



Munich Personal RePEc Archive

National Data Centre and Financial Statistics Office: A Conceptual Design for Public Data Management

Cakir, Murat

Central Bank of the Republic of Turkey

20 January 2014

Online at <https://mpra.ub.uni-muenchen.de/53869/>

MPRA Paper No. 53869, posted 25 Feb 2014 12:52 UTC

National Data Centre and Financial Statistics Office: A Conceptual Design for Public Data Management

Murat Cakir

“National Data Centre and Financial Statistics Office: A conceptual Design for Public Data Management”¹

Murat Cakir²

Abstract

Data processes run by states, governments and the like have been a great deal and as old as the modern human history. Data had always been important. Tons were collected and siloed, but never in the past had its importance been felt as much as it had been when the last crisis broke out in 2008. Because these tons of data either, as some were redundant and occupying large spaces with huge storage costs, were not useful given the processing power and due to outdated mind-sets, or were not even the tiniest portion of the data necessary to do analysis³, the experts realised. With the advances in the digital world dealing with data has become easier. Combined with the urgent needs and demands from the bottom up and top down there now is more enlightened and educated perception of data and whatever its extensions⁴ are, and its / their potential use, though a little bit late. In the late 90s, however, things were not as computerised and Data^{eXve} was not as Big as it is today, and manual operations dominated the automated ones. There were definitely inefficiencies in Data^{eXve}. Still, even then, there were efforts to improve these processes. This work focuses on one of those early efforts.

¹ Presented at the “Data Management Strategies and Technologies Workshop” in New York, 04 February 2014.

² Murat Cakir is a Specialist in the Central Bank of the Republic of Turkey. The original ideas and views expressed herein belong to the author only and do not represent those of the Central Bank of the Republic of Turkey or its staff.

³ Backcasting, forecasting, nowcasting whatever routine you might see fit in there.

⁴ Term was made up by the author, to abbreviate *data exhaustive*; which is all there is about or relating to data, i.e. processes, management, governance etc.

An Early Attempt!

Almost 2 decades ago, in the late 90s, while computerisation was not as spread as it is today, the data needs of government agencies were mostly catered by using paper based forms, as were the reports being prepared mostly manually rather than using office suites, though due to high costs only those agencies were the ones to afford these suites. Therefore, all sorts of processes based on this data feeding format had to be performed almost manually; data collection, data entry, and report production were all alike.

Obviously, these processes were *not user friendly*, data collection and entry were *painful*, and data were more *error prone*. Therefore the data collection and processing in the public sector by the public authorities were *inefficient* and *ineffective*, because of

1. *Redundancies* stemming from the data providers' obligation to report to multiple users, for example banks, tax authority, Central Bank etc., which was in turn an overburden to the data providers This had resulted in multiple and dissimilar reports of the same financials and therefore there were inconsistencies and lower integrity in the same and/or similar datasets.
2. Data collection and IT *costs* that were *higher* than they could be for the whole system, as data of the same category were being collected and stored in more than one places (collection, storage, processing and miscellaneous costs).
3. *Lower speed* of data collection process. Data provision and reporting to authorities and policymakers therefore were slow, resulting in lagged policy reactions.

This inefficiency could be eliminated provided that the *overburden* over the

1. *Data providers*, was lessened by *unifying the reporting* through a unified reporting package on the basis of "*one report to one collector*" principle (later found costlier hence replaced with "*one report to multiple collector*" principle). This package was supposed to be interconnected with the accounting and enterprise information systems if the user chose to use it so, and
2. *Data collectors*, was lessened by data sharing, which was confronted with secrecy and confidentiality reasons!

These reasons, though logical given the legislative framework, could only be overcome and a more and more efficient and effective system could be established, just by *centralising* rather than *sharing* the data. Centralisation of different institutional databases would have been more efficient and effective, as

1. While the lowest cost of sharing would be equal to the cost of data storage in *at least two places*, in centralisation the data had to be stored *only in two places*, in data feeders' database and in the data centre. Theoretically, the highest potential cost for the system would be equal to the storage in two databases,

2. In sharing due to potential coordination pitfalls (e.g. decision changes by the original data providers whether to share or not) there *would be operational risk*, whereas in centralisation as individual data feeders would be responsible by legislation for their datasets, this risk *could not exist*,
3. Individual institutional data providers (data feeders hereinafter) would be held responsible for one or a couple of *unique datasets* or *partial datasets*, *data collection would be faster*, and
4. *Unique instances of each record* for different datasets would be practically possible, with lower total storage costs and more correct data with less processing costs.

Note that the data centre must not have been directly reported to, as data responsibility principle for individual feeders would be breached and expertise of each feeder would be bypassed resulting in less qualified data to be stored in the centre. In order to guarantee the *data consistency and integrity* dynamic cross checking at random times on the Basic Identity Information Datasets (BIIDS) should have been carried out. BIIDS, have the same particular design in all individual data feeder's database where the unique identifiers can be obtained by combining more than one fields, like tax number, street address, name of owner(s) of the entity, etc.

The original design of the data centre was made as early as 2000. Although more crude than this, without some fine tuning and data management terminology added recently, the conceptual basics were almost exactly the same. Unfortunately, it was so early then that, ostensibly, the system and its incumbents were not as ready for such a *radical and aggressively confrontational design* (!). Pronounced in several close surroundings, seen as *utopia* by many, *faced with reactions*, and finally found its place in the *inactive work portfolio*...

After the crisis set in and set out in 2008, a couple of decades old concerns of data and statistics professionals had been finally understood and been remade recognised, again thanks to a small number of people, whether the data issues had to be resolved no matter what it takes and whatever it costs, as otherwise, neither supporting policy decisions of the authorities nor informing the public would be possible whatsoever. Ideas concerning the consolidation, integration and sharing have been and are being pronounced more openly in data and statistics gatherings and more excitedly debated

on in recent years⁵. The good news is politicians had become finally aware of the need, at a faster pace, with steps further than the authorities, researchers and data people⁶.

With the recent urging demands from the politicians and policymakers, this conceptual project design had become again a part of the agenda, and was *resuscitated*; this time combined with the state of the art data and IT theoretical approaches.

⁵ Lately, nearly all data and statistics gatherings had involved data issues like governance, big data, integration, sharing and the like. Biannual “Data Management Strategies and Technologies Workshop” New York 2014, “Integrated Management of Micro Databases” Porto 2013 are the past events where we had such heated discussions (positively), and ESS Big Data Event 31 March-1 April 2014 in Rome and “Workshop on Using Big Data for Forecasting and Statistics” 7- 8 April 2014 in Frankfurt am Main are the coming ones where Big Data shall be a central subject.

⁶ a. Dodd–Frank Wall Street Reform and Consumer Protection Act, effective as of July 21, 2010 is an important motivator for the US authorities to focus on the data issues.

b. President Obama’s National Big Data Research and Development Initiative (Big Data Initiative for short) was unveiled in 2012, involving 6 Federal departments and agencies, together, promise to greatly improve the tools and techniques needed to access, organize, and glean discoveries from huge volumes of digital data (^[1] Kalil, 2012, ^[2] Executive Office of the President, 2012).

These two illustrative developments are clear reflections of US Government’s approach to and intent with data issues in the aftermath of the 2008 financial crisis.

G20 Data Gaps Initiative (DGI) ^[5] that relates to overcome problems in all sorts of data processes and looking for a cure is another huge initiative by international organisations. This alone involves huge responsibilities by individual governments, for them to close these data gaps.

Data Management Concepts and Definitions: Design Terminology

There is a relationship mapping in the current design, which is depicted in Table 1. There are some overlapping definitions, that is some or all of the individual entities (rows in the table) can be covered by more than one definition. It all depends on how one looks at the table.

Row-wise definitions are:

- *Data stakeholders* are the feeders and users.
- *Data custodian* is the one that stores and keeps the data secure, (here the National Data Centre)
- *Data owners* are responsible for *producing, providing, collecting* and *feeding*, from the data provider to data collector. National Data Centre is *not the owner* of any raw data except for those it produces for its internal use.
- *Data processors collect, feed, and keep the data secure.*

Column-wise definitions are:

- *Data feeders* collect the data it is responsible for and feeds it to the data centre.
- *Data users* use the *raw and processed data and reports and queries.*

General principles and definitions are:

- *Data responsibility* is that the collectors / feeders have to *collect data, make sure these data are correct and transfer the data to the centre.*
- *Reporting burden*, by the help of unified reporting tool / package, is the *responsibility* of the data provider(s) to *report to multiple collectors at once.* Multiple reporting to multiple users can be eliminated by this reporting tool.
- *Data and information provision* is the *providing the users with raw and processed data and reports* stored at the centre at pre-defined or dynamically changed levels of authority.
- *Security and Authority:* Security of the data at all levels and at each phase has the highest priority for the centre. The use of data is only *possible* according to the *level of authority* the user has vis-à-vis the centre.

Table.1 Data Relationships

DATA	FEEDER	CENTRE	USER
STAKEHOLDER	YES		YES
CUSTODIAN		YES	
OWNER	YES		YES
PROCESSOR	YES	YES	YES

Proposed Architecture: Design Properties of the Data Centre

In its simplest form the proposed architecture has been designed on three main sub structures; namely, data providers (feeders), data users and the data centre (Figure 1). In this structure,

1. *Data Feeders* are the principal data providers to the centre which are bound to *collect, control, and transfer* the datasets that are in their *expertise* area and under their *data responsibility* from the primary reporters (Firms, banks, etc.).
2. *Data Centre*, regarding and considering the *interrelatedness* or *independence* of the datasets provided by the feeders carries out *internal integrity* and *correctness* tests, and after confirming them, *stores* the raw and processed datasets into its database(s), updates and revisions of which are always possible. With the controls of the cross-checking sets, depending on the frequencies, *processes* the raw datasets and *prepare reports automatically*. This automatic and high speed and batch processing capacity of high volumes of data, is one of the primary and essential features of the data centre, most stakeholders don't own nor could they afford to have. Data and information provision to the data users is subject to authority levels.
3. *Data Users* can utilise raw and processed data and reports prepared by the centre as long as they comply with the *authorisation rules* and according to their predefined / predetermined *levels of authority*. Queries made and ad-hoc reports created by the users can be stored by the centre in a *query* and *reports repository*, and can be used by other users if authorised by the original user and the data centre, as long as the *secrecy* and *confidentiality rules are not breached*, and these reports are considered to be publicised *without causing information asymmetry* nor any *strategic risk* for the economy management; otherwise restrictions apply.

A predefined structure for the whole database isn't appropriate as it would put limits on the *creative possibilities* and *potential use* of the queries and reports with regard to raw data. However, a *data warehouse* and an *operational data store* system where reside *pre-run queries* or *pre-established relations* and user run queries that can be revealed for use shall definitely exist. This is necessary for *quick* and *timely* reporting purposes for medium and longer term analyses; still, not sufficient to explore all the possible relationships. All the data usage shall be on the grounds of *pre-granted* and/or *dynamic* and *stratified authority* levels.

Common characteristics of the data that will be collected, stored, processed, and reported in the data centre are that they are *high frequency* and *very detailed* (granular). Besides, due to these high frequencies, the data inflow will be very fast and high volume (real time), and analysis and reporting, depending on the type of the data series, should be almost real time. Therefore, *software* and *hardware* configurations of the

centre should be as *up-to-date* as possible. Here *no specific IT architecture* can be pointed out to, as it all depends on the database, analysis and reporting architectures.

Structurally, data centre has two main departments, one of which was once not separately classified in the original design. They are Country Statistics Office (CSO), responsible for dealing with non-financial datasets and Financial Statistics Office (FSO), responsible for dealing with financial data.

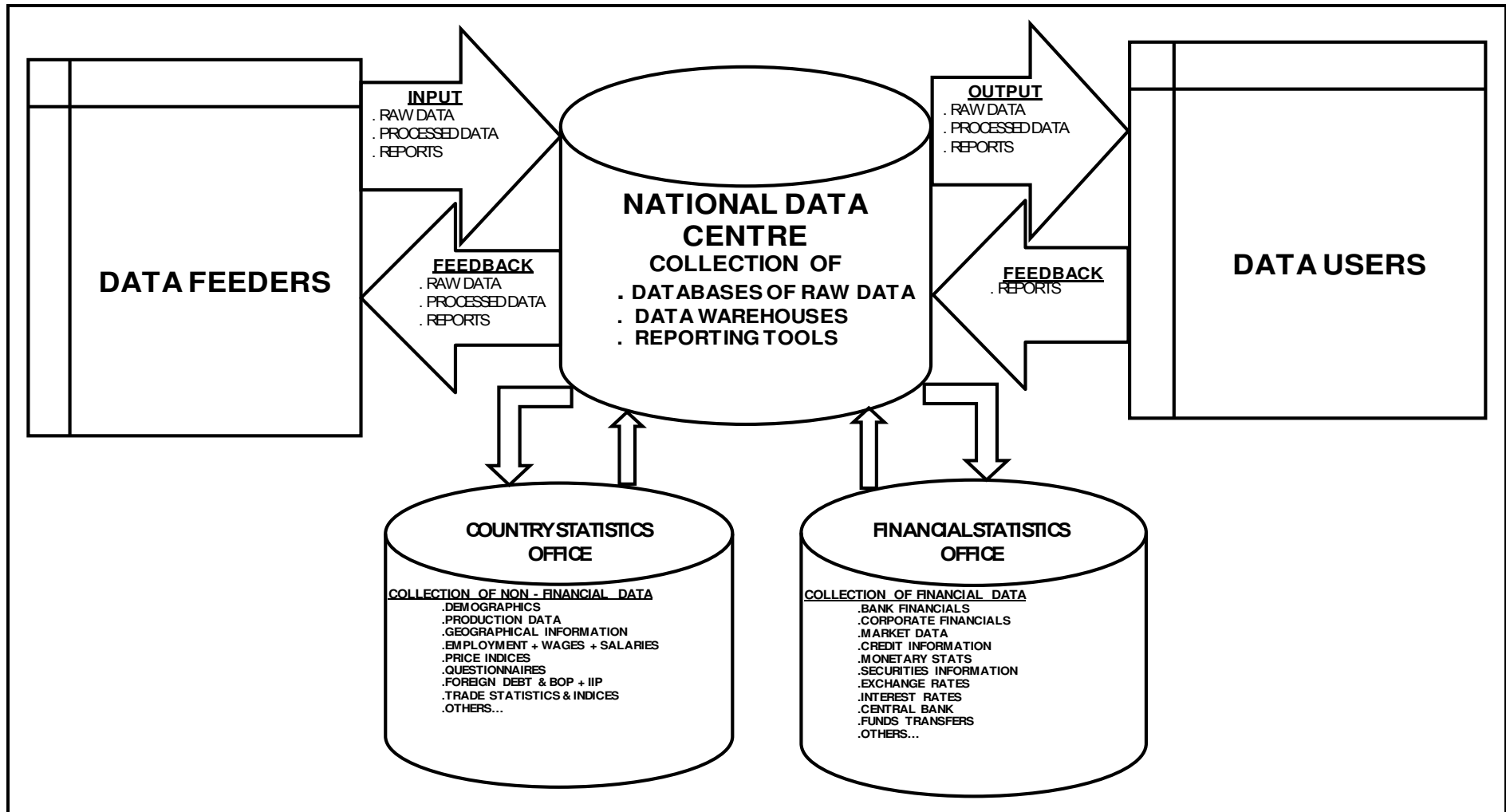


Figure.1 Proposed Architecture of the Data Centre

Core Competencies: What to Expect from the Centre, and What Can “it” Serve?

Data centre, by design, will be the *unique body* and the *sole provider* of the *public datasets*. Given its hardware and software capabilities what data centre is expected, in data management terms, can then be defined as the ability to:

- A. Collect as much data as possible (volume and variety)⁷
 1. At the most *optimal levels of cost* for the whole system, from data feeders to data users,
 2. In the *shortest time* possible (velocity),
 3. With the *lowest levels of error*,
 4. With *no redundancy*,
 5. With the *least effort* possible, and
 6. With the *smallest burden* for the stakeholders.

- B. Process the data
 1. With the *highest it capabilities* most stakeholders couldn't afford,
 2. In the *shortest time* possible (velocity), and
 3. At the *lowest possible cost*.

- C. Provide
 1. Stakeholders with the most *correct* and *timely* raw data and reports (veracity), and
 2. *Custodianship* for the data at the *highest levels of security*.

- D. Increase the *quality* of the data and reports

Thereby, provide the stakeholders with the highest *utility*.

Financial Statistics Office (FSO)

From 2008 on, the need by the economy managements and the policymakers for the *granular* data, with the *highest frequencies*, for their analyses and policy designs, have been felt at a higher pace and intensity. The ever rising demand for such data by the policymakers and the market players led us to, once not classified separately in the original design, come up with a distinctly defined financial database structure added to the original design. Financial Statistics Office has therefore become one of the two main departments of the design of the National Data Centre. Main financial datasets that are planned to reside in this sub database design basically are:

1. Banks' financial data
2. Corporate financial data

⁷ Volume, variety, velocity + veracity are (3Vs+V) the original design to describe Big Data by Gartner and veracity is the 4th V, one that was added to the model later on ^[3] ^[4].

3. Market data
4. Credit data and information
5. Credit card statistics and data
6. Monetary statistics
7. Securities data and information
8. Exchange rates data and information
9. Interest rates data and information
10. Central bank data
11. Treasury activities
12. Funds transfers
13. Others

Granular high frequency micro datasets belonging to unique credits, securities and financial products collected and stored by the FSO are predicted to find intensive use in⁸:

1. The timely and true identification and evaluation of individual credit risks of financial institutions and total credit risks for the whole system by supervisory bodies, and rating agencies,
2. Monitoring trends in capital markets, and timely intervention when the potential risks that'd discomfort/distress economic stability happen to be true (crisis management),
3. Detailed statistical and economic analyses,
4. By establishing proper credit risk rating systems, providing efficiency in credit markets, and better rationing of credit,
5. Better monitoring of financial innovations,
6. Financial stability analyses and reports,
7. Market efficiency evaluations,
8. Using capital flow informations in monetary policy design,
9. Computing the credit/debit positions and other relations of domestic and foreign actors in the markets,
10. Discovering the inter and intra-company fund transfers,
11. Monitoring fraudulent, money laundering, and illegal activities,
12. Tax collection, and designing the fiscal policy-extensions,
13. Determining the buy/sell relationship of the actors for financial products,
14. Calculating statistics related with the activities of the domestic and foreign banks and funds,

⁸ This inexhaustive tentative list has been prepared by using notes taken and presentations made at the IFC Workshop on "Integrated Management of Micro-Databases Adding Business Intelligence to Central Banks' Statistical Systems" at the Bank of Portugal, Porto, on 20-21 June 2013 (Proceedings published on BIS' web site as IFC Bulletin No 37 January 2014, <http://www.bis.org/ifc/publ/ifcb37.htm>) blended with authors own knowledge.

15. BOP and IIP calculations,
16. Systemic risk evaluations related to business (real sector), financial sector, and households, and
17. Macro prudential policy making, financial stability and systemic risk evaluations.

A Potential Use for the Central Banks: “Big Data and Systemic Risk Diagnosis”

As the monetary policymaker and looking for financial stability, the Central Bank is bound to look for most of these potential uses. Especially, the last two items in the list, prudential policy making, and systemic risk evaluation should be one of those with the highest priority. Briefly speaking, the systemic risk can be defined as "financial system instability, potentially catastrophic, caused or exacerbated by idiosyncratic events or conditions in financial intermediaries", and it refers to the risks imposed by inter-linkages and interdependencies in a system or market, where the failure of a single entity or cluster of entities can cause a cascading failure, which could potentially bankrupt or bring down the entire system or market.

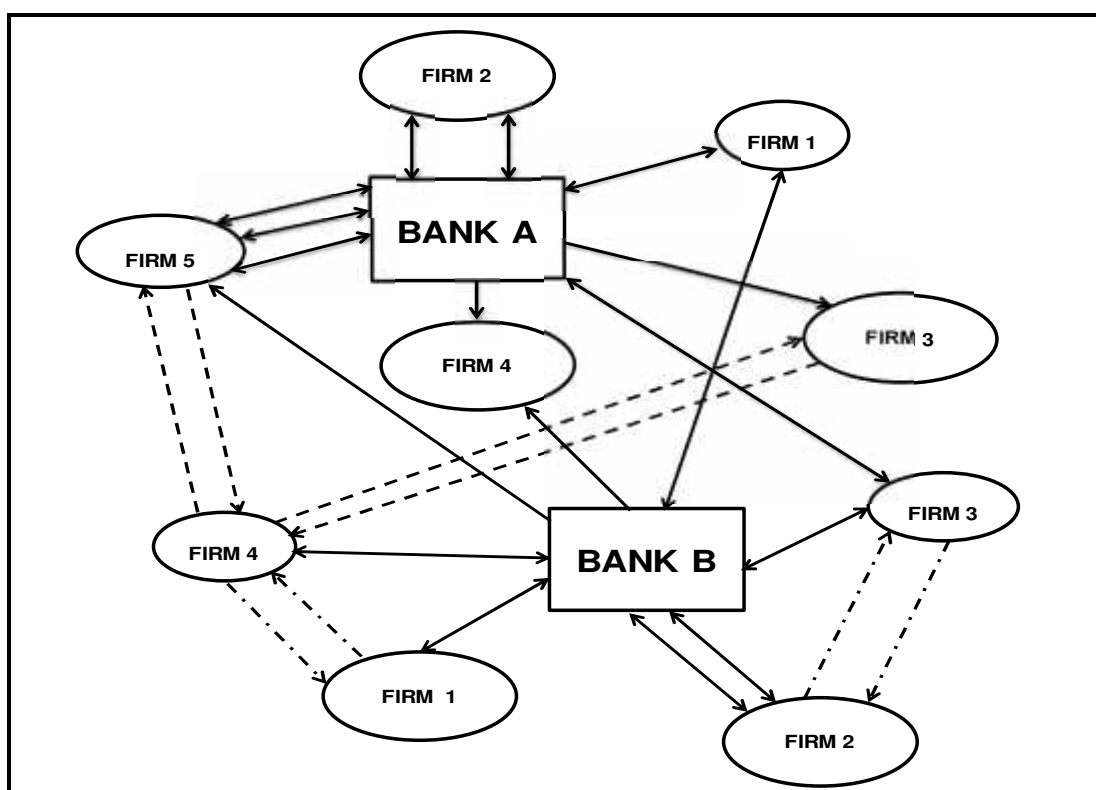


Figure.2 Network Topology of the Real and Financial Sectors via the Credit Relationships⁹

With this in mind, I have come up, a couple of years ago, early in 2011, with a project proposal that was meant to discover the *network topology* of the *real and financial sectors* via the *credit relationships* (either financial loan, or trade credits or both). The *degrees of interconnectedness* and *relations* would mean / define the *overall risk embedded in the whole system* if any one or a couple of large debtors would *fail to honour its / their debt(s)* (Figure.2). This risk cycle / circle here could be defined with a

⁹ Recall that a fully exhaustive systemic risk design should be one that captures all individual and counterparty risks of all entities within the circular flow of income in an economy. This simple model is presented here for illustrative purposes.

series of events, leaving out inter / intra company trade and other credit relationships in order of occurrence:

1. Big firm(s) *default(s)*
2. Bank goes into *financial distress*
3. Bank *fails to honour its debt*
4. Bank *recalls* smaller credit balances, *does not renew old balances* (no rollover funds)
5. Small firm(s) *default(s)*
6. Bank *defaults*
7. There is going to be a painful *bankruptcy procedure* for all three entities resulting in a
8. *Domino effect* for the *whole system* if the total of the relationship in question is a considerably high percentage of the total of the whole system.

This simple framework can be a good example of what we call *systemic risk*. This has to be monitored by authorities and an *early warning system* must be established if possible.

What's the relevance of the FSO to this project

It's the capacity of FSO to keep, store and process as much data as possible and all the reporting and analyses can be carried out in light speed in situ/place and feedback can be given to authorities and policymakers in the shortest time possible.

References

- [1] Kalil, Tom. "Big Data is a Big Deal", White House, (29 March 2012) Retrieved 15 December 2013 <http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal>
- [2] Executive Office of the President (29 March 2012) "Big Data Across the Federal Government", White House, Retrieved 15 December 2013
http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final_1.pdf
- [3] "Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data", June 27, 2011, Retrieved 15 December 2013
<http://www.gartner.com/newsroom/id/1731916>
- [4] "What is Big Data?", Villanova University, Retrieved 15 December 2013
<http://www.villanovau.com/university-online-programs/what-is-big-data/>
- [5] "G20 Data Gaps Initiative (DGI) – background", European Commission - Eurostat, Retrieved 15 December 2013
[http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/G20_Data_Gaps_Initiative_\(DGI\)_%E2%80%93_background](http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/G20_Data_Gaps_Initiative_(DGI)_%E2%80%93_background)