Ref. Ares(2019)7020513 - 13/11/2019

COFFERS
EU Horizon 2020 Project

Journal paper submission on the contribution that company regulation might make to the gap

Document Details

Work Package	WP1
Lead Beneficiary	University of Warwick
Deliverable ID	D1.9
Date	October, 31, 2019
Submission	October, 31, 2019
Dissemination Level	PU – Public
Version	1.0

Author(s)

Yuval Millo
University of Warwick
Yuval.millo@wbs.ac.uk

Nikiforos Panourgias
University of Leicester
np220@leicester.ac.uk

Acknowledgements

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 727145.

What does data quality mean? Defining data quality in the organising of a global identification infrastructure for legal entities

Journal:	*Organization Studies*
Manuscript ID	OS-19-0665
Manuscript Type:	Special Issue on Technology and Organization
Keywords:	Big Data, Financial regulation, Data capitalisation, Identification infrastructures, Data quality
Abstract:	The vast amounts of data generated as a result of the growing digitisation of all aspects of contemporary life underpins a growing interest among practitioners and academics regarding how this mass of data can be utilised to enhance planning, strategizing, decision-making, sense-making, and other forms of cognition through data-driven business analytics and intelligence solutions. Realising this potential,

	however, is not straightforward and requires much sociotechnical work in order to make the data useable and useful for such solutions. We analyse how data can be made useable, useful, and therefore valuable through a study of the development and implementation of the Global Legal Entity Identifier System (GLEIS) as a case of establishing an infrastructure that certifies the accuracy and validity of identification data for corporate entities involved in financial markets transactions. Our findings describe a process whereby an identification infrastructure - including a non-replicable methodology for assessing data quality - is established that contributes to making the developer and controller of that methodology an irreplaceable intermediary for users of the infrastructure; this in spite of the need for an associated reference data infrastructure to be open and widely accessible to all participants in order for the infrastructure to be successful. The findings indicate that in the process assets are created on the basis of openly-accessible data through certifying of a desired set of qualities to be achieved by adopters and the infrastructure.

What does data quality mean? Defining data quality in the organising of a global identification infrastructure for legal entities

Abstract

The vast amounts of data generated as a result of the growing digitisation of all aspects of contemporary life underpins a growing interest among practitioners and academics regarding how this mass of data can be utilised to enhance planning, strategizing, decision-making, sense-making, and other forms of cognition through data-driven business analytics and intelligence solutions. Realising this potential, however, is not straightforward and requires much sociotechnical work in order to make the data useable and useful for such solutions. We analyse how data can be made useable, useful, and therefore valuable through a study of the development and implementation of the Global Legal Entity Identifier System (GLEIS) as a case of establishing an infrastructure that certifies the accuracy and validity of identification data for corporate entities involved in financial markets transactions. Our findings describe a process whereby an identification infrastructure - including a non-replicable methodology for assessing data quality - is established that contributes to making the developer and controller of that methodology an irreplaceable intermediary for users of the infrastructure; this in spite of the need for an associated reference data infrastructure to be open and widely accessible to all participants in order for the infrastructure to be successful. The findings indicate that in the process assets are created out of openly-accessible data through the certifying of a desired set of qualities to be achieved by adopters and the infrastructure.

Keywords

Big Data, Data quality, Data capitalisation, Identification Infrastructures, Financial regulation

Introduction

The recent explosion of digital data across nearly all industries has created several opportunities for decision-makers to extract intelligence from data and translate it into business advantage (Brynjolfsson, Hitt, & Kim, 2011; McAfee, Brynjolfsson, Davenport, Patil, & Barton, 2012). The increased variety and speed of data creation, stemming from the growing digitisation of all aspects of economic life, has provided many opportunities for practitioners as well as regulators to plan, strategize, improve decision-making and enrich sense-making through so-called 'big data' business analytics and intelligence solutions (Chen, Chiang, & Storey, 2012; Glaeser, Kominers, Luca, & Naik, 2018; Günther, Mehrizi, Huysman, & Feldberg, 2017; LaValle, Lesser, Shockley, Hopkins, & Kruschwitz, 2011; Liu, 2014; Obermeyer & Emanuel, 2016; Phillips-Wren & Hoskisson, 2015; Warren Jr, Moffitt, & Byrnes, 2015; Zhang, Yang, & Appelbaum, 2015). These developments are of particular importance in information-based industries, such as financial services, where the calculation or risk, processing of payments, clearing and settlement of transactions, and overall provision of services relies on data-intensive and infrastructure-based processes. In this context, the ability to acquire, maintain, and analyse data is regarded as a key competitive factor, but also as an attractive proposition for regulators to exercise effective supervision and ensure market stability.

Early studies around the development of 'big data' were mostly driven by the idea that what cannot be measured cannot be managed and thus, focused largely on the usefulness of data to help manage businesses (McAfee et al., 2012) and bring strategic advantages to firms (Brynjolfsson et al., 2011; Harris & Davenport, 2007). Further work examined the effectiveness of analytical methods and their application and impact in various business cases and sectors.

In spite of the recognition in the literature about the centrality of data to organisations, there seems to be a lack of attention to the dynamics involved in managing large and varied datasets. In particular, data quality and its validity as a basis for organisational decision-making have until very recently remained unexplored apart from a few notable exceptions (Park, Huh, Oh, & Han, 2012). This relative gap in our conceptual understanding of data quality in organisations (Fox, Levitin, & Redman, 1994; Nagle, Redman, & Sammon, 2017; Redman, 1995, 2013, 2016) is becoming even more crucial in light of recent developments relating to the integrity and use of data in relation to social media and other such platforms (Rubin & Lukoianova, 2013; Zeng, Chen, Lusch, & Li, 2010), as for example with 'fake news' on Facebook, 'engineered' feedback and ratings on e-commerce websites such as Amazon or TripAdvisor, and fake profiles, of so-called ‘bots’ on Twitter.

Data quality is of crucial importance in the context of financial data. For example, when it comes to processes such as customer on-boarding, know-your-customer (KYC), and other functions that involve the initiation or settlement of a transaction or credit service. In such cases, organisational decision-making relies critically on the details of the counterparties being correct, full, and up-to-date. On such occasions, customer identification and/or authentication is pivotal and rests on the assumption that the reference data and identification information held by those institutions involved in the transaction cycle are valid and reliable. Inaccurate, erroneous, and missing data can lead to suboptimal and costly situations resulting from the failure of transactions, the miscalculating of risks, or from not complying with regulatory requirements. This highlights the importance in such a context of information infrastructures that can ensure a high level of identification reference data veracity in order to guarantee the accurate and incontestable identification of market participants.

In this article we examine efforts to establish such an information infrastructure through a regulatory initiative for identifying organisational entities that undertake trading activity in financial markets. We regard this initiative as an attempt to establish an identification infrastructure (Camp, 2004; Clarke, 1994; Eriksson & Agerfalk, 2010; Whitley, Gal, & Kjaergaard, 2014), utilising a standardised global identifier for legal entities (LEI) and, as part of this infrastructure, establish a set of organisational and inter-organisational arrangements aimed at ensuring the reliability and veracity of identification data. We examine empirically how processes aimed at ensuring data-quality were developed as part of the establishment of the Global Legal Entity System (GLEIS). We relate these processes to the concept of data *veracity* in the literature on 'Big Data' (Abbasi, Sarker, & Chiang, 2016; IBM, 2012; Rubin & Lukoianova, 2013). We analyse the emergence of the LEI through the lens of data quality, a concept closely related to Big Data's data veracity. Our analysis challenges the widely held assumption in the Big Data literature that data veracity is conditioned on achieving certain degree of technical proficiency organisationally. Instead, we propose that data quality is an outcome of a broader sociotechnical process that we refer to as the development of identification infrastructures. Importantly, our analysis indicates that data quality is not a one-off achievement, but a dynamic and contested process in which many different views among infrastructure participants have to be reconciled and inscribed into the technical aspects of the infrastructure. Additionally, our case teaches us how high data quality, as part of an identification infrastructure, enables the transformation of data into a valuable commodity.

Our empirical examination focuses on how the infrastructure for data quality is constructed as it reconciles the views of two differing groups of actors with regards to the nature of data quality: the regulators and the users. Our examination indicates that the viability and stability of this infrastructure depends on maintaining an effective dynamic reconciliation between the two groups of actors. This empirical examination also addresses wider interest in the big data and analytics literature (Abbasi et al., 2016; Agarwal & Dhar, 2014). While this literature highlights the importance of having data quality and data veracity, it tells us relatively little about how this is to be achieved and maintained in practice. Our examination addresses this gap in the literature. Building on this, our findings also help to shed light on the dynamics of 'data capitalisation' - the processes through which the value of data is captured and realised in terms of future economic gains and what is the role of information infrastructures in this context (Abbasi et al., 2016; Doganova & Muniesa, 2015; Langley & Leyshon, 2017; Najjar & Kettinger, 2013; Thrift & Leyshon, 2007). Our findings indicate that to capitalise big and varied datasets (i.e. turn them into valuable assets), trust must be obtained, but also actively maintained, in relation to the accuracy and relevance of the data. Whilst the current literature merely notes the importance of such trust, our research documents empirically the conditions necessary for establishing such trust and the dynamics accompanying this process.

Background to the establishment of the GLEIS

To understand the motivations for the development of the LEI, we must return to the immediate aftermath of the global financial crisis of 2008. The collapse of Lehman Brothers in 2008 revealed a significant blind spot in the ability of financial regulators to foresee the concentration of liabilities that were accumulated via subsidiaries of global and systemically important financial entities. In particular, the absence of a standardised identification scheme prevented regulators from being able to determine who are the ultimate owners and/or controllers of financial obligations (Fleming & Sarkar, 2014; Jenkinson & Leonova, 2013; Lai & Mordel, 2012). This blind spot, in turn, highlighted a broader problem: the inability to trace how legal entities are first, identified, and secondly, relate to one another. This realisation called for a way to identify the legal entities and, as a corollary, to ensure the quality of the reference data describing these entities. For example, maintaining high quality identification data called for eliminating duplications resulting from the use of many variants or spellings of a company name and a plethora of different identifiers used for the same entities in relation to different activities in which they were engaged (e.g. FDIC Certificate ID, SEC CIK SWIFT ID, and various systems vendor proprietary IDs).

Regulators decided to address this problem through the instigation of a new market-wide and cross jurisdictional identification infrastructure aimed at developing a universal identification schema for legal entities engaged in any kind of financial markets transactions across asset classes and trading venues. (Financial Stability Board, 2012; Legal Entity Identifier Regulatory Oversight Committee, 2015). At the core of the infrastructure is a standard identifier format (International Standards Organisation, 2012; Legal Entity Identifier Regulatory Oversight Committee, 2015) and a mandating of the use of the identifier in the reporting of financial markets transactions (Legal Entity Identifier Regulatory Oversight Committee, 2015). From a data perspective, the LEI links basic business data items (referred to as Level 1 or '*who is who*' data) with identifying details of the legal entities. Later (as of May 2017) this link also included Level 2 ('*who owns whom*') data about the ultimate and immediate owner of the entity.

Crucially, the set of regulatory requirements was coupled with the establishment of an issuance and maintenance mechanism for the data linked to the identifier and associated governance arrangements.

Operationally, any legal entity that trades financial products and needs an LEI can apply for one from a LEI Operating Unit (LOU) by providing the require identification data and an issuance fee which is charged by the LOUs on a cost-recovery basis but which can vary between LOUs according to their costs. Coordinating the operation of LOUs was assigned to the Global Legal Entity Foundation (GLEIF), a not-for-profit organisation. There are no restrictions as to who can become an LOU, but there are clear requirements as to how LOUs can operate which are set out in a Master Agreement between the GLEIF and the LOUs. GLEIF oversees the issuing and validating of LEI identifiers but also the overall operation of GLEIS.

At the very top of the governance structure of GLEIS is the Regulatory Oversight Committee (ROC) on which representatives of the financial markets regulators from around the world sit and which oversees the work of the GLEIF and provides direction as to the development and running of the GLEIS and the way the GLEIS is integrated with the existing and future regulations it is associated with. This infrastructure for the creation and maintenance of legal entity identification is referred to collectively as the Global Legal Entity Identifier System (GLEIS).

Theory and development of research questions

In the big data and data analytics literature, data-quality is increasingly recognised as an important dimension for datasets that can underpin big data initiatives, with recent attention focusing on the *veracity* of data. Data veracity is defined as 'the level of reliability associated with certain types of data' including]truthfulness, accuracy or precision, correctness' (Rubin & Lukoianova, 2013, p. 7). Data veracity, in turn, relates to the credibility assigned to the data, which is a function, in turn, of the trust the data source has with regards to a potential user of such data (Rubin & Lukoianova, 2013). In the context of identification data, such credibility relates not only to the data itself, but also to the perceived quality of the sociotechnical arrangements responsible for producing and maintaining the data. For example, identification data can suffer from semantic inconsistencies, lack of structure, conflicting evidence, multiple entries, and inaccuracies (Zeng et al., 2010). Many of these issues can be exacerbated when the data held by the infrastructure are user-generated (Zeng et al., 2010) and in such a context, ensuring high veracity of data sources can be a major challenge (Abbasi et al., 2016).

Identification infrastructures (IDIs) are a specific type of information infrastructures (Ciborra & Hanseth, 1998; Constantinides & Barrett, 2005, 2014; Hanseth & Braa, 2000; Hanseth, Monteiro, & Hatling, 1996; Monteiro & Hanseth, 1996) with the specific aim of establishing a recognised identification of a particular actor, such as - in this case - a legal entity, by associating them with a set of specific characteristics (Beynon-Davies, 2016; Eriksson & Agerfalk, 2010; Otjacques, Hitzelberger, & Feltz, 2007; Whitley et al., 2014). These typically incorporate a unique identifier that is matched to a reference dataset and in doing so helps to establish a link between an entity's long-lived temporal attributes and the various occasions and contexts in which the entity is involved (e.g. financial transactions with other entities) (Beynon-Davies, 2016; Eriksson & Agerfalk, 2010; Otjacques et al., 2007; Whitley et al., 2014).

The credibility and usefulness of an identification infrastructure (IDI) depends on the quality of the identification data that underpin its operation and this will affect the value this identification infrastructure provides to the entire ecosystem of users and systems that utilise it. This is particularly important when an identification infrastructure is highly embedded and linked to other infrastructures and practices. Consequently, maintaining high veracity can be seen as a cost that needs to be accounted for when evaluating the effectiveness and overall benefits (or return-on-investment) of the infrastructure (Abbasi et al., 2016).

Following the arguments above, we suggest that to arrive at a working identification infrastructure, a process needs to be established that maintains the quality of the reference data held by the infrastructure and identify what elements of data-quality issues are more important to address. Thus, the establishment of an infrastructure for reference data, such as identification data, calls for intermediated exertion of influence between two or more actors through the development of domain-wide rules that govern how references are to be associated with identification items and how high veracity of data can be ensured. We suggest that the success of the GLEIS project as an identification infrastructure should be examined through the achievement of this goal.

In turn, achieving such success depends on a high level of trust in the validity and reliability of the identification data included in the GLEIS by the various actors with a stake in the infrastructure. As such, we propose that it is vital to examine the dynamics that evolved between the different stakeholders involved in the development of the GLEIS identification infrastructure (regulators, policy makers, infrastructure sponsors and operators, and market participants) around different views they hold regarding data quality and how these different ways in which data quality was viewed and understood were reconciled and incorporated into the technical and organisational arrangements that form the GLEIS identification infrastructure. The way this was done in practice during the development of the GLEIS highlights the central importance of data and data quality in creating value for an identification infrastructure such as GLEIS (Abbasi et al. 2016).

From the above discussion about the importance of data quality to big data initiatives and the need for veracity to complement accepted data quality dimensions such as data volume, variety, and velocity, three key questions emerge:

1. How are data quality – and in particular data veracity – achieved in practice in an identification infrastructure that underpins a 'big data' initiative to regulate financial markets around the world?
2. How is an intersubjective understanding of what constitutes data quality achieved in practice and how does that condition the attitudes of infrastructure participants towards the infrastructure?
3. How do processes intended to ensure and maintain data quality in the GLEIS IDI relate to the capitalisation of data and making the resulting identification infrastructure valuable through greater adoption and the generating of network effects?

Research methods

Data collection

The main forms of data collected for the research were interviews and the documentation relating to the establishment of the Global Legal Entity Identifier System (GLEIS) from the sponsors of the GLEIF identification infrastructure such as the Regulatory Oversight Committee (ROC), the Commodities and Futures Trading Commission (CFTC), the Securities and Exchanges Commission (SEC), the Financial Stability Board (FSB), the G20, and the Global Legal Entity Identifier Foundation (GLEIF).

The aim in relation to the interviews was to speak to key individuals of relevance to the adoption of the LEI and/or establishment of the GLEIS from at least one organisation per stakeholder group. The duration of the interviews was between 35 minutes and 2 hours and aimed to gain an in-depth understanding of the workings of the GLEIS from the perspective of the interviewees and their organisations, in particular how the adoption of the LEI impacted the operations they were responsible for.

Overall, 32 individuals were interviewed in interview sessions totalling just under 38 hours and 36 documents relating to the establishment of the GLEIS and the development of a data quality assurance process by GLEIF read. Figure 1 provides a breakdown of the data collection by infrastructure participant groups.

[Figure 1 about here]

In addition to interviews and documents, the authors also undertook a day of participant observations at one of the LOUs responsible for issuing LEIs to entities applying for an identification number in order to understand better and at first-hand the issuing, verification, and validation process associated with the LEI.

For all the above data, collection followed expected academic research ethics provisions. Research participants were provided with an information sheet outlining the purposes of the research and a consent form which was signed and returned ahead of any further contact with the researchers. The consent forms outlined their rights as participants to anonymity and to withdrawing their participation – and any data resulting from that participation – at any time during the research. All the primary data collected and any secondary data resulting from the processing and analysis of this primary data were stored on and shared over a secure EU-approved academic data management facility to which only the researchers had access.

Data analysis

Interviews and documents content were coded in terms of key controversies/points of contention (Marres, 2004; Panourgias, 2015) using the NVivo software package as well as by type of IDI participant group in order to arrive at a taxonomy of infrastructure participants.

While a number of issues were identified through the first coding, through a second coding, first order coding was consolidated into a few overarching categories comprising of hierarchies of lower level coding clusters. This way it was possible to classify the various top-level controversies in terms of their relative importance according to the number of codes that comprised them, the depth of associated lower level codes that comprised them, and the extent to which a controversy spanned the entire set of IDI participant groups.

Based on this preliminary coding, a second phase of more targeted interviews in terms of participants and questions asked were undertaken focusing entirely on the key points of that controversy and the underlying issues relating to it and identified in the previous interviews from the coding.

Empirical account

Data quality processes in the GLEIS

Through a combination of interviews from LOUs and the GLEIF, relevant documents, and participant observation of the work done at an LOU in relation to the checking of applicant data against authoritative sources, the following description of the processes through which the quality of data in the GLEIS is assured was assembled (Langley, 1999).

The LEI issuing process (Figure 2) provides the first step in turning basic business data into LEI data. The applicant entity provides its legal name, legal form, address and corporate register to the LOU and this data is then checked by operatives at the LOU against at least one local authoritative source, usually the relevant business register of the applicant entity's jurisdiction.

[Figure 2 about here]

Once the quality of the data provided by the applicant has been checked by the LOU, an LEI is issued. From that point onwards, it is the responsibility of the registering entity and the LOU to ensure that the details of the identification data remain current. The LOU may do so of its own accord, as part of its mandatory periodical checks or it can be aided in this process by a challenge process (illustrated in Figure 3) whereby a challenger can use the GLEIF web site to report on an erroneous piece of data. Following this, the LOU is obliged to investigate the case and amend the data if necessary.

[Figure 3 about here]

Both elements of the data quality assurance process described above depend on a chain of relations between the legal entity applying for an LEI, the LOU issuing the LEI, and the GLEIF that oversees the LEI IDI and ensures the adherence of the LOUs to the terms and conditions of their master agreement and subsequent service level agreements with the GLEIS for the provision of LEI issuing services. These relations are illustrated in Figure 4 and are a key aspect of the process through which the usefulness and value of the LEI data is increased but also made dependent on the GLEIF and its view of quality.

[Figure 4 about here]

A key element in this regime is the way that data quality itself is defined. Across the different groups of infrastructure participants interviewed there were very differing views of what constituted 'data quality', how this quality may or may not be measured, and what implications any lack of quality – as they understood it – might have for the use and adoption of the GLEIS IDI. It was, therefore, a key strategic move on the side of the GLEIF to develop and put in place a methodology for assessing data quality into which a particular view and definition of data quality was inscribed.

Quality of data in the approach put in place by GLEIF is measured and then certified through an assessment of the data associated with the issued LEIs in terms of the conformance of this data with set of Data Quality Criteria (see Table 1).

[Table 1 about here]

Underpinning each of these criteria are numerous checks (more than 200 and growing) that measure the data in different ways (qualitative and quantitative) according to these criteria. Each of the quality criteria can have a number of checks associated with them. Furthermore, some of the checks are relational, where a certain check is performed if another check has a certain value. The outcomes of these checks are then aggregated and used to arrive at a quality score for the entire LEI dataset but also the subset of the data according to the LOU it was issued by. Any failure of check is recorded. It is through this process that each iteration of the LEI dataset is certified as being of a certain quality according to the definition assembled by GLEIF. The resulting data quality scores for both the entire LEI dataset, but also the data generated by the various LOUs, is published monthly on GLEIF's website.

[Figure 5 about here]

GLEIF's service-level agreements operationalise its definition of data quality. The incorporation of the measurement methodology into the operational governance of data quality influence the activities of other IDI participants and the sponsors and participants of adjacent infrastructures that may provide inputs to, or use aspects of, the GLEIS IDI.

This operationalisation aims to position GLEIF as the arbiter of the performance of the LOUs, and, more broadly, as a critical hub in the overall identification infrastructure. Firstly, the agreements between the LOUs and GLEIF make LOUs responsible for the data quality of both the LEI applicants (that is, in terms of maintaining their own identification data) but also, indirectly, of the authoritative sources used to verify the data. This way, the influence of the GLEIF starts to extend beyond its own immediate IDI and reaches throughout the entire identification infrastructure and beyond. For example, the service-level agreements are used to incentivise the LOUs to apply pressure on national business registers to improve the quality of their data which is use to check the correctness of legal entity data when it is submitted as part of a request for an LEI number.

Furthermore, the combination of agreements and methodology also enables the GLEIF to provide a degree of intervention into the operations of the LOUs by providing them guidance through relationship managers as to how recurring types of data quality problems (e.g. see 'Top 5 Failing Checks' in Figure 5) can be rectified through process and policy change recommendations which can be made very 'hard' through a possible revocation of the authorisation to issue LEIs. In fact, through forward-looking dimensions of the data quality measures (e.g. 'Quality Maturity Level' in Figure 5), the agreement-based relationship between GLEIF and LOUs even directs future action of the LOUs towards attaining higher data quality criteria. For example, setting up a very high goal that needs a continuous striving and improvement to attain, is conducive to GLEIF maintain themselves as the arbiter of data quality. This configuration of data quality is also aligned with the possible goal of maintaining GLEIF as an obligatory point of passage (Callon, 1984; Latour, 1987) in the nascent IDI.

Competing views on data quality

Overall, the analysis generated 15 high-level codes consolidating 118 lower level codes representing a total of 218 coding points. Of these, 39 of the lower level codes (45 coding points) related to the overall high-level category of 'Data' and of these, 31 (41 coding points) related to the sub-category of 'Data quality'. The top issues within the 'Data quality' category related to 1) the inclusion of Level 2 ownership data, their compilation, and the implications this would have for IDI participants and 2) the maintenance and validity of LEI data.

Taken across the 'Data quality' subcategory, our empirical examination of these codings and the underlying coding points indicates that across the different groups of actors involved in the identification infrastructure there were very differing views of what constituted 'data quality', and what implications any lack of quality may have.

As described in the previous section, quality of data in the methodology that GLEIF developed is measured and then certified in a process whereby the data is examined through assessing its conformance with an array of descriptors, known as Data Quality Criteria, aimed at capturing different dimensions of quality (see Table 1). This view of data quality is informed by an engineering/manufacturing approach to 'quality' along the lines of the Total Quality Management, whereby the ultimate aim is to produce and maintain high data quality that strives for zero errors across the entire dataset. This view can be understood as a process-oriented view towards data quality. In essence, GLEIF's concept of data quality regards it as an indication of the suitability of organisational structures and practices associated with the assembling of the identification data and, following this view, aims to create a trusted infrastructure for data quality; a structure that generates and guarantees reliance on the infrastructure.

On the other hand, financial service providers, such as investment banks, who are both users and producers of the identification data but may also act as intermediaries for the LEI needs of their clients, regard data quality from a commercial perspective. In particular, the service providers focus primarily on the sections of the identification data that their commercial experience indicates are likely to be of high demand and point their resource towards ensuring that this data remains of high quality. As a result, less attention is paid to the rest of the data, whose quality deteriorates over time, as it is not updated and verified.

The focus on data quality through the prism of revenue generation also affected another element in relation to how LEIs are treated – as a way to reduce operational costs. A typical comment from our interviewees follows:

Operationally, the creation and maintenance of a universal identifier for companies' names is going to simplify and reduce the cost of maintaining identifiers [because] in order to create cross platform compatibility, they need to rely on fuzzy matching, which is expensive, not accurate and time consuming.

Interviewees regarded LEIs as a viable solution to the situation where multiple identifiers existed both within their organisation and across organisations. In such a scenario the adoption of the LEI would be of benefit, but its viability relies on maintaining a high quality of data across all identification items, which is a costly process with uncertain revenues. Given this difficulty, interviewees regarded the maintenance of high -quality LEIs primarily as a regulatory requirement and, consequently, focused on the costs it may save once in place, rather than as an opportunity to generate revenues.

In particular, when envisioning the operational environment of LEIs, many of our interviewees preferred to talk about the goals of the LEI system once these goals were achieved, but noticeably leave out of the discussion how these would come about. For example, interviewees repeatedly mentioned how beneficial it would be for their organisations when there is a common language in place for the identification of entities, but when asked about what their organisation can do to help bring about this desired outcome, many interviewees indicate that this is a regulatory challenge and, as such, it cannot be the task of any single organisation to realise this outcome.

Another important area where differences regarding the conceptualisation of data quality are revealed concerns the motivations of the actors involved in operationalising data quality. A difficulty that many of our interviewees envisioned related to the fact that financial institutions typically provide trade execution services to their customers, which makes the customers in need of LEI registration. Such situations would leave the service provides, if they register their customers to the LEI system, responsible for the accuracy and validity of the identification details of, potentially, thousands of trading entities, many of which may terminate, as time passes, their commercial relations with the service providers. This potential obligation was one that virtually none of our interviewees from service-providing financial institutions wished to undertake, as this was seen as 'bottomless pit' of legal and operational burdens. Furthermore, there were also worries about providing third parties (their current clients) with 'a licence to trade' and all the implications this carries in terms of reputational and regulatory risks for them from these clients and their activities.

In summary, GLEIF focuses its design and enforcement efforts on the LOUs, which are rewarded for auditing the quality of identification information. These show that discrepancy between how the regulators see the implementation of LEIs and how financial institutions see the exact set of activities. While the regulators focus on data quality from the perspective of the LOUs, which are focused primarily on registering LEIs and assuring data quality, the end users view the LEI process as an auxiliary of their financial service provision, but one that generates no revenue and carries with it the imminent possibility of significant costs.

GLEIF: from identifier developer to data quality certifier

While studying the development of the GLEIS in conjunction with the unfolding of the 'data quality' controversy, another aspect of 'data quality' surfaced. This was that the development and adoption of an intersubjective way to define and measure data quality within the GLEIS started to have effects beyond the confines of the GLEIS IDI. The definition and measurement of data quality within the GLEIS helped make the GLEIF view of data quality and its assessment an industry yardstick against which the data quality of other existing financial and regulatory IDIs can be assessed. In the face of a lack of alternative definitions and quality assurance methodologies in these related IDIs, GLEIF's definitions and methods are being used to assess data quality in other IDIs where legal entity identification is central (e.g. Business Identification Codes (BICs) in the banking industry and legal entity identification associated in relation to the Foreign Account Tax Compliance Act (FATCA) in the United States).

Looking at the evolution of GLEIF as an institution in parallel to the putting in place of the data quality assurance criteria and processes for the GLEIS, the role of GLEIF has been changing, from one of overseeing the operation of the GLEIS identification infrastructure to one of becoming a wider arbiter of data quality for entity identification in other financial and accounting infrastructures.

[Figure 6 about here]

Discussion

Veracity and the dimensions of data quality

Our empirical account reveals significant differences between the way data quality is defined in the context of the GLEIS, and the way data quality is discussed in the existing big data literature (Abbasi et al., 2016; Rubin & Lukoianova, 2013). In relation to data quality and its definition in the area of 'big data', we argue, drawing from on our empirical findings, that the meaning of 'data quality' is situated in the use-context of the 'big data' initiative examined, raising questions about the applicability of predetermined dimensions of data quality, such as the 4Vs (Abbasi et al., 2016; Rubin & Lukoianova, 2013). We find that data quality and its definition and measurement is drawn from the use contexts which reflect the concerns and interests of stakeholders and, as such does not fit the 'top-down' approach implied in the Big Data literature. While the notion of dimensions of quality is there in the empirical case, those dimensions are more specific to the context of the GLEIS and the concerns and interests of the various infrastructure participants. Rather than the 4Vs of velocity, variety, volume, and veracity (Abbasi et al., 2016; Rubin & Lukoianova, 2013), three high-level 'characteristics' of *openness* (data being accessible to everyone), *reliability* (GLEIS data comparing favourably to other existing standards), and *trust* (continuous improvement of the data quality), were defined by the GLEIF. These three high-level characteristics were then broken down into 12 quality criteria (see Table 1) and then more than 200 checks used to measure data quality and compare the performance, in terms of data quality, of LOUs but also of the GLEIS as a whole.

Giving meaning to 'data quality'

It is a key finding from our empirical study that data quality is a dynamic on-going sociotechnical achievement that reflects the concerns and interests of stakeholders touched by the data rather than a stable final state. It is therefore crucial to understand how an intersubjective understanding of the meaning of 'data quality' is arrived at and what implications this has for the attitudes those touched by this exhibit towards the 'big data' solution developed.

To understand how such alignment of interests may come about, we refer again to data veracity. Data veracity is related to the usefulness of the data in the organization. Indeed, As Abbasi et al (2016) point out 'with organizations treating data as a primary asset, assessing and, in some cases, quantifying the value of volume and variety relative to the costs of veracity becomes of paramount importance to evaluate the effectiveness of big data investments'. The case of GLEIS's data quality shows that in the case of identification data, recognising the data items as valuable comes as a result of multiple stakeholders aligning their incentives and thus facilitating the viability of the infrastructure that maintains it in order to invest in infrastructures that would maintain a high quality of data.

Maintaining such alignment among the actors is difficult, since the different actors follow conflicting logics. We distinguish these as being the followers of either a *commercial logic*, such as the financial service providers, in contrast to GLEIF, which follows a *regulatory* logic. Actors that follow a commercial logic, as our case indicates, wish to minimise costs in relation to data veracity if they cannot identify how increased data quality would lead to economic benefits. Their focus is more on maintaining the high quality of the parts of the data that are likely to generate revenue for them. As a result, only a relatively small subset of the data would be maintained at high quality.

In contrast, actors who follow a *regulatory logic* in relation to identification data aim for the overall data quality to increase so as to reduce the risk stemming from having inaccurate data items. These two logics highlight the conflicting motivations and divergent identification needs of different participants in an IDI and the different views they have regarding identification data quality and how that quality should be achieved and maintained.

In the case presented, we understand the decision of GLEIF to delegate the maintenance of LEI data quality to LOUs as a way to alleviate the threat to the stability of the IDI. GLEIF wished to associate data quality maintenance with actors who follow a commercial logic, but ones whose commercial interests are aligned directly with the maintenance of the entire corpus of identification data at a high quality, the LOUs. The LOUs, in contrast with the end-users of the LEIs, expected to make financial gains from providing verification and registration services to end users. The LOUs' ability to generate revenues were dependent on the end users' trust in the quality of data.

Moreover, GLEIF's restrictive data quality regime and the fact that the LOUs cannot know in advance which section of the identification data is likely to be more in demand, motivated the LOUs to high quality data universally. GLEIF's definition of data quality, once operationalised into the service-level agreements with the LOUs making the latter responsible for the data quality of the LEIs, brought about an implicit agreement about how data quality is to be maintained and contributed to an improved stability of the IDI. In more generalised terms, we can argue, on the basis of our findings, that GLEIF made itself an indispensable part of the infrastructure, what is referred to as an 'obligatory point of passage' (Callon, 1984, 1987; Latour, 1987) by framing the quality-related concerns of the various actors involved in IDI and inscribing this framing into a techno-institutional data quality methodology.

Processes of 'capitalisation through certification'

From both the regulatory and commercial perspectives outlined above, the definition of data quality through the GLEIF quality criteria (Table 1) and the putting in place of a quality assurance process through which the performance of the LOUs is assessed and managed have important future implications.

First, it provides a way of giving value to the basic business data owned by those being identified through the new IDI. Even if substantially the same as the raw company data, the entity data that is certified through the GLEIS is more valuable than the raw company data outside the GLEIS. Furthermore, having in place the GLEIF quality criteria and the GLEIF quality control process makes it possible for future regulators to control the evolution of the infrastructure by adjusting the criteria and/or the quality control and certification processes. Second, the definition and measurement of data quality within the GLEIS also establishes a benchmark against which other IDIs in adjacent areas of financial and regulatory activity can be assessed.

Given the current state where there is lack of alternative definitions and quality assurance methodologies, GLEIF's definitions stand a good chance of becoming a more widely adopted standard. The gap that exists in defining and assessing data quality in other areas of financial and regulatory activity where regulators are seeking data-driven approaches to regulation has enabled the direct or indirect linking of GLEIS with other related informational infrastructures, increasing further the potential value of both the GLEIS and of GLEIF-certified entity data. The development of a process of 'capitalisation by certification', therefore, not only gives usefulness and value to the data of an IDI, but, in the absence of other such quality assurance processes in related areas of identification, can generates potential network effects through an association with other IDIs that do not have similar processes in place.

Conclusion

This article addresses the wider growing interest in the big data and analytics literature (Abbasi et al., 2016; Agarwal & Dhar, 2014) regarding how data quality can be defined and measured and how this relates to the dynamics of how data gains value (Langley & Leyshon, 2017; Thrift & Leyshon, 2007; Millo et al., 2019) and the processes through which the value of data is captured and realised (Doganova & Muniesa, 2015; Najjar & Kettinger, 2013). It relates this to the extensive literature in information systems research on information infrastructures (Ciborra & Hanseth, 1998; Constantinides & Barrett, 2005, 2014; Hanseth & Braa, 2000; Hanseth & Monteiro, 1997; Hanseth et al., 1996; Monteiro & Hanseth, 1996), and in particular identification infrastructures (Camp, 2004; Clarke, 1994; Eriksson & Agerfalk, 2010; Whitley et al., 2014). It also addresses the recent interest in both academic literature (Hu, Zhao, Hua, & Wong, 2012) and among practitioners (Davies, 2019; van Steenis, 2019) about the use of ‘Big Data’ in financial services regulation in what Davies (2019) refers to as ‘regtech’ solutions for the regulation of increasingly digitised and global financial activities.

Due to the way that ‘big data’ can touch on a multitude of stakeholders, both in terms of the use but also the generation of data, being in a position to define in an intersubjective way what is meant by ‘data quality’ and how this is measured can provide a degree of control to the promoters of a ‘big data’ initiative in terms of the design and evolution of the initiative vis-à-vis the resistances and/or buy-in of affected stakeholders. This is because good data quality that is recognised intersubjective as such by those touched by it fosters network effects in terms of adoption and usage that add value to both the data itself but also the infrastructures through which the data in question is compiled and used.

All the points highlighted in this concluding section point towards certain ‘infrastructural’ aspects of ‘big data’ initiatives such as the one presented. This gives new currency to the literature on infrastructures and infrastructures-making built on the conceptualisations developed by Star (1999); (1996) and which can contribute much to the study of phenomena such as Big Data (Kornberger et al., 2019). Along these lines, the GLEIS case shows that organising financial regulation on a global scale using data is likely to be as much about giving meaning to categories and concepts such as ‘data quality’ and developing ways to measure and asses these as it is about putting in place regulations mandating or prohibiting certain activities and practices. Furthermore, such apparently small technicalities, once put in place, can spread, having an influence on adjacent areas of activities with common date needs and uses.

This is particularly the case with IDIs, which can act as linkages between many diverse datasets enabling cognition and action in many other areas beyond those for which the IDI might have been developed for in the first place. It will be paying attention the sociotechnical ‘life’ of data and studying data as such, instead of treating data as a purely technical issue, that new data-driven ways of organising that are entering more and more aspects of modern life can be understood and studied better.

Acknowledgements

We would like to thank all the participants in the research for their time and support with regard to interviews and data gathering and in particular Paul Janssens from SWIFT for his help with setting-up the fieldwork undertaken.

Funding Acknowledgement

The project has received funding from the European Union’s Horizon 2020 research and innovation programme [Grant No. 727145].

References

Abbasi, A., Sarker, S., & Chiang, R. H. (2016). Big data research in information systems: Toward an inclusive research agenda. *Journal of the Association for Information Systems, 17*(2), I.

Agarwal, R., & Dhar, V. (2014). Editorial—Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research. *Information Systems Research, 25*(3), 443-448. doi: 10.1287/isre.2014.0546

Beynon-Davies, P. (2016). Form-ing institutional order: The scaffolding of lists and identifiers. *Journal of the Association for Information Science and Technology, 67*(11), 2738-2753. doi: doi:10.1002/asi.23613

Brynjolfsson, E., Hitt, L. M., & Kim, H. H. (2011). Strength in numbers: How does data-driven decisionmaking affect firm performance? *Available at SSRN 1819486*.

Callon, M. (1984). Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay. *The Sociological Review, 32*(1_suppl), 196-233. doi: 10.1111/j.1467-954X.1984.tb00113.x

Camp, J. L. (2004). Digital identity. *IEEE Technology and Society Magazine, 23*(3), 34-41. doi: 10.1109/MTAS.2004.1337889

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: from big data to big impact. *MIS Quarterly, 36*(4), 1165-1188.

Ciborra, C. U., & Hanseth, O. (1998). From tool to Gestell: Agendas for managing the information infrastructure. *Information Technology & People, 11*(4), 305-327. doi: doi:10.1108/09593849810246129

Clarke, R. (1994). Human Identification in Information Systems: Management Challenges and Public Policy Issues. *Information Technology & People, 7*(4), 6-37. doi: doi:10.1108/09593849410076799

Constantinides, P., & Barrett, M. (2005). Approaching information infrastructure as an ecology of ubiquitous sociotechnical relations *Designing ubiquitous information environments: Socio-technical issues and challenges* (pp. 249-260): Springer.

Constantinides, P., & Barrett, M. (2014). Information infrastructure development and governance as collective action. *Information Systems Research, 26*(1), 40-56.

Doganova, L., & Muniesa, F. (2015). Capitalization Devices. *Making things valuable*, 109.

Eriksson, O., & Agerfalk, P. J. (2010). Rethinking the Meaning of Identifiers in Information Infrastructures. *Journal of the AIS, 11*(8), 433-454.

Financial Stability Board. (2012). A Global Legal Entity Identifier for Financial Markets.

Fleming, M. J., & Sarkar, A. (2014). The Failure Resolution of Lehman Brothers. *Federal Reserve Bank of New York Economic Policy Review, 20*(2), 175.

Fox, C., Levitin, A., & Redman, T. (1994). The notion of data and its quality dimensions. *Information Processing & Management, 30*(1), 9-19. doi: https://doi.org/10.1016/0306-4573(94)90020-5
Glaeser, E. L., Kominers, S. D., Luca, M., & Naik, N. (2018). Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry, 56*(1), 114-137.
Günther, W. A., Mehrizi, M. H. R., Huysman, M., & Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. *The Journal of Strategic Information Systems, 26*(3), 191-209.
Hanseth, O., & Braa, K. (2000). Who's in control: Designers, managers-or technology? Infrastructures at norsk hydro. *From control to drift: The dynamics of corporate information infrastructures*, 41-55.
Hanseth, O., & Monteiro, E. (1997). Inscribing behaviour in information infrastructure standards. *Accounting, Management and Information Technologies, 7*(4), 183-211. doi: https://doi.org/10.1016/S0959-8022(97)00008-8
Hanseth, O., Monteiro, E., & Hatling, M. (1996). Developing Information Infrastructure: The Tension Between Standardization and Flexibility. *Science, Technology, & Human Values, 21*(4), 407-426. doi: 10.1177/016224399602100402
Harris, J., & Davenport, T. (2007). Competing on analytics. *Boston, MA: Harvard Business School Publishing*.
IBM. (2012). The Four V's of Big Data.   Retrieved September 2019, 2019
International Standards Organisation. (2012). ISO 17442:2012: Financial services — Legal Entity Identifier (LEI).   Retrieved 23/09/2018, 2018, from https://www.iso.org/obp/ui/#iso:std:iso:17442:ed-1:v1:en
Jenkinson, N., & Leonova, I. S. (2013). The importance of data quality for effective financial stability policies - Legal entity identifier: a first step towards necessary financial data reforms. *Financial Stability Review*(17), 101-110.
Kornberger, M., Bowker, G. C., Elyachar, J., Mennicken, A., Miller, P., Randa Nucho, J., & Pollock, N. (2019). *Thinking Infrastructures* (Vol. 62). UK: Emerald Publishing.
Lai, A., & Mordel, A. (2012). The Resolution of Systemically important Financial institutions. *Financial System Review*, 37.
Langley, A. (1999). Strategies for Theorizing from Process Data. *The Academy of Management Review, 24*(4), 691-710.
Langley, P., & Leyshon, A. (2017). Platform capitalism: the intermediation and capitalisation of digital economic circulation. *Finance and society., 3*(1), 11-31.
Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*: Harvard university press.
LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Management Review, 52*(2), 21.
Legal Entity Identifier Regulatory Oversight Committee. (2015). The Global LEI System and regulatory uses of the LEI.

Liu, Y. (2014). Big data and predictive business analytics. *The Journal of Business Forecasting, 33*(4), 40.

Marres, N. (2004). Tracing the trajectories of issues, and their democratic deficits, on the web: the case of the Development Gateway and its doubles. *Information Technology & People, 17*(2), 124-149.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., & Barton, D. (2012). Big data: the management revolution. *Harvard Business Review, 90*(10), 60-68.

Monteiro, E., & Hanseth, O. (1996). Social Shaping of Information Infrastructure: On Being Specific about the Technology. In W. J. Orlikowski, G. Walsham, M. R. Jones & J. I. Degross (Eds.), *Information Technology and Changes in Organizational Work: Proceedings of the IFIP WG8.2 working conference on information technology and changes in organizational work, December 1995* (pp. 325-343). Boston, MA: Springer US.

Nagle, T., Redman, T., & Sammon, D. (2017). Only 3% of companies' data meets basic quality standards. *Harvard Business Review* https://hbr. *org/2017/09/only-3-of-companies-datameets-basic-quality-standards*.

Najjar, M. S., & Kettinger, W. J. (2013). Data Monetization: Lessons from a Retailer's Journey. *MIS Quarterly Executive, 12*(4).

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine, 375*(13), 1216.

Otjacques, B., Hitzelberger, P., & Feltz, F. (2007). Interoperability of E-Government Information Systems: Issues of Identification and Data Sharing. *Journal of Management Information Systems, 23*(4), 29-51. doi: 10.2753/MIS0742-1222230403

Panourgias, N. S. (2015). Capital markets integration: A sociotechnical study of the development of a cross-border securities settlement system. *Technological Forecasting and Social Change, 99*, 317-338. doi: http://dx.doi.org/10.1016/j.techfore.2015.02.014

Park, S. H., Huh, S. Y., Oh, W., & Han, S. (2012). A social network-based inference model for validating customer profile data. *MIS Quarterly: Management Information Systems, 36*(4), 1217-1238.

Phillips-Wren, G., & Hoskisson, A. (2015). An analytical journey towards big data. *Journal of Decision Systems, 24*(1), 87-102.

Redman, T. C. (1995). Improve data quality for competitive advantage. *MIT Sloan Management Review, 36*(2), 99.

Redman, T. C. (2013). Data's credibility problem. *Harvard Business Review, 91*(12), 84-88.

Redman, T. C. (2016). Bad data costs the US $3 trillion per year. *Harvard Business Review, 22*, 11-18.

Rubin, V., & Lukoianova, T. (2013). Veracity roadmap: Is big data objective, truthful and credible? *Advances in Classification Research Online, 24*(1), 4.

Star, S. L. (1999). The Ethnography of Infrastructure. *American Behavioral Scientist, 43*(3), 377-391. doi: 10.1177/00027649921955326

Star, S. L., & Ruhleder, K. (1996). Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. *Information Systems Research, 7*(1), 111-134. doi: 10.1287/isre.7.1.111

Thrift, N., & Leyshon, A. (2007). The capitalisation of almost everything: the future of economy and finance. *Theory, Culture and Society, 24*, 79-115.

Warren Jr, J. D., Moffitt, K. C., & Byrnes, P. (2015). How Big Data will change accounting. *Accounting Horizons, 29*(2), 397-407.

Whitley, E. A., Gal, U., & Kjaergaard, A. (2014). Who do you think you are? A review of the complex interplay between information systems, identification and identity. *23*(1), 17-35. doi: 10.1057/ejis.2013.34

Zeng, D., Chen, H., Lusch, R., & Li, S.-H. (2010). Social media analytics and intelligence. *IEEE Intelligent Systems, 25*(6), 13-16.

Zhang, J., Yang, X., & Appelbaum, D. (2015). Toward effective Big Data analysis in continuous auditing. *Accounting Horizons, 29*(2), 469-476.

Data Quality Criteria	Description
Accuracy	Data is free of identifiable errors and conforms with an authoritative source; Represents correctly real-world objects
Accessibility	Ease and legality of access
Completeness	All required occurrences are populated.
Comprehensiveness	All required data items are included and all possible scope of the data is collected
Consistency	Unique piece of data holds the same value across multiple data sets.
Currency	Data are up-to-date
Integrity	Conforms to defined data relationship rules
Provenance	Pedigree of a data property value
Representation	Fit of format of data to its intended use
Timeliness	Data is available when it is required.
Uniqueness	Distinct values of data elements appear only once.
Validity	Data value conforming to its domain value

Table 1. Data Quality Criteria, adapted from GLEIF documents (2017).

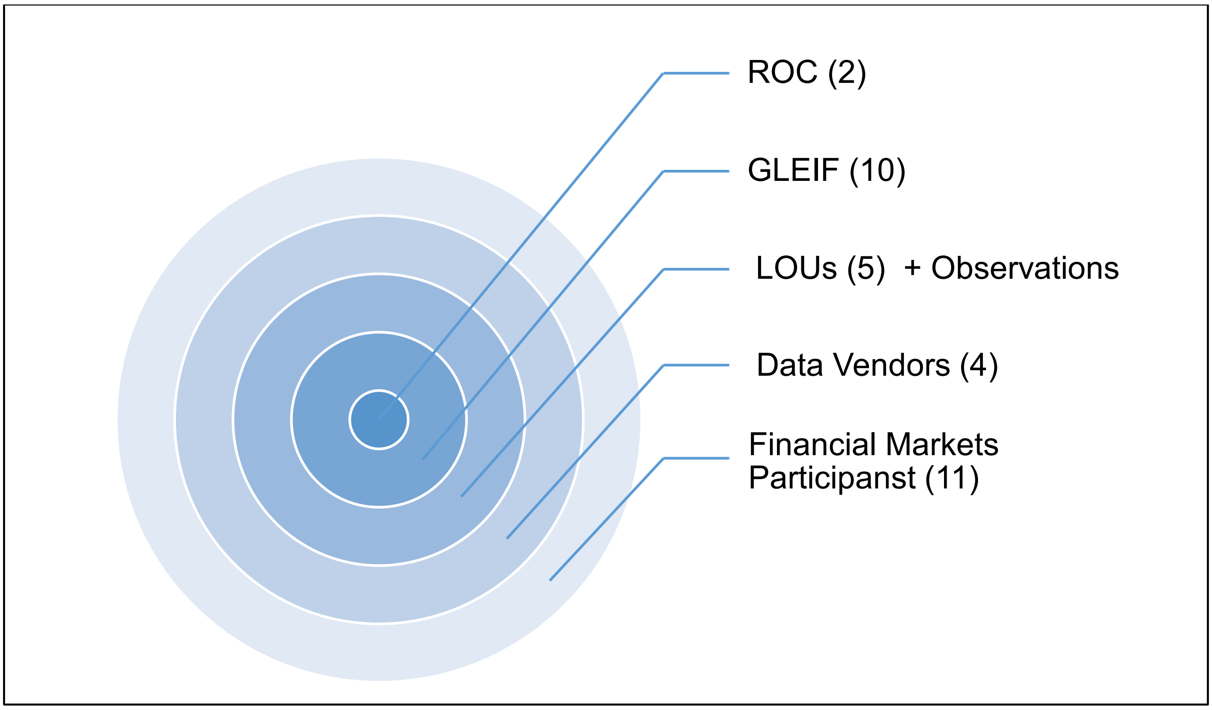

Figure 1. Data collection - breakdown by infrastructure participant group.

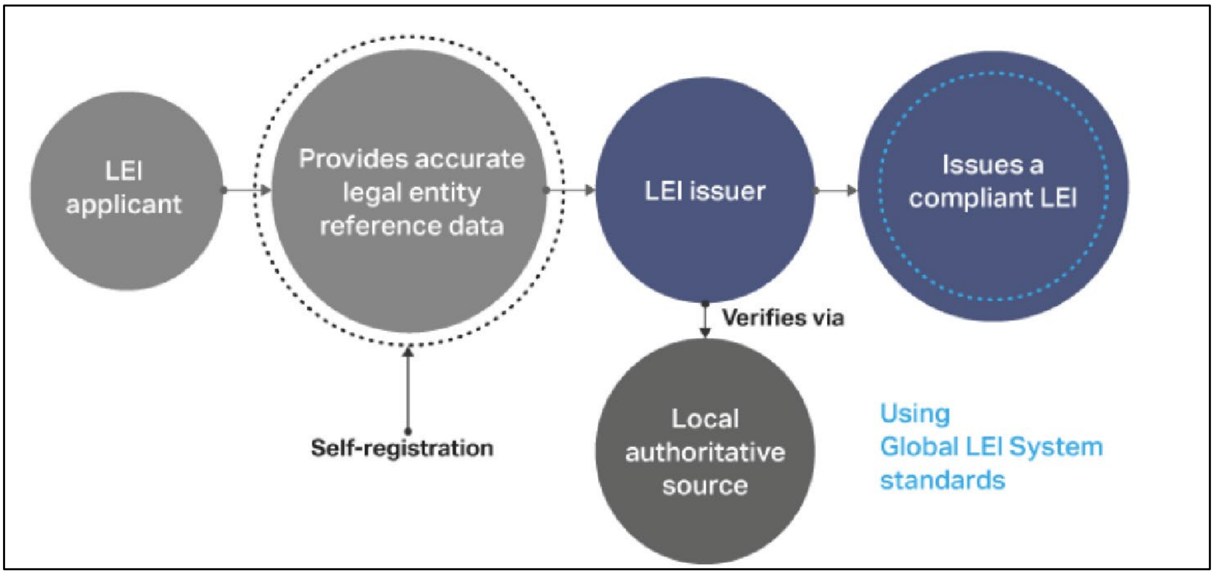

Figure 2. The LEI issuing process (source GLEIF).

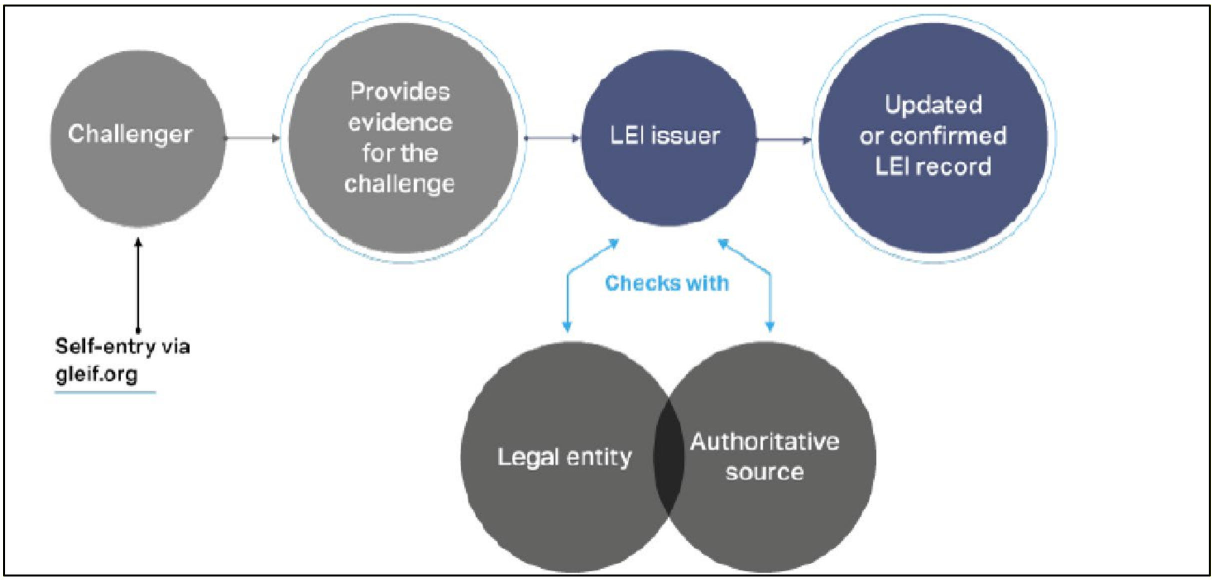

Figure 3. Improving data accuracy through the LEI challenge function (source: GLEIF).

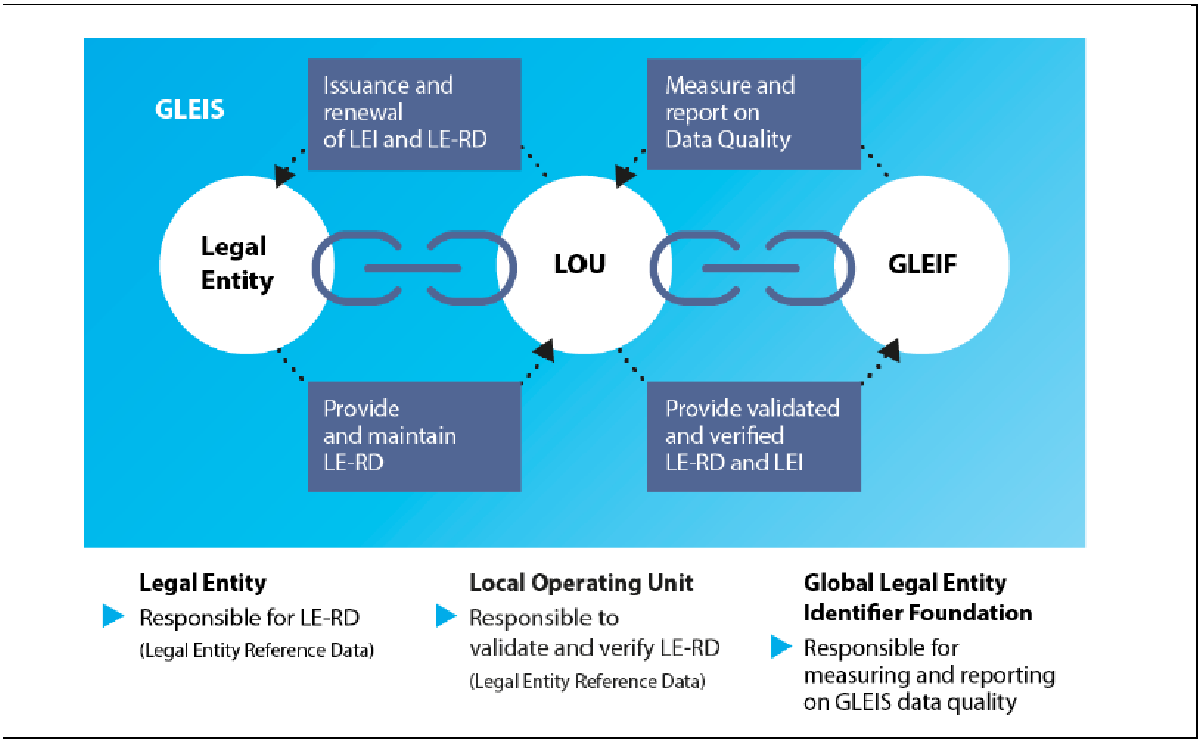

Figure 4. Data quality control process (source GLEIF).

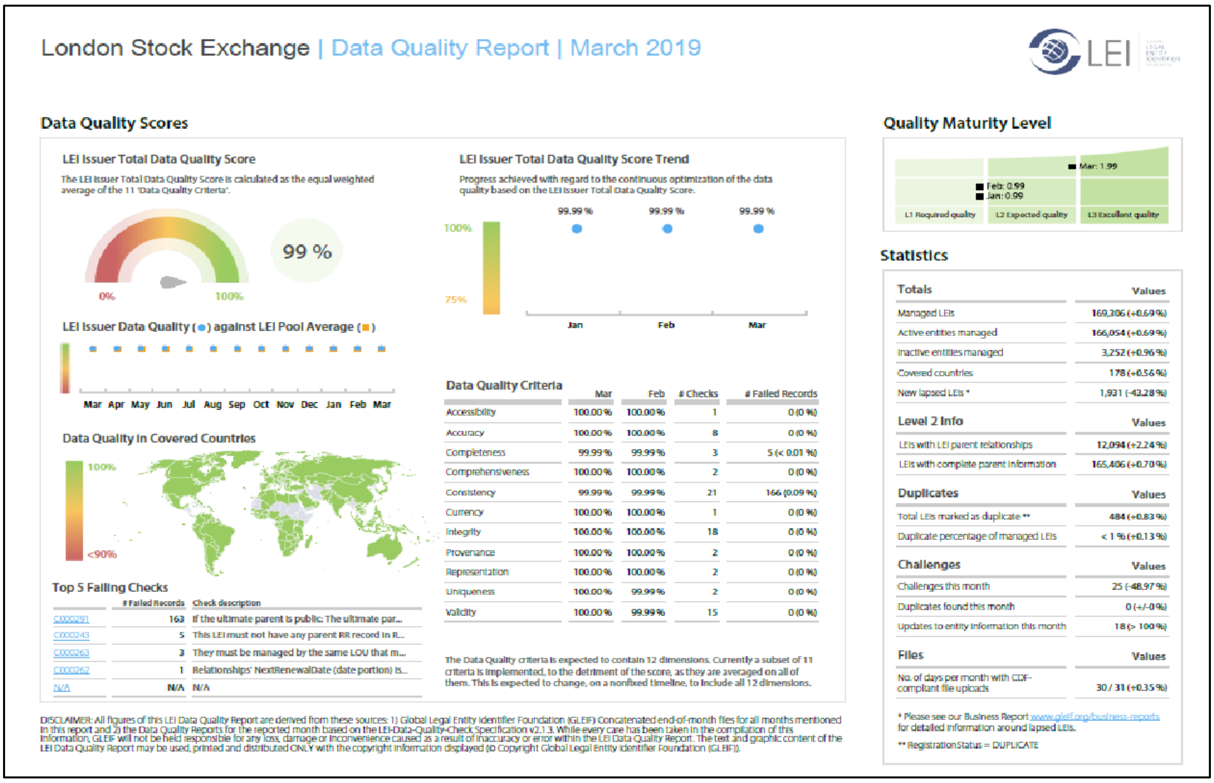

Figure 5. Example of an LOU data quality report (source: GLEIF).

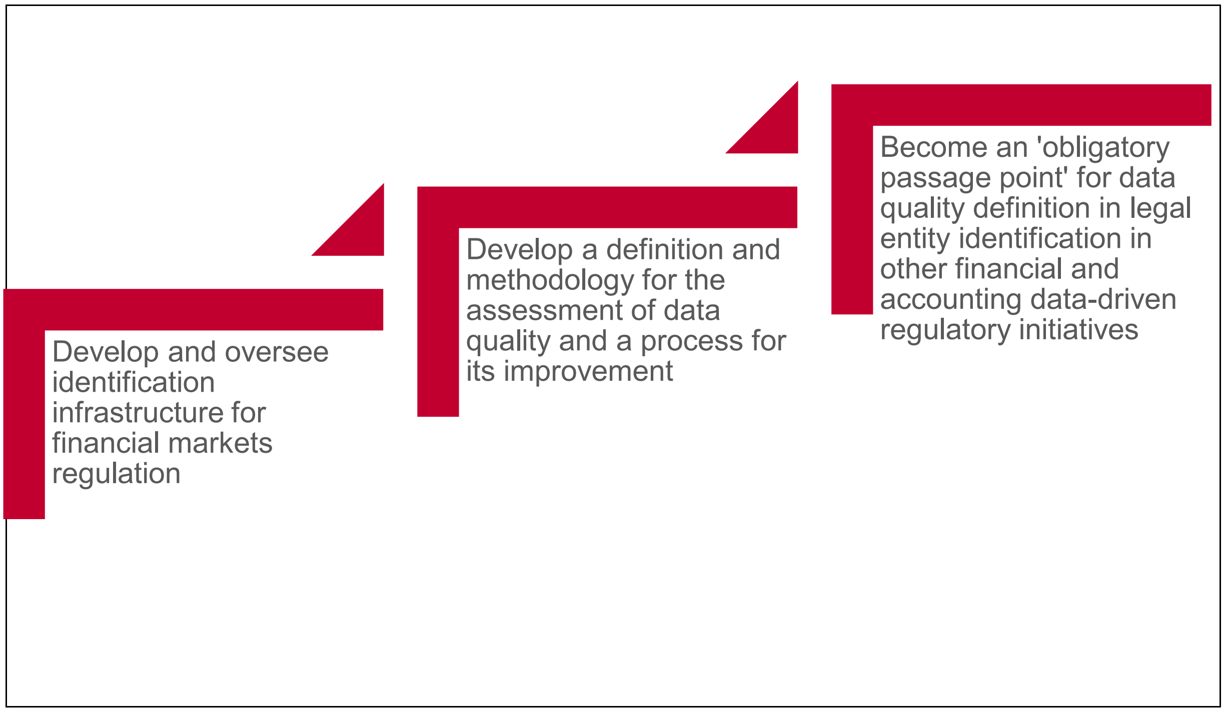

Figure 6. The changing role of GLEIF in financial and regulatory IDIs.