Micro-data as a necessary infrastructure – standardisation of reference data on instruments and entities as a starting point: need for a Reference Data Utility

Francis Gross[1]

I. The ECB's experience with the CSDB

The Centralised Securities Database holds micro-data (so far reference data, prices, income data, amounts outstanding) on around seven million instruments. It was conceived as an infrastructure serving the production of macro-economic statistics, and has now been in use as such for the past four years. Staff of the ECB and the 27 NCBs of the ESCB have online access to the CSDB, mainly for data quality management.

In the absence of comprehensive official sources, the CSDB is mainly fed by several commercial data providers. The CSDB had to rely on several sources for its data in order to reach the coverage required for the production of statistics, i.e. all instruments issued or potentially held by Euro Area residents. However, these sources overlap, as many instruments see data delivered by several sources. This offers an advantage as attributes omitted by one source might well be covered by another one. Conversely, it poses a larger challenge with those attributes that are delivered by more than one source: the CSDB must then select the value that will be stored. That process is called "compounding" and is conducted in a fully automated way, based on algorithms; indeed, the sheer size of the database and of the flows of data (e.g. two million prices per day) would not allow manual processing.

The development of the CSDB was made extremely challenging by the reality of the data delivered by the various providers. Data is characterised by diversity in data formats, taxonomies and definitions, by errors such as typos, by varying levels of maintenance (e.g. updating for changes) and by the use of diverse identifiers, both for instruments and for issuers. In that context, the mere identification of data sets representing the same instrument can be a challenge, let alone the identification of the true value of an attribute.

The experience with the CSDB revealed an industrial production process that has become a liability to a whole industry and to the authorities in charge of regulation, supervision and policy. Indeed, whenever an instrument is issued, relevant documentation (e.g. a prospectus) is used by many data vendors and market participants as a source for creating the data set on that instrument, which will be sold to their clients or used in their own processes. Each one of these data capture processes is conducted using the firm's own proprietary taxonomy, data model and format; each one with its own focus on the specific needs the data set should serve. Often enough, some discretion is left to the analyst on how to describe a specific attribute, for instance the name of the issuer – Deutsche Bank, DtBk, DB, etc. whereby DB could also stand for Dresdner Bank, Deutsche Bahn, or something else. Whereas human experts could deal with such data on the basis of their expertise, IT systems cannot, unless

---

[1] Directorate General Statistics, European Central Bank.

The views expressed in this paper are those of the author and do not necessarily reflect the views of the European Central Bank.

they become unduly complicated, i.e. expensive and unwieldy, for performing otherwise simple tasks.

Dialogue with many data experts in industry showed this experience to be the general rule. All organisations that use data in their processes need to do what is euphemistically called "data cleansing". The practice can be striking when assessed with a mathematician's mind. For instance, in a case where three sources provide three different values for a given attribute, consulting a fourth source of the same type cannot really provide certainty anymore, especially when terminology can vary too: the information is lost and the data needs to be produced again, directly from a credible source, for instance the prospectus. Nevertheless, in the absence of a better, affordable alternative, such processes are relied upon across the data community. One large European clearing house employs well over 100 people to read prospectuses in order to produce the basic data the market is not able to deliver in a reliable quality for their operations.

II. Motivation and drivers in the quest for better data

1. Measuring economy and finance: the context for a data strategy

Analysts and policy-makers in the economy and the financial markets need continuous measurement of the phenomena which they analyse and on which they act. A brief review of the act of measurement sets the context for designing a data quality strategy that could be useful in the context of the development of micro-data for large-scale analysis, as will likely be needed to support work on systemic risk.

A simple view of the measurement process underlies the analysis that follows. In that view, the first step of measuring a complex system such as financial markets consists in structuring that system with concepts represented in a terminology, usually a first source of divergence. That step is driven by one's own theory of the system and is usually driven by economic theory, which can be another source of divergence. Next, elements of the system are identified and classified accordingly in the terminology, data is defined along these lines, collected and used to build statistics that feed analysis and support decision-making. This works better if theory is consistent and complete, if its underlying semantics are rigorous, and if the divergences are kept to a minimum across the communities involved in the process.

In the practice of data management and financial statistics, a number of limiting factors are at work. Firstly, the underlying theory cannot be complete, as it often lags innovation in markets that move faster than theory and methodology can; witness the limitations we experience in measuring derivatives. The theory and its language can also not be as rigorous as mathematics, given that it tries to deal with a complex, uncertain reality and is thus in constant flow and subject to many debates and interpretations. We therefore have a gap between a data practice rooted in mathematics, which would thus require rigorous discipline, and a "messy" theory underlying it. Secondly, policy- and decision-making require a continuous flow of statistics: there will be no time for pause to rebuild statistical or data systems. Thirdly, data collection is sometimes perceived as expensive, thus sometimes resisted or subject to compromise on quality. Fourthly, data collection systems are slow to change, whereas the reality they serve to measure can change fast. These few factors illustrate why micro-data cannot be perfect, and they will guide towards identifying possibly promising areas where focused effort could yield valuable improvement.

2. The reality we face: the Tower of Babel again, this time with data

Over the decades, many have grown accustomed to low quality in data and have grown to accept it as a given: "market data has always been bad, and it will always be bad". Low

quality in data cannot be overcome quickly, but developments in recent years indicate that that attitude is not sustainable and that action is needed. These developments can be broadly linked to two drivers, technological development and globalisation, which seem irreversible, and the crisis has acted as a catalyst in revealing the limits to our data practices.

Technological development and globalisation have led to the coalescence of organisations, processes and systems in the financial system into a single global web, in which data flows without borders, each one of them being exposed to myriads of data sources, which often carry contradictory data on the same object. The number of data sources has also kept growing alongside these developments; whereas efforts at standardisation have developed alongside, their success was limited by the comparatively low energy invested in them compared with the dynamic growth of the markets and data sources. As a result, each data source tended to develop its own "data dialect", in order to serve local needs and constraints. In summary, the need for ever larger data pools collides with increasing fragmentation of the data landscape; this is reminiscent of the Tower of Babel, but this time with data.

3. Obstacles to change in data practices

Addressing the data challenge is rendered more difficult by two factors. Firstly, the very existence of the "data problem" is not even known to the majority of users; most simply rely on the statistics and databases put at their disposal, believe that statistical methods will ultimately give them good aggregates from whatever quality of data material, or have become used to accepting weak data as a given, as fatality. Secondly, the interconnected nature of the "data problem" results in a "first mover disadvantage" for any individual market participant who might try to improve the situation on their own: if the first investor in a new standard is alone in moving, he will have spent money for no gain.

Therefore, little of significance is happening: on the one hand, the usually low-profile standards community is trying to promote the adoption of usually limited standards, which is made difficult by the weight of legacy in many user organisations, whereas on the other hand an industry has grown up which competes to offer palliatives, such as "data cleansing", which tries to extract the truth from various divergent versions of data on the same object. Whereas that is often the only economically viable solution available, strictly speaking, once data offers more than one value for the same item, the information is lost and only producing the data again, from the source, will provide certainty.

4. "Horizontal" transparency versus "vertical" transparency

Another type of obstacle to better data rests on the commonly held belief that disclosure is sufficient to guarantee transparency. This is true for a single item, such as a security, which can be fully described and made transparent by a public prospectus; that type of transparency could be called "vertical" transparency. "Vertical" transparency is a necessary condition for transparency, but it is by far not sufficient. Indeed, the analyst interested in a market as a whole needs to grasp the behaviour of large sets of items, for instance a combination of instruments, legal entities and their portfolios; that requires another type of transparency, which could be called "horizontal" transparency.

"Horizontal" transparency requires (possibly large) pools of standardised data fit for processing in IT systems. It cannot result from the availability of even millions of prospectuses or other documents in natural language, whether on paper or in electronic format, which need to be accessed individually.

"Horizontal" transparency is a necessary condition for enabling the kind of analysis that will be required to fulfil the mission of preserving financial stability, be it for analysing market developments or for monitoring the positions of entities or groups. Progress in IT leads to ever more "horizontal" transparency work being based on simulation models, increasingly of

the sort based on micro-data, which allow finer and more flexible analysis than those based on reported aggregates. “Horizontal” transparency is as good as the data that supports it.

“Vertical” transparency is generally ensured by laws that enforce disclosure on instruments (e.g. the European Commission’s Prospectus Directive). Some “horizontal” transparency is delivered through legally enforced reporting of aggregates (e.g. financial reporting of companies), but the production of micro-data, for instance reference data on financial instruments and legal entities, required for “horizontal” transparency is generally left to industry.

5. Reference data, the technical workhorse of micro-data processing

Usually, industry produces reference data on instruments and entities (i.e. the descriptive data of a more stable nature), which is used for creating “horizontal” transparency, from the information disclosed under “vertical” transparency. It is that production process that is at the heart of the preoccupation of this paper.

Reference data represents the infrastructure for identifying, describing, classifying, labelling and organising other micro-data, which is usually the more interesting and sensitive data for policy analysis or business processes. Reference data is usually public, largely stable over time and non-sensitive.

Reference data is per definition unequivocal and supports communication between systems and their users, as well as between businesses and institutions. It must therefore be unique for all users, to allow reliable reference.

In practice, however, as described above reference data is of low quality and often fails in its mission, because usually many versions of the same data item circulate among users, and none of them can be trusted. Sometimes also, reference data is just not available or its use is made difficult by intellectual property considerations.

6. The crisis has highlighted the need to address data quality at a higher level

The crisis has shown that shocks to the financial system can propagate fast and in unpredictable ways, and that swift, decisive action is required from the authorities. Also, the “flash crash” of 6 May on Wall Street has demonstrated that financial markets can skid out of control very quickly in ways no human can understand anymore.

The systemic risk analysts will need the best technical infrastructure they can get to address these challenges. Better micro-data, fit for flexible, near-time and large-scale analysis, is part of that infrastructure and probably a necessary starting point for any further progress.

Significant improvements in data practices cannot be expected to emerge from the markets. These are far too complex, fragmented and competitive, and focused on other concerns to converge anytime soon on the stable and strict discipline required for large-scale improvements in data quality.

In the USA, the recently passed Financial Stability Act of 2010 (also named Dodd-Frank Bill) addresses that same need by establishing the Office of Financial Research (see box below). The approach chosen there is based on data standardisation imposed through regulation and enforcement where needed.

7. Ad-hoc data collection: another function that depends on good reference data

A necessary condition for the financial stability agencies to fulfil their mission resides in the ability to run fast, ad-hoc collections of micro-data that yield pools of high-quality data immediately fit for analysis in large-scale IT systems to assess market developments with potential implications for systemic risk.

Such ad-hoc data collection will always be necessary, given the complex nature of the financial system, which guarantees by definition that unexpected developments will occur, for which the required data will not have been collected in advance.

The effectiveness and efficiency of ad-hoc data collection depends on the quality of data available: it is easier to collect high-quality additional attributes on a class of instruments on the basis of a complete, high-quality register of that class. In the absence of such a register, the ad-hoc data collection will deliver weak results, which affects the quality, reliability and timeliness of subsequent analysis.

8. Industry needs better data, too

Generally, business processes and ICT systems are not good at handling “data dialects”, so their growth has seen the emergence of costly “data cleansing” activities, conducted both by users and by specialised companies; also the ECB’s complex CSDB is a good illustration of such a “data cleansing” operation. Moreover, the collision between “data dialects” hampers automation and drives cost up. So, for instance, the reality of the vision of “Straight Through Processing” (STP) in banks remains the discussion about “STP rates” and the employment of many to fix failed transactions.

In industry, same as in the public institutions, low data quality at source impacts the whole downstream value chain, creating excess costs and quality losses all along, hampering internal transparency of organisations and increasing their operational risk.

Meanwhile leaders in the IT industry recognise that data quality, especially data standardisation, is becoming the decisive roadblock to effective use of the ever growing power of information technology, be it by industry or authorities.

Many in industry now realise and accept that improvement in data quality will happen for all or for none, and that real progress will require legal compulsion to impose across the market the rigour and discipline which are needed for reliable implementation of data standards, and which the market itself will not be able to deliver.

9. Single source or multiple sources for reference data?

The discussion on whether data should come from a single source or from multiple sources seems to be immanent to the field; the answer probably depends on the nature of the data. It can indeed be beneficial to hear several independent and differing views on a matter of opinion or valuation, or to examine several independent modelling approaches to a complex analytical question. However, where information is unequivocally set, for instance in a prospectus or a contract, it would seem safer to rely on a single electronic source, which would be used by all parties. This is especially attractive for reference data, now that technology makes it possible at low cost, in principle even globally so.

In practice, it remains sad to see that unequivocally and exactly defined information, such as the basic reference data on financial instruments and legal entities, is de facto destroyed through the independent production and circulation of many, often different, incompatible and contradictory data sets on the same information by data vendors who compete in the market. Such differences usually result from the use of different terminologies and definitions, from errors or from diverse levels of maintenance of the data. This is the case, for instance, for basic reference information contained in the prospectus of a financial instrument or the same about a legal entity.

10. The long way to better data needs to begin with a feasible first step

After having been allowed to grow for several decades, the micro-data challenge has become pervasive, and it can be expected that it will require some time to solve. Therefore, a feasible first step is required to start.

The decision to take such a first step should not necessarily be contingent on the whole route being carefully mapped out in advance. Indeed, the complexity of the problem, the number of stakeholders and the time needed for implementation make it impossible to do so credibly in the context of ongoing change in markets and technologies. Much rather, the nature of the problem is more conducive to taking a first step, and to learning from it for the subsequent steps. Subsequent steps will anyway be taken in the new context created by the previous steps.

Such a first step should be designed to be credibly feasible. It should thus focus on a limited yet significant area that is relatively free of “political” ballast and deep-rooted obstacles. The measure should be implementable at low cost, designed to deliver fast benefits for all stakeholders and be market neutral. It should also be designed from the outset with a potential for further growth, in reach and depth, and to provide opportunities for learning and experimenting. Ideally, it would also foster positive developments beyond its formal reach, for instance by acting as a “crystallisation germ”, offering to industry a fixed anchor towards which to converge for effective standardisation beyond the reference data it covers. In that sense, it would reduce the “first mover disadvantage” mentioned above.

Such a measure would likely depend on rigour and discipline among a large group of market participants, which might require legal compulsion. In turn, legal compulsion that serves all market participants could provide a wonderful opportunity for market authorities to develop a showcase for cooperative regulation, a credible win-win proposition. Success could bolster the credibility of authorities in their missions to strengthen market transparency, to improve oversight and to control systemic risk while improving market infrastructure.

Finally, such a measure would need to be integrated into the existing data ecosystem as a basic infrastructure that facilitates the development and work of other data collection mechanisms positioned at a higher level.

A measure that fulfils all these criteria could be built around the basic reference data mentioned above. In that spirit, the article introduces below a sketch of the possible concept of an international Reference Data Utility.

III. The concept of a Reference Data Utility

The idea of a data utility is quite natural and has been floated many times across the data community. The initial idea of the Reference Data Utility (RDU) presented below emerged four years ago from the ECB’s experience with the CSDB. It was initially shelved as too difficult to implement. It was revived as the crisis revealed the need to improve data and seemed to offer a rare opportunity for the kind of consensus for reform such an infrastructure would require.

That idea has been discussed and refined in numerous conversations and conferences with ESCB colleagues, in Statistics and beyond, and with many throughout the financial industry and the regulatory and policy community, in Europe and the USA, especially people involved in the data supply chain and the downstream functions.

I would like to thank all who contributed for their patience, support and candour, and for their constructive contributions; these were indispensable for the mere continuation of that exploration as well as for the emergence of a concept that might now stand a better chance of seeing an implementation one day.

1. "Thin Utility": an infrastructure focused on basic reference data

An RDU would focus on the most basic reference data on financial instruments and legal entities, the kind that is needed by virtually all data users: identification, basic description, interrelations and classifications, as well as, for legal entities, a manned electronic address – the latter could for instance be used to support quick and efficient large-scale ad-hoc data collection by other repositories. Hence a "thin" utility.

2. A single, strategic infrastructure shared by all stakeholders

Reference data on financial instruments and legal entities is needed by all stakeholders, private or public, in financial markets and beyond. Moreover, in daily operations, data flows through the systems and between entities regardless of its origin. Therefore, the coexistence of versions of the same data under several standards will always lead to collisions and local fixes, hence to new breeds of the same data and ultimately confusion.

A single standard on reference data that serves all needs should be the goal. An RDU could serve that goal if it was designed from the outset to offer basic, high-quality reference data on financial instruments and legal entities to all stakeholders, public or private, of course within legal limits, for instance concerning confidentiality.

The development, governance and operation should thus be conceived to involve these stakeholders in an appropriate manner.

3. Legal compulsion

The production of high-quality reference data requires rigour and discipline in the adherence to the reporting and maintenance process and in the application of the data standard. With a high number of market participants involved, this can only be obtained through legal compulsion, as many in the industry recognise.

The need for legal compulsion for reference data standardisation is now acknowledged by most in the industry, recognising that it would provide legal certainty for standard-driven investment and a level playing field for all stakeholders. Many, including some major data vendors, now see reference data as a commodity and a public good that should be held in a public infrastructure. Moreover, industry-driven standardisation has largely failed so far and shows no credible promise in the absence of a dominant party who could impose its practice as de facto industry standard, and in view of the "first mover disadvantage" that would affect any institution being the first to invest in a new standard, hoping that the others would follow.

In the USA, that legal compulsion has now been created by the recently passed Financial Stability Act, which establishes the Office of Financial Research (OFR).

The **Office of Financial Research** (OFR) has been established by the Dodd-Frank Bill recently passed in the USA. The OFR will be an independent entity within the Treasury; its Director will be appointed by the President of the USA for a term of six years.

It will provide technical support to the newly established Financial Stability Oversight Council.

The OFR will have two main entities: a Data Centre and an Analysis and Research Centre.

The Data Centre will collect, among others, reference data on instruments and entities, as well as data on transactions and positions. The OFR will have the power to issue standards and to enforce reporting.

In Senate hearings it has been made clear that the OFR will need international solutions in data standardisation, including on reference data. Indeed, for its analysis on systemic risk and financial stability it will also need to process data on cross-border transactions and positions, i.e. involving instruments and entities based abroad.

4. International reach is required, global reach should be the goal

The riskiest developments in financial markets are likely to be surprising ones, which will require agility (i.e. flexibility and speed) in the tools used to analyse them. Furthermore, the global nature of financial markets and of potential crises or threats to financial stability would suggest the need for analytical tools with the same reach. Macro-economic analysis of financial market developments based on micro-data seems to be a promising avenue in that respect, but it requires data on transactions and positions involving instruments issued abroad and entities based abroad. Such data must be of the same quality as the data concerning purely local instruments and entities. Hence global coverage must be the goal for a Reference Data Utility.

In the course of globalisation, the financial industry has seen the emergence of organisations and processes that are increasingly international and data intensive. Here as well, an international RDU would be good, a global one better.

The European Commission has long ago launched initiatives to improve Europe-wide access to data on entities in the field of business registers. That goal would be facilitated by an international RDU. The debate in the USA has also clearly shown an awareness of the need for internationally standardised reference data for the OFR to succeed. For instance, Gov Daniel Tarullo from the Federal Reserve Board indicated in the Senate Hearing on the Dodd Bill (later the Financial Stability Act) in February that international solutions will need to be sought for data standardisation.

5. A single Reference Data Utility should be the goal

Data quality would require a single RDU; technology makes it possible. However, building such a single RDU harbours legal and organisational challenges.

Therefore, ease of implementation could lead to envisaging national or regional RDUs. A large number of such national, regional or sectoral registers already exist, especially in the field of business registers. Practice shows that such fragmentation is exactly what needs to be overcome, as is documented by numerous initiatives that aim at progressively linking existing registers, not least by the European Commission and CESR. These efforts have so far proved to be slow in their development, at best. One reason might well be that the organisations concerned have very diverse missions, legal backgrounds, technical legacy, limited resources and probably little intrinsic incentive to change. Even Eurostat's Eurogroups Register uses data from a commercial provider, rather than assembling data from national business registers.

Moreover, it is easy to imagine that several, networked RDUs working in parallel on the same standard would unavoidably be exposed to different situations, which could lead them to diverge in their interpretations of the standard, thus to develop local "data dialects", which even if based on the same standard would defeat the purpose.

Finally, parallel RDUs could see overlaps and gaps between their respective coverage.

That suggests that the goal of a single RDU is the one really worth pursuing, as a fragmentation would likely add to the problem it intends to alleviate.

6. A design concept that could reconcile global reach of an RDU with national law

Assuming a single RDU with global reach and backed by legal compulsion, a theoretically feasible design concept could be reached by separating the functions of a technical nature from the functions with a legal character, e.g. enforcement.

Under that design concept, the technical functions would be performed by an International Operational Entity (IOE) with a purely technical focus and without legal powers, whereas in each legal constituency the legal powers and obligations could be conferred upon a relevant local authority. The two components could be linked by a service agreement, under which each participating local authority would outsource the technical conduct of data collection, storage and distribution, as well as the coordination of standards design work, to the IOE.

Such a design concept would enable a modular growth of the RDU's geographical coverage from an initial base. Legal constituencies could join individually, at their discretion. The operation of an RDU could thus start with a smaller number of participants. Chances of success would be larger if the initial participation could represent an attractive critical mass.

Such a design concept could also serve a European solution and could be "upward compatible" to a broader international one.

7. An International Operational Entity

The International Operational Entity (IOE) would perform the technical operations of the RDU on behalf of the national authorities that are mandated by their national law to enforce that data collection at the national level. The IOE would do so under a service agreement passed with the national authorities of all participating countries. It would be designed to operate as lean as possible, to reduce its impact on existing activities in the industry while leveraging them for better reference data.

To ensure acceptance, the IOE would need to operate as a non-profit, self-financing entity, leaving profit opportunities offered by the data supply chain to the private sector. The IOE would work under a business model, a legal form and governance which are yet to be analysed. These parameters should be designed to support the credibility of the IOE's technical competence, market neutrality, efficiency, global reach and acceptability in the market. In that respect, it might be useful to associate in the governance of the IOE a round of well-known international institutions, which would represent market authorities, governments and industry.

The IOE's revenue could be derived from a combination of sale of reference data and registration fees from reporters, whereby the latter could be less well accepted.

In order to fulfil its design objectives, the IOE would use as much as possible existing services from existing suppliers. So, for instance, production and maintenance of RDU data would be conducted by external parties acting in a competitive market, in which providers would offer their services to data reporters (issuers and entities) who would make their "make or buy" decision and possibly select a provider. In order to keep control of data quality, the IOE would accept data sets only from analysts it certifies. Analyst certification would, in turn, rest on analyst training on the RDU's published methodology and standards; such training could also be performed by a competitive market.

The tasks of the IOE could encompass running the RDU's daily operations as well as the development and management of its infrastructure, both technical and organisational.

Running the RDU's daily operations would consist in (1) receiving, storing and distributing reference data on legal entities and financial instruments, (2) running the quality feedback process between data users, data producers and national authorities, which also facilitates quality enforcement by the national authorities (more below) and (3) running the commercial and administrative processes of the RDU, depending on the business model chosen.

Developing and managing the RDU's infrastructure would consist in (1) developing the coverage of data in the RDU to new asset classes or attribute classes through dialogue with the relevant communities: data users, standard designers, technologists, lawmakers and national authorities, to identify new needs and feasibility, (2) developing the geographical coverage of the RDU, (3) steering the development of standards for its data, (4) managing

and developing the IOE's supply chain and its analyst community and, (5) possibly, to foster and sponsor relevant research into data.

8. Design of the RDU's supply chain

Whereas the major part of the data supply chain around the IOE would be left to industry, operating in a competitive market, a monopolistic core limited to essential functions of a utility is recognised as necessary to help embody consistent implementation of a standard.

Whereas data sets would be produced for, delivered to and maintained in the RDU by a competitive industry, the IOE would ensure that each data set is produced once only and that for each data set a single analyst, certified by the IOE, is identified as responsible. The IOE would act as an obligatory point of passage for each data set and would thus offer a unique point of reference to users worldwide for each data set covered, with access through a website.

Such design would enable the Utility to combine the benefits of competition and those of a one-stop-shop. It would also be central to the concept of a lean, user-driven data quality assurance process.

9. Data quality management process

A lean IOE would not employ staff to check data quality. Instead, quality data should be delivered in the first place by the certified analysts who produce it on behalf of the legally responsible reporting agents, either issuers of financial instruments or legal entities.

A community-centric data quality management process could be imagined, which would require very few specialised staff and would rest on feedback from users to the certified analyst through the RDU's website. The principle would foresee that in case of wrong data, users noticing the error could go to the webpage of the data item in the website of the RDU and click the "quality button" of the item, which would create a message automatically addressed to the responsible analyst. In that message, the user could convey their observation, which would lead to correction for all, at the centre.

The message would be copied to a compliance centre of the IOE, from where, failing timely repair, it would be forwarded to the national authority in charge of the reporting agent's compliance, from where the enforcement process could start.

10. Incremental growth of coverage: start feasible and develop from there

The scope covered by the Thin Utility would begin from a small but quickly feasible base with categories of instruments, entities and attributes that are immediately useful to many, thus it would deliver immediate value at low risk; all involved would collect experience in those first steps that would feed further development. The value provided from the start would ideally generate demand for broader coverage. Coverage would then grow from there, over time, driven by demand and feasibility.

11. Data coverage: financial instruments

The scope of instruments covered could start with debt and equity and progressively grow to cover for instance derivatives, perhaps also OTC at some point. The RDU could also grow to cover the basic reference data of ABS and individual loans, at least those involved in securitisation.

Attributes covered for financial instruments would need to encompass an identifier, basic technical descriptive attributes, some uncontroversial classifications e.g. for statistical

purposes, and some interrelations e.g. who is the issuer, what instruments the asset backs, or which assets back it.

The scope of instruments and attributes covered would be determined in a process of dialogue between the stakeholders, which would be managed by the IOE.

12. Data coverage: legal entities

Coverage of legal entities could be determined through a similar mechanism as for instruments.

Attributes related to legal entities would cover the same categories as for instruments: an identifier (see next paragraph), basic technical descriptive attributes, some uncontroversial classifications e.g. for statistical purposes, and some interrelations e.g. what other entity owns it, which ones it owns (within the limits of confidentiality rules), what instruments it issued.

Data about a legal entity could also be imagined to encompass an electronic address manned by a manager responsible for the entity. Such an electronic address could serve two purposes that are not readily served today. Firstly, it could serve for data collection from Special Purpose Entities (SPEs), which are usually very difficult to approach as many have no operational staff, but whose reporting would be essential, for instance to achieve reliable FDI statistics. Ideally, legal compulsion on the SPE could be used to elicit reporting even from such managers based outside the legal constituency in which the SPE is based. Secondly, such an electronic address in the RDU could be used to run large-scale ad-hoc data collection from many entities in a very lean fashion, through a query sent from the RDU (see below).

13. Build on existing standards, cooperate with ISO

An RDU would, as much as possible, build on existing standards and established practices, such as the ISIN code for identifying instruments, and the ISO process for designing standards. It would add momentum to developments that have so far stalled, such as the creation of a standard entity identifier. Indeed, the mere involvement of the ECB and the Federal Reserve in the discussion about data standards and a utility has revived ISO efforts in that very field. By leveraging and catalysing existing infrastructure and resources, the creation of a utility would require relatively little invention.

Conversely, the recently strengthened focus on data standardisation in the data community has led the ISO Technical Committee to review the standard that established the BIC, the Bank Identifier Code, which is now called the “Business Identifier Code”, which opens the way for its development into a universal entity identifier.

14. Ad-hoc data collection: serving financial stability and systemic risk control

Large-scale ad-hoc collection of micro-data could well become central to systemic risk analysis and to financial stability work. For such collections to be useful, it might become essential to avail of a capability to quickly collect, in a targeted way, large pools of data immediately fit for large-scale processing, i.e. standardised. Such data might need to cover well-defined attributes on all instruments in given classes from all entities in certain categories.

An RDU could allow sending a relevant query to the electronic address of each entity concerned. Knowing the complete population of instruments and entities to be surveyed, it becomes easier to reduce cost and increase speed through well-controlled sampling.

That approach could also support the collection of very confidential attributes, whereby the query sent to entities from the RDU would ask for data to be delivered for instance to the (national) supervisory authority entitled to hold such data. The data could be treated there, for instance anonymised and aggregated, and then transmitted to the analysts of financial stability authorities or other authorities that need them.

Such flexible, fast, low-cost and targeted collection of high-quality data fit for large-scale processing is not possible today.

IV. Positioning of a Reference Data Utility

An RDU would appear as a new entity in a complex "data ecosystem".

The RDU would be positioned in the data supply chain upstream of other data repositories or operations, public and private, and provide them with the basic reference data on financial instruments and legal entities that they need to conduct their business, which can be to collect, organise, distribute and/or analyse more specific or dynamic data, for instance specific transaction data or position data.

1. Positioning of an RDU versus issuers of financial instruments

Issuers of financial instruments would be faced with the legal obligation to deliver and maintain in the RDU reference data on the instruments they issue.

The legislators in many countries have long recognised that well-functioning financial markets and investor protection need to be ensured by legislation. So far, that is done by enforcing disclosure by issuers of information on individual financial instruments, for instance by enforcing the publication of a prospectus for certain types of instruments.

As mentioned above, that practice guarantees transparency at instrument level ("vertical" transparency). That "vertical" transparency is itself limited by the fact that updates to certain aspects of prospectuses are usually done through "corporate action" messages that need to be reconciled with the prospectus information, a task only a few organisations can conduct. Prospectuses themselves are usually not updated.

Investor protection and market stability also require "horizontal" transparency, i.e. the visibility of the behaviour of larger sets of instruments and investors taken as a whole, and their interaction with other asset classes or groups. "Horizontal" transparency requires the availability of data fit for analysis in large-scale IT systems; it is useful only if reasonably near-time.

Therefore, the case for investor protection could justify that issuers of financial instruments be required to provide, alongside documents that support "vertical" transparency, the data required for "horizontal" transparency, and to maintain it up-to-date. Some of that data, such as transaction data, is generated on a daily basis by the marketplace, whereas another part, reference data, needs to be produced and would be stored in an RDU.

The costs of producing and maintaining reference data in an RDU would be a legitimate component of the issuer's cost of doing business, but would represent a very limited share of this cost. Just compare the cost of setting up a prospectus, which goes into the tens and hundreds of thousands of euros, with the cost of a few hours of an analyst's time over the life of an instrument.

For many issuers, who are also data users, these costs should be (more than) balanced by the benefits from better reference data throughout their operations. However, these benefits will not be easy to account for explicitly.

2. Positioning versus data providers as clients of an RDU

The RDU would be positioned as a commodity provider, delivering in a standardised electronic format reference data that represents information unequivocally known.

Commercial data providers, who today produce their own reference data or assemble it from various specialised companies or from other sources could choose to simplify their sourcing by becoming clients of the RDU. Thereby they would shift from high-cost in-house production of low-quality reference data to buying a low-cost, high-quality commodity. Such outsourcing and moving to higher layers of the value chain would represent a move that is very common in the development of many industries. Data vendors could then focus on the cutting edge of the data business.

For the clients of existing data providers, this would represent a significant quality improvement in the reference data delivered to them, perhaps associated with a small decrease in prices. Furthermore, that improvement in reference data could provide the basis for significant improvements in these clients' process performance, firstly, because processes that use data from commercial data providers would face fewer data-induced failures (STP rates could increase), and secondly, because data interchange between organisations would be easier if their data sources used the same reference data. Such improvements would of course reduce costs and operational risks in the processes that benefit. They might also allow simpler design of certain IT systems, a further opportunity for reduction of cost and operational risk.

3. Positioning versus data providers as suppliers of an RDU

The entities required to submit data to the RDU and to maintain it there could either do that for themselves or take recourse to services offered by organisations specialised in the field of data. These could be either the established commercial data providers or other organisations, such as CSDs or established registers, which would be well placed to produce data for the RDU.

For such data suppliers, producing data for the RDU could represent a new business line and source of revenue.

4. Positioning versus clients of commercial data providers

In theory the RDU would be a commodity provider mainly supplying to industrial clients such as commercial data providers and larger public or commercial data operations. Yet, it cannot be excluded that certain data users would choose to source reference data directly from an RDU, especially those who build in-house solutions from basic components. The majority of users of commercial data services should however be expected to continue seeking package or turn-key solutions wherever possible. In that sense, the emergence of an RDU should not disrupt the business of commercial data providers in a significant way.

5. Positioning versus public sector data users (e.g. registers, statistics)

An RDU could be a source of standardised, high-quality basic reference data for public sector data users (business registers, credit registers, administrative databases, etc.), which they could use either as an input into their databases or as a benchmark for their data quality management. In that role, an RDU would serve as a commodity provider and not otherwise interfere in the execution of these users' missions.

Sourcing reference data from an RDU would however relieve these users from the tedious task of producing or collecting their own reference data, at least within the scope covered by the RDU, and would allow them to focus their resources on tasks with higher value addition.

Providing high-quality, standardised reference data to such public functions would allow these to more easily exchange and combine their data where they need to, for instance for combined analysis, or it would enable their users to do so. As an example, it would become much easier to combine micro-data from different sources, for instance data on securities holdings, on securities issuers and on transactions, to understand the dynamics of a given market. An RDU could also serve public authorities to conduct swift ad-hoc collection from market participants of micro-data not yet available, at a level of quality that would make them fit for near-time, large-scale processing. Such capabilities would be important for analysis serving functions that monitor financial stability or systemic risk. The potential process is described above.

6. Positioning versus the ESCB's Centralised Securities Database (CSDB)

The CSDB would certainly source its reference data for financial instruments and issuers from the RDU, once available. That would improve data quality on the attributes covered and lift some of the limitations on data usage that result from contracts with commercial data providers. It could also reduce costs and simplify the operation of the system by reducing the need for compounding data from various sources and saving costs on data quality management throughout the ESCB.

7. Positioning versus standards bodies

The RDU would put significant weight on cooperating with standards bodies, first and foremost the International Organisation for Standardisation, ISO, who design, among others, standards for data on financial instruments and legal entities.

For an RDU, this would guarantee that standards it applies are designed by all stakeholders who wish to contribute, including industry.

For the ISO, it would ensure that the design of standards for financial data would attract more attention, thus more active participation from stakeholders, which could strengthen the quality of standards and their acceptance in the stakeholders' community.

8. Positioning in the data standardisation process at large

An RDU would offer standardised data on a limited scope of attributes, with data collection based on legal compulsion. However, standards would be designed to cover a broader scope of data than that covered by the RDU. It could thus be expected that the existence of a core body of highly standardised reference data would provide systems and process designers with a solid anchor of standardisation towards which to converge. That could encourage them to apply the same standards more broadly, beyond the confines of the RDU's scope. Especially designers who build new systems could see an opportunity to adopt data standards backed in their core by the RDU, which could result in a broader migration towards standards.

In that sense, the RDU could be a germ of crystallisation for broader data standardisation across the financial system.