CENTRAL BANKING

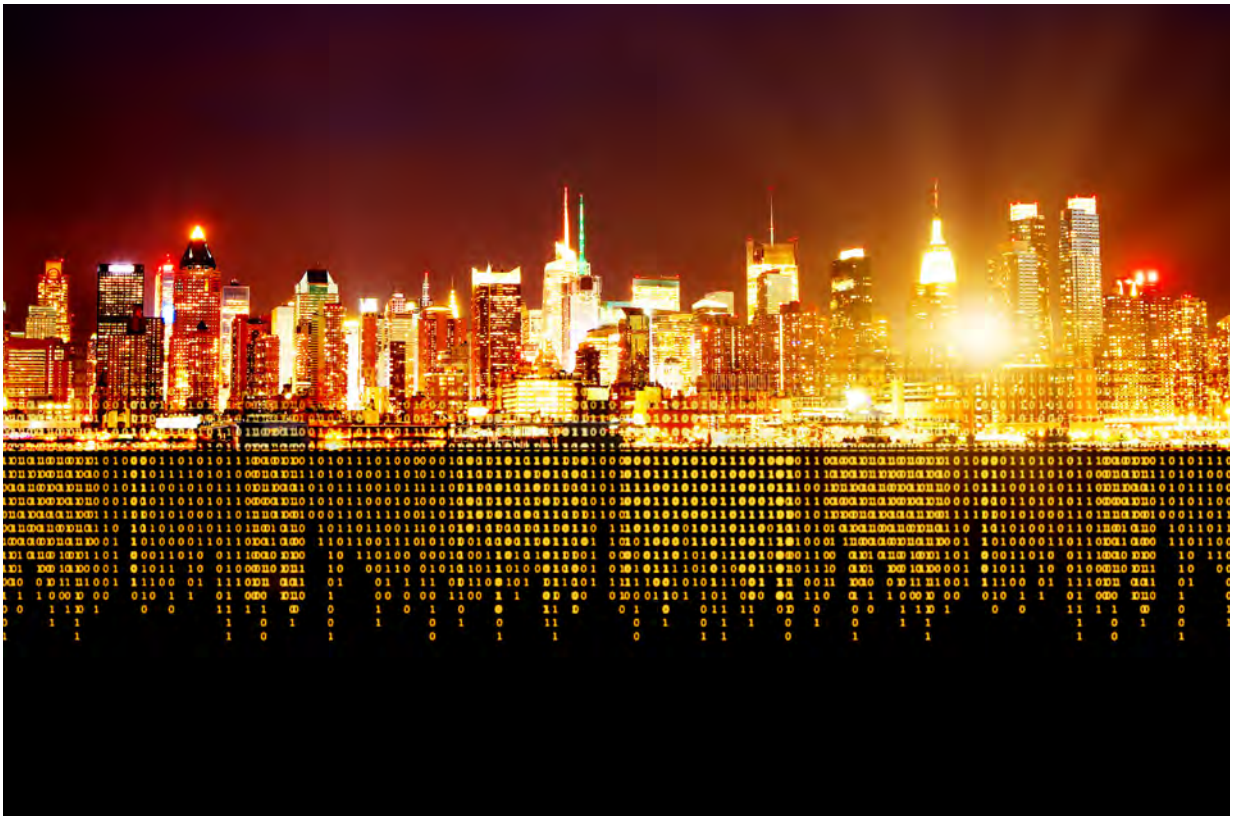
FOCUS REPORT

ECB'S ANACREDIT
Big data on bank risks

BIG DATA SURVEY
Central banks adapt to new paradigm

ONLINE FORUM
Experts discuss emerging trends

Big Data in Central Banks



In association with

BearingPoint®

Contents

The data deluge

Report editor: **Daniel Hinge**
daniel.hinge@incisivemedia.com

Contributors: **Emma Glass**
emma.glass@incisivemedia.com

Chairman: **Robert Pringle**

Editor: **Christopher Jeffery**
chris.jeffery@incisivemedia.com

Publisher: **Nick Carver**
nick.carver@incisivemedia.com

Commercial director: **John Cook**
john.cook@incisivemedia.com

Commercial editorial manager: **Stuart Willes**
stuart.willes@incisivemedia.com

Commercial subeditor: **James Hundleby**
james.hundleby@incisivemedia.com

Global corporate subscriptions manager: **Samima Danga**
samima.danga@incisivemedia.com

Cover image:
Peshkova/Shutterstock

Central Banking Publications
Incisive Media
Haymarket House
28–29 Haymarket
London SW1Y 4RX, UK
Tel: +44 (0)20 7316 9000
Fax: +44 (0)20 7316 9935
Email: info@centralbanking.com
Website: www.centralbanking.com

Published by Incisive Risk Information Ltd

The world's data is growing at an astonishing rate. Increasingly, statisticians are turning to data generated as the by-products of our daily, digitised lives, as well as larger and more fine-grained sets of regulatory data.

Good data, combined with good statistical analysis, has long helped economists overturn weak theories and throw light on dark areas of our knowledge. Learning to channel the deluge of data should only improve our understanding.

But with bigger data come larger and deadlier pitfalls. Messy, unstructured data sets and vast volumes may introduce biases that humans struggle to identify, and those biases must not find their way into policy. The need to handle such large volumes throws up questions of processing power, storage, efficient coding and how to organise the information within an institution so that the right people can make use of it – but no one else.

The aim of this focus report is to offer assistance to central bankers in executing this demanding transition. Our survey of central banks highlights how the move to better data governance is proving a challenge for many, while our Q&A with the St Louis Fed, an established leader in data management, may offer some answers and guidance. The European Central Bank's Aurel Schubert outlines a major European project to harness data for better regulatory outcomes, and our online forum panellists discuss the potential for big data to better shape policy and regulation.

Central banks are uniquely placed to deepen our collective understanding of data. They concentrate within their ranks experts on economics and statistics, but also employ a growing number of computer scientists, mathematicians and physicists. Central banks are also well versed in communication, giving them a sturdy platform from which to encourage a deeper public understanding of statistics and resist the rise of lazy or misguided interpretations.

There is much work to do. We hope our modest contribution can make the way ahead a little clearer. ❑

Daniel Hinge
Editor

AnaCredit: banking with (pretty) big data

There has been a decade-long effort to create a supranational credit database in Europe that will help policy-makers and banks assess cross-border risk when it goes live in 2018. By Aurel Schubert.

In May this year, the governing council of the European Central Bank (ECB) approved a new statistical regulation (ECB/2016/13) establishing, as of end-2018, a common, granular big data database, known as AnaCredit, shared between eurozone member states and comprising harmonised data on credit and credit risk.

The establishment of AnaCredit marks the beginning of a new era for central banking statistics, a genuine paradigm shift triggered by the need to "move beyond the aggregates", to quote the title of the Eighth ECB Statistics Conference in July 2016.

At the same time, it can be seen as the final achievement of a lengthy and, at times, very challenging process, the start of which can be traced back to 2008–2009. At that time, just a few months after the onset of the financial crisis, a small, high-level committee chaired by Otmar Issing[1] identified the creation of a 'global credit register' as one of the necessary elements of a more resilient international financial architecture. In particular, referring to the data from a credit register, the Issing Committee observed that "while the value of such information is appreciated almost universally on a national level, there is nothing commensurate on an international level", thus suggesting that "given the current high level of international lending and exposures, a global credit register will greatly enhance risk management, both at the firm level (improving due diligence of cross-border exposures) and at the systemic

Aurel Schubert has been director-general of the European Central Bank's (ECB) statistics department since 2010. He also chairs the Statistics Committee of the European System of Central Banks, and is vice-chair of the Bank for International Settlements' Irving Fisher Committee on Central Bank Statistics. Schubert spent 25 years working for the National Bank of Austria before moving to the ECB.

Network analysis

Network analysis, or graph theory, is the study of systems of nodes and vertices. In the field of financial stability analysis this can be used to represent financial entities (nodes) and the links they have to counterparties (vertices).

While it need not involve big data, network analysis often draws on very large quantities of data. An obvious source is information on derivatives trades, which is accumulating rapidly in trade repositories.

A recent study by the Bank of England used this data to map the interconnections between dealers, central counterparties and end-users in the derivatives market.[2] Such a graphical representation can immediately shed light on where the most important connections lie, and therefore where shocks are likely to have the greatest contagion. The authors found the links between counterparties were highly correlated with measures of systemic importance.

It is also possible to map networks on the basis of less structured data, as a study published in January 2016 by the European Central Bank (ECB) showed.[3] Researchers focused on how mentions of bank names in news stories could produce a network of interrelations between banks, allowing them to rank institutions by their systemic importance.

Examples of the ECB's interactive visualisations are available online, and show how the network grew denser and more connected as the 2008 crisis developed.[4] Banks such as RBS, which was ultimately bailed out, moved to the centre of the network as the situation worsened. As the crisis faded, the links became weaker and the major players moved to less central zones. Tugging a node shows how ripples spread through the system – RBS impacts every other bank, while less connected players such as Nordea or DZBank have much less effect.

Other forms of text mining (see box on page 57), particularly sentiment analysis, have the potential to offer richer insights into such networks, the authors say.

Daniel Hinge

level (adding a cross-border dimension to financial stability stress testing and to an evaluation of real effects on the economy)."[5] In concrete terms, the committee proposed that "a harmonised approach should be adopted, where harmonisation refers to the standardisation across countries."

Ten years after these recommendations were made, AnaCredit will finally offer policy-makers in the eurozone the possibility of making their decisions based on highly detailed, timely, accurate and standardised information on credit and credit risk.

Bank credit plays a key role in the eurozone economy, where the share of loans in the total external financing of small and mid-sized enterprises (SMEs) is overwhelmingly higher than in the Anglo-Saxon world, where market financing

plays the most important role. Moreover, non-performing loans in the eurozone have surged since the onset of the sovereign debt crisis and have recently been estimated at almost €1 trillion ($1.12 trillion), a value that is similar to the annual GDP of a large EU member state. Good statistics on credit are therefore clearly essential for the ECB – or any other central bank – to fulfil its mandate.

Aggregate statistics failings

At the same time, the recent economic and financial crisis has suddenly revealed the inadequacy of 'traditional' aggregate statistics to support 'unconventional' policy-making. In particular, we have learned that aggregate statistics, although of high quality, are not sufficient for an adequate understanding of the underlying developments in the wake of increased heterogeneity and market fragmentation, which has resulted in developments across different segments of the economy. Information at the granular level is thus becoming increasingly necessary to support policy decisions.

It is worth recalling that the collection of detailed information on credit and credit risk is not new. Several countries within the EU are operating granular credit data sets, either via national credit registers – normally run by the national central bank (NCB) – or via private credit bureaus. Nevertheless, information is very limited or totally absent in some countries, while significant methodological differences prevent meaningful supranational analysis of the information currently available – for instance, data is collected on a loan-by-loan basis in some cases and on a borrower-by-borrower basis in others. The reporting threshold, defining the minimum exposure recorded in the credit register, also varies considerably across countries, ranging from effectively no threshold up to a few hundred thousand euros. The result is that no meaningful insight can be derived from a cross-country analysis of the current national data.

By contrast, the AnaCredit data set will deliver, mostly on a monthly basis, loan-by-loan information on credit to companies and other legal entities (no natural persons will be covered) extended by, at the minimum, eurozone banks and their foreign branches.

A complete credit picture

In particular, the AnaCredit data collection has been designed to obtain a complete picture of the credit exposure of the reporting population. By linking 'observed agents' with 'reporting agents', data is collected on loans granted by credit institutions resident in the eurozone irrespective of whether such loans are provided directly by the credit institutions or, indirectly, via subsidiaries or branches under their control. The information collected comprises almost 100 different 'attributes' covering various aspects of the credit exposure (outstanding amount, maturity, interest rate, collateral or guarantee, information on the counterparty, and so on), and is organised in several 'tables' connected to each other via the 'instrument', which is the centre of the data model.

Besides eurozone member states, other EU countries can join this endeavour on a voluntary basis. Some have already declared their intention to do so, and others will hopefully follow in the coming years. This will result in a complete, detailed, timely and accurate overview of credit and credit risk developments in all participating countries, finally allowing a complete and meaningful cross-country analysis, as urged by the Issing Committee.

The decentralised approach, with NCBs collecting information from their reporting agents and then transmitting it to the ECB, has been designed to allow

Text mining

Text mining is an umbrella term for a range of techniques for translating text into a form that can be analysed quantitatively. Economists at the Bank of England (BoE) recently authored a handbook on the subject, noting that, while other forms of quantitative analysis are more widely used and therefore less costly, investment in text mining could unlock analysis that is useful for policy and cannot be tapped by other means.[6]

One such technique is Boolean text mining, which uses simple operators such as "and", "or" and "not" to form logical search expressions. Its simplicity means it can be scaled to handle very large quantities of data. The researcher chooses combinations of words and the computer program counts how often they appear within a piece of text, which in turn can offer basic measures of sentiment, uncertainty, and so on.

The BoE used this technique in its study of Twitter traffic around the Scottish independence referendum in 2014, searching for combinations that could imply the development of a bank run. The problem with such a method is it fails to capture context or subtlety of meaning, which can lead to mistakes, such as when the BoE's program highlighted instances of "RBs" referring to "running backs" – a North American football position – not the Royal Bank of Scotland (RBS).

More nuanced approaches seek to capture at least some of the meaning in text. One example is latent semantic analysis, which is designed to assess the similarity of language across documents. Miguel Acosta, an economist studying under John Taylor at Stanford University, used the technique to compare Federal Reserve meeting transcripts before and after members of the Federal Open Market Committee (FOMC) were aware that verbatim transcripts were being made.[7]

Acosta showed there was a difference between the two periods. His results imply that, while publishing verbatim transcripts increased transparency, it also reduced the extent to which FOMC members disagreed with one another.

Daniel Hinge

the necessary national flexibility and exploit the synergies with existing credit registers to the maximum extent. A satisfactory balance has been found between the need to take into account country-specific features, such as the degree of concentration of the banking sector, and the strong demand for common rules allowing a level playing field across all participating countries. The relatively low reporting threshold (€25,000 calculated at the borrower level) allows a sufficient coverage of credit to SMEs, which are the backbone of the European economy both in terms of employment and value added. At the same time, to alleviate the reporting burden and in line with proportionality, some discretion has been left to NCBs in granting – partial or complete – derogations to small institutions in their respective countries.

Early estimates point to around 100 million exposures reported every month, representing loans granted by about 5,000 credit institutions to more than 15 million counterparties. In view of these large expected data volumes, a state-of-the-art IT infrastructure is currently under development, which also takes into account the need to ensure an adequate protection of confidentiality.

An unprecedented step

All in all, this is clearly nothing short of a 'jump into the world of big data' for central bank statistics. Although the big data definition might not be fully appropriate in absolute terms – we are still far from the billions of intra-day data points flooding in other contexts – it certainly represents an unprecedented step for the ECB, which traditionally relies on a relatively small set of aggregate statistics for supporting policy analysis and decisions: a true paradigm shift.

Still, in the face of the obvious initial investment and adjustment required by a project of this scale, AnaCredit will bring huge benefits to many stakeholders, reporting banks included.

Collecting very granular information will support the needs of policy-makers in several important fields: monetary policy analysis and operations, risk and collateral management, financial stability and economic research. The AnaCredit data sets will allow a better understanding of the monetary policy transmission channel, particularly with regard to transactions involving SMEs. Although in the initial stage of the project no data will be specifically collected for supervisory purposes, banking supervision will also find the data very useful in many respects, thanks especially to the information on the link between lenders and to the unique identification of counterparties across the entire lender population.

Obviously, potential users – the ECB and participating NCBs, national and European supervisory authorities, national and European resolution authorities, the European Systemic Risk Board and the European Commission – will have access to the AnaCredit information at different levels of detail depending strictly on their proven needs, and in any case always according to strict confidentiality rules as set out under existing relevant European law.

Just as importantly – and more in perspective – we will be in the position to respond to rapidly changing data requests from users in a timely and cost-efficient manner, without having to rely on very costly *ad hoc* data requests or new reporting requirements, and with clear savings and benefits for the banks. It will be truly multipurpose and allow flexibility for analysis. Moreover, more complete and comparable information returned to reporting agents – via feedback loops that might be established at national level by the respective NCB – will also be beneficial for banks in assessing the creditworthiness of new potential borrowers, even when the latter have multiple cross-border exposures.

Being fully aware of the paradigm shift we are confronted with, the ECB statistical function is working, with the involvement of the financial industry, towards designing and implementing co-ordinated data management, comprising information collected under different statistical and legal frameworks. The main workstreams in this field, which have already been under way for some time, relate to: the development of a bank's integrated reporting dictionary (BIRD), defining a sort of 'common language' for the information provided by reporting agents, an ECB Single Data Dictionary and an integrated European Reporting Framework, covering both statistical – monetary policy, financial stability, and so

on – as well as supervisory reporting. They all point in the direction of providing financial institutions with a unified reporting framework based on a consistent and stable set of rules and definitions with a twofold goal: alleviate the statistical reporting burden and, on the side of the authorities, ensure data quality and consistency; and allow a combined use of all granular information – for example, data on debt securities and, in perspective, credit exposures.

The gradual development of a fully fledged master data set with reference data on all counterparties involved in the exposures covered under the various requirements – lenders, borrowers, holders, issuers, protection providers, etc. – is an important example of this effort and of the challenges entailed in the move from aggregates to granular statistics. Such a reference data set is directly functional to the unique identification of counterparties, which is a precondition for calculating the total exposure of a borrower (and/or issuer) vis-à-vis the whole lenders' population. Together with complete and up-to-date reference information on counterparties – for example, sector of activity, size, geographic location, annual turnover – this will allow a very informative analysis per specific segments of the economy.

The definition of the AnaCredit requirement has posed several challenges and now others arise in preparation for the first reporting, due in autumn 2018. Still, the ECB and the Eurosystem are confident that all such challenges were addressed in the best possible way and that the benefits of AnaCredit will definitely outweigh the efforts put into its establishment and continued operation.

Clearly, AnaCredit by itself might not save us from any new financial crisis but, as the Issing Committee observed 10 years ago, a credit register "would capture the longer-term trends that history shows have often posed the biggest threat to financial stability". It will definitely put policy-makers in a better position to mitigate the risks *ex ante* and to limit their potential impact *ex post*. ❑

The views expressed are those of the author and do not necessarily represent the views of the ECB or the ESCB. The author thanks Riccardo Bonci for his help in preparing this contribution.

Notes

1 Besides Otmar Issing, the Committee comprised Jörg Asmussen, Jan Pieter Krahnen, Klaus Regling, Jens Weidmann and William White.

2. Robleh Ali and Nick Vause (Bank of England) and Filip Zikes (Federal Reserve Board of Governors), *Systemic risk in derivatives markets: a pilot study using CDS data*, Financial Stability Paper No. 38 (Bank of England, July 2016) *http://tinyurl.com/z3q5ge9*
3. Samuel Rönnqvist and Peter Sarlin, *Bank networks from text: interrelations, centrality and determinants*, Working Paper Series No. 1876 (European Central Bank, January 2016) *http://tinyurl.com/h68phf3*
4. Bank networks from text: Interactive visualization *http://risklab.fi/demo/textnet/*
5. Issing Committee, *New financial order recommendations*, White Paper No. I *http://tinyurl.com/gojx4k8* and White Paper No. II *http://tinyurl.com/zy82czx* (Center for Financial Studies, Goethe University Frankfurt, February 2009).
6. David Bholat, Stephen Hansen, Pedro Santos and Cheryl Schonhardt-Bailey, *Text mining for central banks*, Handbook No. 33 (Centre for Central Banking Studies, Bank of England, 2015) *http://tinyurl.com/jmetfz6*
7. J. Miguel Acosta, under the direction of Professor John B. Taylor, *FOMC responses to calls for transparency: evidence from the minutes and transcripts using latent semantic analysis* (Department of Economics, Stanford University, 2014) *http://tinyurl.com/htgtpu9*

Harnessing growing data volumes: the St Louis Fed

Huge demand for macroeconomic data that is easily accessible, international and granular is driving the rise of this global data hub. Daniel Hinge meets the Fred team.

The Federal Reserve Bank of St Louis has become a hub of global data, with its Federal Reserve Economic Data (Fred) database growing from small beginnings to a leading example of data management. Across nearly 400,000 time series – and growing – the database provides an easily searchable resource for economic researchers.

With such large volumes of data being produced, gathering the data quickly and then organising it in such a way that people can find what they need is a major challenge. As any user knows, many national statistical websites are showing their age, with poor search functionality and frustrating interfaces. Search for 'headline inflation' and you are often presented with 100 different series – but not the one you are looking for. The St Louis Fed's experience developing Fred may offer a useful example of how to cope as data becomes bigger.

How did Fred get started?

Katrina Stierholz: More than 50 years ago at the bank, back in the print days, there was an interest in making data available. We put out data publications to provide people with information on the current economic conditions, particularly monetary indicators. These became very popular, and when everything went online so did Fred. This year is the 25th anniversary of Fred being accessible via the internet.

It started out as an electronic bulletin board, and it was a fairly small product for many years. In the 1990s, I think we only had 3,000 time series available. It was not until probably 2004 that we grew to more than 10,000 series. There has been an incredible growth rate in the past 10 years.

Keith Taylor is co-ordinator of the St Louis Fed's data desk, which is part of the reserve bank's research division, and manages the collection, organisation and publication of Fred data. Before joining the Fed, Taylor practised law at firms in Missouri and Illinois. He holds a doctorate in law from the Washington University School of Law.

Katrina Stierholz is a vice-president in the St Louis Fed's research division, and director of library and information services. She is in charge of the reserve bank's physical and digital libraries, and oversees the data desk, which posts Fred data. Before joining the Fed, Stierholz worked at the Washington University School of Law Library. She holds an MSc in library and information science from the University of Illinois at Urbana–Champaign.

What changed in 2004?

Katrina Stierholz: A couple of things. We realised that the paper data publication world was disappearing, so we decided to focus on Fred. We hired some new people who had fresh ideas and we surveyed our users. We asked users what they wanted and the answer was overwhelming: "more data".

We doubled down on adding data, especially a lot of international data. We had regional data and continued to work on that, but then we added international data as a focus. You can imagine, when you start adding time series for every country in the world, that adds up pretty quickly.

How is the data represented? Fred charts are one of the standout features.

Keith Taylor: Originally, it was just downloading data. Around the early 2000s we added charts, and since then we have steadily tried to develop better charting software. Much of that was initially driven by an internal interest in being able to represent the more advanced graphics that you would see in our data publications. Many of them are basic time series, but we also add other kinds of interesting graphs. We have various graph types – line charts, bar charts, pie charts, area charts and scatter plots.

We also have GeoFred. We realised that, once we have this regional data along with the international data, mapping it is one of the best ways to represent it. And we have dashboards, which allow you to put a bunch of different charts all on one webpage, where it all updates automatically.

Have you seen demand for the data change over time?

Keith Taylor: The big change was that users really wanted a lot of international data and more granular data. If you ask any researcher: "Would you like this data?" they say: "Yes, of course." No one ever says no to data.

Our core data is macroeconomic indicators such as GDP, inflation and unemployment. But there has been great interest in indicators that impact the economy but maybe aren't traditionally thought of as macroeconomic indicators. We are exploring crime statistics, poverty data, health data and health insurance – there has been a lot of interest in that kind of data.

In addition, we have seen our user base move from super-sophisticated people – bankers, academics, economists – towards, especially over the past five years, less sophisticated users who are much more interested in data that is easily accessible.

You've touched on some non-conventional data sources – are you interested in big data?

Keith Taylor: We're not doing anything in the sense of, for instance, the Billion Prices Project. But I think I have a big data problem any time I have more data than I can handle. For certain kinds of users, a spreadsheet with 10 columns is a big data problem, because they just do not have the right knowledge. Then, at the other end of the scale are Google, Amazon or Facebook, which are dealing with petabytes of data.

Fred is really about solving the big data problem for the people at the other end of the spectrum. If you wanted to get GDP from the Bureau of Economic Analysis (BEA), you could do analysis and compare that with, say, the Bureau of Labor Statistics' (BLS) non-farm payrolls. For many people, that is a big data problem because they have to go to two different sources, navigate two different sets of websites, download relatively large files, parse out what they want and make certain kinds of conversions. That's really where our focus is: streamlining that process and ensuring that someone can do their analysis much more quickly.

Are you seeing that kind of problem appear more often now that data volumes are growing?

Keith Taylor: Yes, definitely. I've been here for about five years, and when I started we would typically work with a release that had several hundred series, maybe a few thousand series. Now, let's say you want all the counties for the US – that's 3,000 series, and that's going to be in a data set of 60,000 series. It is not super complex, but many people don't have the skills to to tease that stuff out.

Are you hiring new specialists to manage these challenges?

Keith Taylor: Not yet, but we recognise that day is coming. However, we are already hiring more people with a high degree of fluency with metadata.

Can you explain the idea of metadata more?

Keith Taylor: Metadata is the data about the data. From the Fred perspective, you have the series-level information, and that describes what the indicator is. Take GDP. Which country is it? What are the units? Is it seasonally adjusted or not? What other adjustments have been made? Is it inflation adjusted?

Having well-defined metadata allows users to have a high degree of confidence that, when they search for something, they find what they are looking for. It also gives them a greater degree of understanding if they have additional questions. We're focusing a lot on metadata to improve search functions and to help people understand the data.

It seems like metadata is a big issue, even if it doesn't grab headlines in the way standard data does.

Keith Taylor: That's especially true when you're talking about big data. For example, on Fred, if you search for GDP, you get perhaps 27,000 results. They all measure GDP in some way, or they are a component of GDP, but how do you sort through that and then know that what you have really is what you were looking for? Metadata is a good way to do that.

Machine learning

Machine learning is the process of teaching a computer to act in certain ways without being specifically programmed. Examples in broader society include self-driving cars, speech recognition and web searches. Two major areas of machine learning in use among central banks are regressions and classifiers – the former extrapolates trends from a set of 'training data', while the latter sorts items into categories by recognising patterns.

One use of regressive learning is to produce 'nowcasts' for fast estimates of macroeconomic variables, derived from less traditional but higher-frequency sources of data than those typically employed by national statistical authorities. Computers can extract patterns from large sets of noisy data, producing a model with good predictive power, even if humans struggle to explain why.

A recent International Monetary Fund working paper applied this approach to nowcasting Lebanese GDP,[1] which tends to be published with a long lag. Author Andrew Tiffin, however, observed there are "pitfalls" to such an approach. A major problem is "overfitting", where the computer produces such a good fit to the training data that it captures noise as well as signals, meaning its predictions are likely to be poor.

By contrast, researchers from the Central Bank of Colombia recently made use of a classifier approach to categorising banks according to features of their balance sheets. Employing a "neural network", they found balance sheets are "unique and representative", meaning banks can be identified on the basis of their balance sheets.[2] The researchers hope their approach could improve supervision and contribute to the design of early warning systems.

Daniel Hinge

Does the St Louis Fed have research streams looking to use the data in innovative ways?

Keith Taylor: Michael McCracken, an economist here, is working on something we call Fred-MD. The relatively famous Stock and Watson data set has not been updated since it was originally compiled. It is very difficult for other researchers to obtain that data.

Michael worked to identify related series that existed in Fred; then he wrote a paper demonstrating that the data set he created was basically the same as Stock and Watson on a number of different measures.[3] We now update that routinely; we have been adding vintages to it as well. You not only have the latest values for these 90 series, but also the monthly vintages, going back about a year now, which allow a new kind of analysis.

How does the St Louis Fed handle data governance?

Katrina Stierholz: Everything is in the research department. The Fred data team handles it well and has many processes and procedures for uploading

the data and checking it. We also have a data librarian who looks at the data to ensure it complies with copyright and licensing. If it is not public data, sometimes data series require permission, although around 99.9% of data is publicly available.

Is security a challenge? Where do you store the data?

Keith Taylor: Like all large organisations, we take security very seriously. We cannot really discuss where we keep the data.

What other challenges are thrown up by data management?

Keith Taylor: Our biggest challenge is keeping the database going, from a processing standpoint.

We can have high levels of traffic on our website, which ends up really pulling on the database a lot. We have done a number of things in terms of virtualising servers, and we have put in place caching software to lessen the burden.

We are an aggregator, so when the data becomes available at the source we go out and grab it. We want to get it into Fred as quickly as possible, so if someone is making some market decisions they can use Fred to do that. Another constraint is how quickly we can get the data down, process it, run it through all of our checks and then load it into Fred.

Can you explain more about how the aggregation process works?

Keith Taylor: We do it in a variety of different ways. In a small number of situations, we partner with the other organisation. It will normally tell us it would like us to host the data on Fred, but that tends to be only very small agencies. For the big agencies – such as the BEA, BLS, Organisation for Economic Co-operation and Development and the World Bank – we go out and use the interfaces they provide to the public. Whenever possible, we use application program interfaces (APIs) or bulk downloads, which many sites now have.

In the cases where they do not have that we 'scrape' the data, or we download a file that is really meant for human consumption. We have written custom scripts that will deconstruct these files and analyse the data before loading it into Fred. In an ideal world, they just say "here's the data", which is perfectly formatted, but that is a very small percentage of the time.

Where the interface is provided, is that quick enough for your needs?

Keith Taylor: It depends. Some sources have interfaces for pulling the data down very quickly and easily, but do not have all the data on the release. The BEA set up an API recently, which is awesome, but it is not updated as quickly as its public consumption files, so we are not able to use it.

There has been a move to put many of these things in APIs or in bigger bulk download programs. We have to keep looking out there to see if there is a more efficient way to get it.

And does the data match up as well as you would like? Is that a question of metadata?

Keith Taylor: Yes, we spend our days trying to get the data to match up. One challenge is that the metadata is revised over time, so even if you get the metadata

to match up initially, that may change. We track those revisions in another product called Alfred. We have all of the revisions of both the series-level metadata, and the observations themselves. That takes up a huge amount of time.

Right now, observations are handled automatically in terms of tracking revisions to them. The series metadata is kind of half-automated – and so we are looking to improve our metadata schema here and trying to work with others to establish standards around metadata, which would allow us to build an automated process.

What are your plans for the future? What improvements would you like to implement?

Keith Taylor: I am really interested in looking at how we can improve our metadata, which would lead, I think, to improvements in searching. According to user feedback, this is an area we can do better in. Because we have about 400,000 series, it's a case of finding it and having a high degree of confidence that what you found is what you want. The other area of focus is to continue to expand the data into other areas.

Katrina Stierholz: We have a growing base of novice users, so we're doing things like adding related content to series to help people understand what the series is. We're adding links to our economic education materials, to historical publications and to articles written on the subject. If you search for GDP, and you don't know what GDP is composed of, then you can look at some of that information and understand it more. We are trying to help those fairly new users understand what they are getting.

Many central banks are testing new communication methods. Do you have educational programmes?

Katrina Stierholz: We have a very comprehensive economic education programme here at the St Louis Fed. We have materials based on the curriculum that have been written to support teachers. They are designed to be given to teachers, who then deliver them to students. The curriculum provides basic economic concepts, is matched up to curriculum standards and is available in several formats – print, online, podcasts and videos. There is a ton of really powerful economic education, and it is widely used in the US.

Do you have ways of measuring the impact?

Katrina Stierholz: The programmes include pre-tests, and post-tests for the online courses. Students are required to take a pre-test, and then a post-test to see if they have learned anything. They consistently, regularly and, to a degree of confidence, show improvement in their scores. ❑

Notes

1. Andrew Tiffin, *Seeing in the dark: a machine-learning approach to nowcasting in Lebanon*, (International Monetary Fund, 2016) *http://tinyurl.com/hvwnyjx*
2. Carlos León, José Fernando Moreno and Jorge Cely, *Whose balance sheet is this? Neural networks for banks' pattern recognition,* Borradores de Economía 959 (Central Bank of Colombia, 2016) *http://tinyurl.com/he4rcmr*
3. Michael McCracken and Serena Ng, *Fred-MD: a monthly database for macroeconomic research* (Research Division, Federal Reserve Bank of St Louis, August 2015) *http://tinyurl.com/hqujf4c*

Big data in central banks

A new survey of central banks highlights active interest in big data, governance concerns and a split over support from policy-makers, writes Emma Glass.

This article sets out the results of a survey of the way in which central banks view big data and data governance in their institutions. The survey was conducted by Central Banking, in association with BearingPoint, during August and September 2016. The work has only been possible with the support and co-operation of the central bankers who agreed to take part, and who did so on the condition that neither they nor their central banks would be named in the report.

Key findings

- Central banks have an active interest in big data. This is manifested in improving processing technology, adapting institutional strategies and increasing staff awareness of the area.
- Central banks typically see big data as unstructured data that is sourced externally, although this view is not universally held.
- Overwhelmingly, central banks develop their own data platforms to handle regulatory data collection, a role that has taken on greater significance since the financial crisis as central banks have expanded their involvement in financial stability.
- Big data is predominantly regarded as useful for research, but a significant minority sees its immediate involvement in policy-making, or scope for this.
- Lack of support from policy-makers is seen as the most significant challenge to increase the use of big data.
- Central banks do not in the main have a dedicated budget for handling data, including big data, though many are seeking one.
- A little more than 80% of respondents said they have no intra-departmental or divisional bodies dedicated to big data.
- More broadly, central bankers have concerns over the arrangements in place for managing data in their institutions. Many are looking to improve data governance.
- More than three-quarters of respondents indicated they had a shared internal platform, typically in the form of reporting frameworks and data warehouses.

- Central banks generally source their own big data sets, though a significant minority increasingly look elsewhere. Overwhelmingly, they process the data sets themselves, and there is no indication of a desire for this to change.
- Monetary policy is seen as standing to benefit most from big data, though it is expected to have a significant impact on macro-prudential policy as well.
- Support from the executive level and policy-makers divided respondents: 35% saw it as the highest priority for investment within the central bank, while 38% saw it as being the lowest.
- Central banks have developed their own data platforms to deal with regulatory data analytics, often used in conjunction with other options. Excel remains popular.
- Central bankers broadly welcome the idea of self-assessment of data management using an adapted version of the Basel Committee on Banking Supervision's principles for supervisory data aggregation (BCBS 239).

Profile of respondents

Responses were received from 42 central banks, of which the average staff size was 1,652; three-quarters of respondents had fewer than 2,000 employees. Just over half of respondents were from central banks in Europe. Those individuals taking part in the survey were drawn primarily from the statistics function: 32, or three-quarters of respondents, were located in this area; three were from information technology; and responses were also received from research, banking supervision, international relations, data management, infrastructure and technology, and data collection departments.

Geography	% of respondents
Europe	53
Americas	21
Africa	14
Middle East	9
Asia-Pacific	2
Total respondents	42

Staff size	% of respondents
<500	24
500–999	31
1,000–1,999	24
2,000–4,999	16
>5,000	5
Total respondents	42

Economic classification	% of respondents
Industrial	36
Developing	31
Emerging-market	29
Transition	5
Total respondents	42

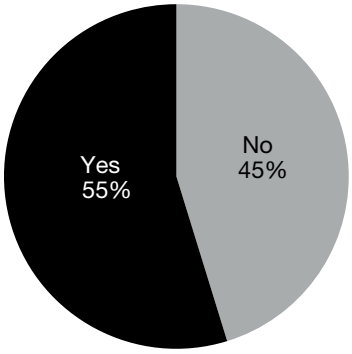
Department	% of respondents
Statistics	76
IT	7
Other	17
Total respondents	42

Percentages in some tables may not total 100 due to rounding.

Detailed responses

Has your central bank given any active consideration to big data in the past 12 months?

Central banks have an active interest in big data, particularly in improving processing technology, adapting institutional strategies and increasing staff awareness of the area. Just over half of respondents to the survey said they had given the area active consideration in the past 12 months. This group of 23 central banks was drawn from around the world, although European central banks figured prominently.



Seven central banks referred specifically to technological upgrades needed for, or focused on, big data – in particular updating infrastructures and embarking on data warehouse projects. A statistician from Europe detailed a recent innovation at the central bank: "In 2015, the central bank opened a new platform to share individual data." An African central banker said their institution was building a data warehouse, and one central bank chief data officer said: "We are looking to adopt technologies that have evolved in the big data industries."

Five central banks made reference to big data in strategic planning, either at a data departmental or bank-wide level. A European central banker explained how big data was playing a part in a broader strategic initiative: "The optimisation of the data we collect, to derive insight and enhance decision-making, is a key element of the bank's strategic direction. Big data is a component part of this conversation." A central banker from the Americas highlighted the secular increase in data: "This development is expected to continue at an ever-quicker pace and the central bank has reviewed its data and information strategies in order to benefit." Another respondent, also from the Americas, indicated the strategic plan of the bank had been altered in the past year, adding: "We are also running projects using unstructured data sets."

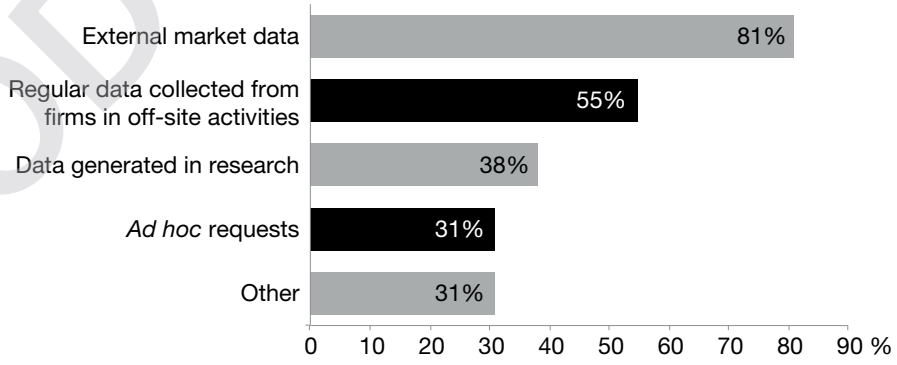
Several respondents described how their central banks were looking to increase awareness and skill sets in the area, notably through shared platforms, internal forums and training for employees. Another central bank had launched a big data forum, for internal discussion and "encouraging the development" of big data projects. A statistician from a developing country said it had initiated "training specific personnel in data science and R programming", and officers from that central bank were actively discussing big data in the context of creating a new autonomous national statistics institution. One large central bank has created a specific big data team: "In the last two years we have been conducting some experimentation on both business and technological profiles. This year we set up a team to pursue the first statistical results from the application of big data technology from structured and unstructured data."

In contrast, 19 respondents said that they were not actively considering big data. They included seven European central banks and two central banks from

the Middle East. One European central bank said that, as it did not have a "predefined strategy", big data was not a priority, and one said simply: "We do not intend to use big data." However, several respondents who did not perceive big data as a priority in the short term commented that it would figure in plans in the medium term. A European central banker explained why it had not impacted their central bank to any significant extent: "We can say that the central bank's actual involvement in the use of big data is currently rather limited, as we are relying mainly on the data from 'official administrative' sources. We are certainly aware of the growing worldwide interest in using new, private, big data sources such as Google search data, different commercial data vendors' data sets, mobile positioning data and news media."

Although big data was not included in its strategic plan for 2016–2021, a respondent from the Caribbean said it would move up on the list of priorities in the future and, as a result, the central bank has introduced training initiatives.

Which of the following types of data would you classify as big data?

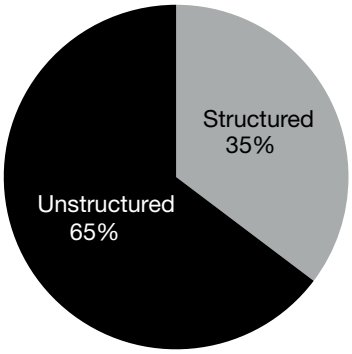


All respondents answered this question; most gave multiple answers

Which statement, in your view, represents the most accurate definition of big data?

While there is no single, agreed definition of big data,[1] industry has gravitated towards viewing it as large volumes of data, both structured and unstructured, which a standard desktop computer cannot handle. Central banks typically see big data as unstructured data that is sourced externally, though this view is not universally held.

Just under two-thirds of respondents believe big data should be defined as unstructured data sets. In a comment, a European central banker said: '"Unstructured' seems to be a more accurate definition of big data because at least it requires unstructured data processing, which is more challenging." A central banker from the Caribbean stressed the typical sources of big data as driving their view: "I perceive big data as data collected in the course of current business: for example, data automatically collected by social media, by scanning machines and by CCTV cameras. As such, it is in general unstructured, and it is the job of data scientists to move in and extract structured, pertinent information."



Forty-two central banks answered this question; eight of which checked both 'structured' and 'unstructured'.

Few of the dozen respondents who view big data as structured gave comments explaining their thinking. However, one central banker in Europe said: "Central banks are mainly concerned with the management of structured data collected from reporting agents. Unstructured sets of data from social networking websites do not play a prominent role for the time being."

Eight central banks said they recognised big data as both structured and unstructured. Big data lies in both categories, a European statistician suggested, saying the need to join together "many gigabytes" of tables in a few seconds is an example of structured data, while "textual information from the web" counts as unstructured.

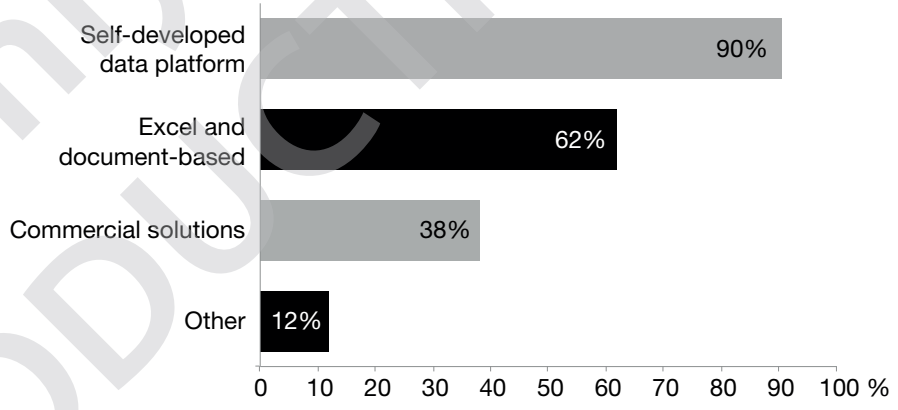
More than 80% of respondents said an example of big data was 'external market data', but only two respondents, both from Europe, checked this option exclusively. Nineteen of the 34 also chose 'regular data collected from firms in off-site activities'. A statistician from a central bank in Africa included all five classifications in their answer and added "data from ministries, departments and agencies". A European central bank ignored the suggested classification and offered the category of "commercial and administrative databases", as indicated in the United Nations Economic Commission for Europe list of classifications.

How does your central bank deal with regulatory data collection?

As central banks have expanded the depth and breadth of their involvement in financial stability policy-making since the financial crisis, so the importance of collecting accurate, complete and timely data from institutions has increased. Overwhelmingly, central banks develop their own data platforms to handle regulatory data collection. This was the view of 38 – or 90% – of the 42 respondents. A European statistician said the central bank has an "online portal" to support the collection of regulatory data, while a respondent from an advanced economy central bank commented that even though it does not handle the data collection itself, it nevertheless created the platform.

While own-built was the most popular response, only a minority said they use this solution exclusively. Half a dozen central banks from industrial economies and three from emerging markets also chose 'Excel and document-based handling'. A further nine central banks checked those two options and 'commercial solutions from the market'; four of these were African central banks. A respondent from an emerging market explained this was due to a multiplicity of sources and methods of handling data: "All are used due to historical differences in available technologies and regulatory processes." An African central bank listed the commercial solutions it employed: "ETL DataStage and ERP Oracle."

Commercial solutions from the market provide a viable alternative for central banks and were chosen by nearly 40% of respondents. These were typically used in conjunction with self-developed platforms, as was the case for three-quarters of this group. Ten central banks use all three of the given solutions, while one central bank in the Americas uses four, hiring "companies that develop solutions according to the specific needs of our institution." Two central banks in the Americas indicated that commercial solutions were their only method of regulatory data collection.

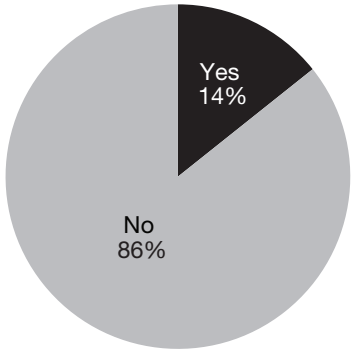


All respondents gave at least one answer.

A Caribbean officer explained: "The central bank is in the process of incorporating a commercial solution to replace our current system." A respondent from the Americas said commercial solutions cater directly to the requirements of the central bank: "We hire companies that develop solutions according to the specific needs of our institution." One central bank from Europe noted that, although it uses open-source software, it also employs Excel as a stopgap: "We use an open-source technology based on Hadoop and Hbase. Sometimes, in order to speed up the implementation process and to save costs, we use Excel as a temporary solution for data collection limited in scope."

Has this approach to data collection changed in the past 12 months?

Arrangements for regulatory data collection display a high degree of continuity. Thirty-six respondents said their approach to regulatory data collection was unchanged in the past 12 months. Few respondents volunteered a comment explaining their view, though a handful indicated a transition is under way. A statistician from the Caribbean said his bank's regulatory data collection approach



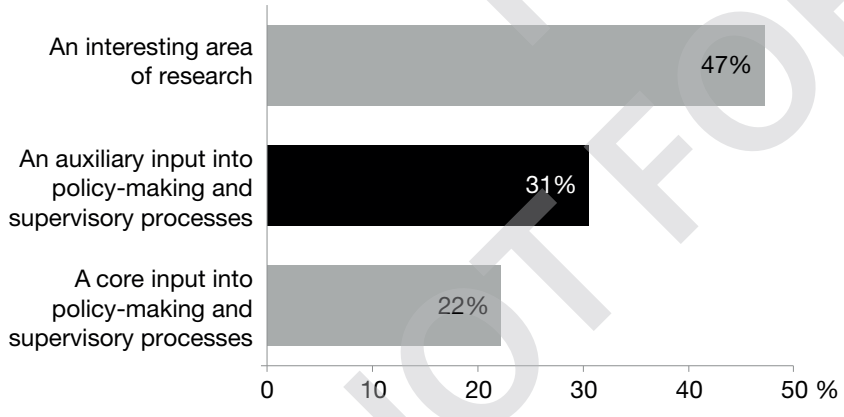
is expected to change over the next six months. Similarly, a respondent from the Americas said: "We are in the process of creating better foundations for policy decisions and have therefore changed strategies for managing data and statistics." One European central bank noted it is reviewing future plans, while another said it was implementing software for both the banking supervision and statistics departments, demonstrating the collaborative nature of big data projects. As a statistician explained: "We are in the process of implementing common reporting software at the central bank level for Banking Supervision and Statistics Department. This new project scheme was financed by the World Bank."

New software and big data techniques are also under development. A central bank with just over 500 staff is implementing a new data capture and analytics solution, and a respondent from the Americas said the central bank is looking into "scraping" data. Six central banks have changed how they deal with regulatory data collection in the past 12 months. This group was dominated by central banks based in Europe, with one adding that it was necessary to change its data collection process "for technological and business reasons".

Which best represents your central bank's view of big data?

Big data is predominantly regarded as useful for research in central banks, but a significant minority sees immediate involvement in policy-making, or scope for this. Nearly 50% of respondents chose 'an interesting area of research' as the best match for their central bank's view of big data. A central bank with fewer than 500 staff noted: "This topic is a potential area for research of future possibilities and the pros of using big data for statistical and analytical work." A European respondent echoed several others, saying: "The central bank's view of big data is going to change over time. Right now, it is an active area of research." A respondent from an advanced economy implied big data would influence policy: "The optimisation of data is of key importance in driving decision-making."

The second most popular choice was 'an auxiliary input'. One European central bank said it could see the potential "in the future" of big data as an



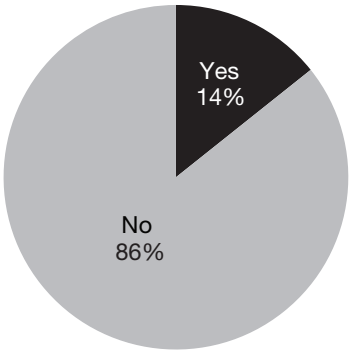
All respondents answered this question; six provided multiple answers.

auxiliary input in policy and supervision. A central bank from a developing economy said its statistics department manages a handful of micro-databases that are useful for cross-checking data: "The statistics department manages a set of micro-databases whose primary goal is to produce and disseminate high-quality statistics. In that sense this production chain benefits from a business intelligence architecture where data from different micro-databases (and sometimes different statistical domains) are intersected and validated. The use of microdata also brings flexibility to data management in a way that we can readily adjust and satisfy *ad hoc* requests, in some cases tailor-made to our customers' needs. Finally, the establishment of protocols with other institutions gives us access to external and complementary information to our own sources, which is one of the primary keys to ensure data quality."

Eight central banks indicated big data was a core input into policy-making and supervisory processes. This included the largest central bank that participated in this survey, as well as two respondents from the Middle East. One commented: "An efficient data management is obviously needed in the course of all missions for which the central bank (including the supervisory authority) is responsible. Big data techniques are useful in that respect."

Has your central bank's view changed in the past 12 months?

Views of big data's role in a central bank are largely unchanged. More than 85% of respondents said their view had not changed in the past 12 months. "We expect this to change in the medium term," observed a European central banker, who said the bank has dedicated a team to this work and therefore its view may change "as a consequence of our research conclusions".



The six central banks that have made a change in the past 12 months highlighted how big data was having a catalytic effect on their institution. A statistician from Europe said that data management is part of the central bank's new strategic plan. Another European central bank is advancing its work in this area, with dedicated teams within divisions outside of the statistics department: "There has been a change over the past 12–24 months that has seen the creation of dedicated analytics teams within supervisory divisions and a move towards creating a more data-centric organisation, which is reflected in organisational level objectives."

An officer from the Americas looked to the importance of an information strategy for the future development of big data: "To ensure that information gathered is handled in an appropriate manner, a vision for information supply is therefore required, along with an accompanying strategy that guides how data is required and processed." A European statistician commented: "Data management is part of the new strategic plan launched this year by the central bank."

Which, in your view, present the greatest roadblocks or challenges to increased use of data sets in your central bank?

Lack of support from policy-makers was most commonly identified as the greatest challenge and, intriguingly, also received the most votes as the least challenging. Twenty-eight per cent of respondents – seven Europeans, and four from the Americas – scored this as the most significant hindrance. "There are no problems with security, but there is not enough support," one European said. Conversely, 21, or 54% of, respondents saw a lack of support from policy-makers as the least significant challenge. This group was largely made up of developing and emerging market economies.

Concerns about skill sets figure prominently in central bankers' thinking. Just over half of respondents placed a lack of trained staff as the first or second most significant challenge. This group contained a significant number of central banks from advanced economies. A respondent from Europe noted: "Specific skills are needed for an efficient management of large data sets, both in IT and statistical departments." One respondent sounded a despondent note: "We expect additional human capital costs to be higher than acceptable compared to the possible benefits from using big data."

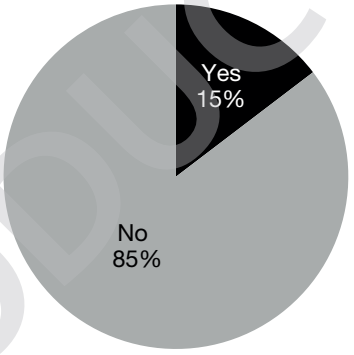
	1		2		3		4		5		Total	
	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%
Security risks	7	18	5	13	5	13	15	38	7	18	39	100
Confidentiality/ access concerns	6	15	8	21	13	33	8	21	4	10	39	100
Under-resourcing in IT systems	9	23	10	26	15	38	4	10	1	3	39	100
Lack of trained staff	6	15	14	36	4	10	9	23	6	15	39	100
Lack of support from policy-makers	11	28	2	5	2	5	3	8	21	54	39	100
Total	39	100	39	100	39	100	39	100	39	100	–	–

Votes were cast using a scale of 1–5, where 1 denotes the most significant roadblock, and 5 the least significant. Three respondents did not reply.

Just under half of respondents placed under-resourcing of IT systems in first or second place. A respondent who ranked this as their bank's greatest concern said: "The IT infrastructure is currently not sophisticated enough to support any use." A third of respondents ranked confidentiality and access concerns in first place. Security risks were typically ranked in fourth place. A European central banker offered another challenge: "The key initial issues were co-ordinating a unified organisational approach with the differing divisional needs and balancing implementing organisational change while still meeting mandatory regulatory requirements."

Does your central bank have a single allocated budget for the handling of data, including big data?

The vast majority – 85% of respondents – said their central bank did not have a single allocated budget for data. This group was largely made up of central bankers from advanced and developing economies. In comments, respondents typically attributed this to budgeting being divided on departmental or project-specific bases rather than by resource, function or output. In this way, data was often included as a component of the technology budget. One respondent commented: "The central bank has a general budget with a fixed share allocated to IT projects as a whole, which can only be partly related to handling of data but is not exclusive to this subject."



One respondent did not reply.

One central banker commented: "Several entities are involved in the handling of data." A statistician from Europe explained how data projects were referred to a central body for approval, but that a group had been created to look across these: "Data projects are handled as per all other projects (through an investment committee). A dedicated function has been put in place to handle data initiatives within the organisation to ensure there is a joined-up and consistent approach."

A budget for data was the preserve of only a handful of central banks. On the subject of agreeing a budget, one central bank from Asia commented: "The budget is formed once a year by every division separately." The allocated budget for one European central bank is used to cover the cost of the commercial solution: "The initial three-year software maintenance is included in the overall price of the contract. Additional charges in the software will be charged variably." One central bank in the Middle East confirmed they did have an allocated budget, but declined to give more detail.

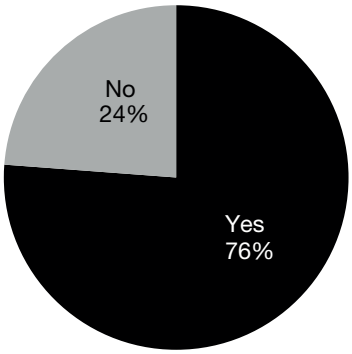
It is interesting to compare the results here with those from question 1, regarding active consideration. These six respondents were, naturally enough, among the 23 who said their central bank was actively considering big data. However, 17 central banks are giving active consideration but do not have a dedicated budget. At the policy-making level, of the eight respondents who said big data represented a core input into policy-making and supervisory processes, only one said they had a dedicated budget for data.

Does your central bank have a shared internal platform to enable different areas of the central bank to access data resources?

More than three-quarters of respondents indicated they had a shared internal platform, typically in the form of reporting frameworks and data warehouses. A respondent from a European central bank described the two-pronged approach

used there: "We have a shared internal platform for accessing most of our supervision data. For other macroeconomic data we have another platform contributed by the economic department and used by everyone inside the bank."

A respondent from a large central bank explained how it granted access to the data: "The first key element in the data strategy was to allow users with an appropriate business case to view all data relevant to them across the organisation. A technology solution was implemented to allow for this."

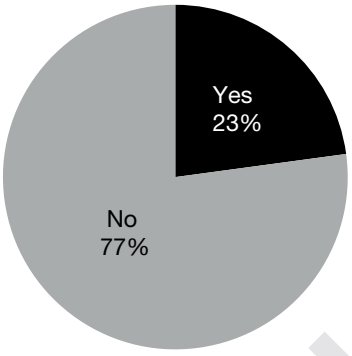


Ten respondents do not have a shared internal platform at their central bank. This group included emerging-market and developing economies from across the globe. Within this group, three central banks – one each from Europe, the Americas and Africa – are still working towards creating their own internal platform. One described the project as "a single reference platform (register) for the identification and the characteristics of financial and non-financial companies". Similarly, a central bank in the Americas is "in the process of developing a data warehouse to ensure data is stored in a standardised and secure way that is easily accessible to different internal data users."

Commercial solutions also provide an alternative to self-developed platforms. Four examples were given by an African central banker: "PIF, SPTR, Sigma and SAP." Respondents also made reference to reporting systems used for off-site regulatory reporting and an integrated reporting platform at a European central bank that is used for Finrep and MonStat.

If your central bank has a shared internal platform, can this platform be accessed externally (for researchers or dissemination purposes)?

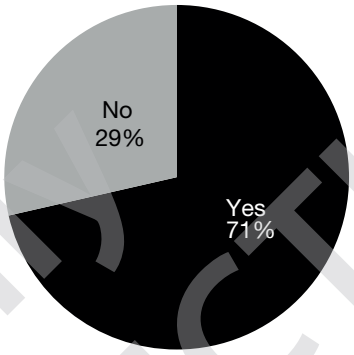
Typically, data-sharing platforms in central banks are built internally by the central bank and are not available for external use.



Seven respondents did not reply.

Of the eight central banks that allow external access to their self-developed platforms, five are European central banks, one is from the Americas, one from the Middle East and one from Africa. A European central banker said: "The data software is externally accessed by banking supervisors when examining the financial institutions. This platform is restricted for researchers and dissemination purposes." Of the eight central banks that allow external access, six built their own platform independently.

Was this shared platform built by the bank?



Seven respondents did not reply.

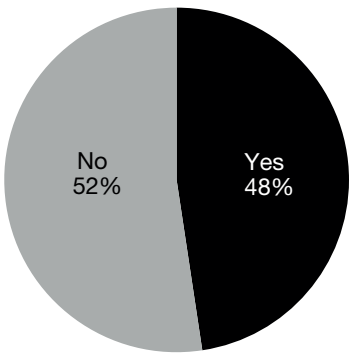
Several respondents from developed countries, however, said they were establishing platforms that would be available for external use. One central banker said "external access is still under development", while another commented that some parts of the database were being made available to registered users.

An independent platform has been launched and provides anonymised data to external researchers, on a need-to-know basis: only data sets deemed useful for a research project will be provided on this platform.

In the main, central banks build rather than buy shared platforms: this was the experience of just over 70% of respondents. One European central bank said it used "in-house developments with standard big data infrastructures". Ten central banks indicated they did not build their own platforms. Their reasons ranged from seeking assistance from another central bank in the development to subcontracting the process. A central banker in the Middle East said the institution had been "assisted by AbacusConsulting (Middle East) LLC".

Does your central bank have clear data governance with defined roles and responsibilities, such as chief data officer and data stewards?

For many central bankers, data management arrangements are a concern, and just over half of respondents said their banks did not have clear data governance. European central banks featured prominently in this group, as did a significant number of respondents from the Americas. One said: "We have no chief data officer or equivalent."



This is clearly an area of intense activity, however, as central banks are striving to improve data governance frameworks. A central banker from the Caribbean said the framework is "in its infancy" but is "expected to mature" in the coming years. An officer from a central bank in Africa commented that a clear data governance structure "is now being put in place".

Several European central bankers said data governance is being discussed, but it had not been formalised, a view typified by this statistician's comment: "Certain relevant people are more aware of their roles and responsibilities as data owners or data stewards, but it has not been fully formalised and implemented yet."

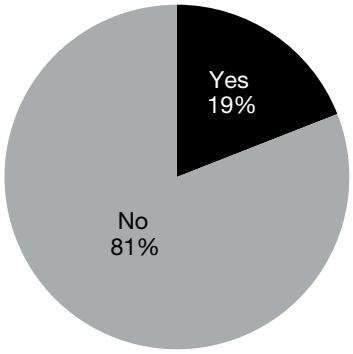
Another European respondent set out the structure of their bank's data governance: "Within the central bank, we don't have the mentioned type of personnel. However, we have a sort of data governance, defined by the security policy." A developed-country central banker noted: "We have a chief data officer and data stewards identified in place; however, more robust governance is being put in place over the coming months."

Twenty respondents were more confident in the arrangements for data governance. One European central banker commented it has "data owners", who can grant others the access to the data.

A statistician from the Americas said: "For each information-specialised area, there is a work team." A central bank with more than 5,000 staff has committed a whole team to the supervision of big data: "Through the project, we have built a specific organisation in charge of security policy and the high-level supervision. A dedicated committee chaired by the director of general statistics including high-level representatives of all business areas is responsible for the high-level monitoring of the platform."

Does your central bank have intra-departmental or divisional bodies, such as committees or working groups, dedicated to big data?

Big data is generally confined to departments or divisions in central banks. Just over 80% of respondents said they do not have any intra-departmental or divisional bodies dedicated to big data. This group was composed of five African central banks and all nine central banks from the Americas, along with two central banks from the Middle East.



The following are comments made by the 19% of central banks that have intra-departmental or divisional bodies centred around bank-wide data teams and committees dedicated to big data. A respondent from a European central bank with more than 2,000 staff said: "We have committees dedicated to data, which include big data." Another central bank commented: "This year we have started an inter-departmental team dedicated to big data issues", while another made reference to big data being discussed within a team of people focused on data in general.

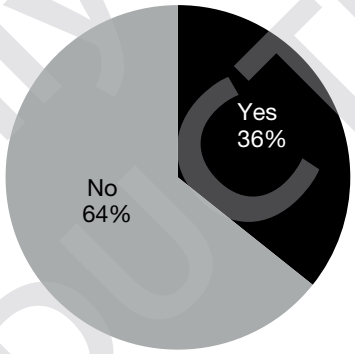
A central banker from an advanced economy said big data plays a part throughout the central bank; however, it is governed by the statistics department director: "Many projects deal with big data issues: the corresponding steering committees are temporarily chaired by the director of general statistics."

Although 34 central banks indicated they do not have an intra-departmental or divisional body at their central banks, three said it was a field they are looking to develop. One such central bank in the Americas commented: "We have some employees working on big data."

A respondent from a central bank with more than 1,000 members of staff

explained how their institution was nevertheless bringing people together: "Even though we do not have an intra-departmental or divisional body, we do have the big data forum with a multidisciplinary team involved in roundtables on a temporary basis."

Does your central bank use external data providers for big data sets?

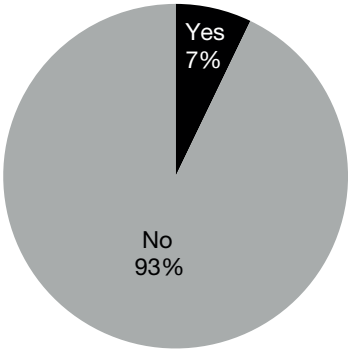


Central banks generally source their own big data sets, though a significant minority look elsewhere. Respondents that do not use external data providers were predominantly from emerging-market economies, including 17 central banks from Europe.

Those that used external providers tended to turn to commercial banks and firms, social media and Google blogs, and mobile phone operators. "We employ data from blogs, social media and private websites," said one respondent. A statistician from Europe commented on the techniques used to collect the data from websites: "Currently, we are doing web-scraping projects, collecting data from different websites."

Does your central bank outsource any data processing for big data sets?

Overwhelmingly, central banks process their own big data sets, and there is no indication of a desire for this to change. More than 90% of respondents said they do not outsource any data processing.



A central bank in the Americas said, straightforwardly: "Data processing is run in-house", while one central bank in Europe commented: "The data management and statistical analysis of big data sets is taken on by highly qualified staff in the directorate general of statistics with the support of the dedicated team in the IT department." Three central banks outsource data processing for big data sets; however, they all declined to comment.

Of the 39 central banks that did not outsource data processing, only three noted this was likely to change in the near future. One central bank in the Americas is currently hiring in this sector to cater for in-house data processing: "The central bank is now recruiting an expert in the field of data architecture in order to develop intra-departmental systems and working routines of data."

In your view, which of the following areas stands to benefit most from big data in practical terms?

Respondents saw monetary policy as likely to benefit most from big data, though they said it will have a significant impact on macro-prudential policy as well. As a central banker from an industrial economy explained: "Monetary policy, macro-prudential and micro-prudential policies can benefit from big data. Monetary policy can benefit from better and timelier nowcasts of macroeconomic variables. Macro- and micro-prudential policies might benefit as well."

	1		2		3		Total	
	No.	%	No.	%	No.	%	No.	%
Monetary policy	18	45	9	23	13	33	40	100
Macro-prudential policy	12	30	24	60	4	10	40	100
Micro-prudential policy	10	25	7	18	23	58	40	100
Total	40	100	40	100	40	100	–	–

Votes were cast using a scale of 1–3, where 1 denotes the most likely to benefit, and 3 the least likely. Two respondents did not reply.

Interestingly, 90% ranked macro-prudential policy in either first or second place. Those that ranked it in first place typically chose monetary policy as their second choice. Micro-prudential policy was ranked third by 58% of respondents but 10 respondents placed it first. One noted: "Big data applications shine at the most detailed level."

Which area do you consider the priority for investment to increase big data use in your central bank?

Support from the executive level and policy-makers again divided respondents: 35% saw it as a top priority, 38% saw it as the lowest. Those that considered it a top priority were mostly central banks from Europe and the Americas. A European officer stressed the importance of having a budget: "Obtaining a budget is of fundamental importance in an initiative, so therefore is a key priority."

Conversely, many saw executive-level support as a minor challenge. A statistician from a small central bank in the Americas explained the focus there: "Big data also poses considerable challenges for the central bank, both technically and methodologically. In addition, new considerations need to be made in terms of strategy, organisation, skills, budgets and risks."

Sixty per cent of respondents ranked trained staff in either first or second place. This group included half a dozen emerging markets. However, these central banks declined to comment. Resourcing IT systems was most commonly ranked

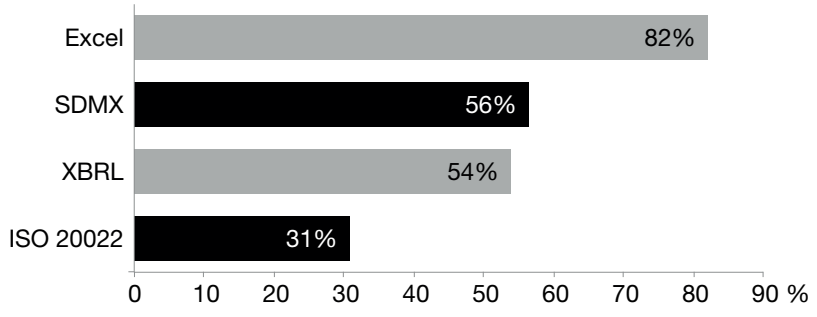
in third place. One European central bank commented: "Ensuring that IT infrastructure can handle additional demands is a key component of where this investment will be required." Network for internal sharing was largely ranked in fourth and fifth place by respondents, as was security.

	1		2		3		4		5		Total	
	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%
Security	2	5	3	8	6	16	17	46	9	24	37	100
Network for internal sharing	7	19	4	11	5	14	10	27	11	30	37	100
Resourcing IT systems	8	22	11	30	15	41	2	5	1	3	37	100
Trained staff	7	19	15	41	6	16	7	19	2	5	37	100
Executive-level support	13	35	4	11	5	14	1	3	14	38	37	100
Total	37	100	37	100	37	100	37	100	37	100	–	–

Votes were cast using a scale of 1–5, where 1 denotes the top priority, and 5 the lowest. Five respondents did not reply.

Which of the following standards does your central bank use, or plan to use, for dealing with data exchange and collections?

Excel is the most popular standard for central banks when dealing with data exchange and collection, but it is typically used in conjunction with another solution. Of the 39 respondents, 82% use Excel, of which around 40% use it exclusively. Of the remainder, most used either SDMX or XBRL as well, and three-quarters use both. "The central bank collects data from financial institutions using the XML format combined with Excel format", noted one statistician from a developing economy. An officer from a developed country commented: "XML



Three respondents did not reply. Respondents checked multiple answers.

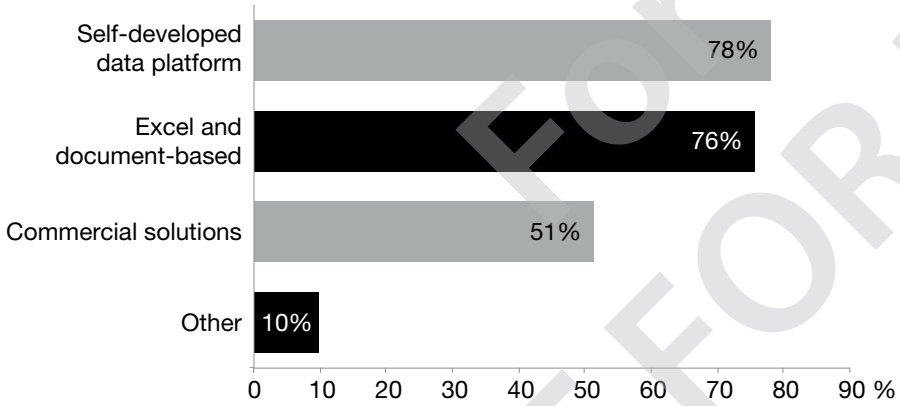
with structured data prescribed by the central bank is also used for data collection and data exchange."

SDMX was the second most popular standard choice for the majority of developed, European respondents. One central banker commented: "SDMX is used for data exchange of statistical data, primarily in data exchange applications with the European Central Bank." The statistics department of an African central bank is working towards implementing this standard: "SDMX is an ongoing project."

The standards of XBRL and ISO 20022 are typically used for specific functions: supervision and payments, respectively. Twenty-one respondents indicated they use SDMX at their central bank. A central bank from an advanced economy noted: "XBRL is only used in the transactional system component of Register of Institutions and Affiliates Database application." Although the least popular standard, ISO 20022 was also described by respondents as useful for specific functions. One European officer noted: "Currently we are using the three standards mentioned: SDMX and XBRL are widely used, whereas ISO 20022 is only being used in very specific exchanges." An African central bank said: "ISO 20022 is being considered for payments."

How does your central bank deal with regulatory data analytics?

Central banks have developed their own data platforms to deal with regulatory data analytics, but these are often used in conjunction with Excel and document-based handling. This was the answer from more than three-quarters of respondents, the majority of which were from developed countries. Only four of these 32 central banks use purely self-developed regulatory data analytics, while the remaining 28 combine it with another standard. Nearly all combined this with Excel and document-based handling, and just under half reported a commercial element to their regulatory data analytics workflow.

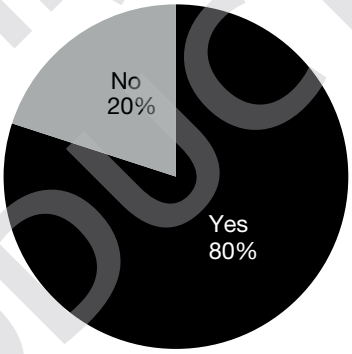


Respondents were invited to provide multiple answers to this question. One respondent did not reply.

Commercial solutions from the market proved popular among respondents, with over half choosing this option. Interestingly, more central banks use commercial solutions from the market for regulatory data analytics than they do for regulatory data collection, with only 16 central banks.

As of January 2016, the global systemically important banks (G-Sibs) have been regulated to meet the BCBS 239 principles, after carrying out self-assessments regarding data collection, data aggregation and dissemination capabilities. Would it be useful for central banks to self-assess using an adapted version of these principles?

Central bankers largely welcome the idea of self-assessment using an adapted version of the BCBS 239 principles for G-Sibs.[2] Of the 35 respondents, 80% said such self-assessment would be useful.



Seven respondents did not reply.

As databases increase in size and complexity, there is naturally a concern that they are policed and governed properly. Furthermore, one central banker noted existing systems were under pressure "due to increasingly large and complex data that is now challenging traditional database systems".

Several central banks drew attention to the internal processes already in place to regulate big data processing. One European respondent from an advanced economy noted: "There is a benefit that can be accrued from this. It is important to note, however, that our internal audit function already undertakes full audits in this area." Another European respondent said BCBS 239 was handled by the supervision department: "However, if the objective of these principles is to strengthen banks' risk data aggregation capabilities and internal risk reporting practices, it might be relevant to make a similar self-assessment in central banks."

Conversely, seven central banks disagreed that it would be useful for central banks to self-assess based on an adapted set of principles. All were European except one from the Middle East. The BCBS 239 principles are not applicable to data collection, a European central banker noted: "BCBS 239 addresses issues pertaining to risk data aggregation and reporting capabilities for banks so as to achieve full compliance with regulatory expectations. The principles thereof focus on a specific use of data and do not take into account other dimensions of data valuation (for example, statistical analysis for other needs than those underlying the collection of these data)." ❑

Notes

1. For examples of different definitions, please see David Bholat, *Big data and central banks,* Bank of England Quarterly Bulletin (2015 Q1) *http://tinyurl.com/zz22mo5*; and Daniel Hinge, *The big data revolution and central banking,* Central Banking Journal (November 2015).
2. The BCBS 239 principles were created by the Basel Committee on Banking Supervision and put in place in January 2016. They consist of 14 principles for supervisors to follow when considering aggregation in the banks they supervise.

Data as a critical factor for central banks

Central banks must industrialise data-handling processes to deliver better decision-making, says Maciej Piechocki of BearingPoint.

Data remains a critical factor for central banks. The financial crisis revealed that some of the deepest fissures were caused by gaps in data and exposed the need for high-quality comparable and timely data on the global financial network. Since then, policy-makers, supervisory authorities and standard-setters across the globe have been collaborating to better harmonise and standardise regulatory data in financial services. According to a recent BearingPoint Institute paper, urgent debate is still needed on how the world's financial services industry could be better and less onerously supervised via a smarter approach to regulatory reporting and data exchange.[1]

Financial supervision and central banks' momentary statistics and financial stability functions are vastly driven by data. In the aftermath of the financial crisis, a 'regulatory tsunami' flooded the financial services industry. Particularly after the adoption of the Basel III framework, regulatory requirements have significantly increased and new regulations such as AnaCredit, the Basel Committee on Banking Supervision (BCBS) 239, Solvency II, Dodd–Frank and International Financial Reporting Standard (IFRS) 9 have posed new challenges to the banking and insurance sector on global, regional and local levels. Moreover, regulations such as the European Market Infrastructure Regulation (Emir), money market statistical reporting, the Markets in Financial Instruments Regulation (Mifir) and the Securities Financing Transaction Regulation (SFTR) oblige the major monetary financial institutions to report derivatives or money market data on a daily basis.

However, in the central banking area, while no single agreed definition exists, big data has already been heralded as offering a wide range of central banking applications: from 'nowcasting' to modelling, to early warning systems and systemic risk indicators. For some it opens a new chapter in policy-making. A number of central banks are currently rethinking their data infrastructures, which today are rather siloed, and demonstrating the legacy of the past decades with no central approach to data handling.

Notwithstanding the huge potential of big data, decision-making is now even harder than before, and businesses need adequate solutions to analyse this data.[2] A crucial point is how to mine all this information from the different sources

Fotolia/chombosan

exhaustively and at reasonable cost. Despite innovative tools and technologies such as blockchain, cloud computing and machine learning, even today plans often fail because the required processing power outweighs the potential returns or computing time is too long.[3]

The specific challenge for central banks in the sense of an effective 360-degree risk-based supervision is to rapidly access, effectively manage and process and analyse in a timely manner the increasing amounts of supervisory, statistical and markets (big) data. The near- or real-time access and efficient processing are especially regarded as critical factors due to limitations in human and IT resources.[4] According to a report by the Institute of International Finance (IIF), some regulators still use outdated portal solutions and methods that are inefficient and increase chances of introducing error.[5] The IIF recommends automated secure data transfer mechanisms based on standards such as XBRL. But even with the use of such standards as XBRL or SDMX, central banks must abandon a paper- or document-oriented world and think of data in an integrated and interlinked way.

Current systems do not meet today's requirements when regulators have to deal with large amounts of data of various kinds – collected from supervised entities for statistical, prudential or stability purposes, provided by information providers or obtained from internal research and analysis. Such data ranges from granular micro information on single mortgage loans, securities traded and counterparties affected, to macroeconomic analysis of countries or regions and form-based collections of financial and risk data or *ad hoc* supervisory exercises.

Some of this data will remain only within the perimeter of the central bank, while some will be remitted to other stakeholders such as the European supervisory authorities, national governments, the International Monetary Fund and the Bank for International Settlements, and some will be disseminated to the wider public or research community.

Therefore, it is mission-critical for regulators to:

- Effectively handle the large amounts of increasingly granular data from various sources: that is, rethink existing IT system architectures and landscapes.
- Gain transparency on the status of the reporting agents in the collection and dissemination process.
- Consider interlinkages between micro and macro data sets in 'going beyond the aggregates' from macro and financial stability perspectives.
- Have a timely overview of relevant micro and macro developments in the financial markets.
- Execute reliable trend analyses on key performance indicators and key risk indicators based on validated collected data.

In view of the developments described above, it is undisputable that it is mission-critical for central banks to reshape their data management and further automate and industrialise processes of handling data. Automation helps to minimise risk, reduce errors, increase transparency and thereby deliver a better basis for decision-making.

According to a BearingPoint Institute article,[6] a new information value chain is needed for reporting that helps to increase the efficiency of supervisory processes, minimise risk, allocate resources effectively and improve the basis for decision-making by higher transparency and faster availability of data. We further notice a trend to shared utilities – a kind of 'regulatory-as-a-service'.

A prominent example is the Austrian solution, where the National Bank of Austria and the supervised banks joined forces to stepwise replace the template-driven model and use innovative technologies to create a new regulatory value chain. The initiative is based on greater harmonisation and integration of data within banks as well as greater integration of the IT systems of the supervisory authority and the supervised entities. It works through a common data model developed by the central bank in co-operation with Austrian banks and a shared utility, Austrian Reporting Services, which is co-owned by the largest Austrian banking groups. This model allows cost-sharing of compliance as well as standardisation of data collection.

To summarise, it is clear that data is and will remain a critical factor for central banks across the globe. But first, combining data with the right people, technology, processes – and also collaboration models – will allow central banks to leverage it for their missions and objectives. ❑

Notes

1. Maciej Piechocki and Tim Dabringhausen, *Reforming regulatory reporting – from templates to cubes* (Bank of International Settlements, 2016) *http://www.bis.org/ifc/publ/ifcb41o.pdf*
2. BearingPoint Institute Issue 002, *Seeing beyond the big (data) picture*, pp. 3–4.
3. *Ibid*., p. 6.
4. Irving Fisher Committee, *Central banks' use of and interest in 'big data'* (Bank of International Settlements, October 2015), p. 11 *http://www.bis.org/ifc/publ/ifc-report-bigdata.pdf*
5. Institute of International Finance, *RegTech in financial services: technology solutions for compliance and reporting*, (March 2016), pp. 22–23 *http://tinyurl.com/jaa9c8n*
6. BearingPoint Institute, *Reforming regulatory reporting: are we headed toward real-time?* (2015).

Evaluating big data capabilities in central banking

Central Banking convened a panel of experts to discuss how big data can add value to central banking in the near term, the need for good governance, and balancing the pros against the cons.

The Panel

David Bholat
Senior Analyst, Advanced Analytics Division, Bank of England

Maciej Piechocki
Financial Services Partner, BearingPoint

Iman van Lelyveld
Statistics Division, The Netherlands Bank and VU Amsterdam

Many observers believe big data has the potential to open up new possibilities for monetary policy-making, financial supervision, and economic research. Nowcasting, text mining, machine learning and other new techniques have been made possible by improvements in processing technology and new, larger, more granular data sets. Nevertheless, questions remain over how central banks make best use of the new methods.

This forum draws on the viewpoints of three experts who discuss financial stability and supervisory applications, direct uses in economics and modelling, who should 'own' big data, e-sourcing and budgets, future developments, as well as the operational challenges of gathering, structuring, storing and processing data.

Central Banking: What does big data mean to you?

David Bholat, Bank of England: Defining big data, you could do worse than follow the standard schema of the 'three Vs': volume, velocity and variety.

From left: David Bholat, Maciej Piechocki and Iman van Lelyveld

For a central bank like the Bank of England (BoE), regarding data with more volume, it's about complementing our traditional focus on macroeconomic time series and data extracted from financial statements with having more granular financial transaction data.

Data that has greater velocity, such as Twitter data, complements the data that flows into the central bank traditionally on a slower cycle. Finally, for variety we're talking not only about structured numerical data, but also unstructured data, like text.

Iman van Lelyveld, The Netherlands Bank and VU Amsterdam: I would add that the data is large but that means it's approaching the population, which is different from the traditional statistics where you would sample part of the population. So you have to look at different analytical tools.

Maciej Piechocki, BearingPoint: What was mentioned about the three different aspects of big data is really interesting. I believe there are also different maturities across central banks – big data is within the industry and serving not only the central banks but also the wider financial and non-financial services industries.

Central Banking: Some within central banks, IT departments especially, view big data as "just another fad". Do you view big data as a fad?

Iman van Lelyveld: No, it is here to stay. What might be faddish is the claim that it gives you the tools to analyse everything – in the sense that you now know everything, therefore you can analyse everything. That's pushing it because what you can't see are things that people want to reveal. Big data, as in totally unstructured and scraped from somewhere, has a limit.

Maciej Piechocki: I have a different view because, for many IT staff at

Maciej Piechocki Financial Services Partner, BearingPoint
For the past 10 years, Maciej Piechocki has specialised in the regulatory value chain, designing and implementing digital solutions for regulatory clients, including central banks and supervisory authorities. He also supports supervised entities in implementing regulatory mandates, particularly regulatory reporting requirements. Piechocki's main focus is on analytics, XBRL and oversight, and he has been particularly involved in CRD IV, Solvency II and IFRS. A member of the IFRS Taxonomy Consultative Group of the IASB and a board member of XBRL Germany, Piechocki has written several books, including XBRL for Interactive Data: Engineering the Information Value Chain. *He is a frequent contributor to independent journals on the subject of regulatory reporting.*

central banks, it's actually a great topic. There are several completely new technologies that some of the more innovative IT personnel at central banks love to work with.

I see a danger there as well because big data shouldn't be treated solely from an IT perspective – it's extremely important to analyse it from the use-case perspective. Big data has to be user-driven and not just what we were getting before – a huge warehousing project where data was merely stored but not really used.

David Bholat: At one level it is a fad for some of the nomenclatures, while at another level, it is not – the tools are here to stay. What's quite hot now is 'data-driven analysis', though I'm not entirely sure how that's different from what we used to call 'evidence-based policy-making'.

Central Banking: Looking at some of the main applications, there's a lot of activity going on around financial stability and supervisory. What do you see as the main applications for big data?

Maciej Piechocki: As far as supervisory statistical stability applications are concerned, the big drivers are for derivatives data; so if you look at Dodd–Frank and European Market Infrastructure Regulation (Emir) requirements, they are bringing great granularity and the possibility of definitely going beyond the aggregates. Also, for continental Europe, there is a commotion about loan-level data. The loan-by-loan view of the borrowers and, on the whole, markets and flows are there, which is interesting if you start combining this with the aggregates the central banks are holding. It allows much more flexibility in finding out where the risks are really sitting without the industry actually incurring the burden.

Iman van Lelyveld: The key is the ability to link up data sets. For large data sets, this starts in payment systems where this kind of data is readily available, and you can match up the information from the payment systems and interbank lending with supervisory data, for instance.

Central Banking: What about the direct usage for big data when it comes to the economic side, the modelling, and so on?

Maciej Piechocki

David Bholat: Big data is not only applicable in the financial stability domain, but also when performing real-economy analysis. For example, in the UK labour market, we have been pleasantly surprised that, coming out of the financial crisis, the employment numbers have been quite strong. At the same time, there are some worries about automation increasing unemployment, and we are using vacancies data from websites to understand this.

Maciej Piechocki: Returning to the issue of volume and velocity, that's a good use case for the supervision of financial stability, where central banks have a good grip on structured data. Unstructured data is coming along extremely well. Having very different data sources, you can pull the most from the unstructured big data and support the structured data sets.

Central Banking: How are some of the governance rules developing with regard to big data? How are people managing all this data in terms of governance?

Maciej Piechocki: For many central banks, it is a matter of staffing and skill sets, and there is a huge demand for data scientists, for example, everywhere in the industry. Another important topic is governance within central banks. There are a few central banks that have a designated chief data officer (CDO), while others are looking at the statistical function to fill this role, so it's a question of who has ownership of big data within central banks.

Iman van Lelyveld: It's very important to keep the ownership of the data within the business. As soon as it becomes a statistics or – even worse – an IT thing, there is a distance between the users and the definition of the data, solely between the way it's stored and the way it's accessed.

David Bholat: Many outside observers of central banks focus mostly on policy decisions, our forecasts, and so on, and don't actually realise that, at very senior levels, data is discussed. It's taken very seriously, and increasingly so.

Central Banking: Some central banks have set up data officers, some focus on the tech side, while others are looking at it from the statistics side. Is there a correct approach or a desirable approach people can take?

Maciej Piechocki: It doesn't really matter as long as it's a user-driven function. The danger with having too much tech is that it will become a purpose in itself, which it shouldn't be. It doesn't really matter as long as there is board-level ownership of the big data, which does happen at some central banks. If it is

David Bholat Senior Analyst, Advanced Analytics Division, Bank of England

David Bholat leads a team of data scientists and researchers in Advanced Analytics, a big data division in the Bank of England that he helped establish in 2014. The division is recognised as a leader among central banks in the area of big data. A former Fulbright fellow, Bholat graduated from Georgetown University's School of Foreign Service with highest honours. He subsequently studied at the London School of Economics, the University of Chicago and the London Business School. Publications in 2016 include Modelling metadata in central banks; Non-performing loans: Regulatory and accounting treatments of assets; Peer-to-peer lending and financial innovation in the United Kingdom; *and* Accounting in central banks.

driven by statistics, stability or supervision depends on the central banking governance, with strong support from the technology departments.

Iman van Lelyveld: It depends to some degree on which part of the data, and whether it is really something operational. For instance, we look at how banknotes are checked, a very process-driven, huge-volume data set. It's easy because nobody else needs this data, whereas for other data sets there might be multiple uses – and then it is much more important to perhaps have a management layer that co-ordinates these requests for the same data from different parts of the banks.

David Bholat: In terms of data, the concept of 'ownership' can be quite unhelpful because it leads to silo thinking. The real push, at least inside the BoE, is to be very 'one bank' and cross-organisational – we're trying to set up a kind of ecosystem. The BoE owns the data. We have three divisions spearheading different efforts: our Statistics and Regulatory Data Division, which collects much of the regulatory report data; the Chief Data Officer Division, which puts in place the IT infrastructure to facilitate sharing; and Advanced Analytics, into which we're bringing some data science machine-learning techniques.

Iman van Lelyveld: This is an important point – try to make everyone the owner of this data and make this kind of data as widely available as possible because that's the only way to unlock the potential of combining different pieces of information in ways you haven't thought about before. If you are in a silo organisation, it's the death of data since you're not using it. And, if you don't use it, it will become polluted, you can't use it any more and it will die.

David Bholat: A key initiative that our CDO division has been spearheading is the creation of a data inventory, a central register of all of the data sets – whether proprietary or purchased or open-source public – that are actually in the building. Previously, there would be someone working on financial stability issues who might have access to an interesting data set that their colleague – for example, an economist working on monetary analysis – didn't even know existed in the building. If you have a central register, you can share more effectively.

David Bholat

Central Banking: Are there any special considerations regarding some of the confidential data? In certain functions, there's information that is given for those purposes that obviously has great value in other parts of the central bank but, perhaps for confidentiality or even legal reasons, cannot be shared or needs to be shared in a way that doesn't pinpoint where it's coming from.

Maciej Piechocki: Certainly there are consequences, especially for central banks with monetary as well as supervisory functions. There have been conversations about how governance should be organised around data sharing between these functions to ensure they are sufficiently separated.

Also, if you look at the large volumes of data, they're often related to more granular contract data, down to the personal-level data. There has been much discussion around loan-by-loan collections, and the extent to which central banks are allowed to reach this level of information.

Iman van Lelyveld: These are important issues, but most of this personal data is already shared somewhere, and we need to sort them out in a proper way. What you would like to do from an analytic point of view, for loan-level data or mortgages, is also to have some information on the tax situation.

Central Banking: Now that central banks have access to all this data from trading and clearing platforms, are all the reporting requirements on major market participants necessary? They are costly and could put large firms at a disadvantage.

Iman van Lelyveld: If you look at most of this data, specifically the trade repository data, it's not up to scratch. We cannot replace other surveys or reports with this data just yet. Naturally, if that was possible, I'd be all for it. So we need to flesh out the areas where other reports have the same bits of information and see if we can replace that with trade repository data.

Maciej Piechocki: The first central bank to trial this approach is in Austria. It's an interesting example, where instead of collecting the aggregated data, which is quite a costly burden, they're trying to obtain the contract-level data, and are transforming the whole regulatory value chain. As for logistics becoming much more granular, I can well imagine in the future that plugging into the trading systems even earlier could lead to a situation in which you could remove the whole value chain and the classical current one by pulling the information that a central bank needs.

Iman van Lelyveld Statistics Division, The Netherlands Bank and VU Amsterdam

Iman van Lelyveld is a senior policy adviser with The Netherlands Bank (DNB) and professor of banking and financial markets at the finance group of VU Amsterdam. At DNB, he has been involved in many regulatory policy issues including interest rate risk in the banking book, deposit guarantee pricing and capital valuation charges. Van Lelyveld is a member of the BCBS Research Task Force and has chaired several international groups, most recently on liquidity stress testing. He has worked for Deutsche Bank, the Bank of England and the International Data Hub at the Bank for International Settlements. Van Lelyveld holds a PhD from Radboud University.

Central Banking: People have been working on it for some time, so why are the trade repositories not up to scratch?

Iman van Lelyveld: In any data set that you build up, you need some time to get the quality up. With a trade repository it's such a huge set and the governance is so dispersed that it's difficult to get a feedback loop between the users of the data and the people submitting the data. As you're trying to get a very wide coverage, relatively small non-financial firms need to report, but they're not matched up to the regular reporting frameworks.

Central Banking: If there are so many challenges in getting the structured data right, how can one hope to get all the unstructured data?

David Bholat: In a sense, the structured/unstructured divide is artificial. This is what makes the Austrian central bank approach quite innovative – when you go back to what a financial instrument or a product is, it's a contract, something unstructured. Then, as it proceeds through different parts of an individual firm and that data is disseminated to regulators and the regulators publish the aggregate data, it becomes structured. We want to get more granular because then that reduces regulatory reporting costs in the long term.

Central Banking: Databases are relatively easy for IT departments to centralise, but what about derived data series and data produced and modified by analysts on the business side? Do good enough tools exist to achieve this without the involvement of IT departments?

Maciej Piechocki: This is a very important topic because, if you are aiming at flexible working with data, you are automatically switching to self-service-oriented aspects of the business that can dive into data.

We are beginning to see the flexibility of systems that can handle this, but the aim is also to provide self-service for the business departments, not only to construct new analyses, but to provide aggregated reports with the possibility of drill-down and to broadcast them to other departments across the organisational boundaries.

Iman van Lelyveld

Iman van Lelyveld: It puts the burden on the analyst to create an analysis suite in a more structured way. It will do more that just load things onto an Excel spreadsheet, build a series of graphs and say: "here's my analysis". Over time, we will move to coding it up in some way and having a paper trail of analysis. Then, with some sort of centralised IT function that keeps the source data – a 'golden copy' – you can trace back your analysis.

David Bholat: Spreadsheets have traditionally been a workhorse, but with the problem of that when changes are made you don't know who made them, at what time and in what direction. What is valuable about some of these coding technologies is that you can actually trace how that data was brought into the data frame.

The principle is to have one golden copy of data, not multiple versions of the data in different spreadsheets. With a golden source of data you have a coding interface on top that is traceable and auditable in terms of what kind of analysis was done.

Maciej Piechocki: A single data dictionary is a great idea, with all terms defined across the central bank. But this goes against the object of velocity because the time taken to create this dictionary, or to enter something new into it, is time that you are losing from working with the data.

Iman van Lelyveld: I disagree, I don't think it is time lost. For instance, what is an entity? Are we looking at a consolidated entity or a sub-consolidated entity? You need a legal entity identifier, some sort of unique identifier of the bits you need to put together. Generally, this has not been given much thought, so people make their own constellation of a series of legal entity identifiers and then, two months later, you have to do the same analysis.

Central Banking: What are some of the security issues related to big data, whether it's cyber security, physical security, confidentiality when using external providers, and so on?

David Bholat: When we went through our strategic review and Mark Carney took over as governor, one of the outputs was to set up an Information Security Division. They use some machine-learning approaches, and information security has ratcheted up the agenda because I'm constantly doing compliance exercises.

Iman van Lelyveld: I'm in the same position. Security is being ratcheted up, and I certainly can't share confidential data outside of our systems, so any kind of innovation needs to be a local installation; I cannot go through the internet, for example.

Maciej Piechocki: As a solutions service provider to central banks, over the last decade I have found security has improved vastly, and the requirements of providers in terms of certifications and security screening are definitely growing.

Central Banking: How can central banks attract the correct sort of talent to work in this area – a combination of economists, statisticians and computer scientists? They're still not that common, so how do central banks recruit the people they need? They're competing against some of the technology pioneers, so it can't be easy.

David Bholat: Where we can compete, I think, is if you're intellectually interested in working on really tough problems. Often, what we find when we're recruiting tends to be people who are just coming out of university, who are interested in trying to address these big-picture problems and also have a sense of service. We exist at the BoE to serve the people of the UK, to promote their common good and that's the kind of service ethos that attracts a lot of people to us.

Iman van Lelyveld: It's the same pitch we make, but we also offer something regarding the balance of work. We have interesting data and questions, but also space to actually investigate. There are important things to come up with solutions for, which for inquisitive people straight from university is fantastic.

Maciej Piechocki: That's interesting because we are aiming for exactly the same profiles of the market and industry. I've seen approaches from several central banks to sponsor universities or high schools. Also, economics is losing some competitive edge and we are aiming more at mathematicians, those with majors in physics and natural sciences, because many already have a good grip on data.

Iman van Lelyveld: It's important to give them time to learn the language. I've been working with theoretical physicists, and they know a lot more about the structure in data, etc., but we need to learn to talk to each other. For instance, having algorithmic learning without actually knowing how the rules work doesn't generally find anything. We need a combination of the two, at least in the same analysis team.

Central Banking: There is the whole aspect of causal reasons for correlations – if an online retailer sees a correlation between selling two different items, they'll flag the item and ask you if you want to buy it, whereas a central bank taking policy actions based on data correlation isn't necessarily a great communication strategy, is it?

David Bholat: At central banks we have to tell stories. Correlation is never enough and, even if you get a really good correlation, that doesn't of course mean causation.

Iman van Lelyveld: In terms of supervisory policies, we can't go to a bank and say, "you have to build down your portfolio because there's some correlation with the sale of a certain product". You need a causal story of how that product affects the exposure on your derivatives portfolio. So that's a challenge because,

if there's a high correlation, it might lead to further investigation in that area, but certainly not to direct policy actions.

Central Banking: If the data is perfect, to what degree can central banks depart from traditional economic methods and use or combine them with new machine-learning methods?

David Bholat: Assuming there are no data quality issues, then the value of using a machine-learning approach is if you think the pattern in the data is non-linear in form.

Central Banking: What could be the big breakthrough in big data in coming years?

Iman van Lelyveld: I'm very excited about the Emir data because we're looking to make big steps. But at the bank we're also looking at other big data sets that are going to be helpful, one being older chamber of commerce data that tells us a lot about individual non-financial firms.

David Bholat: Over the next 12 to 18 months, we want to add the most value through the supervisory arm of the BoE, the Prudential Regulation Authority. There has been a lot of central bank research produced, but not much on the micro-prudential function. We're text mining the letters that are sent from our supervisors to firms they regulate to understand whether we're being consistent or systematically biased in our communication. We're also going to work with all these new, large, granular data sets, as well as changing how supervisors go about their jobs by building more interactive and visually appealing dashboards.

Maciej Piechocki: I share the view that for the next 12 to 18 months the central banks will be exposed to Emir derivatives data and money markets statistics; the Markets in Financial Instruments Directive II is also coming with another large set of very granular data sets. Also repo data; and credit registers are being revamped, not only in Europe, but also outside Europe. There has been an exposure to large, granular data volumes that are then driving different analytical approaches and statistical applications, which are transforming the data functions in central banks. ❑

This forum was convened by Central Banking, and moderated by Central Banking's editor, Christopher Jeffery. The commentary and responses to this forum are personal and do not necessarily reflect the views and opinions of the panellists' respective organisations.

Watch the full Central Banking webinar proceedings, Evaluating big data capabilities in central banking*, at www.centralbanking.com/2474740*