and control tools. Hence, improves the cost/benefit ratio of the statistical production chain (payback of the reporting burden!).

Sources

Supervisory and Central Credit Register (including CCR interest rates).

Link: Bank of Italy - Statistical Return Flows

2. Case studies
1. Statistical return flows

Information provided

Granular information on total values for the whole Italian credit system or segments of it, by phenomenon, residence, currency (euro / non euro) and maturity.

Additional details are provided for specific phenomena of particular interest, e.g. for loans and deposits: sector of economic activity, territorial distribution, financial instrument, counterpart country, etc.

127.000 time series each month (**150 tables**), for **2** reference periods (most recent and revised T-6).
Timeliness: T+ 50 days

Metadata and documentation

The data are accompanied by:

- metadata (elementary and aggregate data dictionaries, respectively, on codifications and aggregation rules)
- technical documentation.

2. Case studies
1. Statistical return flows

User needs
The contents of the flows are agreed and periodically reviewed in cooperation with the financial industry, also via national category associations.

How are data made available to users ?
Via the BoI web-site (INFOSTAT portal), only to authorized users (who registered to access the portal), with appropriate access control protection.

Dissemination strategy
Distributed as a set of databases with 'raw' analytical information. Intermediaries perform elaborations using software developed in-house or acquired from the market.

Planned improvements
A project of revision of SRF is underway (to end in 2018).

Main changes:
- Implementation of FINREP and consolidated data
- Revision of aggregations
- Reports on Loss Given Default (LGD)
- Extension to non-bank financial firms
- Operational enhancements: revised T-3, integration of EBA data template (DPM)
- Possible development of a "data inquiry" application

2. Case studies
1. Statistical return flows

Confidentiality protection criteria

- The BoI Statistics Committee sets the general criteria for confidentiality protection and the circularity of information (who can see what), consistently with legal provisions.
- General rule: to only disclose information resulting from the aggregation of at least 3 subjects (reporters or reported).
- This criterion, when technically feasible and cost-efficient, is applied on a "continuous" basis, i.e. by verifying for each instance that the condition holds true.
- In other cases, an approriate aggregation level is identified based on historical data and periodically verified (this offers the advantage of a certain stability over time of the published aggregations).

2. Case studies

2. Case studies
2. Bank of Italy Remote access Data Base

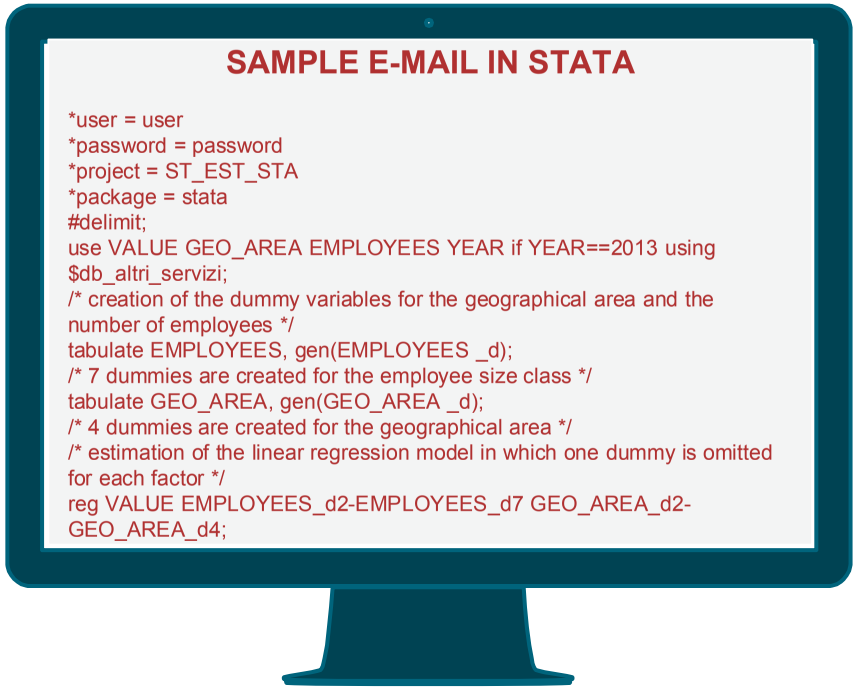
BIRD

A system of data dissemination and elaboration allowing users to run personalized remote econometric/statistical analyses of BoI's micro data in specific datasets while preserving confidentiality.

Objective

Enhancing processing/analysis flexibility for users, compatibly with existing privacy protection legal provisions.

Users carry out their statistical analyses without having direct access to the micro data: they send an e-mail containing a program written in one of the prescribed languages (STATA, R) and the system sends back an e-mail with the results of the calculations.



How BIRD works in practice

700 requests per month

REGISTRATION

User register by filling in a form, in which they provide ID data and accept conditions of use, to be sent by e-mail.

PROGRAM SUBMISSION VIA EMAIL

Once authorized, user can carry out elaborations by sending an e-mail to the system (with a heading identifying the sender) and the program to be run written in one of the prescribed languages.

CHECKS AND DELIVERY OF RESULTS

BIRD carries out a series of formal checks on the commands used. If successful, runs program in batch and, after additional automatic and/or manual checks on the content, results are sent to users' e-mail addresses.

2. Case studies
2. Bank of Italy Remote access Data Base

BIRD DATA SETS

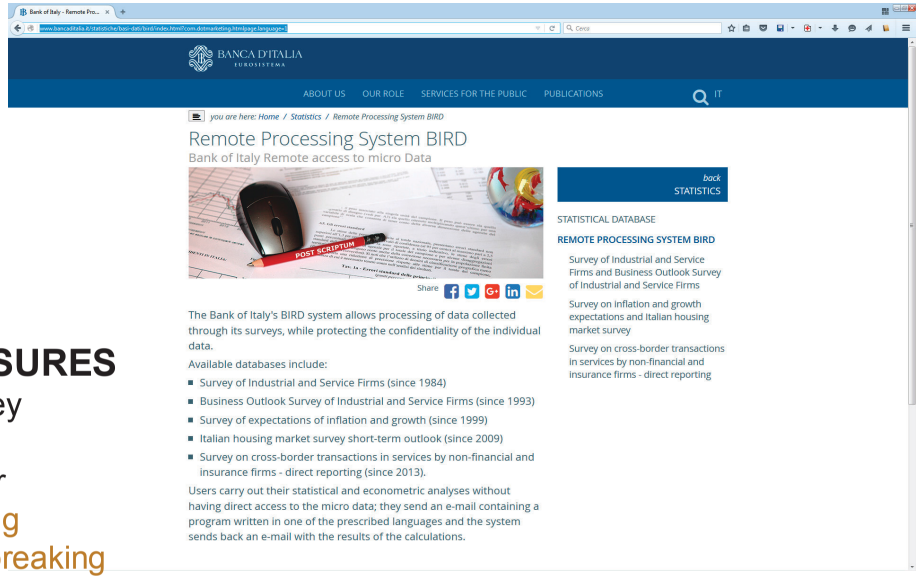
- Survey of Industrial and Service Firms
- Business Outlook Survey of Industrial and Service Firms
- Survey of expectations of inflation and growth
- Italian housing market survey short-term outlook)
- Survey on cross-border transactions in services by non-financial and insurance firms - direct reporting
- *Bank's balance sheet micro data (by February 2017)*

CONFIDENTIALITY PROTECTION MEASURES

Data are anonymized: key identifiers removed from database. Privacy further safeguarded by forbidding potential confidentiality-breaking programme statements (e.g. the "list" STATA command). A series of automatic and manual checks further ensures confidentiality.

TARGET USERS

Mainly external (and internal) researchers, but the tool is potentially able to serve the needs of other national and international institutions



Link: Bank of Italy - Remote Processing System BIRD

IT INFRASTRUCTURE

Based on the LISSY platform (also adopted by the Luxembourg Income Study and other research centers), driven by plain-text e-mails.

3. Conclusive remarks

- A cultural shift is taking place: leaders in regulatory agencies, industry and academia recognize the value of data sharing, also to enhance transparency. They are now focusing on how — instead of why — data should be shared.

- The BoI primarily relies on the high level of integration of its statistical system, which ensures standardization. A relevant issue also at the international level, in particular in the ESCB context, with a view to forthcoming developments in the field of statistical / supervisory information.

- In addition, according to BoI's experience, an effort is needed to develop and improve targeted inter-institutional agreements and technical arrangements. As to the latter, feedback data flows and remote access proved to be effective strategies.

- Further work and the exchange of experiences in the sharing of granular information, also about possible organizational changes, is crucial to improve scope and possibility of integration of the data and process efficiency.

“If you torture the data long enough, it will confess.”

Ronald Coase, Economist


THANKS!
Any questions?
BANCA D'ITALIA
EUROSISTEMA
PALAZZO KOCH
SEDE STORICA DELLA BANCA D'ITALIA
DALLA SUA FONDAZIONE (1893)
PROGETTO DELL'ARCHITETTO GAETANO KOCH
giovannigiuseppe.ortolani@bancaditalia.it

TÜRKİYE CUMHURİYET MERKEZ BANKASI

Data Sharing

CBRT Practice

Erdem Başer

Statistics Department

Data Security

Data Security

Data Protection

Data Sharing

Data Sharing

Benefit:

Adress the need of accessing micro data for growing academic research

G-20 Data Gaps Initiative 2 (DGI-2)

Recommendation II.20: Promotion of Data Sharing by G-20 Economies

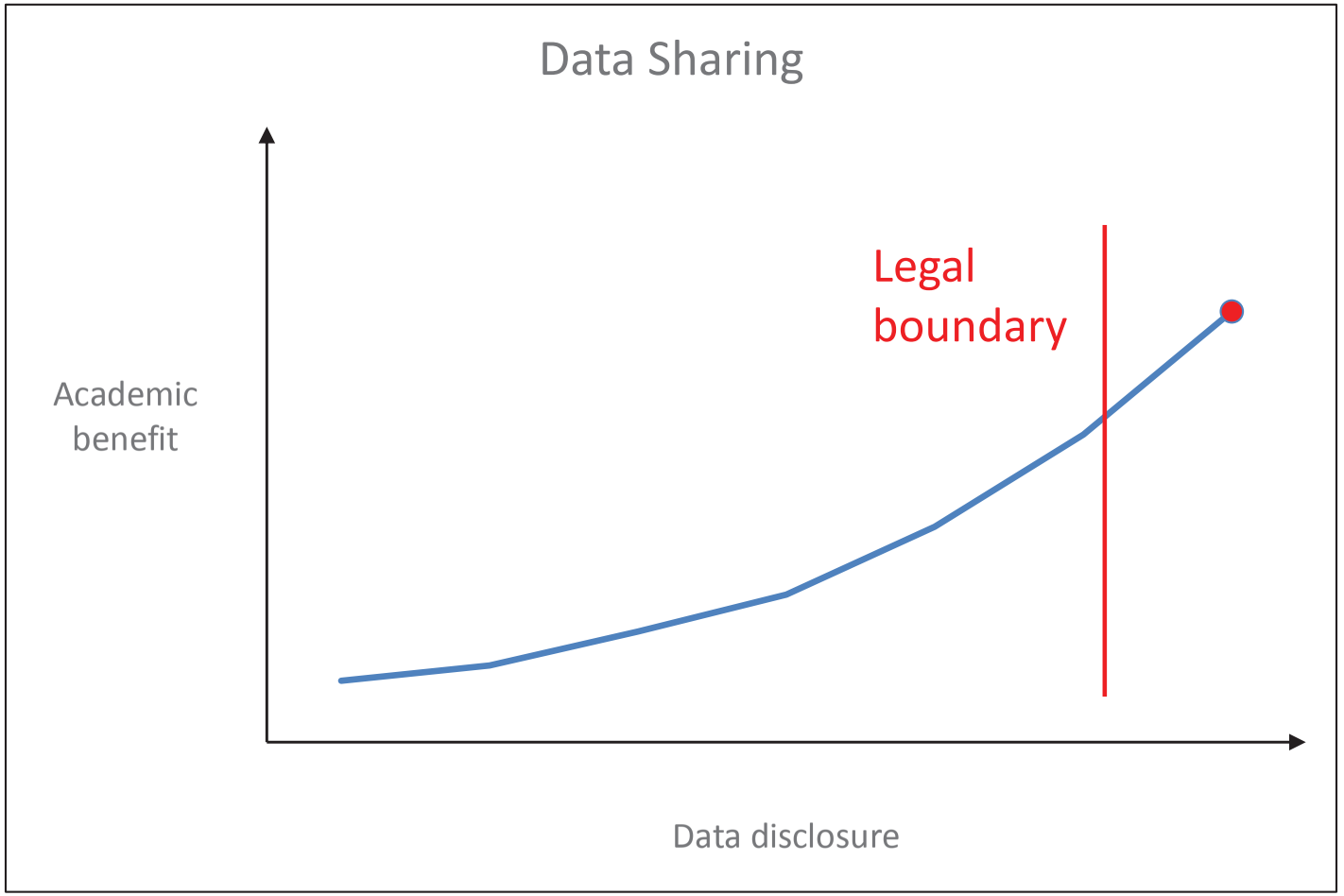
Share information and ideas on ways to apply confidential rules/arrangements in a manner that would allow sharing of more granular data

Eurostat Peer review report on the compliance with the Code of Practice and the coordination role of the National Statistical Institute in Turkey

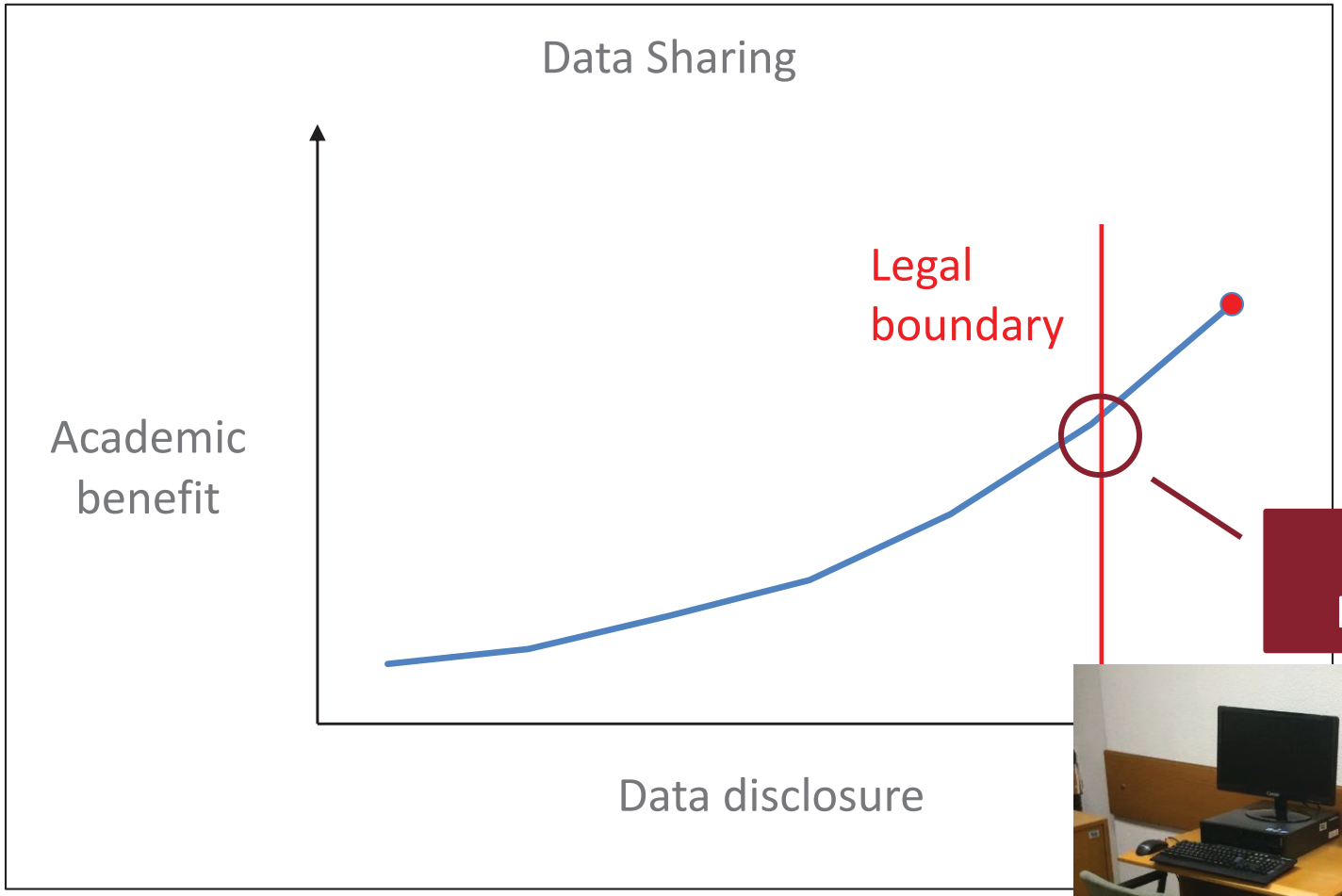
Recommendation 22:

TurkStat should introduce remote access facilities for researchers, who are permitted to use its anonymised microdata for research purposes (European Statistics Code of Practice, indicator 15.4)

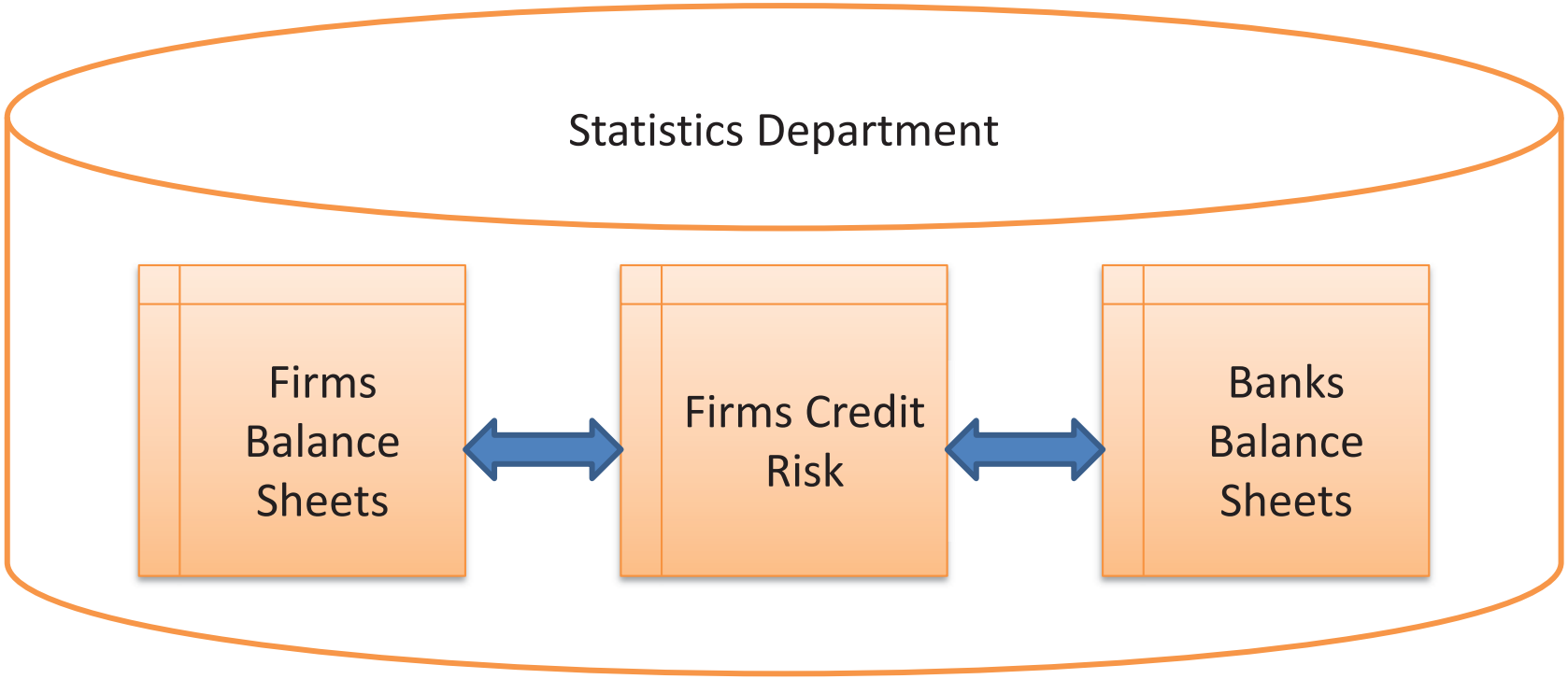
Data Sharing



Data Sharing



CBRT Practice



63 M X 7
16 M X 7

339 M X 17

15 K X 56
15 K X 51

CBRT Practice

	VHESNO	FUNVAN	ILKODU	FADRES1	POSTKOD	TELNO1	TSNO	ANASEK1	ANASEK2	BYIL	BHESNO	BTUTAR
1			10	ORGANİZE SAN. BÖL.	10100	2811400	37707	C	2370	2014	1	10410285
2			16	HASANPAŞA KÖYÜ	.	7234546	37708	C	1624	2014	1	24267606
3			10	BALIKESİR ASFALTI	10200	7338347	37709	C	1061	2014	1	43935636
4			34	OSMAN BEY MAH.	34360	3216884	37710	G	4641	2014	1	11683054
5			27	KÜSGET MAH. NO:1	27000	2413335	37713	G	4631	2014	1	11989910
6			77	RÜSTEMPAŞA MAH	77100	8113315	37717	C	1411	2014	1	20370926
7			6	İSTİKLAL MAHALLESİ	68015	2362525	37718	A	0142	2014	1	9999310
8			6	CEYHUNATIF KANSU CD	6520	4737380	37719	C	2020	2014	1	20451805
9			34	SÜMER MAHALLESİ	34025	5476331	44720	C	1511	2014	1	39825047
10			27	BAŞPINAR 2 OSB VALİ	27007	3374111	44724	C	1393	2014	1	66167895
11			34	TEKNO PLAZA ŞEHİT ŞA	34710	5776300	44725	G	4663	2014	1	29639201
12			34	KOCASİNAN MH.	34180	6390818	44728	G	4634	2014	1	22671439
13			34	SANCAKTEPE EMEK MAH	34785	6356622	47706	C	2932	2014	1	22937943
14			27	AYDINLAR MAH 03040	.	2413534	47707	C	2363	2014	1	5631060
15			34	YENİKÖY MAH KÖYBAŞI	34464	2993000	47708	G	4676	2014	1	25371425
16			36	İNÖNÜ MH	36000	4134123	47710	G	4532	2014	1	19740534
17			42	FEVZİÇAKMAK MA	42250	3453725	47711	G	4639	2014	1	21569118
18			46	KARACASU KARAZİYARET	46100	2363800	47712	C	1320	2014	1	1520795
19			34	İ.O.S.B. İSTEKS	34303	4862004	47713	G	4752	2014	1	18093833
20			0	YAKUPLU MAH.	34524	4227600	47714	C	1812	2014	1	16265369
21			33	YENİ MAH.GMK B	33180	3598600	47715	I	5510	2014	1	5441220
22			6	İVEDİK CD. NO:123/C	6100	3947464	47721	G	4642	2014	1	29170886
23			6	İSTANBUL YOLU	.	2261210	47722	G	4730	2014	1	115974108
24			38	MELİKGAZİ MH.	38050	2214040	47723	F	4120	2014	1	13356292
25			7	MOLLA YUSUF MH.	7000	3465533	47725	F	4120	2014	1	5382386
26			41	SANAYİ MAH. D130	41000	3355118	47726	G	4778	2014	1	11727015
27			41	İSTASYON MAHALLESİ	41000	3734466	47727	C	2512	2014	1	7854867
28			0	5 NİSAN MH	21100	2244042	47728	G	4632	2014	1	41732403
29			31	ATATÜRK MAH İNÖNÜ	31800	2855036	47729	G	4631	2014	1	21174986
30			38	OSB 12. CADDE	38040	3204581	47730	G	4615	2014	1	14234536
31			34	MESİHPAŞA MAH	34000	5169021	47731	I	5510	2014	1	19881
32			58	YENİDOĞAN MH	58580	6548290	47732	G	4639	2014	1	3392538
33			34	İKİTELLİ OSB MAH.	34490	4860505	47733	C	1413	2014	1	141845294
34			34	MAHMUTBEY MAH	34000	4469292	47734	C	2012	2014	1	14152724
35			35	CUMHURİYET MH	.	4238262	47735	F	4120	2014	1	2786106
36			0	SERÇEÖNÜ MAH AHMET	38110	2221919	47737	F	4120	2014	1	16686157
37			34	TOPÇULAR KIŞLA CAD	.	6132330	47739	G	4672	2014	1	12165582
38			16	TAVŞANLI MAH.	16700	6761219	47741	C	2932	2014	1	20833428
39			6	KENTKOOP.MAH1868SK.	.	2573333	47742	S	9609	2014	1	661328

Firms Balance Sheets

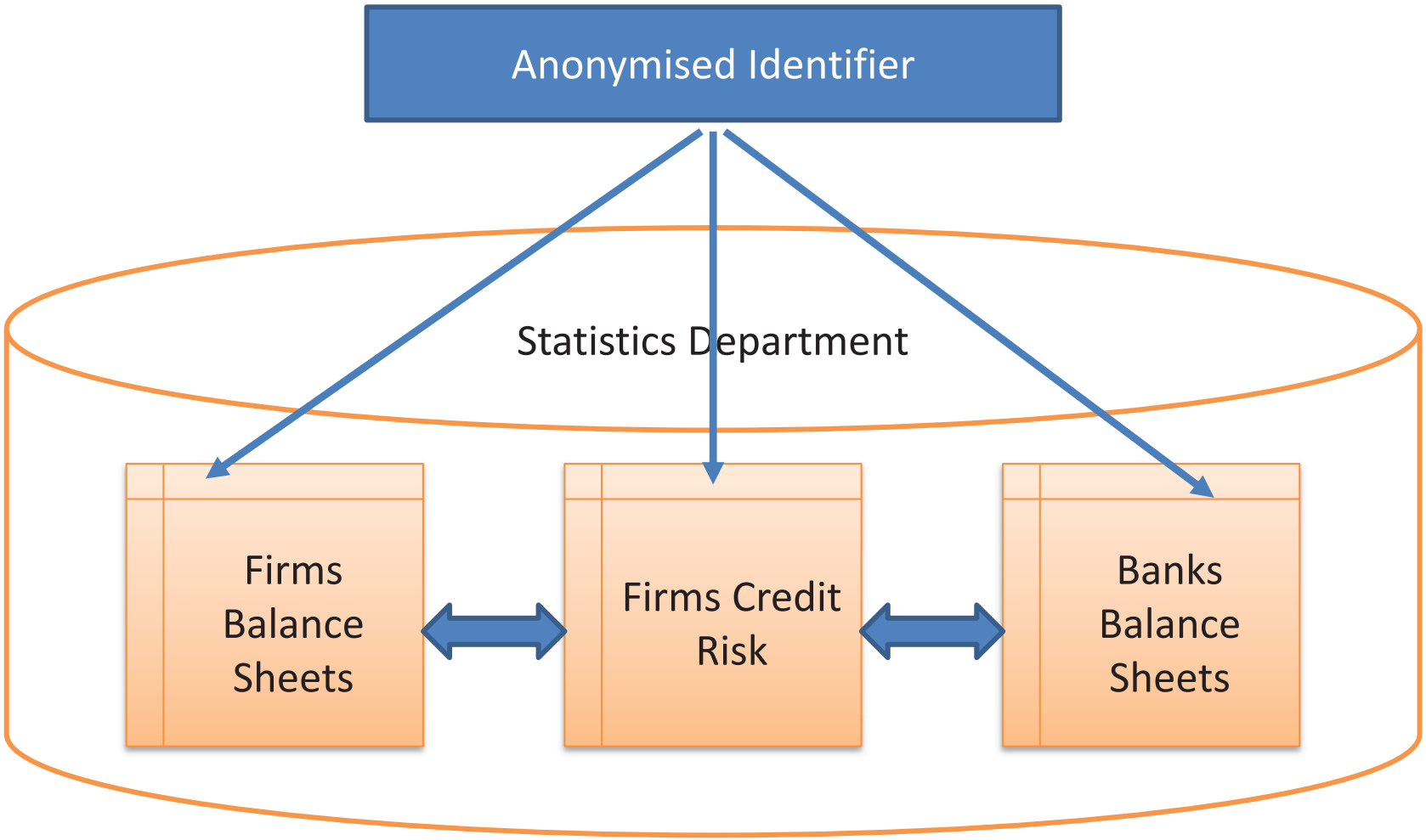
CBRT Practice

Anonymised Identifier

Firms Balance Sheets

TSNO	ANASEK1	ANASEK2	BYIL	BHESNO	BTUTAR
37707	C	2370	2014	1	10410285
37708	C	1624	2014	1	24267606
37709	C	1061	2014	1	43935636
37710	G	4641	2014	1	11683054
37713	G	4631	2014	1	11989910
37717	C	1411	2014	1	20370926
37718	A	0142	2014	1	9999310
37719	C	2020	2014	1	20451805
44720	C	1511	2014	1	39825047
44724	C	1393	2014	1	66167895
44725	G	4663	2014	1	29639201
44728	G	4634	2014	1	22671439
47706	C	2932	2014	1	22937943
47707	C	2363	2014	1	5631060
47708	G	4676	2014	1	25371425
47710	G	4532	2014	1	19740534
47711	G	4639	2014	1	21569118
47712	C	1320	2014	1	1520795
47713	G	4752	2014	1	18093833
47714	C	1812	2014	1	16265369
47715	I	5510	2014	1	5441220
47721	G	4642	2014	1	29170886
47722	G	4730	2014	1	115974108
47723	F	4120	2014	1	13356292
47725	F	4120	2014	1	5382386
47726	G	4778	2014	1	11727015
47727	C	2512	2014	1	7854867
47728	G	4632	2014	1	41732403
47729	G	4631	2014	1	21174986
47730	G	4615	2014	1	14234536
47731	I	5510	2014	1	19881
47732	G	4639	2014	1	3392538
47733	C	1413	2014	1	141845294
47734	C	2012	2014	1	14152724
47735	F	4120	2014	1	2786106
47737	F	4120	2014	1	16686157
47739	G	4672	2014	1	12165582
47741	C	2932	2014	1	20833428
47742	S	9609	2014	1	661328

CBRT Practice



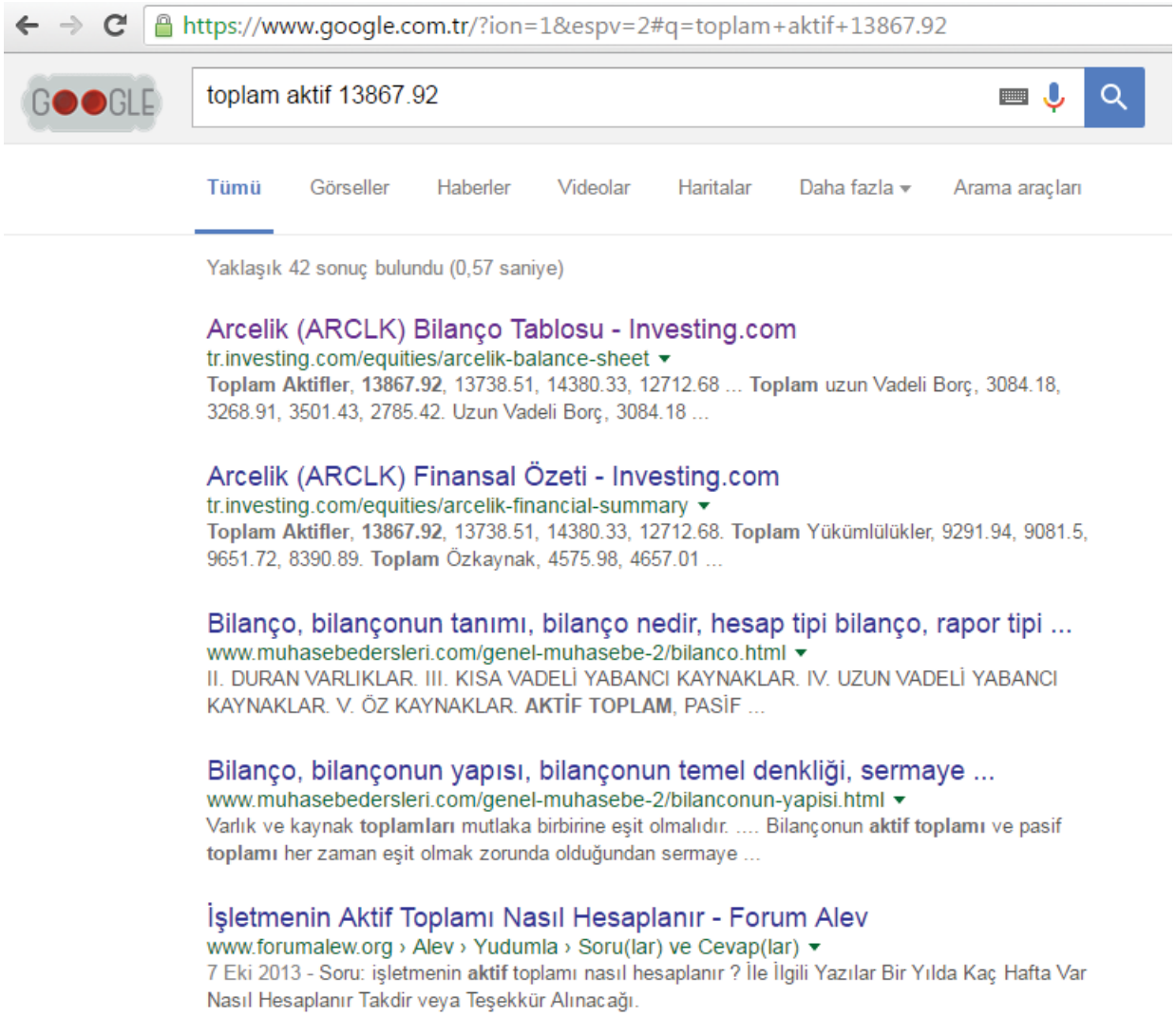
CBRT Practice

Anonymised Identifier

Firms Balance Sheets

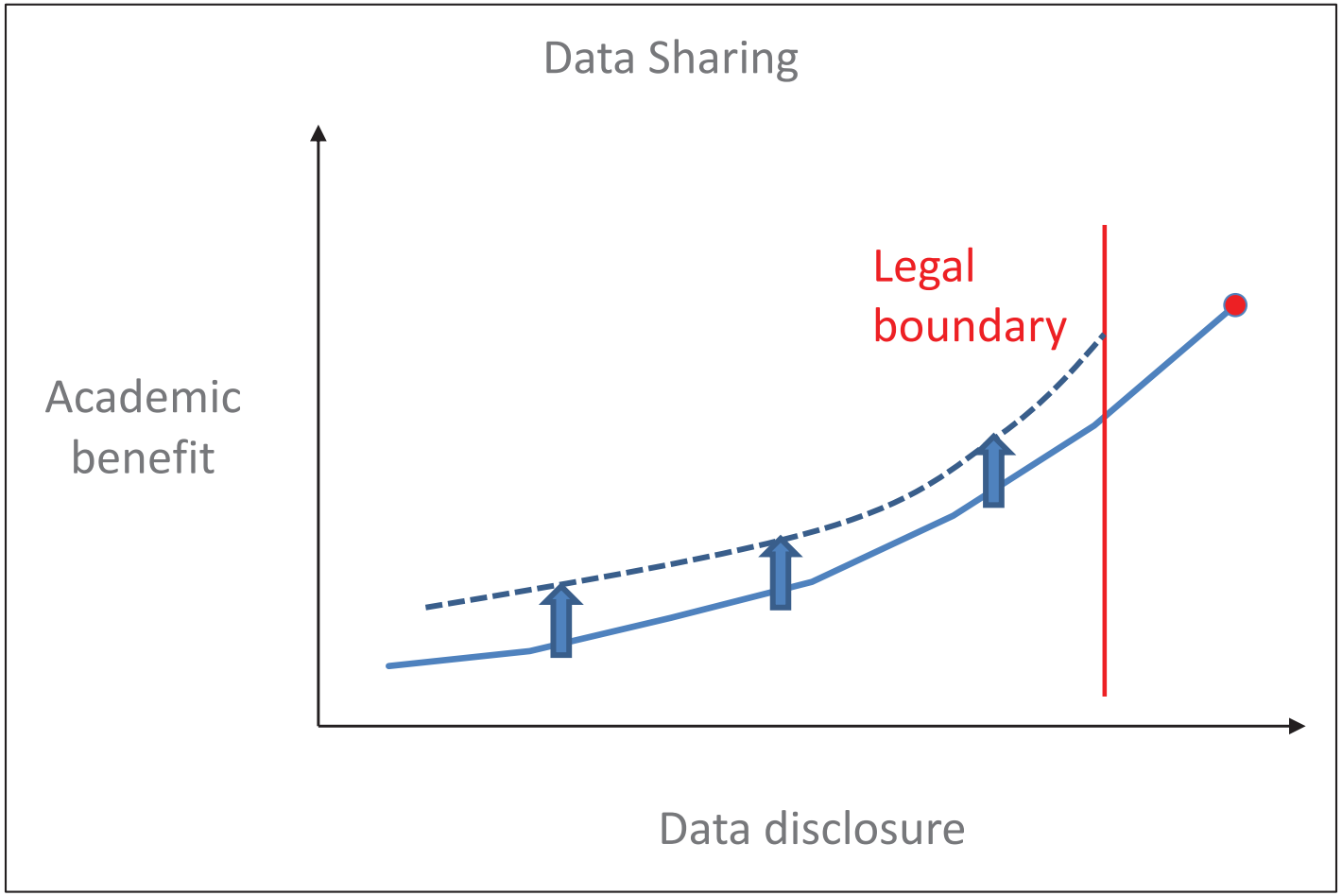
TSNO	ANASEK1	ANASEK2	BYIL	BHESNO	BTUTAR
37707	C	2370	2014	1	10470885
37708	C	1624	2014	1	24267605
37709	C	1061	2014	1	43935636
37710	G	4641	2014	1	11683054
37713	G	4631	2014	1	11989910
37717	C	1411	2014	1	20370926
37718	A	0142	2014	1	9999310
37719	C	2020	2014	1	20451805
44720	C	1511	2014	1	39825047
44724	C	1393	2014	1	66167895
44725	G	4663	2014	1	29639201
44728	G	4634	2014	1	22671439
47706	C	2932	2014	1	22937943
47707	C	2363	2014	1	5631060
47708	G	4676	2014	1	25371425
47710	G	4532	2014	1	19740534
47711	G	4639	2014	1	21569118
47712	C	1320	2014	1	1520795
47713	G	4752	2014	1	18093833
47714	C	1812	2014	1	16265369
47715	I	5510	2014	1	5441220
47721	G	4642	2014	1	29170886
47722	G	4730	2014	1	115974108
47723	F	4120	2014	1	13356292
47725	F	4120	2014	1	5382386
47726	G	4778	2014	1	11727015
47727	C	2512	2014	1	7854867
47728	G	4632	2014	1	41732403
47729	G	4631	2014	1	21174986
47730	G	4615	2014	1	14234536
47731	I	5510	2014	1	19881
47732	G	4639	2014	1	3392538
47733	C	1413	2014	1	141845294
47734	C	2012	2014	1	14152724
47735	F	4120	2014	1	2786106
47737	F	4120	2014	1	16686157
47739	G	4672	2014	1	12165582
47741	C	2932	2014	1	20833428
47742	S	9609	2014	1	561328

CBRT Practice



TSNO	ANASEK1	ANASEK2	BYIL	BHESNO	BTUTAR
37707	C	2370	2014	1	10470885
37708	C	1624	2014	1	24267605
37709	C	1061	2014	1	43935636
37710	G	4641	2014	1	11683054
37713	G	4631	2014	1	11989910
37717	C	1411	2014	1	20370926
37718	A	0142	2014	1	9999310
37719	C	2020	2014	1	20451805
44720	C	1511	2014	1	39825047
44724	C	1393	2014	1	66167895
44725	G	4663	2014	1	29639201
44728	G	4634	2014	1	22671439
47706	C	2932	2014	1	22937943
47707	C	2363	2014	1	5631060
47708	G	4676	2014	1	25371425
47710	G	4532	2014	1	19740534
47711	G	4639	2014	1	21569118
47712	C	1320	2014	1	1520795
47713	G	4752	2014	1	18093833
47714	C	1812	2014	1	16265369
47715	I	5510	2014	1	5441220
47721	G	4642	2014	1	29170886
47722	G	4730	2014	1	115974108
47723	F	4120	2014	1	13356292
47725	F	4120	2014	1	5382386
47726	G	4778	2014	1	11727015
47727	C	2512	2014	1	7854867
47728	G	4632	2014	1	41732403
47729	G	4631	2014	1	21174986
47730	G	4615	2014	1	14234536
47731	I	5510	2014	1	19881
47732	G	4639	2014	1	3392538
47733	C	1413	2014	1	141845294
47734	C	2012	2014	1	14152724
47735	F	4120	2014	1	2786106
47737	F	4120	2014	1	16686157
47739	G	4672	2014	1	12165582
47741	C	2932	2014	1	20833428
47742	S	9609	2014	1	661328

Data Security



1. Differential privacy
2. Homomorphic encryption

Data Security

1. Differential privacy: Y=f(X)=X*Rand1+Rand2

DP Parameters			X					Y			
Rand1	0.72593	Id	Net Sales	Net Profit	Correlation	Id		NS2	NP2	Correlation
Rand2	1498.969	1	24,899	4,629	0.81	1		19,574	4,859	0.81
		2	46,765	8,902	Intercept	2		35,447	7,961	Intercept
		3	26,896	3,723	18,615	3		21,024	4,202	10,080
		4	103,227	12,758	Slope	4		76,435	10,760	Slope
		5	48,206	12,486	3.29	5		36,493	10,563	3.29
		6	48,716	7,852		6		36,863	7,199	
		7	11,101	1,760		7		9,558	2,777	
		8	107,145	10,975		8		79,279	9,466	
		9	34,887	7,317		9		26,824	6,811	
		10	39,791	7,482		10		30,384	6,930	
		11	98,958	26,113		11		73,336	20,455	
		12	90,297	23,828		12		67,048	18,796	
		13	61,305	7,610		13		46,002	7,023	
		14	42,445	7,757		14		32,311	7,130	
		15	41,898	10,716		15		31,914	9,278	
		16	100,374	28,508		16		74,363	22,194	
		17	27,810	5,467		17		21,687	5,468	
		18	14,355	2,069		18		11,920	3,001	
		19	53,057	12,331		19		40,015	10,450	
		20	54,466	11,756		20		41,037	10,033	

Data Security

2. Homomorphic encryption:

DP Parameters			X					Y			
Rand1	0.72593	Id	Net Sales	Net Profit	Correlation	Id		NS2	NP2	Correlation	
Rand2	1498.969	1	24,899	4,629	0.81	1		19,574	4,859	0.81	
		2	46,765	8,902	Intercept	2		35,447	7,961	Intercept	
		3	26,896	3,723	18,615	3		21,024	4,202	10,080	
		4	103,227	12,758	Slope	4		76,435	10,760	Slope	
		5	48,206	12,486	3.29	5		36,493	10,563	3.29	
		6	48,716	7,852		6		36,863	7,199		
		7	11,101	1,760		7		9,558	2,777		
		8	107,145	10,975		8		79,279	9,466		
		9	34,887	7,317		9		26,824	6,811		
		10	39,791	7,482		10		30,384	6,930		
		11	98,958	26,113		11		73,336	20,455		
		12	90,297	23,828		12		67,048	18,796		
		13	61,305	7,610		13		46,002	7,023		
		14	42,445	7,757		14		32,311	7,130		
		15	41,898	10,716		15		31,914	9,278		
		16	100,374	28,508		16		74,363	22,194		
		17	27,810	5,467		17		21,687	5,468		
		18	14,355	2,069		18		11,920	3,001		
		19	53,057	12,331		19		40,015	10,450		
		20	54,466	11,756		20		41,037	10,033		

Data Security

Methods / Analysis	Simple DP	Improved DP	Improved DP + HE
Simple regression	✔	✔	✔
Multiple regression, Logistic regression, Log-data	✘	✔	✔
Security	**Low**	**Middle**	**High**

OBJECTIVE

Data Security

Thank you…

BANQUE DE FRANCE
EUROSYSTÈME

Directorate General Statistics

Data sharing and data management: the Banque de France experience

Renaud Lacroix
Director, Statistical and IT engineering division

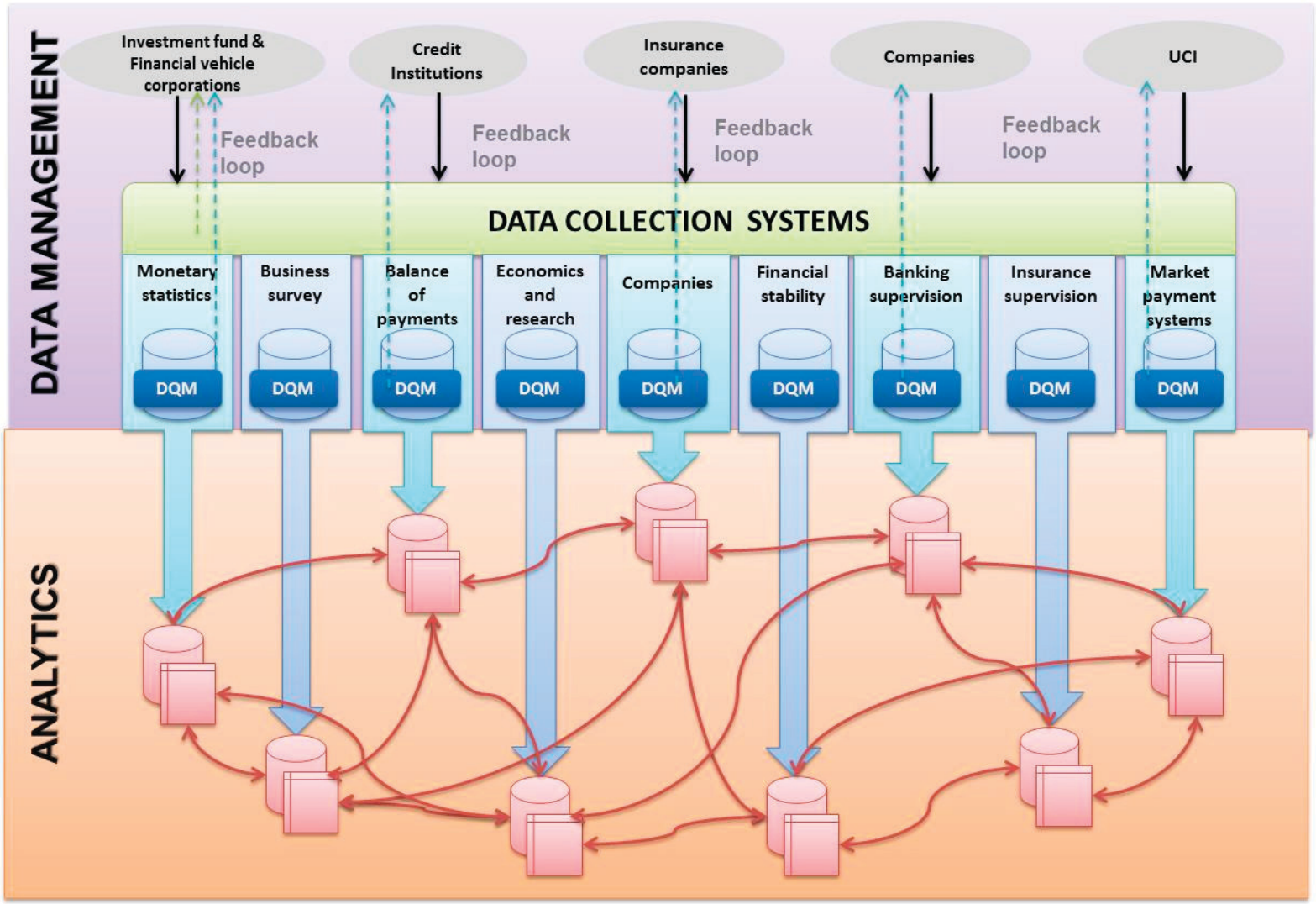
G20 workshop on Data Sharing
Frankfurt, 31 January 2017

OUTLINE

1. Data management at the Banque de France
2. An internal data sharing platform: the Pooling and Sharing Statistical Series project
3. Sharing granular data with external users and the academics: the Open Data Room
4. Data exchanges with the National Statistical Institute

DATA MANAGEMENT AT THE BANQUE DE FRANCE

Where we come from : the Information System in 2009



DATA MANAGEMENT AT THE BANQUE DE FRANCE

A considerable growth of request for data

- **All national stakeholders, the media, the general public as well as researchers ask for more and more data, in particular granular data**
  - Our own staff and researchers have also growing requests
- **The multiplication of statistical requests in the future is not hypothetical but certain**
- **Manage, manipulate and leverage on terabytes of data: paradigm shift : we cannot work as before, otherwise we will be snowed under with data**
  - DG Statistics is where to answer the challenge
  - **Innovation** is necessary to develop cost-saving systems
  - **The integration and mutualisation of data management tools** provides a pragmatic approach able to meet short-term challenges

DATA MANAGEMENT AT THE BANQUE DE FRANCE

First step : mutualisation of data collection tools

ONEGATE

- **Since 2010, a single point of entry (portal One Gate) for all data collected by the Banque de France and NSA (national supervisory authority) from financial institutions, non financial corporates, households, insurance companies,…**

- **A dedicated shared platform for data collection :**
  - Various formats accepted (XML, XBRL, CSV)
  - Management of high volume (size limit > 2 Gb per file)
  - Ability to manage up to 20 000 users and 1500 files / day
  - **200 000 files received and processed / year**

- **DGS chairs the Onegate Governance Committee**
  - The committee comprises representatives from all business lines
  - Addresses both technical issues (maintenance portfolio) and strategic orientations

DATA MANAGEMENT AT THE BANQUE DE FRANCE

Second step : Pooling and Sharing Statistical Series (P3S)

Goals of the set-up :

- **Pooling data ….**
  - To gather data on financial institutions, non-financial corporations and households
  - Collected by the Banque de France and the Autorité de contrôle prudentiel et de résolution
  - While respecting confidentiality rules
- **…to allow enhanced analysis for all involved departments including the supervisory authority**
  - Offering access to internal users on a 'need to know' basis
  - Fostering synergies and economies of scale
- **Classification grid for business opportunities**

Data collection **Access to big collections**	Data analysis **A supercomputer for everyone**	Giving back the results **Data immersion**
• Breaking silos • Add more internal information • Open to external information	• Intensive statistical computing • Real time computing • Non structured data analysis	• Search • Data Discovery • Data Visualization

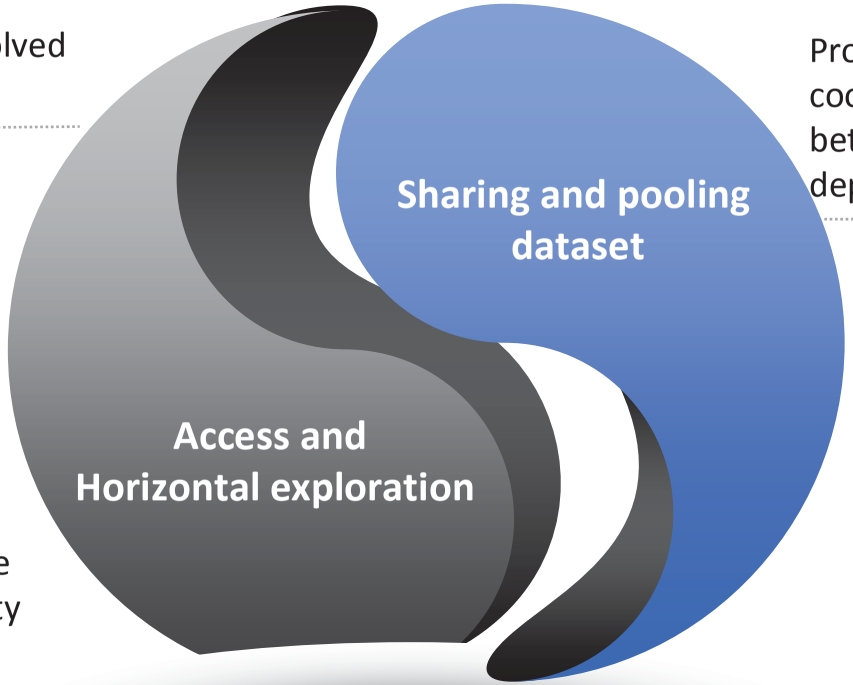
Pooling and Sharing Statistical Series (P3S)

Goals of the set-up

Gather data on financial institutions and non-financial corporations in **a common repository**

Improve analysis from each involved department

Promote **transversality** and cooperation by **sharing dataset** between Banque de France departments



Enhance the influence of Banque de France thanks to higher quality of studies

Add a new d... ...formation System by sharing information beyond organisational silos

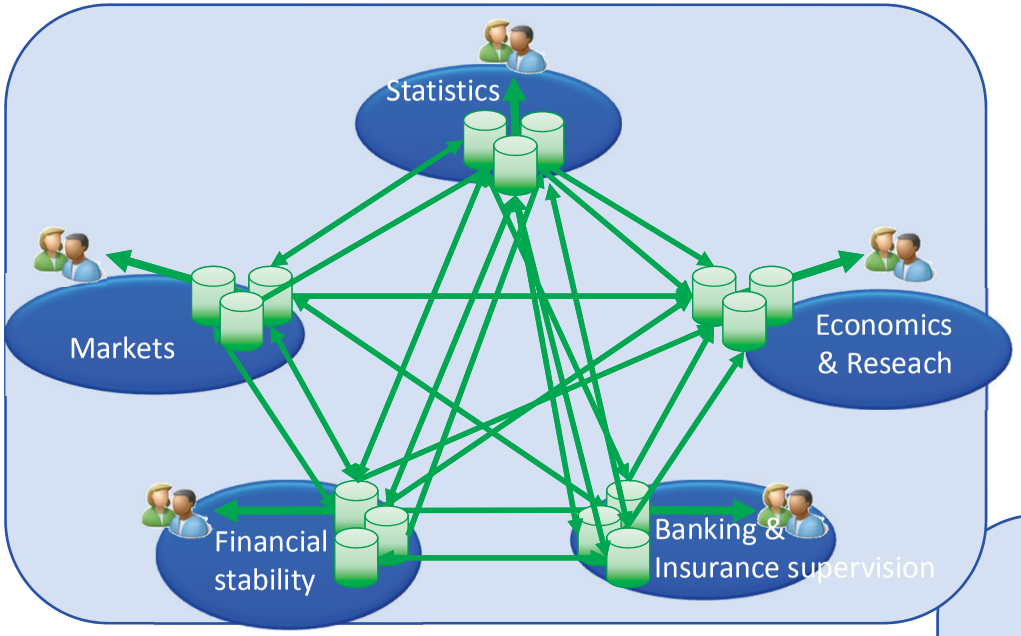
Pooling and Sharing Statistical Series (P3S)

Collaborative work

- A collaborative work involving 5 DGs
- Work-streams on confidentiality issues with representatives of all stakeholders and of the Legal department
- Data supply and demand expressed by each DG
- P3S data typology compliant with legal constraints

Pooling and Sharing Statistical Series (P3S)

A new approach

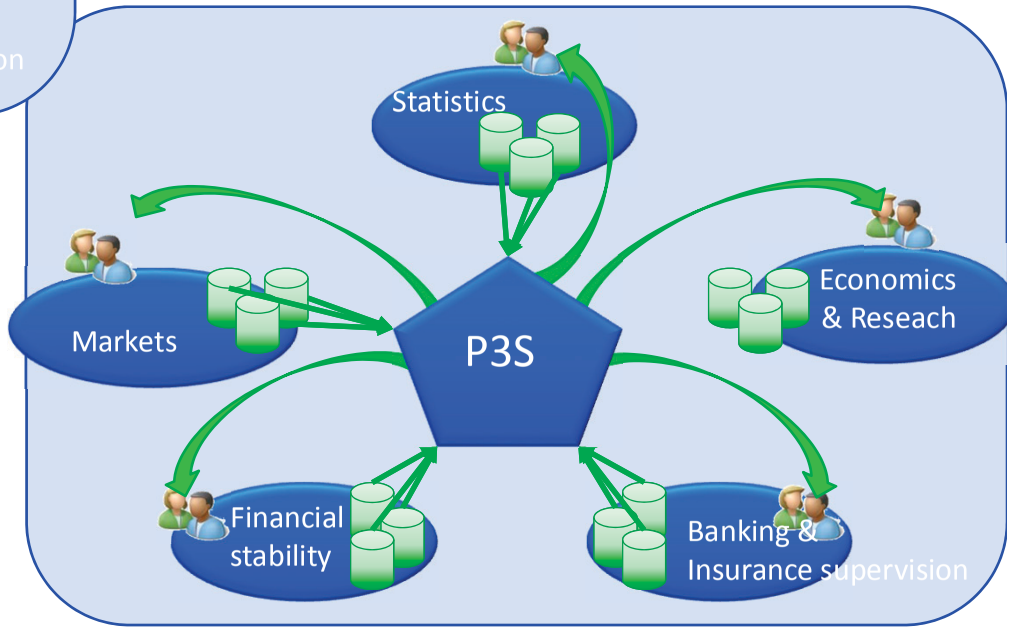


Main objective :

- Foster synergies between directorates through a better access and a common production of statistical economic assets for Banque de France taken as a whole

Expectations :

- Ability to manage huge quantities of data
- Capability to aggregate heterogeneous data flows
- Performant search tools
- Strong ability to evolve



Pooling and Sharing Statistical Series (P3S)

Governance

- **DGS has the budget proposal and implementation responsibility for P3S**
- **DGS carries out in operational terms the application and evolution process, in coordination with IT and Legal departments (DGS chairs the steering committee)**
- **P3S Validation and Monitoring Committee (PVMC) :**
  - **Committee at DG level, co-chaired by DG-S and BDF CIO, including all stakeholders and the Legal department**
  - **Validating the lists of accredited agents:**
    - **Most (individual) data can be shared** by personal accreditations updated every 6 months (167 subsets of generally shareable data), decided according to the governance scheme, with veto power of Legal Director and escalation process to the Governors
    - For a subset of more sensitive data, access will be precisely limited and granted on a need-to-know basis subsets of non generally shareable data). Decisions are made on a case-by-case basis under the same procedure
    - Access rights are defined through an "Accreditation Matrix" which crosses sub-families and users
  - **Monitoring P3S functioning**
  - **Addressing unforeseen issues**
  - **Seeking for consensus**

PVMC

Pooling and Sharing Statistical Series : Technological background

Data storage

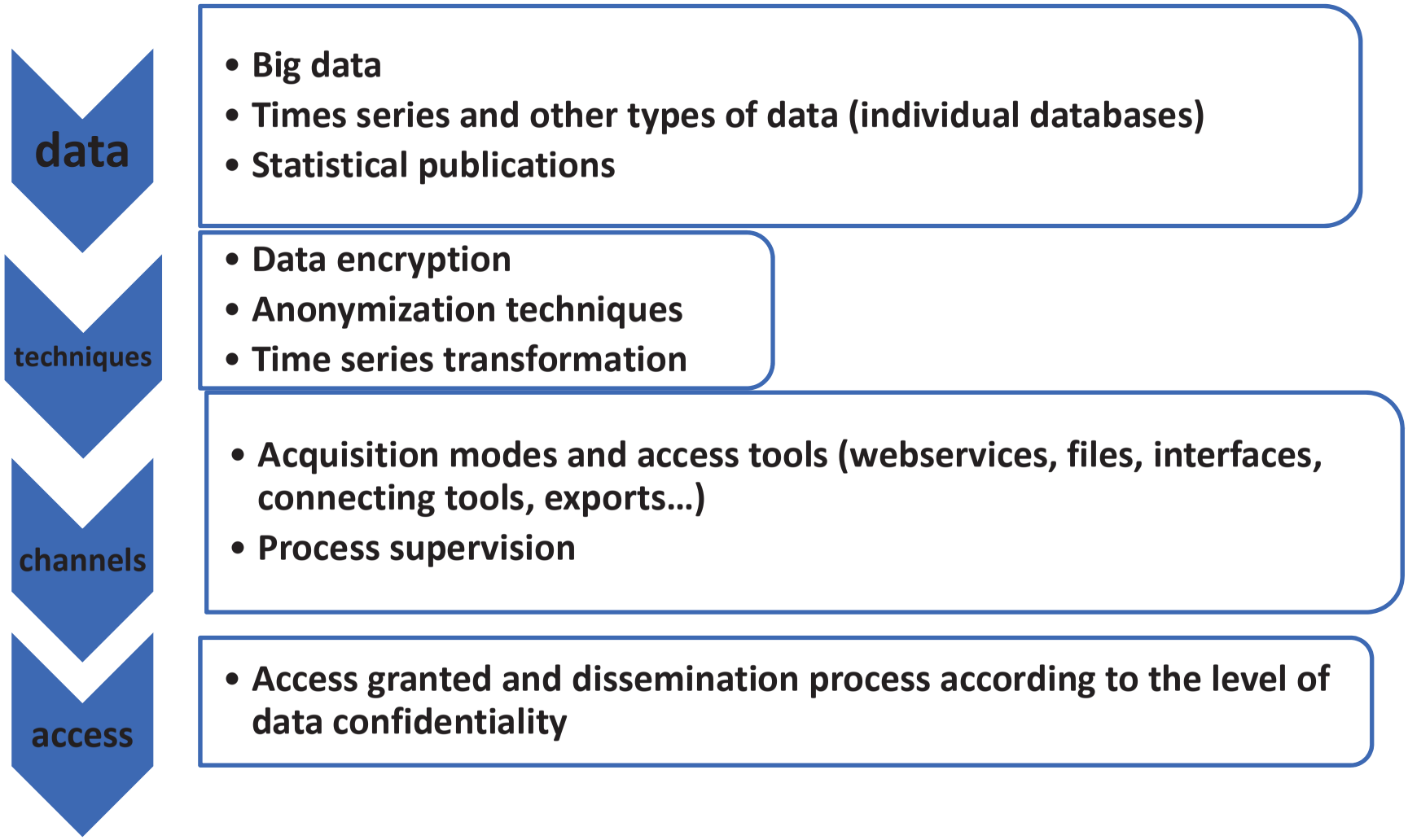
- **2000 GB of data stored into P3S,** generating a storage need of 10000 GB after having taken into account the auxiliary volumes required for the organization, search and data identification
- **530 million series**
- **15 main datasets**

Data from credit institutions
Data from securitization bodies and investment firms
Insurance data
Consolidated prudential data
UCITS
Household over-indebtedness
International banking data
Money and interbank market
Data from payment institutions and electronic money issuers
International activities of firms
Business survey
Securities holding and issue
Data from corporates
National data in TARGET 2
Means of payment

- **The BDF macro-economic database (TSB) stores 7,8 million series - i.e. 28 GB of data ; these series will be made available in P3S**

Pooling and Sharing Statistical Series : Technological background

Management of a wide diversity of data

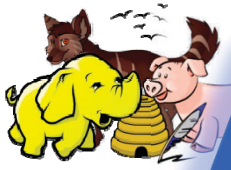


Pooling and Sharing Statistical Series : Technological background

Technology overview

Open Source Communities

Neo4j



elasticsearch.

python

mongoDB

hadoop

HIVE

Marketed Open Source

•P3S Project

Hortonworks

talend

Dataiku

Application vendors Offers

sas Analytic Server

TERADATA

NETEZZA

SAP HANA

ORACLE Big Data Appliance

Greenplum

ORACLE EXALYTICS IN-MEMORY MACHINE

Microsoft SQL Server 2008 R2 Parallel Data Warehouse

Pooling and Sharing Statistical Series : Technological background

Business capabilities

- **Ability to manage large volumes of data and metadata**
- **Secured access rights**
- **Flexibility to integrate and handle heterogeneous data (NoSQL technology):**
  - **Codification** of series is specific to each dataset
  - **Technical formats** : XBRL, SDMX, SAS,…
- **Powerful research engine (ElasticSearch)**
- **Scalability and interoperability (integration of new data sets, connection to additional analysis tools)**

Pooling and Sharing Statistical Series : project plan

- **A two-year project**

- **The platform is open since July 2015 :**
  - Integration of an highly significant group of data covering all business areas
  - Access granted to a first group of 60 users from all business areas (including Supervisory Authority)

- **New features in 2016 :**
  - Extension up to 300 users
  - Integration of additional datasets in P3S (300 subsets)
  - **Access granted to external users at DGS premises** (anonymised data for researchers): **the Banque de France Open Data Room was inaugurated by the Governor in November 2016**

Pooling and Sharing Statistical Series : first lessons

- **Do not forget the human factor !**

- **Need to break down cultural barriers :**
  - Not so easy for users to manipulate data produced in other departments : closer relationships between producers and users must be developped
  - New data call for new ideas : innovation is the key driver
  - Strong involvement of the management at all levels is required

- **Need to invest more in statistical training**
  - New skills are required to be able to handle large amounts of data available : data science, computer science
  - The Directorate General Statistics has launched training cycles intended for its staff:
    - *Data analyst (basic statistical techniques, elementary econometrics)*
    - *Data scientist (advanced econometrics, machine learning techniques)*

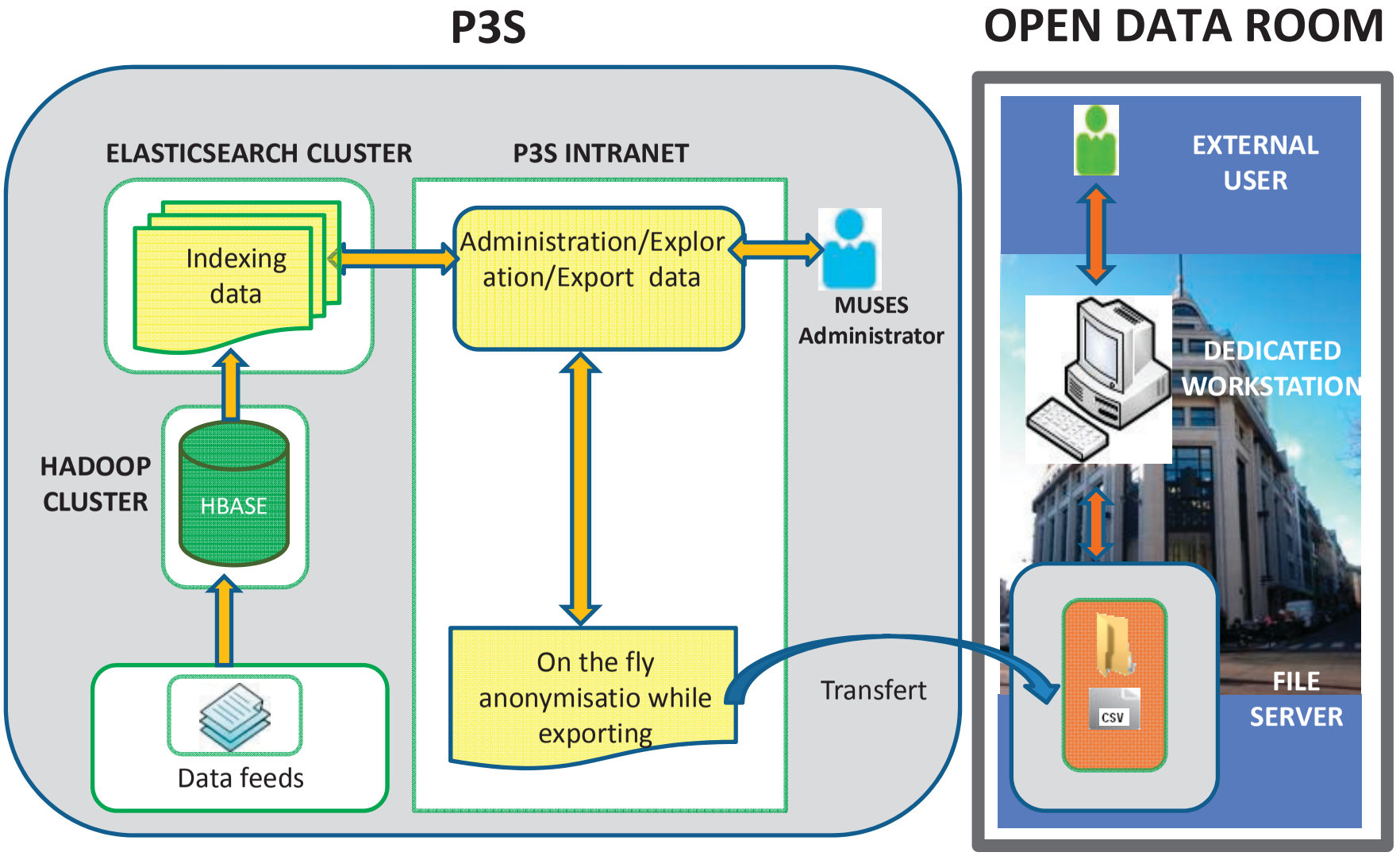
Open Data Room

The procedure

- The applicant(s) fills in a detailed application form describing the research project and the team organisation
- A confidentiality agreement is signed by each member of the research team
- **Applications are collectively reviewed by a decision body ('Committee for the instruction of data requests ') chaired by DGS**
- The Committee is composed of representatives from all business areas of the Banque de France, the legal department + two external academics
- The Committee decides on the approval of the request based on the legal framework under which the data have been collected
  - **Strict adherence to the European regulation Ref. 2533/98 for data collected according to an ECB regulation**

The Open Data Room

The technical set-up



The Open Data Room

A broad set of tools

- Securised access to enhanced workstations in Banque de France, Directorate general of Statistics
- Statistical software available

- Possibility to include external datasets brought by the researcher
- Possibility to get in contact with data producers
- Methodological and IT support provided by a dedicated team in DGS

The Open Data Room

Planned extensions

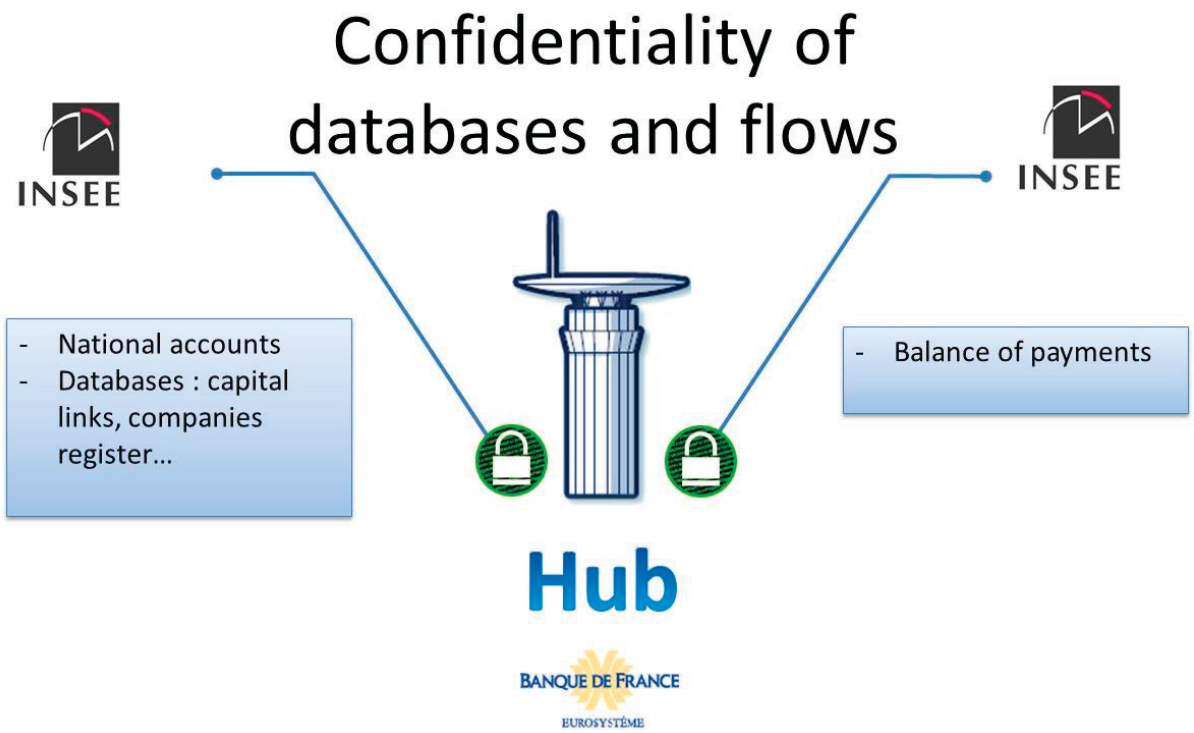
- **A remote access planned for 2018 : project « ODR 2.0 »**
  - Access through secured extranet
  - Special attention is given to information security
  - Main target : the worldwide research community

- **The solution will pave the way to data sharing with other statistical bodies, when necessary and legally possible**

- **Possibility to host in the ODR confidential data of the French national statistical institute is currently being explored**
  - Objective : offer researchers access to databases produced by both institutions, anonymised in the same way
  - Exchange of anonymisation keys prior to the anonymisation phase ?
  - Secured channel for the transmission of data between the ODR and the INSEE Research data center ?

Data exchange with the National Statistical Institute (Insee)

A flexible tool for day-to-day business



- Management of embargoes
- Management of any kind of files (CSV, owner format, XML…)
- SDMX standard

5

Thank you for your attention

renaud.lacroix@banque-france.fr

Linking different data sets and the role of common identifiers

Matteo Piazza, FSB Secretariat

G20 Thematic workshop on data sharing

Frankfurt, 31 January - 1 February 2017

FSB FINANCIAL STABILITY BOARD

Overview

- **A global view is growingly required to interpret many financial and economic developments**. Also national/regional authorities may sometimes need to access global data to conduct their own monitoring activities.

- **Obtaining this comprehensive view is a challenging objective, also due to obstacles to data sharing, but may strongly benefit from internationally agreed common identifiers** (e.g. LEI, UTI, UPI). They may serve several other purposes, for both industry and regulators, in addition to statistical production.

- **Significant progress has been made on common identifiers but there are still challenges ahead**. Jurisdictions participating to the DGI-2 could investigate possibilities to foster the use of common identifiers.

- "Access rights to the global datasets obtained by linking national sources, which would include new information, may need to be decided, as well as the form of anonymization". **How data can be shared for building and using global datasets is an important issue that we should address**.

Data sharing as a multi-faceted concept

- Data sharing is not limited to the dissemination of some statistical outputs to other authorities or to the general public but it is also part of a process to compile more comprehensive and accurate statistics by linking different datasets.

- This dimension of data sharing has clearly gained importance following the financial crisis. The evidence of increasingly global financial transmission mechanisms and strong feedbacks between the financial sector and the real economy indicated that a more complete and correct view of financial and economic developments may need progress along two dimensions:

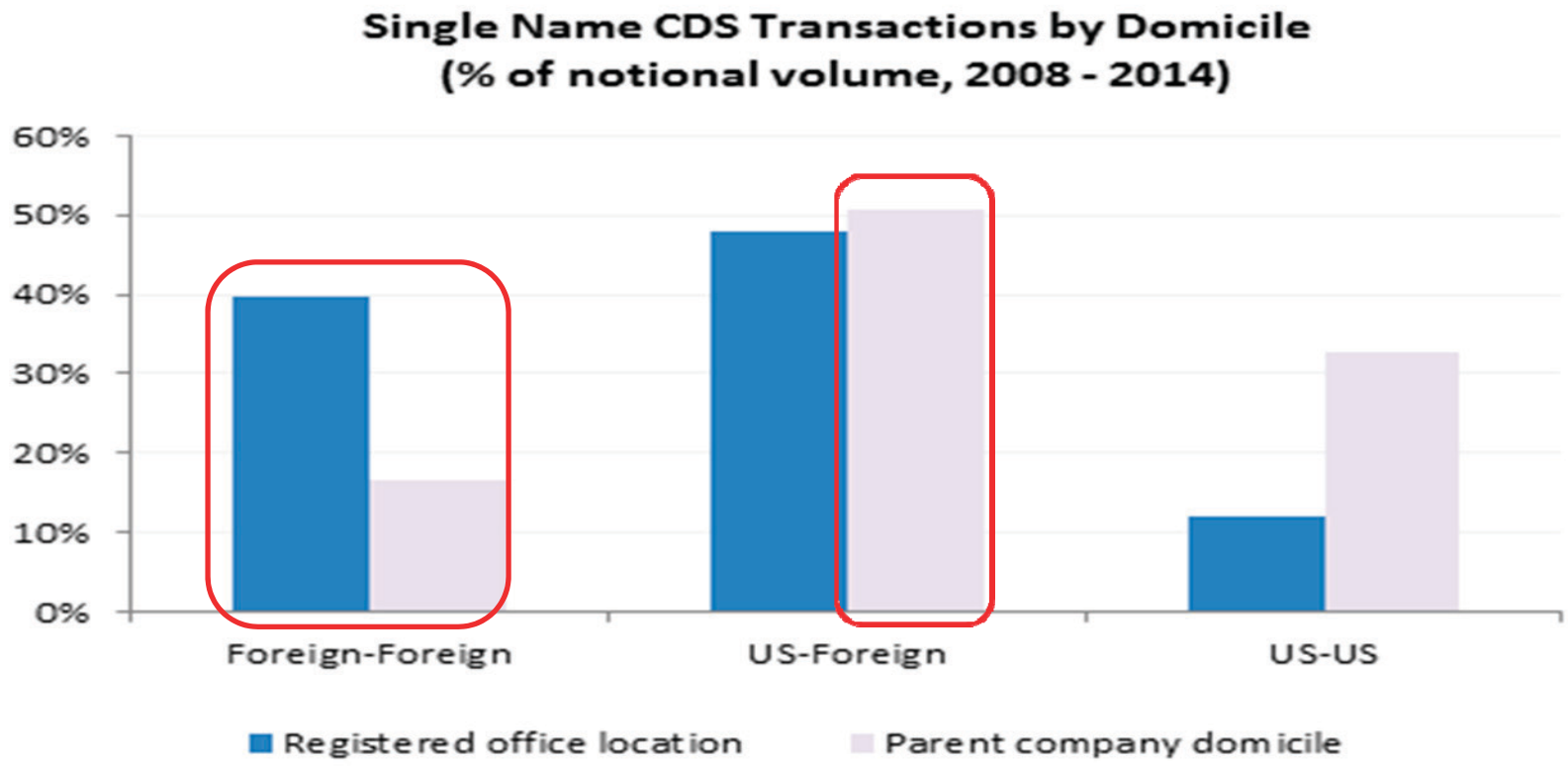
  – data helping to bridge the divide between **micro and macro analysis** (FSB-IMF, Data Gaps Initiative Progress Report 2015; Tissot, 2016);

  – data allowing a **global** view where needed (“Purely domestic efforts risk resulting in inconsistent data sets”; Borio, 2013).

FSB FINANCIAL STABILITY BOARD

A global view

- Aggregation of the data reported across Trade Repositories (TRs) will help authorities in obtaining a comprehensive view of the OTC derivatives market. In global markets this may be necessary even from a national authority's standpoint: the chart below shows the fraction of notional volume in North-American corporate single-name CDS between differently domiciled accounts:



Source: SEC

Linking of datasets and common identifiers

- Granular (enough) data are an essential ingredient for progress along the two dimensions above but a proper aggregation of these data is key to effectively deliver all the expected benefits and may require the availability of common identifiers.
- The wide recognition of the benefits brought by common global identifiers has favored remarkable progress over the past few years. Some of these initiatives are now part of the recommendations of the G20 DGI-2.
- Standardising the identifiers of counterparties, transactions, products and reporting requirements would produce benefits, for both the industry and policymakers, that extend well beyond statistical production.

Current work on common identifiers

- Participant/ party Identifier
  - Legal Entity Identifier (LEI)
    - Global LEI Foundation (GLEIF) – www.gleif.org;
    - LEI Regulatory Oversight Committee - www.leiroc.org
- Transaction Identifier
  - Unique Transaction Identifier (UTI)
    - CPMI-IOSCO Harmonisation group http://www.bis.org/cpmi/publ/d131.htm
- Product Identifier
  - Unique Product Identifier (UPI)
    - CPMI-IOSCO Harmonisation group http://www.bis.org/cpmi/publ/d151.htm

FSB FINANCIAL STABILITY BOARD

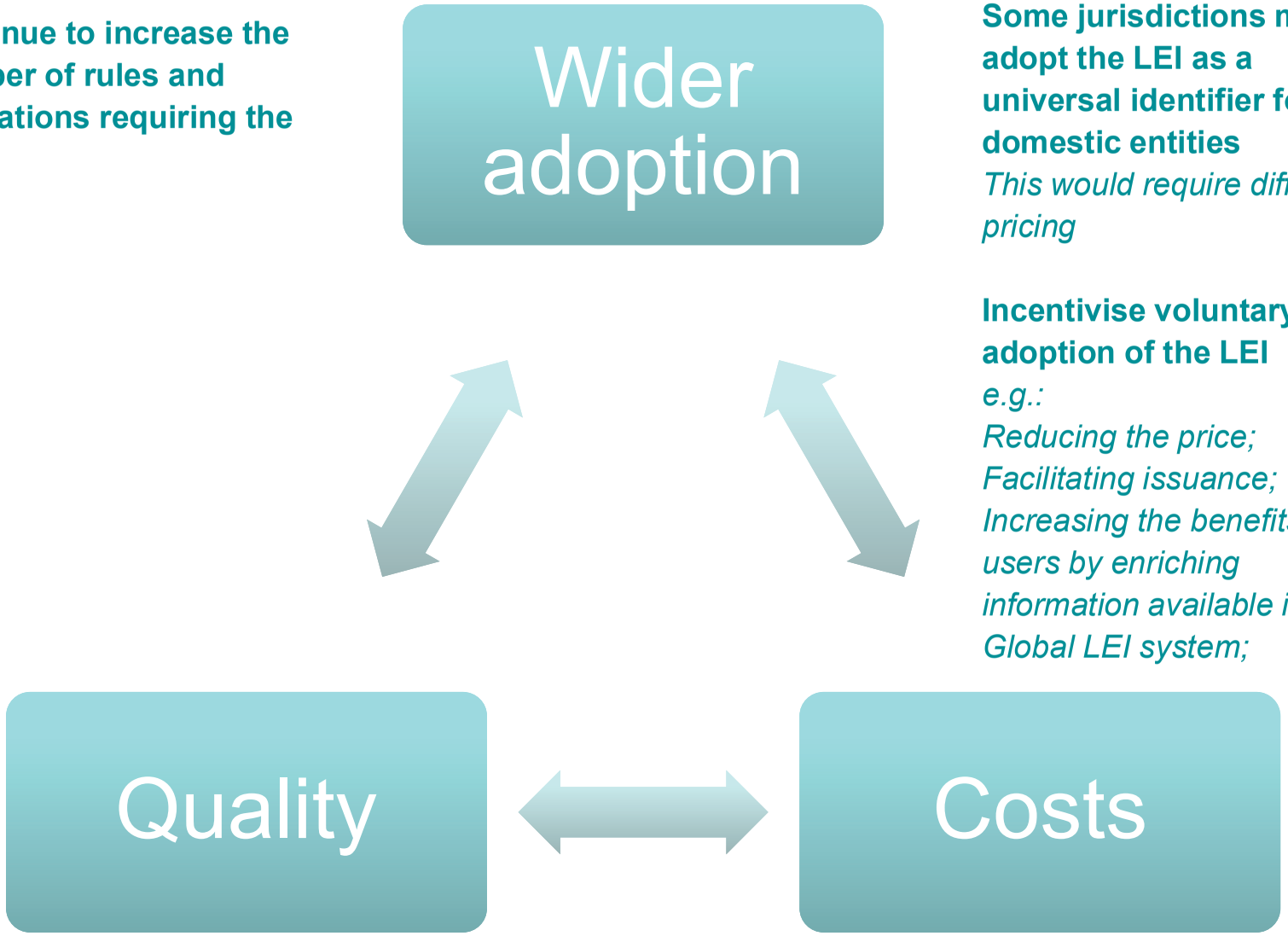
Recent progress on the LEI system

Increasing benefits for users:

- Level 2 data: it is expected that all LEI issuers will have developed by early May 2017 the capacity to record **relationships with direct and ultimate parents**. Mandatory collection will start at that date and the GLEIF expects that parent information for the entire LEI data pool might be available early in 2018. It is a challenging data collection but it serves many purposes: e.g. in the EU, the collection of parent data for commodity derivative reporting is expected to start in early 2018; in the DGI-2 framework, identification of foreign subsidiaries for cross-border exposures and MNEs data may be helped by this enrichment (rec#14);
- data on **international branches**. Implementation is also expected to start in early 2017;
- Expansions considered on individuals licensed or authorised by a financial regulator, on corporate actions, on funds' relationships;
- ongoing work by GLEIF and others to develop free mapping of LEI with other identifiers of entities (e.g.: BIC) or financial instruments (ISIN), embed LEI in XBRL Taxonomy,…

Promoting the LEI adoption

Continue to increase the number of rules and regulations requiring the LEI



Some jurisdictions may adopt the LEI as a universal identifier for their domestic entities
This would require different pricing

Incentivise voluntary adoption of the LEI
e.g.:
Reducing the price;
Facilitating issuance;
Increasing the benefits for users by enriching information available in the Global LEI system;

UPI, UTI and other data elements

- A common product identifier would allow the identification of pockets of risk on specific products. A common transaction identifier facilitates: (i) avoiding double counting if transactions are reported by different parties; (ii) linking transactions when a life cycle event occurs and different events are reported to different TRs; and, (iii) linking an original bilateral transaction to the resulting cleared transactions.
- Rec #6 of the DGI-2: *The CPMI—IOSCO Harmonization Group to define technical guidance on uniform transaction and product identifiers (UTI and UPI)) as well as on other data elements (ODE) to be reported. The FSB to implement the governance of the UTI and UPI classifications and codes."*
- For both UPI and UTI consultative reports on harmonisation have been already issued. Final guidance on both identifiers is expected in the course of 2017. The final guidance on all three batches of ODE is expected by end-2017.
- FSB's Group on UTI / UPI Governance (GUUG) was established in 2016 with a mandate to make recommendations to the FSB on the governance of each of these harmonised identifiers. It is expected to consult on governance of these identifiers in 2017.

FSB FINANCIAL STABILITY BOARD

Data sharing

- A key expectation for the G20 DGI-2 is that it will deliver data fit for policy use, with a global focus on financial stability. Rec. #20 asks "the IAG and G-20 economies to promote and encourage the exchange of data and metadata among and within G-20 economies [..]. G-20 economies are also encouraged to increase the sharing and accessibility of granular data, if needed by **revisiting existing confidentiality constraints.**"
- Revisiting confidentiality constraints is indeed part of what jurisdictions have been asked to do with respect to reporting to different TRs of OTC data. The November 2015 FSB peer review of OTC derivative trade reporting has identified a number of legal barriers in FSB member jurisdictions to reporting to TRs and impediments to authorities' access to TR-held data.
- Following that report, jurisdictions have committed, by June 2018 at the latest, to:
  - remove barriers to full reporting of trade information (including counterparty information) to trade repositories (domestic or foreign);
  - have a legal framework in place to permit non-primary domestic and foreign authorities' access to data in accordance with their mandates.

FSB FINANCIAL STABILITY BOARD

Approaches to data sharing

- "*Access rights to the global datasets obtained by linking national sources, which would include new information, may need to be decided, as well as the form of anonymization*".

- Different level of access according to the users may be a dimension along which data-sharing solutions can be differentiated. Global policymakers may look for a comprehensive, granular but still aggregate picture of financial stability developments rather than for individual entities data.

- There may also be cases where micro-data (e.g. transaction-level data) may be a step in constructing meaningful and comprehensive (granular) aggregates, i.e. aggregates with a fair amount of detailed information but non presenting, apart from occasional instances, confidentiality issues.

FSB FINANCIAL STABILITY BOARD

Approaches to data sharing /2

- An issue is then if data that are not by construction related to individual entities may be dealt with in a different way than data that are related by construction to individual entities, e.g.:
  - Are the reasonable means for identifications the same?
  - Do they need to be shared in the same way (community of policy users vs. one global aggregator and/or TTPs) and does this have implication for the access to these data (e.g. with reference to the concepts of trust and maturity used in the OECD paper for international collaboration on micro-data access)?
  - Are they market sensitive in the same way?
- Approaches limiting the sharing of confidential data to a "global aggregator" as a step in the production of global aggregates can be explored. These approaches may alleviate confidentiality issues while still allowing data sharing of granular ("*less aggregated*") data across jurisdictions.

FSB FINANCIAL STABILITY BOARD

Going forward

- In the recommendations for the G20 it could be flagged the **importance of promoting the use of common identifiers as a key step, along with improvements in data sharing, for delivering (global) data fit for policy use.** Economies can be encouraged to investigate possibilities to foster the adoption and use of common identifiers. Countries could also consider including and using the LEI in their own data administrative and statistical databases so to maximize the benefits of existing LEIs and increase awareness on such benefits.
- As far as possible solutions on data sharing are concerned, we need to recognize that we are moving along a continuum with highly aggregated data at one end and individual, non-anonymised data at the other end of the spectrum. This would allow **to consider and propose a broader spectrum of possible approaches to data-sharing and could make easier revisiting confidentiality constraints**.
- As recalled in the DGI-2 2016 Annual Progress Report, data sharing is a cross-cutting issue. We could plan a **stock-taking of the data-sharing issues that need to be addressed for each DGI-2 recommendation, if any, and the proposed plans to do so**. This could be one of the deliverables for recommendation #20 going forward.