Grant Agreement Number: 825225

Safe-DEED

www.safe-deed.eu

D4.3 Report on context-aware and context-unaware valuation

Deliverable number	*D4.3*
Dissemination level	*Public*
Delivery date	*30 November 2020*
Status	*Final*
Author(s)	*Mihnea Tufiș*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825225.

Changes Summary

Date	Author	Summary	Version
23.10.2020	Mihnea Tufiș	First draft	0.1
12.11.2020	Dieter Decraene, Alessandro Bruni	Internal Review	0.2
16.11.2020	Alexandros Bampoulidis	Internal Review	0.3
17.11.2020	Mihnea Tufiș	Integrated reviews	1.0
18.11.2020	Mihnea Tufiș	Rearrange annex	1.1

Executive summary

This document is deliverable D4.3 – Report on the context aware & context-unaware valuation. It is an extensive review of the literature on the topic of data valuation methods. It starts from a tentative definition of data value around several key areas: contexts, data quality, privacy, aggregation and reporting. It also discusses the properties that make data difficult to assess and brings valuable examples from data valuation applied to personal data.
Focusing on the central notion of data quality, the document reviews a number of data quality assessment methodologies, discussing the diversity of data quality dimensions that they employ and the metrics that support their operationalisation. The report concludes with a discussion on the challenges of aggregating these aspects under a composite measure, and how reporting through certification or impact-based narratives can be a feasible alternative.

Table of Contents

List of Figures

Abbreviations

CRM	Customer Relationship Management
DNL	Data Nutrition Label
DQA	Data Quality Assessment
DQM	Data Quality Measurement
EU	European Union
FNET	Forthnet
GDP	Gross Domestic Product
GDPR	EU General Data Protection Regulation
HMRC	Her Majesty's Revenue & Customs
ICT	Information and Communication Technologies
IFAG	Infineon AG
MIT	Massachussetts Institute of Technology
NSI	National Statistics Institute
p2p	peer-to-peer
RSA	Research Studio Austria
US	United States of America
VAT	Value Added Tax
WP	Work Package

Abbreviations of data quality assessment methodologies:

AIMQ	A Methodology for Information Quality Assessment
AMEQ	Activity-based Measuring and Evaluating of Product Information Quality
CDQ	Complete Data Quality
CIHI	Canadian Institute for Health Information
COLDQ	Cost-Effective of Low Data Quality
DaQuinCIS	Data Quality in Cooperative Information Systems
DQA	Data Quality Assessment
DWQ	Data Warehouse Quality
IQM	Information Quality Measurement
ISTAT	Italian National Bureau of Census
QAFD	Quality Assessment of Financial Data
TDQM	Total Data Quality Management
TIQM	Total Information Quality Management

1 Introduction

Up until 2009, energy and oil companies were dominating the top-10 most valuable firms; a decade later data-centric companies, such as Microsoft, Amazon, Apple, Alphabet (Google's parent company) are almost exclusively sharing the top-5, with Facebook, Alibaba or Tencent trailing not too far behind in the top-10[1]. The global economy is relying increasingly more on data, with businesses adopting data-enabled decision-making practices in the form of analytics or machine learning. So much so, that this has reshaped the paradigm for data production and consumption – including the perception of data as an asset, subject to transactions, the subsequent appearance of new stakeholders whose activity is based on the acquisition, re-packaging and selling of data sets, and finally the steady emergence of data markets. In this context, a question is becoming increasingly pervasive: *what is the value of my data*?

To answer this question, we first need to have a proper understanding of what is "data value". While the ranking just presented makes it clear that data generates value, the mechanisms in which this happens are still very much unclear. Organisations are only starting to think about the necessity to formalise these concepts. Up until now, preoccupations around the value of data were only triggered by large impact events – mergers and acquisitions, bankruptcy, data transactions, data breaches – which is perhaps why comparisons between data and other commodities (oil, gold) or intangible assets have become so common.

A tentative definition by Short and Todd [76] refers to the value of data as the composite between the value of the asset itself, the value resulting from its use and its expected or future value.

Our view is that a practical definition of data value refers to four elements:

i. the dependency to the context in which data is used;

ii. the qualitative assessment of data (both intrinsic and contextual);

iii. the performance / usability of data given its purpose, as stated in the context; and

iv. a method for aggregating or reporting on the value of data, such that the result is actionable.

In this report, we set to explore each of these components that contribute to the value of data, resulting from our proposed data valuation process. Assigning a price tag to a dataset is the ultimate goal of business stakeholders. In Section 2 we take a look at previous approaches to assigning an economic value to data, including a separate discussion on data brokers and data markets for personal data. Section 3 is dedicated to the abstract notion of valuation contexts and potential approaches that assisted us with our context establishing process and may ultimately allow a proper formalisation of the notion. In Section 4 we present in great detail the central topic of data quality assessment, explore a variety of methodologies for data quality assessment and build a powerful toolkit consisting of data quality dimensions and data quality metrics, to support our quality assessment goal. Section 5 is dedicated to a brief discussion about the challenges of factoring data privacy in a data valuation process. Section 6 circles back to data quality measures and discusses the difficulties of aggregating data quality measures and ultimately reporting the value of data.

1.1 The Difficulties of Data Valuation

First, we would like to acknowledge the work that the Open Data Watch has put into compiling the Value of Data Inventory[2], a list of articles and reports on the topic of data valuation. Referring to the Inventory in her report on the "Value of Data" [78], Slotin observes how "striking [it is] that among these diverse perspectives, each author is grappling in their own way with the implications of data as a new economic asset, and yet there appears to be little consensus on how best to measure its value. One

[1] The rankings we refer to are retrieved from Wikipedia's compilation of Financial Times Global 500 lists from 1996-2019. URL: https://en.wikipedia.org/wiki/List_of_public_corporations_by_market_capitalization

[2] https://docs.google.com/spreadsheets/d/1QRNZUKIrwKxq7J6EEfA6fRLpjYUevaNDpXMbwqx_Ogw/edit\#gid=37279104

thing they can agree on is that measuring the value of data – and making [a] case for investing in data – is very difficult".

Attempts to put a price tag on data have failed thus far, since analogies with either tangible (oil) or intangible assets (patents, intellectual property) break at the point where the mapping between features and assigned value becomes less clear. And perhaps this is normal since rules that apply to old commodities possibly don't even apply to this new kind of resource.

In order to understand the difficulties of assessing the value of data, Mawer starts from breaking down the elements of the data value chain [26]. Mawer [61] shows the progression from raw data to action and potential value, by going through the sequential stages of the data lifecycle (discover, ingest, process, persist, integrate, analyse, expose) (see Fig. 1).

Figure 1 : Data value chain [61]

The author is then able to identify the following reasons for which assessing data is difficult:

1. The necessity to complete the value chain to observe the value of data.

   Obviously, the data becomes increasingly valuable as it moves through the value chain towards the end-product. However, it is difficult to establish its intermediate value and report it with respect to the estimated value of the end-product.

   As in the case of traditional "goods", there are "first-copy" costs attached to the production of the initial "information good"; however, unlike the former, the marginal cost for replicating data is approaching 0. This means there is an upfront investment in discovering, collecting, and processing the data before one can even evaluate an often-undefined end-product. Therefore, there is a risk attached to the production of data products and quantifying this risk is difficult.

2. Many possible value chains can spring from the same raw data. These are inherently dependent on the user and their intention and can vary over time. This makes it difficult to develop generic methods for monetizing data.
3. Different valuation chains require different levels of data quality.
4. Different raw data can lead to the same action and potential value. Having alternative data sources that may lead to the same insights will impact the value of the raw data sources.

A common thread of these difficulties seems to be the complexity arising from the contextual nature of the valuation activity.

It appears that applying data valuation (or monetization) methods derived from paradigms such as data-as-a-commodity, data-as-an-asset, data-as-a-product, are inherently flawed or seem at least rushed over-simplifications. In the end, it took decades to negotiate trade agreements for tangible goods or to converge to solutions for intellectual property rights; maybe the short timespan since our interest in data value has occurred was not even sufficient to gauge the extent of the problem.

For instance, from a market perspective, immediate challenges that need consideration are: how to set the price of asymmetric information (selling data points vs. data bundles), how to set a price with minimum information leakage, how to detect fraud, how to avoid an initial period of mispricing (which could lead to an economic bubble).

As the world moves towards a universal online presence and open data systems, it is important to keep discussing the issues surrounding open data and work towards resolving them, as opposed to viewing and evaluating data through the sole economic lens of a commodity [86].

Personal data itself raises many additional challenges of social, legal and ethical nature: what ownership model to adopt, should we even adopt one – since this would involve selling a form of identity [70], how do we adapt to different legal frameworks and different interpretations of privacy across cultures, to enumerate just a few.

When discussing data valuation, we need to consider two paradigm shifts – consequences of the big data "revolution". First, there has been a change in the data production-consumption cycle. Historically, data was produced when it was needed, tailored to the needs of those who would use it and often consumed precisely by those who enabled its generation. For example, a scientist who would need to measure the levels of a certain pollutant in a river would design the data requirements and model, organise, and perform the data collection and eventually process and analyse the collected data. Today, however, a large portion of the data produced every instant is a byproduct of activities, behaviours or processes which are not always the primarily intended focus of data observations. We have shifted from dedicating resources to identifying and processing the data that supported a subset of activities, to generating data about nearly every aspect of our lives.

The second paradigm shift concerns the structure of the data-transactions ecosystem. The 'classic' model of Internet and data-centric companies was to offer a so-called 'free' service in exchange for their users' personal data. But the data deluge from the past decade and the gradual shift of businesses towards data-driven decision making, has created a fertile ground for the so-called data brokers. These intermediary enterprises exist "*solely to collect personal data and sell it as a commodity to retailers, advertisers, marketers, even other data brokerages and government agencies*" [57].

The difficulty of assessing the value of data stems from these two shifts; both centred around the ubiquity of data. To better explain this, we refer to the notion of "surplus data" – introduced by Shoshana Zuboff in her book "The Age of Surveillance Capitalism" [90]. We consider this a paradoxical term considering that on the one hand such data are the fuel of data-driven businesses, and on the other, these same businesses have perpetuated this perception of "surplus", hence diminishing the value of each individual's data. Obviously, there is a legitimate discussion to be had about what is considered valuable data. Data from an individual at the "centre of the distribution" might not be as informative as data in the "tail of the distribution". However, the same individual data might distribute differently in various contexts, leading to different valuations for the same data. Therefore, what might constitute "surplus" for one entity could be an important piece of the puzzle for another. Or, put in more practical terms: "[i]t's like the moment when companies realize they're sitting on patents that they don't really need, but actually have value to someone else" [5].

1.2 The Safe-DEED Data Valuation Process

Here we simply put the schema and a review of the data valuation process.

If the increasing reliance on data and its subsequent role as a fuel of today's economy are indisputable, we are faced with an almost complete lack of tools and methods for assessing the value of data. Indeed, given the particularities of the value chain for data and the lack of maturity of data markets, it is difficult to treat data as any other commodity and develop pricing methods. Understanding and overcoming these difficulties should lead to the design of transparent tools and methods for data valuation, thus favouring their adoption by stakeholders in the data economy.

The development of such a tool is at the focus of our work in WP4 of the Safe-DEED project, a tool relying on a **data valuation process**, that works along the following lines.

1. a method for formalising the context in which data is evaluated. Drawing from data-sheets for datasets [32] and ongoing research on mapping data set properties to data value [48], we synthesized a process for defining the context and bootstrapping the valuation of a dataset, taking into account several facets: systems & economics (including a definition of purpose),data tools, data properties, and business impact (including legal and ethical aspects);
2. data quality assessment. We are currently working on developing a flexible process that can filter the data quality dimensions suitable to a given context and then performing data quality

assessment through the lens of the selected dimensions. The result will be a multi-dimensional measure (rather than aggregate) of the quality of data, which will factor in the final valuation method;

3. once a dataset meets the quality criteria, then its fitness for use in the defined context is assessed. The measures for such an assessment will be derived from the statement of purpose, collected as part of the context (i.e., if the declared purpose is to use the dataset fora classification task and the accuracy for the task is under 80%, the data set will be valuated accordingly);
4. a final valuation method will synthesize the outputs from the 3 previous steps into a measure of value for data.

The details of each step of this process will be discussed in the following sections. We also invite the reader to check the detailed description and architecture of our data valuation component (see Deliverables 4.1 and 4.2).

2 Economic Value of Data

There is an interesting comparison that is usually made when illustrating the economic value of data (of which everyone agrees it exists), as well as the difficulty in estimating it, especially when perceiving data as an economic asset, much like petrol: "Facebook is now worth about $200 billion. United Airlines, a company that actually owns things like aeroplanes and has licenses to lucrative things like airport facilities and transoceanic routes between the U.S. and Asia, among other places, is worth $34 billion" [3] [5].

The US Department of Commerce found that between 2004-2014, data-driven businesses created a $17 trillion economy, according to estimates of their revenues, while the costs on data collection, processing and dissemination amounting to only $3.7 billion annually – a mere 0.02% of the value created [6].

A McKinsey study from 2013 [59] estimated that public open data could help unlock between $3.2 - $5.4 trillion in economic value across seven economic areas (education, transportation, consumer products, electricity, energy, healthcare, consumer finance), together with five actions to achieve that: promoting transparency, exposing variability and encouraging experimentation, segmenting populations, automation, defining new products and services.

In the context of a CRM, Lehr et al. estimate the costs of operating with bad data (duplicates, missing, inaccurate or outdated information) at $100/record[4], cleaning the data at $10/record and maintaining it clean at $1/record. Considering a 100,000 records database, with about 20% of its records dirty[5] and average size growth of 40% per year[6], the authors estimate approximately $8.5 million in data quality savings over three years [53].

Concerning the use of personal data, a 2012 report by the Boston Consulting Group was estimating that its quantifiable benefits could reach €1 trillion / year by 2020 (approximately 8% of EU's GDP), a number that could well be an underestimation as its calculation was made based on the primary use cases from data at the time [71]. Similarly, the global economy based on personal data was estimated to be around $3 trillion in 2017 [86].

[3] The market valuation in the quote refers to the year 2015.

[4] $100/bad record is attributed to impacts such as: printing and mailing to bad addresses, emailing wrong addresses, losing unhappy customers, extra storage space for duplicates, sales conflicts over the same leads, inability to track lead source, incorrect marketing segmentation, unnecessary marketing automation [53].

[5] Sirius Decisions – the impact of bad data on demand creation.

[6] Ebiz – Integration on the edge: data explosion and next-gen integration.

2.1 Data as an Intangible Asset

Kannan et al. propose a method for mapping data properties (which they classify as intrinsic or extrinsic) to a quantitative value [48]. This is based on their view of data as an intangible asset, which acquires value once it is put to use; this value increases once the data moves through the processing pipeline. The authors do recognise that unlike other assets, data only incur a production cost and once it exists, it only requires marginal costs for using it in other applications. Alternatively, if created at an intermediate step of another value chain, there may be a transformation cost associated with adjusting it for the current needs.

Slotin et al. adopt the same view of data [78], having unique properties that make it difficult to assess its value: they are non-rival, and when open, they are public good so it is difficult to assign them a market price. Despite these, she identifies 5 methods for measuring the value of open data in public policy, all of which can be extended to multiple types of data and applications.

1. Cost based approaches. Value is determined based on the full cost to produce data and subsequent statistics. While valuable as an exercise, it doesn't provide any indication about the benefits of such an investment.

1.1 Market-based approaches. Value is determined based on the market price of equivalent products or users' willingness to pay.

1.2 Market equivalent pricing. The value is based on an approximation of similar products in the market. The drawback is that it is difficult to estimate the value of the differences between the data to valuate and the proxy data-products (e.g., public data sets vs. privately enriched data sets that make use of public data).

1.3 Stated/revealed preferences. Users are asked to value the impact of the lack of data (stated preference) or to assess the value of a subset of data attributes given a budget (revealed preferences). While credible, these methods are difficult to validate outside a given context and are dependent on the user's capacity to articulate the value of the data.

2. Income-based approaches. Value is given by the estimated productivity improvements and cash flows derived from using the data. Such methods can be applied in a top-down macro-economic or bottom-up micro-economic fashion, with each of them either over-, respectively under-estimating the value of data. A drawback of such method is that known studies have been performed only in developed economies, with an abundance of quality open data.

3. Benefit monetization. Value is established by defining and quantifying the benefits resulting from the uses of data.

4. Impact based approaches. Value is determined from the causal effect between the availability of data and resulting and resulting economic and social costs. They reveal the importance of efficient reporting of results and show that a blend between showcasing both the human and economic impact could be the "winning combination". Their main drawback is the context-specificity, which may limit their influence.

The author also underlines the fact that these methodologies have been applied in context-specific settings and while their extension to other contexts should be feasible, their complexity and level of abstraction makes this difficult.

2.2 Data as Commodity

It is likely that the major changes in the global economy, with data companies challenging energy companies, inspired comparisons between oil and data, including the over-used "data is the new oil". While this is useful to get across the point that data is a valuable resource in today's world, the comparison can easily break along multiple perspectives:

- data is becoming increasingly more available with time, as opposed to fossil fuel, which is becoming scarcer.
- raw data comes in many different flavours (text, image, video, sound), across a variety of formats which require a variety of methods for extraction; raw oil is all the same and extracting it is done in the same way.

2.2.1 The Ecosystem of Personal Data Exchanges

At this point, we need to make the distinction between personal and non-personal data. According to Article 4 of the EU General Data Protection Regulation (GDPR) [69], personal data refers to information relating to natural persons who can be directly or indirectly identified from the data in question[7]. Estimating the value of personal data is an emerging topic that has not received much attention from the scientific community, because of the difficulty of obtaining such data (unavailability, companies' fears of legal responsibility, legal requirements – such as the GDPR).

The view of data as a commodity gathered momentum with the advent of personal data exchanges, themselves enabled by the advent of targeted online advertisement and data-driven companies. In this model, a company would offer a seemingly free product or service and "in exchange" it collects a variety of data generated by the users' interactions with a digital frontend (a website or an app). It took a while for individuals to understand that the blurred lines between data and data-enabled-service are what enabled the data deluge and the ensuing financial changes, mentioned in the beginning of this report: from a service perspective, individuals are the avid consumers of maps, delivery, booking or communication applications; however when looking from the perspective of the data that fuels these services, the producer-consumer roles flip and it is the companies that become the avid harvesters of digital "selves".

With increasingly more companies realizing the potential of data-enabled business cases, naturally, there is a gap between their new aspirations and their data know-how. This gap has created the opportunity for a new group of stakeholders to join an already unbalanced ecosystem: data brokers. The projected revenue of this industry in the United States alone was of $250 billion in 2018 [57]. Typically, data brokers add value to personal data which individuals often give up during various online activities, by aggregating it, generating user profiles, and enriching them with valuable (and often free) data compiled by National Statistics Organisations. It is these bundles of repackaged data that are then sold back to different companies to satisfy their data necessities.

This complex landscape is finally completed by the presence of governments, which up to this point are yet to find their role. On one side, governments are expected to assume a regulatory role with respect to data transactions in general, and data brokers in particular. In Austria, a reported discussion about applying VAT on revenues resulting from big data transactions by social media companies was abandoned, citing difficulties in assigning a value to such transactions [57]. Similarly, the United States Senate is supposed to discuss the 2019 DASHBOARD Act, a piece of legislation designed to protect individuals' privacy by forcing companies to disclose to the users the "true value" of the data that concerns them [88]; this legislation however, doesn't seem to grasp the complexity of establishing such a value based solely on "revealing the revenue obtained from obtaining, collecting, processing, selling, using or sharing user data".

Beyond its role as a regulator, there are reported instances in which the government seeks to act as a data broker itself. In 2014, citing the abundance of data it amasses, UK ministers attempted to pass legislation that would allow HM Revenue & Customs (HMRC) to sell anonymised taxpayers' data to 3rd parties. This has come under scrutiny since the British government's track record in terms of data security and data anonymization practices is far from clean [60]. More worrisome is that despite restrictions and criticism, the HMRC went ahead and, as part of a pilot project, quietly released VAT

[7] Such information can refer to "*an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person*".

registration data "for research purposes" to three private credit rating agencies (Experian, Equifax, Dun & Bradstreet).

Finally, the most recent addition to this landscape are online platforms for monetizing personal data. These platforms claim they are giving back to the users the control over personal data and enable them to sell it themselves, ideally choosing the shape and the buyer. There doesn't seem to be much separating these platforms from large data brokers (and in fact there is nothing to prevent such platform from growing into one), but where they do set themselves apart is that they acknowledge the value of personal data and are open to sharing a piece of the revenues with those who generate it. Just how much? We will see in the next section.

2.2.2 Putting a Price on ersonal Data – a review

Missing information about the value of data is one of the barriers to establishing pricing models for data-as-commodity. In this section we look at several attempts to monetise personal data, which we were able to collect from three types of sources: review of personal data monetization platforms (platform assessment, online tech articles), industry reports based on user studies and peer-reviewed research papers. We have yet to come across any attempts involving non-personal data, and we might think of at least two reasons for this:

1. personal data is more ubiquitous, and due to poor online behaviour people have been giving up (and still are) on it fairly easily;
2. non-personal data is usually data referring to industrial, financial, or business processes, generated by the activity of a commercial entity, which has little to no interest in sharing it.

We begin with the platforms for selling personal data, briefly discussed in the end of the previous section and review a synthesis of them, particularly looking at the type of data they monetize and what are the rewards they offer.

Data Trader	Reward	Data Type
Datum	0.01 USD / month	Location
CitizenMe	0.1 GBP	Personal data and preferences via online quiz (10 questions)
Datacoup	8 USD / month	• Social media feed; • Credit card transaction details.
Luft Research	100 USD / month	• Browser and search history; • Location; • Twitter usage; • Filled questionnaires; • Device microphone audio recordings.
Permission.io	Non-exchangeable token	Watch ads.
Wibson	vouchers for points; ~ 0.02 USD / pts.	• Location (15 pts.) • Facebook, LinkedIn (20 pts.); • Device information (25 pts.); • Google accounts.

Shawn Buckles on Kickstarter	350 EUR (entire bundle)	Everything: personal data, location, medical, train travel patterns, personal calendar, emails, social media, consumer preferences, browser history, personal "thoughts".
Cambridge Analytica	0.75 USD / record	• Name, gender, location; • Behavioural indicators; • Political views and involvement; • Political quizzes.
Proximus	700 EUR / report	• Market research report: • Location, movement; • SIM card country of origin
AT&T Gigabit	-29 USD / month	• Visited webpages; • Interaction with links and ads.
Telefonica	NA user has full control	Databank of user activities on the network.

Table 1 : Personal data monetisation platforms and rewards offered to individuals

There is a wide range of personal data collected by data brokers: identification, demographic, location, behavioural, online activity, psychological, product and political preferences. Most of the times, this data is sold in bundles, which prompts several questions: are all these equally important to a buyer, are they equally sensitive for a seller and how do each of these stakeholders value them? A reward as low as ¢1/month for sharing exclusively location data might not convince a user to give it away; however a bundle of several data types that can amount to as high as $100/month could prompt individuals to invest time in building, managing and selling personal data portfolios. A second observation concerns the wide price range at which the same type of data is sold. For example, Luth Research pays $100/month for a bundle containing location, social media activity and browsing activity [72], whereas Datacoup pays $8/month for a similar package [77]. We believe this discrepancy is due to the lack of established data markets, data trading rules and, as we will see next, a significant gap between the monetary value expected by individuals and what is actually paid by data brokers.

Much like the data brokers that they claim to be replacing, many of these platforms are still unable to enforce any control on who acquires the personal data, their purpose and their processing methods. A particular area to keep an eye on is the telecom industry, which long realized its data collection potential and is now exploring ways to monetize the troves of data it collects. When Belgian telecom operator Proximus sells bundles of SIM-traces for a minimum of €700/report [80], this raises questions about data ownership. More disturbing practices come from the United States, where AT&T collects and uses all their Gigabit subscriber's activity via deep-packet inspection [21], allowing them to opt-out only by paying an additional $29/month, essentially enforcing a cost on privacy.

There are examples of good practices in terms of dealing with user data. We have already seen that Wibson is trying to enforce transparency, by stating who the data is generated for and to which purpose. Spanish company Telefónica proposes the establishment of a data bank which allows their service users to log all their activity on the network; this is somewhat similar to AT&T's Gigabit, with the major difference that the former would give users full control over their data[8].

[8] At the time of submitting this text, there are no mentions as to what the price of such a service would be and how would Telefónica benefit from it.

A study by telecom company Orange, covering 2023 mobile phone users balanced across age categories and countries of origin (France, Poland, Spain, UK) [54], suggests the existence of three factors that influence the perceived value of personal data:

1. the usefulness of the data to the beneficiary organisation
2. the type of data and
3. the risk associated with sharing it.

The study also revealed that users are aware that their data is valuable to organisations, which can benefit from it. Users' responses also revealed an ordering relationship of how likely they are to share types of personal data (demographic $\succ$ activity and behavioural $\succ$ third party or financial data)[9]. When considering the third factor, the study simply refers to the "perceived" risk associated with sharing personal data and does not go into details about how such a risk might be quantified.

The same study also points to a paradox in consumers' understanding of sharing personal data: while a majority of respondents (77%) declare that privacy and transparency of data usage are important and identify the risk attached to sharing as an important factor influencing data value, they also indicate demographic data as the type they would most likely share – despite the clear risk of identity theft and online fraud attached to it [81].

Data type	familiar organisation	unfamiliar organisation
full name or date of birth	£12.16	£15.22
mobile number	£13.96	£16.20
location (via mobile GPS tracking)	£13.35	£16.02
annual income	£14.61	£16.50
marital status	£9.63	£12.83
sexual orientation	£11.38	£13.85
job	£11.11	£13.83
children's details (sex, age)	£12.44	£14.53
details of family members' preferences	£14.07	£16.21
email addresses of 5 people in a close personal network	£14.46	£16.67
history of purchases made on mobile phone	£13.25	£16.31
postal address	NA	£15.67
main personal address	NA	£15.11
MEAN	**£12.77**	**£15.30**

Table 2 : Self-assessment of expected reward for sharing personal data with a familiar / unfamiliar organization [54]

[9] Third party data: email, personal preferences of other contacts; Behavioural data: location, mobile purchase history; Demographic data: name, date of birth, phone number.

In a 2016 survey, credit comparison site Totally Money [24] asked 1000 UK consumers to estimate[10] the economic value of different categories of personal data. The results revealed interesting attitudes and different data-sharing practices, spread across demographic groups and types of data alike: young people (18–24 years old) value their data the most, while millennials value theirs the least (£1773); men value data about their online activity higher than women do (£1112 vs. £859 for email data, £1056 vs. £817 for browsing data, £951 vs. £778 for location data). Perhaps the most surprising result of this study is the difference between the average self-estimate of respondents' personal data (£2031) and how much brokers are paying for it (£0.45). Both the magnitude of this difference and the paradox mentioned in [81] point to an important challenge of pricing personal data:

> *Building digital literacy, together with legal frameworks to bridge the gap in understanding i) the permeability of our digital traces and ii) the ease with which data companies are able to collect and monetise them.*

Age	0.03p	Home address	0.03p
Gender	0.03p	Home ownership	7.12p
Ethnicity	0.34p	Bank	NA
Job	5.00p	Credit rating	NA
Marital status	7.00p	Email	5.40p
Health condition	17.70p	Phone	0.30p
Children (y/n)	2.37p	Browsing history	0.14p
Schools children are enrolled at	NA	Actual location	0.03p

Table 3: Personal data valuation example. Amount paid by data brokers for a datum of each type (currency: British pence). The four quarters of the table reflect the four categories as defined in the survey: demographics, family & health, finance & property, online information [42]

An interesting perspective comes from credit reporting company Experian, which in 2017 compiled a list of prices for which the most common pieces of personal data are sold over the Dark Web[11] [81]. The report also mentions the methods in which this data is sold (bundles, also called "warez", are very popular) what are the factors that drive personal data prices in the context of the Dark Web: type of data, demand and supply of a certain data type, the balance of the accounts (the more money/points on an account, the better), expiration date and reusability.

[10] TotallyMoney.com conducted research in June 2016 to identify the prices third-party companies pay for data to utilise in marketing campaigns: Financial Times, The Telegraph, McAffee, CostOwl.com, OnePoll.com.

[11] https://en.wikipedia.org/wiki/Dark_web

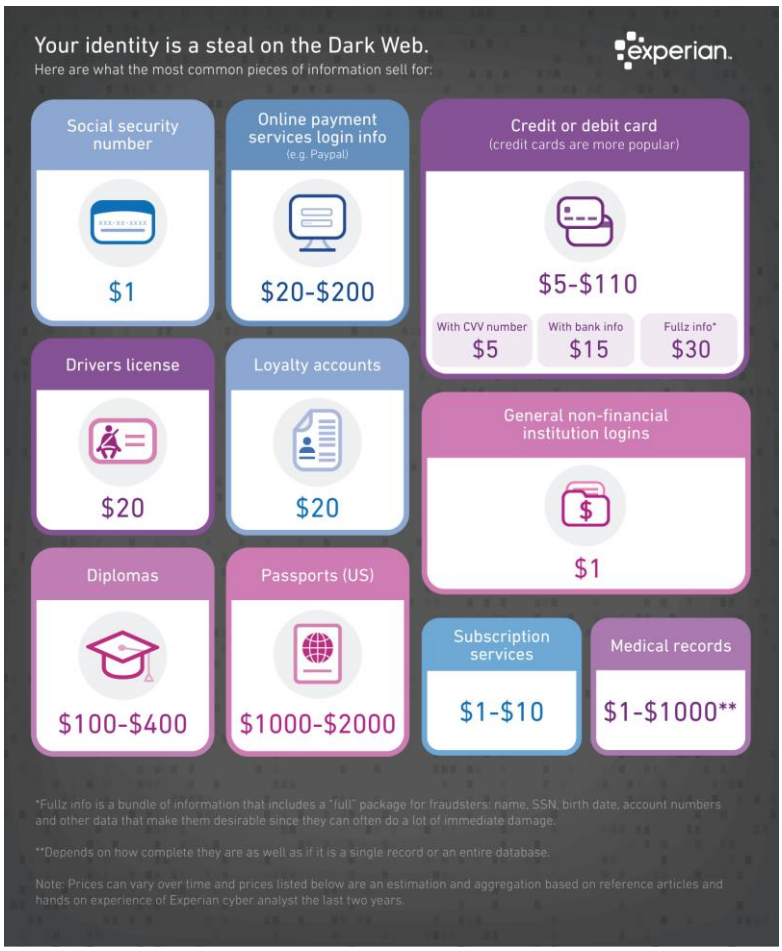


Figure 2 : Value of personal data on the Dark Web [81]

Interesting results are also coming from academia, with a recent increase in the study of methods for valuating user-generated data, particularly geolocation and online behaviour. In one of the most relevant experiments Staiano et al. [82] simulated a data market for personal data transactions. Participants were equipped with devices gathering various types of data (calls, applications usage, location and media usage) at three levels of aggregation (individuals, processed and aggregated). They were then expected to sell the data to the Research Laboratory during auctions (reverse second price[12]), initially running weekly and then daily. In the near 600 auctions organized, participants received rewards totalling approximately €270, with a median price of €2 across categories. The auctions were also able to cast a light into the self-valuation of personal data, and just like in the study conducted by Orange [54] it revealed an order of perceived value among the data types: location ≻ communication ≻ apps ≻ media; not surprisingly, processed data was held to a higher value than raw data. Two additional observations may be important take-aways in designing data valuation methods:

1. Increasing the frequency of auctions (from weekly to daily), decreased the value of the bids; an indication that the data market may play by the rules of supply and demand.
2. The value of data increased when unexpected situations arose (traffic jams caused by either a weather event or a local holiday); this suggests that the value of the same data is highly dependent on the context.

We conclude here our review about the monetary value of personal data. Our analysis of various sources (platforms for selling personal data, technical and industry reports, academic research) results in a better understanding of several aspects:

- the monetary value assigned to different types of personal data;

[12] The lowest bidder wins, but the reward will equal the second-lowest bid.

- the discrepancy between its value as perceived by individuals and the brokers, which buy and resell it;
- reveals a preference ordering between the different types of data;
- confirms the contextual nature of data valuation.

We plan to use pricing information for items of personal data as a reference for the evaluation of the results of the Data Valuation Component applied to similar data sets.

3 Contexts for Data Valuation

Let's consider a data set containing GPS traces of taxis in a city. For a ride-hailing application, such data would provide a way into estimating the customer needs in different areas of the city, at different times, allowing them to develop Machine Learning solutions for load balancing and trip planning and eventually maximising their revenues. The local administration could use this data to understand road congestion and travel times and plan infrastructure interventions (repairs, extensions, restrictions), modify public policies (congestion taxes) or plan connected services (public transportation). A retailer could look at this data in conjunction with other sources and understand behavioural patterns of people living in different areas of the city and thus plan opening schedules, logistic operations or decide to open new branches.

Data can have different value for different roles within the same organisation. Data containing the flow of passengers through a certain area might be enough for a planning manager and his team who decide to build a larger shelter or a new bus stop, but for the R&D department working on a new routing algorithm, such information might be overly-aggregated and useless for their necessities.

Even within the same department, different tasks might impose different requirements from the same data set. The data science team might be able to provide a good enough analysis of travel patterns from data which contains trips aggregated over 30 minutes intervals, but such a data set will not be useful if the task is to create an accurate traffic prediction model.

A common thread across our review of data valuation (Section 2) and data quality assessment (DQA) (Section 4) methods is the dependency of these processes on the context in which they are performed. We have seen how the contextual nature of data is often cited as one of the main reasons for which assigning value to it is difficult. Mawer notices how different value chains (see Figure 1) can be completed from the same raw data or how the same value chain can be completed with different raw data [61], all depending on the purpose of the data processing. Slotin extends that observation and concludes that context-specific, impact-based methods might be the most suitable for communicating data value, despite this specificity being their main drawback. In their data quality principles, the US National Institute of Statistical Science (NISS) cite contextual factors (purpose, user, time) among those that influence data quality.

Today, we are often reminded about the 4 Vs of big data (Volume, Variety, Velocity, Veracity). Considering the motivations of most organisations for processing such data (improving business processes, finding new revenue streams, supporting policies and decision making, enacting change), the interplay of these Vs is critical to the generation of value (a 5th V!). This interplay, together with the purpose creates the context in which the value of the data is assessed.

We know contexts are central to the data valuation processes, and we have also seen that they are one of the reasons for which these are complicated. Thus, building a solution that takes contexts into account has first to surmount the challenging aspects of defining, formalising, and encoding them. Unfortunately, the literature dedicated specifically to this is virtually inexistent. However, the literature on metadata for datasets and data quality assessments provides several useful directions to answer the previous questions.

Cai and Zhu note that “data quality depends not only on its own features but also on the business environment using the data, including business processes and business users” [16]. Pipino et al. differentiate between task-independent and task-dependent assessments of data, with the latter

consisting of organization business rules, company and government regulations and technical constraints [67]. Askham et al. talk about the dependency between data quality assessment and the context in which it occurs, and mention that organisations should consider not only quality dimensions but also organisational requirements for data and the impact of non-compliance [4]. Even and Shankanarayanan suggest that contexts are often disregarded when designing data quality frameworks, observe that the value of the same data may ultimately depend on "contextual factors, such as the organisational level at which the data is used, the specific task, and/or the personal preferences of the decision makers" [30].

These observations do more than confirming the context-dependant nature of data value. They bring a first level of clarity concerning the layers that compose a context. We summarise them here:

- organizational profile
- business user profile
    - a specific task, personal preferences
- business rules/processes, organisational and government regulations

In a sense, defining contexts is akin to understanding users, identifying use case scenarios, and deriving user requirements.

Recent work focusing on data profiling and valuation of metadata offered valuable leads into how data valuation contexts could be established and quantified. Among them, we distinguish a questionnaire-based method for mapping data properties to data value [48], the creation of data-sheets for datasets [32] and the Dataset Nutrition Label [37], a diagnosis framework providing critical information at the point of data analysis.

Kannan et al. posit that the value of data is better quantified when used in an application and similar to Mawer [52] they conclude that data increases in value as it advances through a processing pipeline [48]. In their framework, data is characterized by multiple intrinsic and extrinsic facets, with some gaining precedence depending on the context. They propose a questionnaire-based method to gather objective responses (binary, quantitative, categorical) that characterise a data set. In the simplest of cases, all facets are treated equally; then, to account for the many possible applications, the composition of the questionnaire and the relative values of the responses could be adapted to the context or the role of the respondent. In the current version of the Data Valuation Component (see Deliverable D4.4), we have adapted this approach to our own process of establishing contexts.

A similar approach is that of Gebru et al. [32], who promote the creation of data sheets for data sets, with the main goal of increasing the communication between creators and consumers. Their work derives from the IEC Datasheets, common in the electronics industry, and is somewhat akin to the documentation that accompanies clinical trials in the United States. Moreover, their use was subsequently extended to factsheets for AI systems [3]. The creation of datasheets is the responsibility of the data set creator, through a process of reflection on the creation, distribution and use of the data set, including assumptions, potential risks and implications of use. This should equip consumers with the tool they need for making an informed decision about using the data set, taking into account its content, collection process, recommended uses, restrictions and including the stated assumptions, risks and implications. As opposed to work on mapping data properties to value [48], data sheets do not set to quantifying the responses; on the contrary, the authors are trying to avoid binary responses and encourage creators to be as detailed as possible. The questions themselves, mirror a typical workflow around a data flow and focus on motivation, composition, collection process, pre-processing-cleaning-labelling, uses, distribution, maintenance. The current version of the DVC (see Deliverable 4.4) is using this methodology to refine the questions included in our approach for establishing valuation contexts; and while our questions are trying to be as exhaustive as possible, the answers that we process are meant to be easily quantifiable and mapped to a contextual value.

Finally, Data Nutrition Labels is a diagnostic framework which provides a concise, robust, and standardised view of the core components of a data set [37]. The goal of the DNL is to inform and improve the selection and interrogation of data sets and is primarily aimed at data specialists. The

solution is composed of several loosely coupled modules, covering various facets of a data set, some with a degree of context-dependency: metadata, provenance, variables, statistics, pair-plots, probabilistic models, ground truth correlations. Some of this information is covered by the current version of the DVC either through the context creation process (in the case of provenance) or by the automatic profiling of the data set (in the case of variables statistics, distribution, correlations), while the next version of the DVC will focus on assessing the usability of machine learning models.

4 Data Quality Assessment

The earliest preoccupations towards a formal understanding of quality date back to its application to assembly-line production and manufacturing in the beginning of the 20th century and accelerated later in the 1950s and 1970s with its adoption to business practices. Along the years, various definitions have been put forth, referring to quality as "conformance to requirements" [22], Joseph Juran's famous "fitness for use" [47] or the "degree to which a set of inherent characteristics fulfils requirements" [40]. One particular definition refers to quality as the "value to some person" [16], recognising the intrinsic value derivable from data quality, as well as its potentially contextual nature.

With the development of information technologies, interest in quality of data has sparked during the 1990s. The democratisation of the Internet and the advent of big data and data-centred solutions generated more interest in the topic and laid the ground for a currently mature and dynamic research field. In 1996, the Total Data Quality Management Group at MIT adopted the "fitness for use" definition and acknowledged its dependency on the consumers. The principles of data quality by the US National Institute of Statistical Sciences (NISS) adopt the view of data as a product and as such, consider that its quality results from the process that generates them. Later, data quality was enacted at the governmental level, as was the case of the US Data Quality Act [66] or the Welsh Data Quality Initiative Framework [64]. In Europe, Bergdahl et al. report on the successful integration of data quality assessment in the activities of several National Statistics Organisations: Statistics Sweden, Statistics Norway, CBS in the Netherlands, the Austrian Quality Concept (an In house quality reporting system), the ONS Guidelines for Measuring Statistics Quality (a grading scheme for statistical products), Slovenian Statistical Office (data quality measurement for short-term statistics) [10].

In Section 2.2 we have seen how the interplay of the 4 Vs of big data, together with the processing purpose, influence the generated value. We have also seen how data quality can be regarded as the ability of data to serve its purpose – generally seen as the needs of an organisation in terms of operations, planning and decision-making [50]. Therefore, in order to evaluate the quality of data, a plethora of data quality assessment methodologies have been developed over the recent years, adopting different perspectives in their attempt to encompass the multitude of assessments that gather under the data quality umbrella.

The aforementioned contextual nature of data quality assessment is likely the root of the divergent research directions in the field, resulting in multiple methodologies, covering an even larger amount of quality dimensions, each quantifiable by means of even more quality metrics.

To clarify, "a Data Quality Dimension (DQD) is a recognised term used by data management professionals to describe a [property] of data that can be measured or assessed against defined standards in order to determine the quality of data." [4] Dimensions focus on measuring and communicating the quality of data, as opposed to describing what the data represents.

The Total Data Quality Management Group at the MIT defines 15 quality dimensions [87], the Data Management Association for the UK focuses on 6 primary dimensions [4], Statistics Netherlands mention 49 factors that influence the quality of secondary data and group them into 5 focus areas [85]. In fact, an overview of all dimensions and subsumed metrics [8] (see Table 4) allows us to confirm the complexity and multi-dimensionality of the concept of data quality.

4.1 Methodologies

A data quality methodology is a set of guidelines and techniques which given an application context, provide a rational process to assess the quality of data.

The definition before comes from the well-documented work of Batini et al. who provide a systematic and comprehensive state of the art review of methodologies for assessing the quality of data [8]. The authors are first drawing the perspectives from which the analysis of the methodologies is conducted. They then proceed to select 17 methodologies and analyse them through each of these perspectives.

4.1.1 Methodology Phases

In general, methodologies consist of the following activities, usually occurring in sequential order:

- State reconstruction is akin to establishing the context for the DQA.
- Measurement and assessment. In detail, these involve the following activities: data analysis, data quality requirements analysis (usually needed to perform the assessment), identification of critical areas, process modelling and measurement.
- Improvement. Apply the results of the assessment to define and implement actions that lead to data of better quality.

Data analysis and quality measurement seem to have good support across many of the reviewed methodologies, whereas activities such as data quality requirements analysis, identification of data quality issues or the collection of new quality targets have less coverage.

4.1.2 Strategies and Techniques for Assessment

These are either data-driven (e.g., standardisation, normalisation, record linkage, data/schema, source trustworthiness, error detection) or process-driven (e.g., process control) strategies. On the long term, the latter can be the better option, but they are also more difficult to implement. However, if data quality practices are already in place, a methodology that can alternate between either process based on context-dependent variables should be preferred.

4.1.3 Dimensions and Metrics

The definition of dimensions and metrics is central to any DQA methodology. Once the challenge of defining a quality dimension is surmounted, creating the metrics to measure it follows easier [30]. There is currently little consensus both in terms of data quality assessment methodologies and what are the preferred quality dimensions. Table 4 summarises the variety of dimensions and metrics and is testimony to the complexity of the data quality concept. We refer the reader to Section 4.2 for a detailed discussion on the topic.

4.1.4 Costs Associated to Data Quality

The common wisdom is that the cost of data quality is the sum of direct and indirect costs. Indirect costs are those that are associated to poor quality of data (e.g., processing, missed opportunity costs), are context-dependent and are very difficult to assess, as a certain quality level can have a different value depending on the recipient. Direct costs are associated with data quality assessment (and improvement).

Only 3 of the reviewed methodologies analyse the cost of data quality: TIQM, COLDQ and CDQ. The data quality cost assessment supports the selection and prioritisation of data quality activities (in the

case of TIQM), focuses on the economic impact of bad data (COLDQ) or minimises the cost between alternative improvement processes (CDQ).

4.1.5 Types of Data

This perspective refers to the degree of structure within the data (structured, unstructured, semi-structured); the more flexible the structure (schema) the more complex the data quality issues. The type influences the choice of dimensions and metrics (even for the same dimension) to include in the DQA. Most of the reviewed methodologies deal with structured data, and a handful of them have are either designed or have adaptations for semi-structured data (DaQuinCIS, D^2Q, CDQ).

4.1.6 Types of Information Systems that Process Data

This criterion refers to the degree of integration between data, process and management and exist on a spectrum consisting of monolithic systems, data warehouses, distributed systems, cooperative systems, web systems, peer-to-peer systems.

Methodologies for monolithical systems (AMEQ, COLDQ, QAFD) typically consider structural data and ignore data exchange issues. Some of these methodologies can also be applied to distributed systems (TIQM, CIHI), while others offer support up to the cooperative level (ISTAT, DaQuinCIS, CDQ). Amongst the specialised ones, we distinguish DWQ (for warehouses) and IQM (for web information).

4.1.7 Summary Comparison

In addition to the 6 previously discussed perspectives, there are three others which are mentioned but not included in the analysis of Batini et al.: organisations that process the data, the processes for processing the data, the services created by these processes.

The authors try to summarise their analyses in one single categorisation, which encompasses most of the differences between these methodologies, and separates them into four categories (see Figure 3):

1. complete methodologies – address both assessment and improvement phases and deal with both technical and economic aspects of data quality. These can fit very well in organisations which have already invested in a data quality program. However, their high level of generality may make them difficult to adapt to specific domains or technological contexts.
2. audit methodologies – focus on assessment and provide only limited support for improvement. These are more accurate in the assessment phase, and since they are not constrained by the improvement phase, they are able to identify a wider range of techniques and make their selection clearer.
3. operational methodologies – focus on the technical side of both assessment and improvement phases. As opposed to audit, they put the assessment phase in service of the improvement one and in doing so they become more effective, at the cost of narrowing down their applicability to a certain context.
4. economic methodologies – focus on the evaluation of costs, taking into account both the cost of performing data quality (just like some of the audit/improvement methodologies) and the cost of doing nothing (i.e. the cost of poor data).

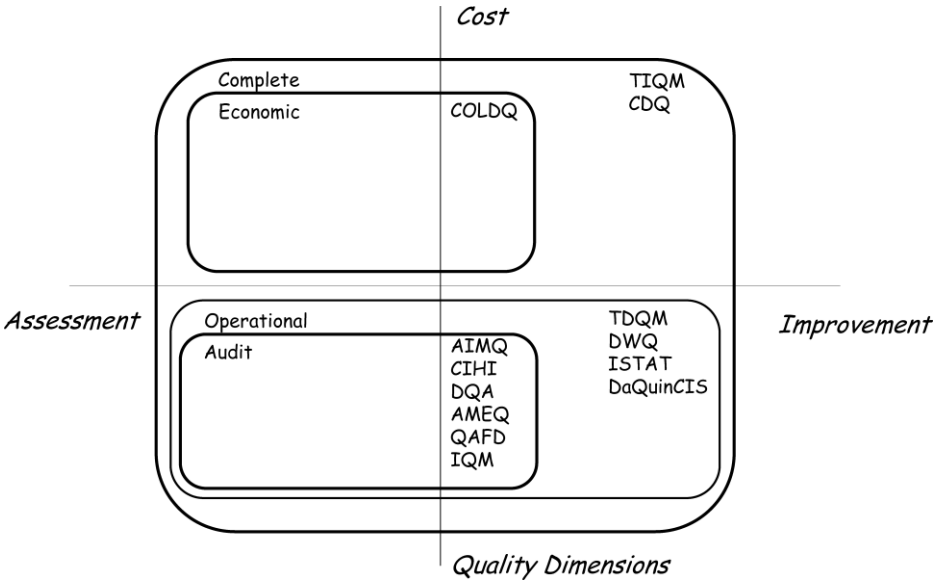


Figure 3 : A summary classification of DQA methodologies

4.1.8 Evaluation

One of the most delicate points when promoting the application of a DQA methodology is its evaluation in real application contexts. The authors of the review conclude that in most cases, empirical validation is either missing or it is based on case studies, which typically take place in industrial settings and fail to be published as large-scale scientific experiments. Moreover, some methodologies are proposed only at a theoretical level and lack any supporting tool or implementation.

Following its use in several US departments, authors of TDQM claim to be able to evaluate how representative and comprehensive the selected data quality metrics are, by considering target payoffs, critical DQ issues and the corresponding types of data. CIHI – used with health administration systems in Canada – appears to have been successful in identifying and ranking critical data quality issues, from an improvement perspective. The authors of QAFD – applied to financial data quality – provide a real case scenario for their evaluation and cite security reasons for not disclosing additional evidence.

4.2 Data Quality Dimensions and Metrics

Historically there is a correlation between the development of information and communication technologies (ICT) and that of data quality assessment methods. The early systems were monolithical, usually consisted of a single data source, simple data flows, and the only source of errors would come from data entry. Data quality would, therefore involve accuracy, consistency, completeness, and time-related metrics. The evolution towards network-based systems involved a readaptation of these dimensions; with the later advent of the Web, data sources have become more numerous and more varied and as a consequence, new dimensions such as accessibility and reputation had to be considered. Currently, p2p systems force a new rethinking of these dimensions and more importantly and increase the pressure for the consideration of privacy issues. This evolution of ICT systems is itself one of the causes for the number of methodologies, some of which specialised on subsets of data quality issues.

4.2.1 Data Quality Metrics

Recall that each data quality dimension is operationalised by defining one or more data quality metrics (DQM), which provide the operational procedure for assessing a DQD from a certain perspective (see Figure 4). Batini et al. make the distinction between data (extension) quality metrics and the schema

(intension) quality metrics and notice that most DQA methodologies focus on the former [8], a view that we also adopt in our development of the DQA module of the Data Valuation Component.

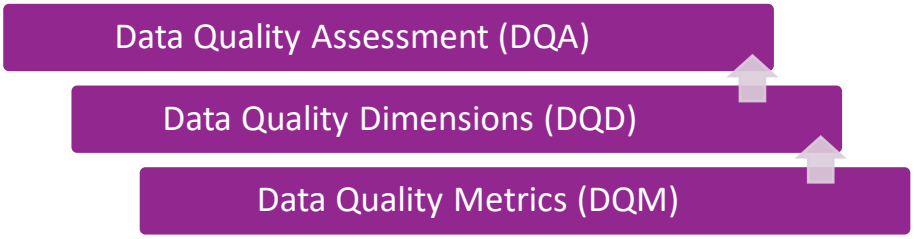


Figure 4: Data quality framework - DQMs support a DQDs, which supports the DQA process

For a better perspective, consider Table 4, a synthesis of all DQDs and DQMs in the methodologies reviewed by Batini et al. Notice how the Accuracy dimension, can be assessed by 3 different measures: syntactic accuracy, number of delivered accurate tuples, user survey.

4.2.1.1 Requirements for Data Quality Metrics

Pipino et al. observe that often the most difficult task is identifying the suitable dimensions to support DQA in a given context, and that formulating the metric is fairly straightforward [67]. They then distil the three most pervasive functional forms for DQMs:

1. simple ratio – typically obtained as a ratio between desired outcomes and total possible outcomes. This is very common when measuring accuracy, completeness or consistency.
2. min/max – used to preserve interpretability when attempting aggregation. One computes the minimum / maximum between similar normalised values at one level to output an aggregate for the next level. The minimum is more conservative, as it outputs the value of the weakest data, whereas maximum is used for more liberal interpretations. This form is used when measuring trust, appropriate amount of data or timeliness.
3. weighted average – ideal in multivariate situations when there is an understanding of the importance of each component. The result should be normalised between 0 and 1 and the weight of each component should add up to 1.

These forms and the motivation behind them have subsequently evolved in a set of requirements necessary for the operationalisation of data quality metrics [34][36]:

1. Normalisation. A measure must be adequately normalised and expressed using a bounded state. This is done in order for the measure to be comparable (e.g., across data sets, time, organisations, in different contexts etc.). Thus, metrics are usually defined between 0 (very poor) and 1 (perfectly good) [23][67].
2. Interval scaled. A measure must be expressed in an interval scale to support monitoring (e.g., over time) and economic assessment of the measure.
3. Interpretability. The results of metrics should be comprehensible and easy to interpret by business users [30].
4. Adaptivity. Measures should adapt to the context in which DQA is performed, for example, via parameters.
5. Aggregation. The value of a metric should be quantifiable at different aggregation levels: datum, tuple, relation, database. The metric should be consistently interpretable (requirement 3) across all levels of aggregation.
6. Feasibility. A measure must be based on input parameters that are determinable and should seek a high level of automation.

While useful one must also be aware of the shortcomings of the requirements above. The most important stems in the lack of procedures for properly assigning a numeric value to data such that it reflects an empirical perception of quality [13]. This may lead to situations in which requirements 1 and 3 are impossible to satisfy. While apparently useful for interpretation purposes, a [0,1] scale can be misleading when interpreting the same value for different DQMs (how is accuracy = 0.9 comparable to timeliness = 0.9). This also means it is difficult to establish a unit of measurement, which makes it difficult to explain quality degradation or to recommend measures for quality improvement.

4.2.2 Data Quality Dimensions

We remind our reader that – due to the contextual nature of data quality assessment – there is little to no consensus as to what might be a subset of necessary data quality dimensions to consider. But is there a subset of "basic" dimensions and metrics that should always be considered when assessing data quality?

Our review of recent DQA methodologies points towards a set of four such DQDs, namely: completeness, validity, accuracy, timeliness [4][9][30][34][50][67][68][74]. This is confirmed by the ISO/IEC 25012, which identifies the same, as well as "credibility" as inherent characteristics of data quality [39].

Our results are confirmed in the review of Batini et al., who employ a quantitative approach for determining the importance of a DQM. They are measuring the degree of consensus on dimension metrics between methodologies, as the ratio between the number of methodologies which use a metric (#DQM) and the number of methodologies which mention that dimension (#DQD), as illustrated in the last column (Usage) of Table 4. Such consensus is high for the same DQMs mentioned earlier – accuracy, completeness, consistency.

In the following, we focus our discussion precisely on these DQDs. For each of them we will highlight some of the many definitions, we will discuss their main properties and explain how they are or will be included in the Data Valuation Component.

Dimensions	Metric	Metric Definition	Usage (#DQM/#DQD)
Accuracy	Acc-1	Syntactic accuracy: it is measured as the distance between the value stored in the database and the correct one. Syntactic Accuracy=Number of correct values / number of total values	9/13
	Acc-2	Number of delivered accurate tuples	1/13
	Acc-3	User Survey – Questionnaire	2/13
Completeness	Compl-1	Number of not null values / total number of values	7/12
	Compl-2	Number of tuples delivered/Expected number	2/12
	Compl-3	Completeness of Web data = (Tmax – Tcurrent) ∗ (CompletenessMax – CompletenessCurrent) / 2	1/12
	Compl-4	User Survey – Questionnaire	2/12

Consistency	Cons-1	Number of consistent values / number of total values	6/10
	Cons-2	Number of tuples violating constraints, number of coding differences	1/10
	Cons-3	Number of pages with style guide deviation	1/10
	Cons-4	User Survey - Questionnaire	2/10
Timeliness	Time-1	(max (0; 1 – Currency / Volatility))^{S}	3/7
	Time-2	Percentage of process executions able to be performed within the required time frame	2/7
	Time-3	User Survey – Questionnaire	2/7
Currency	Curr-1	Time in which data are stored in the system – time in which data are updated in the real world	2/8
	Curr-2	Time of last update	2/8
	Curr-3	Request time – last update	1/8
	Curr-4	Age + (Delivery time – Input time)	1/8
	Curr-5	User Survey – Questionnaire	2/8
Volatility	Vol-1	Time length for which data remain valid	2/2
Uniqueness	Uni-1	Number of duplicates	1/2
Appropriate amount of data	Appr-1	min ((Number of data units provided / Number of data units needed); (Number of data units needed / Number of data units provided))	1/2
	Appr-2	User Survey – Questionnaire	1/2
Accessibility	Access-1	max (0; 1 – (Delivery time – Request time) / (Deadline time – Request time))	1/4
	Access-2	Number of broken links - Number of broken anchors	1/4
	Access-3	User Survey – Questionnaire	2/4
Credibility	Cred-1	Number of tuples with default values	1/2
	Cred-2	User Survey – Questionnaire	1/2
Interpretability	Inter-1	Number of tuples with interpretable data, documentation for key values	1/2
	Inter-2	User Survey – Questionnaire	1/2
Usability	Usa-1	User Survey – Questionnaire	1/1

Derivation Integrity	Integr-1	Percentage of correct calculations of derived data according to the derivation formula or calculation definition	1/1
Conciseness	Conc-1	Number of deep (highly hierarchic) pages	1/2
	Conc-2	User Survey – Questionnaire	1/2
Maintainability	Main-1	Number of pages with missing meta-information	1/1
Applicability	App-1	Number of orphaned pages	1/1
	App-2	User Survey – Questionnaire	1/1
Convenience	Conv-1	Difficult navigation paths: number of lost / interrupted navigation trails	1/1
Speed	Speed-1	Server and network response time	1/1
Comprehensiveness	Comp-1	User Survey – Questionnaire	3/3
Clarity	Clar-1	User Survey – Questionnaire	3/3
Traceability	Trac-1	Number of pages without author or source	1/1
Security	Sec-1	Number of weak log-ins	1/1
	Sec-2	User Survey – Questionnaire	1/1
Correctness	Corr-1	User Survey – Questionnaire	1/1
Objectivity	Obj-1	User Survey – Questionnaire	1/1
Relevancy	Rel-1	User Survey – Questionnaire	1/1
Reputation	Rep-1	User Survey – Questionnaire	1/1
Ease of operation	Ease-1	User Survey – Questionnaire	1/1
Interactivity	Interact-1	Number of forms – Number of personalisable pages	1/1

Table 4 : Data quality dimensions and data quality metrics [8]

4.2.2.1 Accuracy

- The degree to which data correctly describes the "real world" object or event [4]
- The degree of agreement between a set of values and another set of values which are assumed to be "correct" [68]
- How data representation (or value) reflects the true state of the source information [16]

This dimension is highly regarded with respect to others and has historically received a lot of attention from the data quality research community and practitioners. Often a clear separation is made between syntactic and semantic accuracy [8][9]. Syntactic accuracy refers to whether a value is part of a set of possible values or how far is it from one of them, making it essentially equal to structural validity (see Section 4.1.3).

The definitions before, as many in the literature, refer to semantic accuracy. This notion relies on the existence of a set of trusted/accepted "real values" against which the assessment is being performed, a necessity which is often either impractical or can result in a high production cost.

An interesting approach is that of Gorz and Kaiser [34] who postulate that accuracy is the closest descriptor of data quality. Other approaches also notice the dependence of accuracy on other dimensions (e.g., an item which is incomplete or invalid is also inaccurate) [4], however, this approach takes the concept a step further and proposes to aggregate the other main dimensions (completeness, validity, currency) into an estimate for accuracy, which in turn is used as a proxy for data quality. We will come back to this in Section 6 when we describe approaches to aggregation.

The strategy of the DQA module of the DVC will be to implement semantic accuracy only if a trusted reference data set is available; otherwise, the component will only be performing a syntactic accuracy / domain validity check.

4.2.2.2 Completeness

- The proportion of stored data out of the total data entries [4]
- The degree to which a data collection includes data describing the corresponding set of real-world objects [8]
- Information having all required parts of an entity's description [12]
- Ratio between the number of non-null values in a source and the size of the universal relation [63]

Completeness is thus often related to the concept of null or missing values. There are several reasons for which a value is missing: either it doesn't exist (a non-US citizen will not have a state information), or it exists, but its value is not known (the zip code was not collected) or there is no knowledge about its existence (whether a customer has an email address or not).

Consequently, in order to measure completeness, one has to define a set of values that will be semantically equal to NULL, for each attribute of a data set [34].

The DVC implements this DQM by interpreting the schema of a structured data set and generating a fill-in field for each attribute, in which the user will define the set of values that will be assimilated to NULL. In the next version, this will be available for semi-structured data sets (JSON, XML).

4.2.2.3 Validity

- All values of an attribute must be drawn from a specified domain [51].
- Conformance of data values to the syntax (format, type, range) of its definition [4].

Just like for completeness, validity depends on a set of context and attribute dependent parameters (format, range, type) which need to be defined, usually by interpreting business or domain-specific rules. Also, validity is tightly connected to consistency – the violation of semantic rules defined over a set of data items [8] (e.g., integrity constraints, data edits).

The DVC implements two flavours of this DQM – domain validity and format validity (similar to consistency) – and does it in a similar fashion to completeness. After interpreting the schema of a structured data set, the component allows for the definition of rules of each type:

- domain rules – the values that an attribute can take, or the range in which it should fall;
- format rules – several types of these rules can be defined so both a type and an actual value for the rule need to be defined:
    - regex rules (e.g., *^\\d-[A-Z0-9]{5,}$*)
    - datetime rules (e.g., *%d/%m/%Y %H:%M:%S*)
    - more to follow in the next version of the component

4.2.2.4 Time-related Dimensions

This encompasses several different dimensions that refer to temporal aspects of data sets: currency, volatility, timeliness, freshness, readiness. Here are some of the most compelling definitions across literature.

- Timeliness
  - The time delay from data generation and acquisition to utilization [62]. This requires data to be available within a certain time in order to be still fit for its purpose.
  - The degree to which data represents the reality at the required point in time [4].
- Currency
  - The age of information, based on how long ago it was recorded [12][30].
  - The degree to which the data set is current and up to date [30].
  - The probability that an attribute which was accurate at the moment of its storage is still congruent with its real-world counterpart at the moment of assessment [34]
- Volatility
  - The period for which information is valid in the real world [43][67].
  - As a measure of information instability, it's the frequency of change of the value of an attribute [12].

These definitions show that there is little agreement on the definition of time-related dimensions. Very often, methodologies focus on two aspects: the age and the volatility (or currency of data). While the age can be reasonably measured by the simple use of timestamps (one for the entry and one for the time of assessment), volatility requires a certain degree of modelling, in many cases for each attribute. Some of these approaches employ a distribution function to estimate the probability that the age of an attribute is larger than its "shelf life" [34], while others are proposing various decay functions for estimating whether an attribute is still current or not [30]. One such approach is described in Figure 5 and shows the comparison between an exponential decline function (using a decay factor as a parameter to the exponential function), linear decline and steep decline (instant decline).

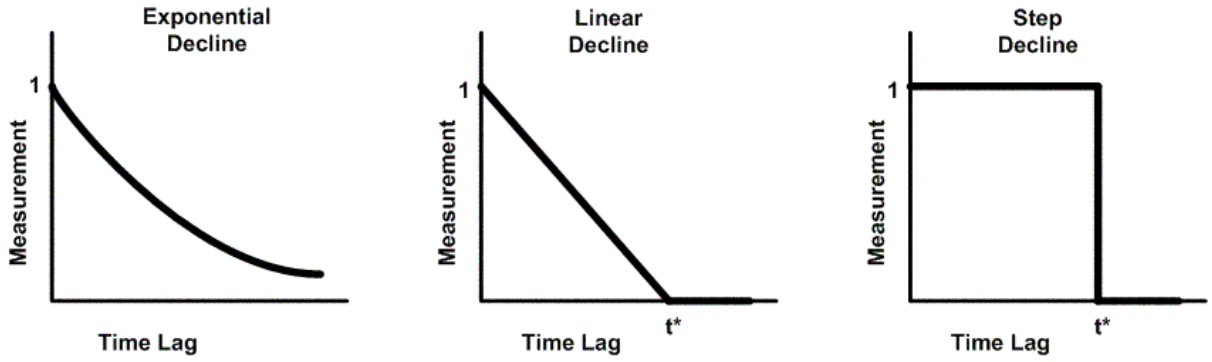


Figure 5 : Currency decline functions [30]

Due to its complexity, an implementation of a combined approach based on [30] and [34] will be available in the next version of the DVC. At the moment, the component deals with timeliness only in the context setting process, when users are required to declare the age of the data set and its next envisioned usage.

5 A Word on Valuating Privacy

The literature contains evidence that both sharing and protecting personal data can have positive and negative consequences at individual and societal levels [1]. Clearly leveraging the power of personal data results in benefits for the consumer (geo-location services, personalised e-commerce etc.), but at the same time they can incur financial loss and disutilities from violations of their privacy. There must be a balance between the extraction of economic value from data and the protection of individual

privacy, as in all the other aspects of data valuation, is context-dependent and can be achieved through the right combination of regulatory interventions, technological solutions and economic incentives [1].

As the world moves towards a universal online presence, we need to augment the debates surrounding our digital "selves", how are they built and subsequently used. This past decade has revealed the inability of our current legal frameworks to keep pace with the rhythm of technological development. Updated legal frameworks must address the tension between the global amplitude of objectives and operations of data-centric companies on one side, and the highly diverse regional and national legislation governing privacy and data protection, on the other. This starts with the very definition of "private data" (which varies with different jurisdictions) and the very important issues surrounding its ownership: from a technical perspective this could be addressed with a model akin to digital rights management; however, from ethical and social perspectives it centres the conversation on the digital extension of human identities and the implications of attaching them a sense of property, which can be bought, sold and owned by others. And even when / if data ownership will be resolved, the question then further extends to the "trade of behavioural futures", as Shoshana Zuboff [90] characterises the prediction products developed with such data. She mentions two solutions to be pursued: first, the need for new forms of collective actions, akin to the 20th century institutions of a strike and collective bargain; second, creating the opportunity for competitive solutions and supporting them if they play by improved rules.

Context base outcomes from surveys and empirical studies on the evaluation of data by individuals

Data has now become a resource of its own and raises a lot of legal, psychological, ethical, and economical question linked to the value given by individuals to their data. However, many interrogations remain unanswered. Privacy paradox and data ownership constitute two illustrations of these debates. While extensive analysis of data ownership debate is offered in WP3, deliverable D3.4, in the next section, we have focused on the discussion concerning the economic evaluation of personal data by data subjects.

Lack of clear trends (privacy paradox)

Data protection and technologies trigger some interesting behaviours, and one of them is the privacy paradox hypothesis. In recent years, research, surveys, barometers have shown that privacy is a significant concern for citizens in the digital age [49][79]. However, in parallel, citizens are still willing to divulge, give up or trade their data for a very small benefit not just for being able to use some services (social network) or popular apps. This dichotomy demonstrates an inconsistency between the values individuals carry and their behaviours to put these values in practice and was named "privacy paradox" [14][65].

Researchers attempted to substantiate this hypothesis and find explanations to this privacy paradox, but contradicting results and incomplete explanations were discovered [49]. There is a lack of a clear trend in the research outcomes. "Several studies have shown a dichotomy between privacy concerns and attitudes and actual privacy behaviour, whilst other studies indicate that individuals' privacy behaviour is in line with their concerns and attitudes" [49]. Some argued that despite the interest in privacy and data protection, the knowledge stays theoretically limited and not translated in actual protective behaviour [45]. These researches are de facto interlinked with studies on the value given by individuals to their data [19][38][58][89]. Here are some elements attempting to explain the lack of a clear trend about this topic.

Firstly, there is a lack of clarity regarding the exact object of privacy paradox studies. Some scholars will look alternatively into privacy attitude, privacy concerns, privacy behaviours and privacy intention. All these concepts are fundamentally different, which influence the research outcomes considerably. Furthermore, some scholars have another perception of the privacy paradox and refer to the tension between personalisation and privacy [84].

Secondly, there is a problem of interpretation. Indeed, when individuals provide price estimation for their data (for instance, seven euros), some may think it's too low and that individuals do not value enough their privacy [19]. Still, others argue attributing a price testifies that they value their privacy [49].

Thirdly, privacy behaviour is a highly contextual phenomenon [49]. Economic research on the subject showed that individuals might not be able to behave as an economically rational agent when it comes to privacy as their privacy-related decisions. In fact, individuals' behaviour is affected by "incomplete information, bounded rationality and psychological biases, such as confirmation bias, hyperbolic discounting and others" [1][33][49]. The context highly influences individuals' behaviour towards privacy; thus, similar studies conducted in a different context may lead to entirely different results. Even the legal framework has an impact on the individuals' valuation of their data as it was observed that more protection is given to data through the legal framework, more individuals are becoming aware of the value of their data [89].

In conclusion, the privacy paradox has already been the object of multiple types of research which discovered several pieces of this complex phenomenon. Nonetheless, a complete picture linking the different puzzle pieces is still missing. Regarding the law, this lack of clear trend has an impact on the design of data protection policymaking. Concerning data markets, the absence of consensus on privacy paradox shows the difficulty of assessing data value.

6 Aggregating and Reporting

The success of a data valuation platform depends on its adoption by data practitioners, which given the multi-dimensional nature of data quality and the complexities of data valuation, is dependent on: the capacity of the platform to promote the transparency of the assessment processes, the interpretability and replicability of results, and the degree to which such results can be used by practitioners. The latter issue leads to a discussion about the necessity of aggregating the results of the processes that support data quality and data valuation into a single measure that can be easily understood at different levels of organisations and based on which ultimately, decision of economic nature can be made.

Obviously, notions such as an "energy label" for data or a "price tag" for data are appealing, especially for those operating at commercial or executive levels of organisations. However, as we will see in this section, such aggregate measures are both difficult to construct (at least for now) and may lead to confusing or inaccurate interpretations, undermining the whole data valuation effort.

Interest in developing a single measure to characterise data has first appeared in the context of DQA: is there a single measure, able to aggregate different DQMs (corresponding to DQDs) [67]?

6.1 The Challenge of Aggregating DQDs and DQMs

Pipino et al. draw attention on the fact that a single-value aggregated measure – a quality index – could be subject to the same deficiencies that affect other commonly used indices: Dow Jones Industrial Average, Consumer Price Index[13] [67].

Similar shortcomings are also noticed by Bronselaer et al. who warn about the difficulty in interpreting an aggregation of DQMs referring to very different meaning. And while Pipino et al. signal the difficulty of aggregating DQMs operating on different scales, Bronselaer et al. believe that standardizing all DQMs in the [0,1] scale can make the interpretation of an aggregate result even more misleading. Their solution proposes the construction of "a set of [quality] predicates that can be evaluated against data and a capacity function that expresses the contribution of each combination of predicates with respect to overall quality" [13].

Bergdahl et al. mention that previous attempts to compile composite indicators for data quality by NSIs have failed and refer to the contextual nature of DQA as the main constraint for selecting the right subset of indicators and assigning them suitable weights [10].

[13] Some of these controversies refer to the statistical methods used for estimations, the interpretation of different components, their contribution weight to the final index, the methodologies used for choosing these components etc.

6.2 Proposed Solutions

Pipino et al. believe that creating an aggregate measure for data quality is feasible over time within separate industries, once DQA becomes the norm. Thus, each industry could develop its own set of DQMs which would become its de facto quality standard and over time aggregated performance measures could be generalised across industries [67].

One of the first designs of an aggregate measure proposes a linear combination of the composing DQDs (accuracy, completeness, validity, interpretability, accessibility) [18].

$$Q_{Agg} = w_{Accu} \times Q_{Accu} + w_{Comp} \times Q_{Comp} + w_{Valid} \times Q_{Valid} + w_{Inter} \times Q_{Inter} + w_{Access} \times Q_{Access}$$

Nevertheless, this approach is exposed exactly to the problems presented before, regarding the difficult, highly contextual choice of values for the weights.

The CIHI methodology (see Annex) offers a compelling solution to the problem of aggregating a large variety of DQDs, by proposing a four levels hierarchical model. 86 basic *quality criteria* that form the base level are aggregated into 24 *quality characteristics*, and further into 5 *quality dimensions* – accuracy, timeliness, comparability, usability, relevance – which are finally aggregated into an overall *database evaluation*. This hierarchical approach using successive aggregations may strike a good balance between the need for composite measure and the interpretability coming from the composing dimensions.

In their DQA assessment of web portals, Calero et al. use Bayesian networks and fuzzy logic in order to aggregate several DQDs (e.g., applicability, availability, believability, flexibility etc.) into a 3-levels assessment (good, medium, bad) [17].

An improvement to the linear combination method presented before is that of Even and Shankanarayanan [30], who propose that the aggregate DQM be based on the algebraic product of the DQMs composing it:

$$Q_{Agg} = Q_{Accu} \times Q_{Comp} \times Q_{Valid} \times Q_{Curr}$$

This formula may be a bit too conservative, but its interpretation is intuitive: the aggregate quality of data is 0 if at least one of the composing DQMs is 0; it only reaches the value of 1 (perfect quality) if all composing DQMs are equal to 1.

Gorz et al. propose an adaptation to this approach, by introducing an indicator function, built upon the requirements for DQMs presented in Section 4.2.1. This function is meant to measure accuracy, which in the authors' view is the best proxy for overall data quality (in absence of "gold standard" data) and can be expressed as a function of the other basic DQDs:

$$Q_{Ind}^{v_I(t_n.a_m)}(Q_{Comp}(v_I(t_n.a_m)), Q_{Valid}(v_I(t_n.a_m)), Q_{Curr}(s^{v_I(t_n.a_m)})) = Q_{Comp}(v_I(t_n.a_m)) \times Q_{Valid}(v_I(t_n.a_m)) \times Q_{Curr}(s^{v_I(t_n.a_m)})$$

This formulation means that currency will only be computed if data is both complete and valid and, like before, if either of these checks fails the aggregate measure will be 0.

In the current version of the DVC we are using this approach to aggregate the quality measures. Next, this carries into the aggregate value of data, which uses a simple average between the quality and the contextual scores (i.e., a linear combination of equal weights). This approach will be extended in the next version of the component when machine learning applicability measures will also be included.

6.3 Reporting Data Value and Data Quality

Finally, in this section we look at several approaches to reporting on data quality and by extension on data value. Reporting is paramount in promoting the adoption of innovative platforms and this may be crucial in the case of complex evaluation processes like those comprised in the DVC.

A first component of reporting is data profiling, which is usually performed as an entry point to data quality management [50], right before data analysis. This results in an initial insight into the data (ranges, distributions of attributes, pair-wise correlations etc.) and will support in gathering the data quality requirements [46]. In the DVC we perform our own data profiling, which is presented to the user as soon as the data set is uploaded. This serves as a base for the definition of data quality rules that support the DQA according to the implemented DQDs: completeness, domain and format validity.

Once DQA is performed, there are several approaches to reporting an often-multi-dimensional result and eventual aggregates. The COLDQ methodology evaluates the cost associated with poor quality data and summarises it in a data quality scorecard. Similarly, the DaQuinCIS methodology issues a certificate of quality associated to data or a quality alert, depending on whether the quality requirements are satisfied or not.

A similar approach is discussed by Bergdahl et al. when discussing labelling and certification of data to promote more accurate user quality assessments and credibility gains [10]. The label message could be related to the result of the DQA, the data itself or the provider of a certain data set. It is recommended that only a small number of self-explanatory and recognisable labels (e.g., "sufficient quality". "experimental data") are created and that once introduced to stay in circulation for some time. The authors document various experiments with NSIs around the world (United Kingdom, Sweden, Finland, New Zealand) and observe that two labelling / certification methodologies are usually employed: commitment-in-advance (ex-ante) and attachment-after-checking (ex post). Finally, an important issue concerns old data which once certified under a certain set of quality requirements, will need to be re-evaluated when these are changing. A possible solution to this would be to consider expiration dates for labels, together with a recertification mechanism.

Finally, when reporting the value of data, we recall the observations made of Slotin et al. with respect to the efficacy of impact-based approaches to data valuation [78]. The success of these approaches consists in the fact that they are able to tell compelling stories based on data and connect them to clear outcomes and contexts. This is also echoed by the Data Narratives approach [35], which acknowledges that "the value of big data is not data, but the narrative that it generates and supports[14].

7 Conclusions and Future work

Throughout this report, we tried to create a complete picture of the challenges to overcome to design and implement a solution for data valuation. We tried to shed more light on the fuzzy notion of "data value" and found that its complexity stems in:

i. potentially flawed comparisons to other currencies (oil) or intangible goods, which leads to considering the wrong economic models for valuating it
ii. a complex data value chain, with many possible ramifications from each step of the data processing pipeline

We defined a data valuation process, which highlights the main components contributing to data value: context, data quality (intrinsic and contextual), applicability in context and privacy. We turned to a review of economic models for data valuation, with the aim of finding a method that allows putting together the previously mentioned components. The limitations of such models – highly contextual, complex, and fairly abstract – makes it difficult to proceed to their implementation. Analysing the landscape of personal data transactions provided us with some concrete economic valuations of a wide range of data types. Put together, we were able to draw important conclusions about key properties of data, generalising beyond personal data:

[14] Their entire approach is very interesting. They start from the story (the communication) itself, and define the information needs, which in turn defines the kind of analyses that can be performed with the facts at hand. Finally, the required facts define how you are going to derive these elements of information from the data you have.

i. Information is infinitely shareable, and more data does not necessarily mean more value. It is likely that given a context, there exists a tipping point after which collecting more data will not make the data set more valuable.
ii. Processed data is more valuable than raw data. This is connected to how data moves through a processing value chain, from raw data to information, supporting an action and eventually creating value.
iii. There is a trade-off between the applicability of data, resulting from its level of aggregation and privacy. In some contexts, aggregation decreases the value of data, but it partially addresses privacy issues. Nevertheless, these trade-offs are very hard to measure.
iv. Combining different data sources increases the value of data. Otherwise put, the value of the combined data is better than the sum of values of each individual data set. Personal data in particular, has this cumulative property: the more data is collected about an individual, the more valuable it becomes.
v. In the case of processing personal data, several preferences can be observed:
    a. behavioural data is preferred to demographic data;
    b. time series and location data are considered to bear more value;
    c. dynamic data is preferred to static data.

Besides these, there were some equally surprising conclusions:

i. The considerable gap between the price for which brokers are buying and selling personal data, and how much individuals believe this data is worth.
ii. The fact that individuals are willing to allow for the harvesting of the most sensitive of their personal data, despite understanding its importance and the fact that it is being under-valued by brokers.

We proceeded to an analysis of each element contributing to data value. First, we looked at contexts, which despite their overwhelming recognition as a source for the complexity of DQA and data valuation, were never approached from a formal perspective. To bridge this gap our approach puts together results from research on methods for mapping data properties to value, the development of datasheets for data sets or data set nutrition labels. This enabled the implementation of our own process for establishing contexts, in which users answer a set of questions focused on five main areas: systems & economics, legal & obligations, data science, data properties, business impact.

Next, we proceeded to an in-depth analysis of data quality. Our review of literature confirmed the contextual dependence of data quality assessment; even the several DQA methods that claim to be context free – value based DQA [30] or building indicator functions for DQA [34] – still require a certain level of parametrisation with respect to the data quality requirements, which in our conception is a form of contextualising the process. The review of Batini et al. [8] provided us with in-depth insights into the various perspectives that weigh in the construction of DQA methodologies: structural phases, strategies for assessment, dimensions and metrics, costs, types of data, types of information systems.

The most practical part of our work is the review on data quality dimensions (DQD) and metrics (DQM). We consider a framework in which the DQA process relies on its composing DQDs, which in turn are operationalised by multiple DQMs. As Pipino et al. observe, defining a DQD is challenging, but once it is done, the definition of its underlying DQMs quickly follows suit [67]. At this point, the reader will not be surprised to learn that this process is highly contextual, which results in the high number of DQDs and DQMs that we documented. While there is little consensus on which of them are recommended, both our literature review and an empirical evaluation by Batini et al. [8] uncovered a set of core DQDs: accuracy, completeness, validity, consistency and time-related. With respect to the latter, we have seen it comes in many “flavours”, each of them measuring a certain aspect related to time – timeliness, currency, volatility – and that depending on the context these may overlap to certain extent. The procedures for implementing time-related measures may require parametrisations (shelf life of data, decay parameters) and assumptions about the statistical distribution of the decay functions (e.g., exponential, linear, step).

All these considerations, together with the review of individual methodologies (see Annex) resulted in valuable theoretical and practical lessons that prompted us to refocus our work within this larger framework.

With respect to the role of privacy in data valuation, we discussed the challenges that need to be overcome to achieve the desired balance between enabling economic value of data and preserving privacy. Thanks to a contribution from our colleagues from KUL (WP3), we reviewed the privacy paradox as it appears in personal data markets.

Currently, privacy is factored into our solution only through the context gathering process, in which we assess – based on user statements – the degree to which a data set respects EU GDPR's privacy principles. We plan to expand this work in the next version of our platform, by including the privacy preserving techniques developed by our colleagues at RSA (WP5).

Finally, we bring all these perspectives (context, data quality, privacy) together, using aggregate metrics. While practical in terms of communicating a final result (and thus very appealing!), members of the data quality community advise to proceed with caution. Aggregate measures may over-simplify the complexities behind DQA – the same observation clearly extends to data valuation – and opens them to the criticism of other aggregate indices: difficult to interpret, difficult to choose the weights of the aggregating components, difficulty in agreeing on an aggregation method. While current approaches lean towards linear models or algebraic products, once data valuation is integrated in data markets and more concrete information is collected, we are very optimistic about the possibility to refine these aggregation models. In the meantime, the recommendation is to focus on reporting the results either by well-documented subsequent aggregations, by implementing a labelling / certification system [10] or through the creation of narratives, as suggested in impact-based valuation methods [78].

With a set of DQMs already implemented (completeness, validity, consistency), our work will now focus on the following several areas:

1. extend the capabilities of the current dimensions. This is the case of consistency where we need to develop more data quality rules (string regex and datetime format validation are currently supported).
2. implement additional dimensions.
    a. **Time-related measures** are on top of our priorities (age, currency), together with the underlying decay models.
    b. **Availability and accessibility** are currently declared by the user as part of the context gathering process and quantified within. We want to study the possibility of automatising this (probably once the platform is deployed in a data market) and considering their scores as independent DQMs.
    c. **Security and privacy**. We are looking forward to the integration with the work in WP5 which could open the possibility of estimating these DQMs and include them in the final score.
    d. **Performance**. This measure will refer to the usability of the data set in intended contexts (e.g., advance analytics, training machine learning models).
3. aggregate measures. Improve the currently basic aggregate measures, by using economic models for data value, which make use of the declared context.
4. validation is an area that still raises difficulties, which we hope to overcome once the component will be deployed and used in a data market. Until then, our strategy will focus on three alternatives:
    a. evaluate personal data sets, for which the review in Section 2.2.2 provides us with a satisfactory "ground truth";
    b. evaluate open data sets [11][15][20][29] [37][75], which were previously used in case studies for other valuation methods [32][37];
    c. work with our colleagues from FNET and IFAG (WP6, WP7) and use their in-house data sets and perform a qualitative evaluation of the results.
5. integrate the work on privacy preserving techniques and chanc estimators, developed by our colleagues at RSA and KNOW.

8 References

[1] Acquisti, A. and Heinz, H.J. (n.d.). Privacy in Electronic Commerce and the Economics of Immediate Gratification.

[2] Acquisti, A., Taylor, C., and Wagman, L. (2016). The Economics of Privacy. Journal of Economic Literature, 54(2), 442–492.

[3] Arnold, M., Bellamy, R. K. E., Hind, M., Houde, S., Mehta, S., Mojsilovic, A., Nair, R., Ramamurthy, K. N., Reimer, D., Olteanu, A., Piorkowski, D., Tsay, J., and Varshney, K. R. (2019). FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity. http://arxiv.org/abs/1808.07261

[4] Askham, N., Cook, D., Doyle, M., Fereday, H., Gibson, M., Landbeck, U., Lee, R., Maynard, C., Palmer, G., and Schwarzenbach, J. (2013). The Six Primary Dimensions for Data Quality Assessment—Defining data quality dimensions. DAMA UK.

[5] Baldwin, H. (2015). Drilling Into The Value Of Data. Forbes. Retrieved from https://www.forbes.com/sites/howardbaldwin/2015/03/23/drilling-into-the-value-of-data/

[6] Ballivian, A. and Fenohasina, R.M. (2015) Measuring the Value of Data. Available at: https://statswiki.unece.org/download/attachments/117772954/World%20Bank_Ballivian_Mare_MeasuringtheValueofData_20151202.pdf?version=1&modificationDate=1473158675433&api=v2

[7] Batini, C., Cabitza, F., Cappiello, C. and Francalanci, C. (2008). A comprehensive data quality methodology for Web and structured data. In: International Journal of Innovative Computer Applications 1, 3, 205–218.

[8] Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. (2009). Methodologies for Data Quality Assessment and Improvement. ACM Computing Surveys, 41(3).

[9] Behkamal, B., Kahani, M., Bagheri, E., and Jeremic, Z. (2014). A Metrics-Driven Approach for Quality Assessment of Linked Open Data. Journal of Theoretical and Applied Electronic Commerce Research, 9(2), 11–12.

[10] Bergdahl, M., Elvers, E., Földesi, E., Kron, A., Lohauß, P., Mag, K., Morais, V., Nimmergut, A., Viggo Sæbø, H., Timm, U., and Zilhão, M. J. (2007). Handbook on Data Quality Assessment Methods and Tools. European Commission - Eurostat.

[11] Bolukbasi, T., Chang, K.-W., Zou, J.Y., Saligrama, V., Kalai, A.T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Advances in Neural Information Processing Systems. pp. 4349–4357.

[12] Bovee, M., Srivastava, R., and Mak, B. (2001). A conceptual framework and belief-function approach to assessing overall information quality. In Proceedings of the 6th International Conference on Information Quality.

[13] Bronselaer, A., De Mol, R., and De Tre, G. (2018). A Measure-Theoretic Foundation for Data Quality. IEEE Transactions on Fuzzy Systems, 26(2), 627–639.

[14] Brown, B. (2001). Studying the Internet Experience.

[15] Buolamwini, J., Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Friedler. S.A., Wilson, C. (eds). New York, NY, USA: PMLR; 2018. pp. 77–91.

[16] Cai, L., and Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14(2), 1–10.

[17] Calero, C., Caro, A., Piattini, M. (2008). An applicable data quality model for web portal data consumers. World Wide Web. 4, 465-484.

[18] Cappiello, C. and Comuzzi, M. (2009). A Utility-Based Model to Define the Optimal Data Quality Level in IT Service Offering. In: Proceedings of the 17th European Conference on Information Systems (ECIS), pp. 1062-1074. Verona (Italy).

[19] Carrascal, J.P., Riederer, C., Erramilli, V., Cherubini, M., and de Oliveira, R. (2013). Your browsing behavior for a big mac. pp. 189–200.

[20] Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W., Choi, Y., Liang P., and Zettlemoyer, L. (2018). QuAC: Question Answering in Context. CoRR.

[21] Cope, S., and Gillula, J. (2015). AT&T is putting a price on privacy. The Guardian. Available at https://www.theguardian.com/commentisfree/2015/feb/20/att-price-on-privacy

[22] Crosby, P. B. (1988). Quality is Free: The Art of Making Quality Certain, New York: McGraw-Hill.

[23] CSO Insights. (2005). 2005 Executive Report: Target Marketing Priorities Analysis.

[24] Davies, J. (2016). Consumers price their data at £2k – Companies pay 45p. Telecoms.Com. Retrieved from https://telecoms.com/474623/consumers-price-their-data-at-2k-companies-pay-45p/

[25] De Amicis, F., Barone, D., and Batini, C. (2006). An analytical framework to analyze dependencies among data quality dimensions. In: Proceedings of the 11th International Conference on Information Quality (ICIQ). 369–383.

[26] Dumbill, E. (2014). Understanding the Data Value Chain. IBM Big Data & Analytics Hub. https://www.ibmbigdatahub.com/blog/understanding-data-value-chain

[27] English, L. (1999). Improving Data Warehouse and Business Information Quality. Wiley & Sons.

[28] Eppler, M. and Munzenmaier, P. (2002). Measuring information quality in the Web context: A survey of state-of-the-art instruments and an application methodology. In: Proceedings of the 7th International Conference on Information Systems (ICIQ).

[29] Erkut Erdem. 2018. Datasheet for RecipeQA.

[30] Even, A., and Shankaranarayanan, G. (2006, November 10). Value-Driven Data Quality Assessment. Proceedings of the 2005 International Conference on Information Quality. MIT IQ Conference, MIT, Cambridge, MA, USA.

[31] Falorsi, P., Pallara, S., Pavone, A., Alessandroni, A., Massela, E., and Scannapieco, M. (2003). Improving the quality of toponymic data in the Italian public administration. In: Proceedings of the ICDT Workshop on Data Quality in Cooperative Information Systems (DQCIS).

[32] Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé III, H., and Crawford, K. (2020). Datasheets for Datasets. http://arxiv.org/abs/1803.09010

[33] Gilovich, T., Griffin, D., and Kahneman, D. (eds) (2002). Heuristics and Biases. Cambridge University Press.

[34] Görz, Q., and Kaiser, M. (2012). An Indicator Function for Insufficient Data Quality – A Contribution to Data Accuracy. In H. Rahman, A. Mesquita, I. Ramos, and B. Pernici (Eds.), Knowledge and Technologies in Innovative Information Systems (Vol. 129, pp. 169–184). Springer Berlin Heidelberg.

[35] Hammond, K. J. (2013). The Value of Big Data Isn't the Data. Harvard Business Review.

[36] Heinrich, B., Kaiser, M., and Klier, M. (2007, July). Metrics for measuring data quality – Foundations for an economic data quality management. 2nd International Conference on Software and Data Technologies (ICSOFT), Barcelona.

[37] Holland, S., Hosny, A., Newman, S., Joseph, J., and Chmielinski, K. (2020). The Dataset Nutrition Label: A Framework to Drive Higher Data Quality Standards. In D. Hallinan, R. Leenes,

S. Gutwirth, and P. De Hert (Eds.), Data Protection and Privacy: Data Protection and Democracy (pp. 1–26). Oxford: Hart Publishing.

[38] Huberman, B.A., Adar, E., and Fine, L.R. (2005). Valuating privacy. In: IEEE Security and Privacy, vol. 3, no. 5. pp. 22–25, Sep. 2005.

[39] International Organisation for Standardisation. (2008). ISO/IEC 25012:2008 Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model.

[40] International Organisation for Standardisation. (2015). ISO 9000 Family for Quality Management Systems.

[41] ISTAT. (2004). Guidelines for the data quality improvement of localization data in public administration (in Italian). Available at: www.istat.it

[42] Jardine, A. (2016). What's your personal data worth? ITProPortal. Rertieved from https://www.itproportal.com/2016/07/24/whats-personal-data-worth/

[43] Jarke, M., Lenzerini, M., Vassiliou, Y., and Vassiliadis, P. (1995). Fundamentals of Data Warehouses. Eds. 1995. Springer Verlag.

[44] Jeusfeld, M., Quix, C., and Jarke, M. (1998). Design and analysis of quality information for datawarehouses. In Proceedings of the 17th International Conference on Conceptual Modeling.

[45] Joinson, A.N., Reips, U.D., Buchanan, T., and Schofield, C.B.P. (2010). Privacy, trust, and self-disclosure online. In: Human-Computer Interaction, vol. 25, no. 1, pp. 1–24, Jan. 2010.

[46] Jones, D. (2016). Data Profiling vs Data Quality Assessment – Let's Explain The Difference. Data Quality Pro. Retrieved from https://www.dataqualitypro.com/data-profiling-data-quality-assessment/

[47] Juran, J.M. (1951). Quality Control Handbook. 4th ed.

[48] Kannan, K., Ananthanarayanan, R., and Mehta, S. (2018). What is my data worth? From data properties to data value. http://arxiv.org/abs/1811.04665

[49] Kokolakis, S. (2017). Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. In: Computers and Security, vol. 64. Elsevier Ltd, pp. 122–134.

[50] Lebied, M. (2018). Guide To Data Quality Management & Metrics for Effective Data Control. Datapine. Retrieved from https://www.datapine.com/blog/data-quality-management-and-metrics/

[51] Lee, Y. W., Pipino, L., Strong, D. M., Wang, R. Y. (2004.) Process-Embedded Data Integrity. Journal of Database Management. 1, 87-103.

[52] Lee, Y.W., Strong, D. M., Kahn, B. K., and Wang, R. Y. (2002). AIMQ: A methodology for information quality assessment. In: Information Management. 40, 2, 133–460.

[53] Lehr, S. (2015). The True Cost of Bad (And Clean) Data. RingLead. Retrieved from https://www.ringlead.com/blog/true-cost-of-bad-data/

[54] Loudhouse. (2014). The Future of Digital Trust. A European study on the nature of consumer trust and personal data (Industry No. 2; The Future of Digital Trust, p. 7). Orange.

[55] Long, J. and Seko, C. (2005). A cyclic-hierarchical method for database data-quality evaluation and improvement. In: Advances in Management Information Systems-Information Quality Monograph (AMISIQ).

[56] Loshin, D. (2004). Enterprise Knowledge Management – The Data Quality Approach. Series in Data Management Systems, Morgan Kaufmann, chapter 4.

[57] Madsbjerg, S. (2017). It's Time to Tax Companies for Using Our Personal Data. The New York Times. Retrieved from https://www.nytimes.com/2017/11/14/business/dealbook/taxing-companies-for-using-our-personal-data.html

[58] Malgieri, G. and Custers, B. (2018). Pricing privacy – the right to know the value of your personal data. In: Comput. Law Secur. Rev., vol. 34, no. 2, pp. 289–303.

[59] Manyika, J. et al. (2013) Open Data: Unlocking innovation and performance with liquid information. McKinsey Global Institute, McKinsey Center for Government, McKinsey Business Technology Office.

[60] Mason, R. (2014). HMRC to sell taxpayers' financial data. The Guardian. Retrieved October 28, 2020 from https://www.theguardian.com/politics/2014/apr/18/hmrc-to-sell-taxpayers-data

[61] Mawer, C. (2015). Valuing Data is Hard. Silicon Valley Data Science. https://www.svds.com/valuing-data-is-hard/

[62] McGilvray, D. (2010). Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information, Beijing. Publishing House of Electronics Industry.

[63] Naumann, F. 2002. Quality-driven query answering for integrated information systems. Lecture Notes in Computer Science, vol. 2261.

[64] NHS Wales. (2004). Data Quality Initiative Framework. Project Report.

[65] Norberg, P.A., Horne, D.R., and Horne, D.A. (2007). The privacy paradox: Personal information disclosure intentions versus behaviors. In J. Consum. Aff., vol. 41, no. 1, pp. 100–126, Jun. 2007.

[66] Office of Management and Budget. (2006). Information quality guidelines for ensuring and maximizing the quality, objectivity, utility, and integrity of information disseminated by agencies. Available at: http://www.whitehouse.gov/omb/fedreg/reproducible.html

[67] Pipino, L. L., Lee, Y. W., and Wang, R. Y. (2002). Data quality assessment. Communications of the ACM, 45(4), 211–218.

[68] Piprani, B., and Ernst, D. (2008). A Model for Data Quality Assessment. In R. Meersman, Z. Tari, and P. Herrero (Eds.), On the Move to Meaningful Internet Systems: OTM 2008 Workshops (Vol. 5333, pp. 750–759). Springer Berlin Heidelberg.

[69] Regulation (EU) 2016/679 of the European Parliament and of the Council – Of 27 April 2016 – On the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/ 46/ EC (General Data Protection Regulation), no. Regulation (EU) 2016/679, European Parliament, 88 (2016). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

[70] Renieris, E. M., and Greenwood, D. (2018). Do we really want to "sell" ourselves? The risks of a property law paradigm for personal data ownership. Medium. Retrieved from https://medium.com/@hackylawyER/do-we-really-want-to-sell-ourselves-the-risks-of-a-property-law-paradigm-for-data-ownership-b217e42edffa

[71] Rose, J., Rehse, O., and Rober, B. (2012). The Value of our Digital Identity (p. 65). Boston Consulting Group.

[72] Ross, W. (2014). Is Your Smartphone Privacy Worth $100 a Month? MIT Technology Review. Retrieved October 18, 2019, from https://www.technologyreview.com/s/529686/how-much-is-your-privacy-worth/

[73] Scanapiecco, M., Virgillito, A., Marchetti, M., Mecella, M., and Baldoni, R. (2004). The DaQuinCIS architecture: a platform for exchanging and improving data quality in Cooperative Information Systems. In: Information Systems. 29, 7, 551–582.

[74] Sebastian-Coleman, L. (2010). Data Quality Assessment Framework. The Fourth MIT Information Quality Industry Symposium.

[75] Seck, I., Dahmane, K., Duthon, P., and Loosli, G. (2018). Baselines and a datasheet for the Cerema AWP dataset. CoRR. http://arxiv.org/abs/1806.04016

[76] Short, J. E., and Todd, S. (n.d.). What's Your Data Worth? MIT Sloan Management Review. Retrieved November 8, 2020, from https://sloanreview.mit.edu/article/whats-your-data-worth/

[77] Simonite, T. (2013). Coming Soon: Take Your Own Personal Data to Market. MIT Technology Review. Retrieved from https://www.technologyreview.com/s/517356/if-facebook-can-profit-from-your-data-why-cant-you/

[78] Slotin, J. (2018). What Do We Know About the Value of Data? Global Partnership for Sustainable Development Data.

[79] Smith, H.J., Dinev, t., and Xu, H. (2011). Information privacy research: An interdisciplinary review. In: MIS Quarterly: Management Information Systems, vol. 35, no. 4. University of Minnesota, pp. 989–1015.

[80] Smith, M. (2016). Proximus starts selling customer data reports for €700 a time. European Communications. Retrieved October 21, 2019, from https://eurocomms.com/industry-news/11968-proximus-starts-selling-customer-data-reports-for-700-a-time

[81] Stack, B. (2017). Here's How Much Your Personal Information Is Selling for on the Dark Web. Retrieved from https://www.experian.com/blogs/ask-experian/heres-how-much-your-personal-information-is-selling-for-on-the-dark-web/

[82] Staiano, J., Oliver, N., Lepri, B., de Oliveira, R., Caraviello, M., and Sebe, N. (2014). Money Walks: A Human-Centric Study on the Economics of Personal Mobile Data. Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '14 Adjunct, 583–594.

[83] Su, Y. and Jin, Z. (2004). A methodology for information quality assessment in the designing and manufacturing processes of mechanical products. In: Proceedings of the 9th International Conference on Information Quality (ICIQ). 447–465.

[84] Sutanto, J., Palme, E., Tan, C.H., and Phang, C.W. (2013). Addressing the personalization-privacy paradox: An empirical assessment from a field experiment on smartphone users. In MIS Q. Manag. Inf. Syst., vol. 37, no. 4, pp. 1141–1164.

[85] van Nederpelt, P., and Daas, P. (2012). 49 Factors that Influence the Quality of Secondary Data Sources. In: Quality and Risk Management (12). Statistics Netherlands. The Hague.

[86] Vasudha, T., and Arvind, G. (2017). The value of data. World Economic Forum. https://www.weforum.org/agenda/2017/09/the-value-of-data/

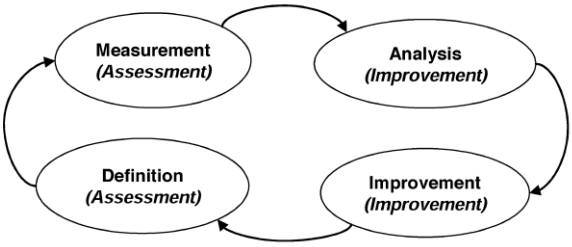
[87] Wang, R. (1998). A product perspective on total data quality management. In: Communications of ACM. 41, 2.

[88] Warner, M.R., and Hawley, J. (2019). Designing Accounting Safeguards To Help Broaden Oversight and Regulations on Data. Retrieved January 17, 2020 from https://www.congress.gov/bill/116th-congress/senate-bill/1951/text

[89] Winegar, A.G. and Sunstein, C.R. (2019). How Much Is Data Privacy Worth? A Preliminary Investigation. In: J. Consum. Policy, vol. 42, no. 3, pp. 425–440.

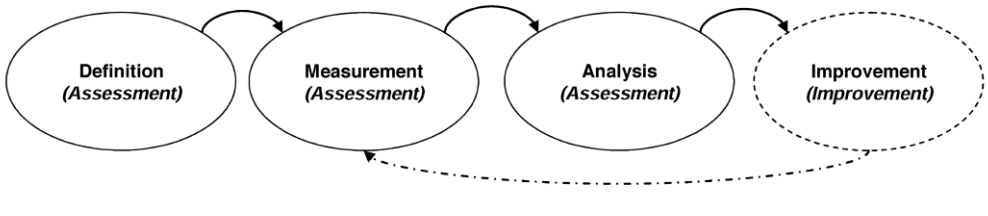
[90] Zuboff, S. (2019). The Age of Surveillance Capitalism. Profile Books Ltd.

9 Annex



Phase 1: **Definition** .
Input → Information product
Data Analysis: Definition of the characteristics at two levels: (1) high level: description of functionalities for information consumers (2) low level description of the basic units and components of the information product and their relationships
DQ Requirements Analysis: Definition of the IQ requirements from the perspective of IP suppliers, manufacturers, consumers and managers.
Process Modeling: Definition of the Information Manufacturing System
Output → Logical and physical design of the Information Product with the necessary quality attributes.
A quality entity-relationship model that defines the IP and its IQ requirements
An information manufacturing system that describes how the IP has been produced
Phase 2: **Measurement**
Input→ IQ dimensions
Measurement of quality: Definition of the IQ metrics
Output → IQ problems
Phase 3: **Analysis**
Input → IQ problems
Identification of the causes of errors: Determination root causes of discrepancies
Output → Actions for improving data quality
Phase 4: **Improvement**
Input → IQ metrics
Selection of strategies and techniques: Identification of key areas for improvement
Output → IQ improvement techniques

Figure 6: Total Quality Data Management (TQDM) [87]



Phase 1: **Definition**
Input → Data Warehouse project and context information, operational databases, stakeholders' perspective
Data Analysis: Identification of relevant data quality dimensions in the data warehouse context and identification of the relations between data quality dimensions and data warehouse objects
DQ Requirements Analysis: Definition of the House of quality in which stakeholders assign weights to quality dimensions in order to express their relevance
Output → List of the quality dimensions and related assessment methods, quality requirements
Phase 2: **Measurement**
Input→ List of the quality dimensions and related assessment methods
Measurement of quality: Hierarchical Quality Assessment and identification of dependencies among data quality dimensions
Output → DQ values, dependencies among data quality dimensions
Phase 3: **Analysis**
Input → DQ values, dependencies among data quality dimensions, DQ Requirements
Identification of critical areas: Comparison between data quality values and data quality requirements
Output → List of dimensions that need improvement, effects of dependencies
Phase 4: **Improvement (only mentioned)**
Input → List of dimensions that need improvement, effects of dependencies
Output → Improved data

Figure 7: Data Warehouse Quality (DWQ) [44]

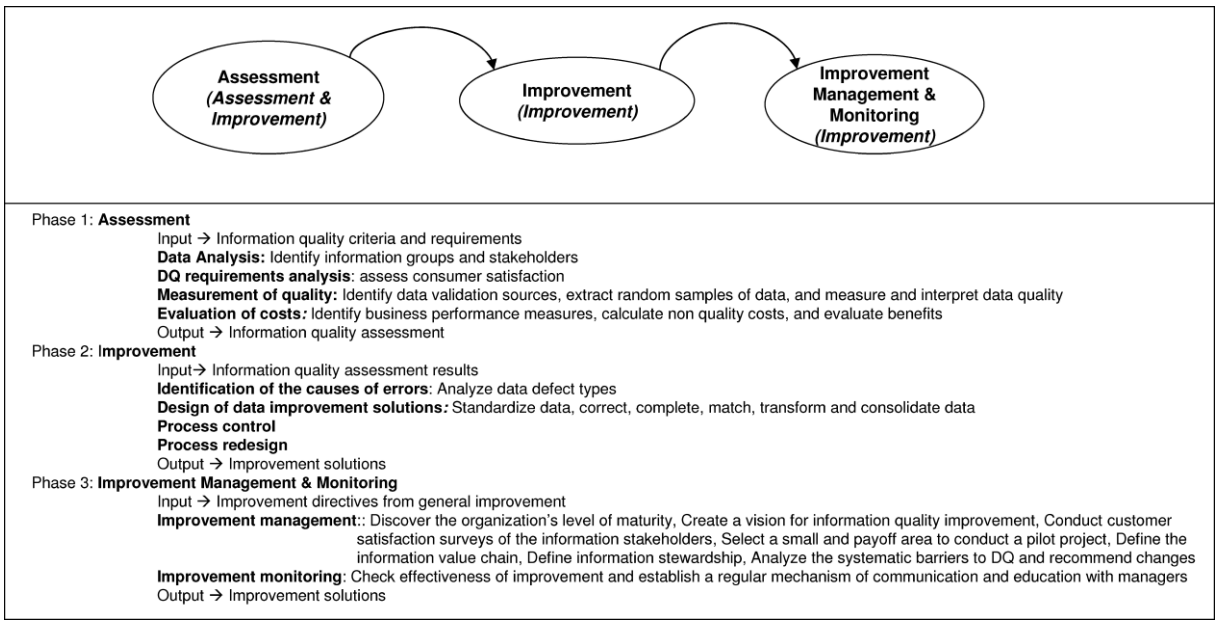


Figure 8: Total Information Quality Management (TIQM) [27]

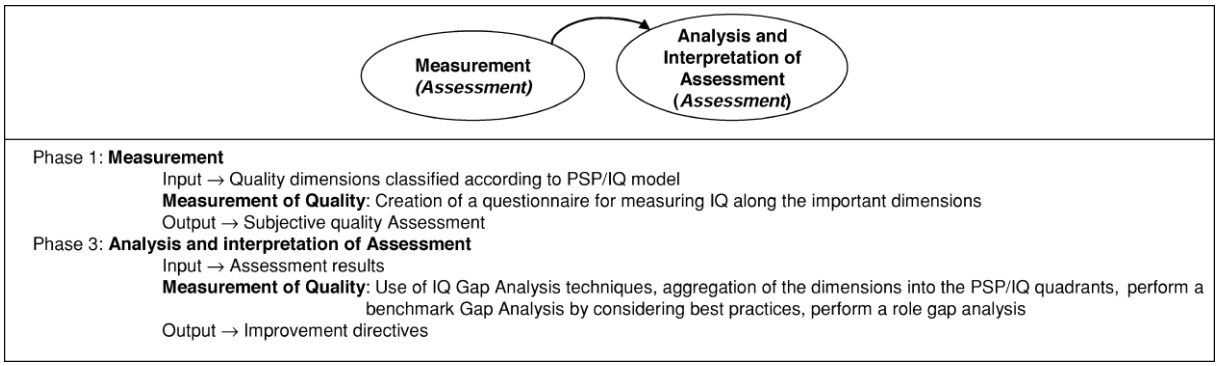


Figure 9: A Methodology for Information Quality Management (AIMQ) [52]

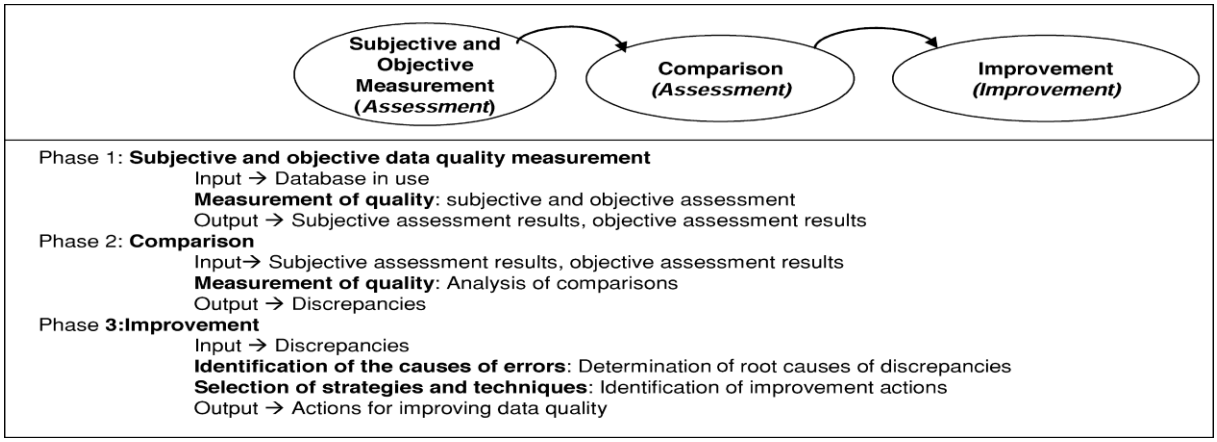


Figure 10: Data Quality Assessment (DQA) [67]

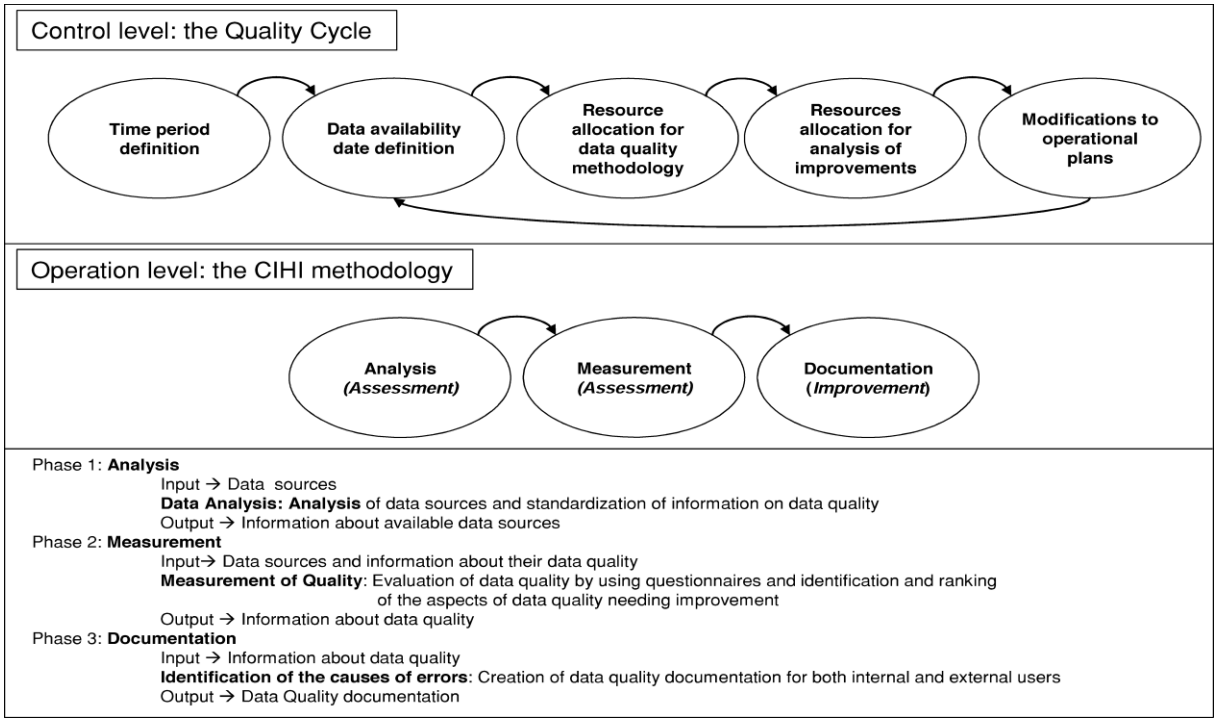


Figure 11: Canadian Institute of Health Information (CIHI) [55]

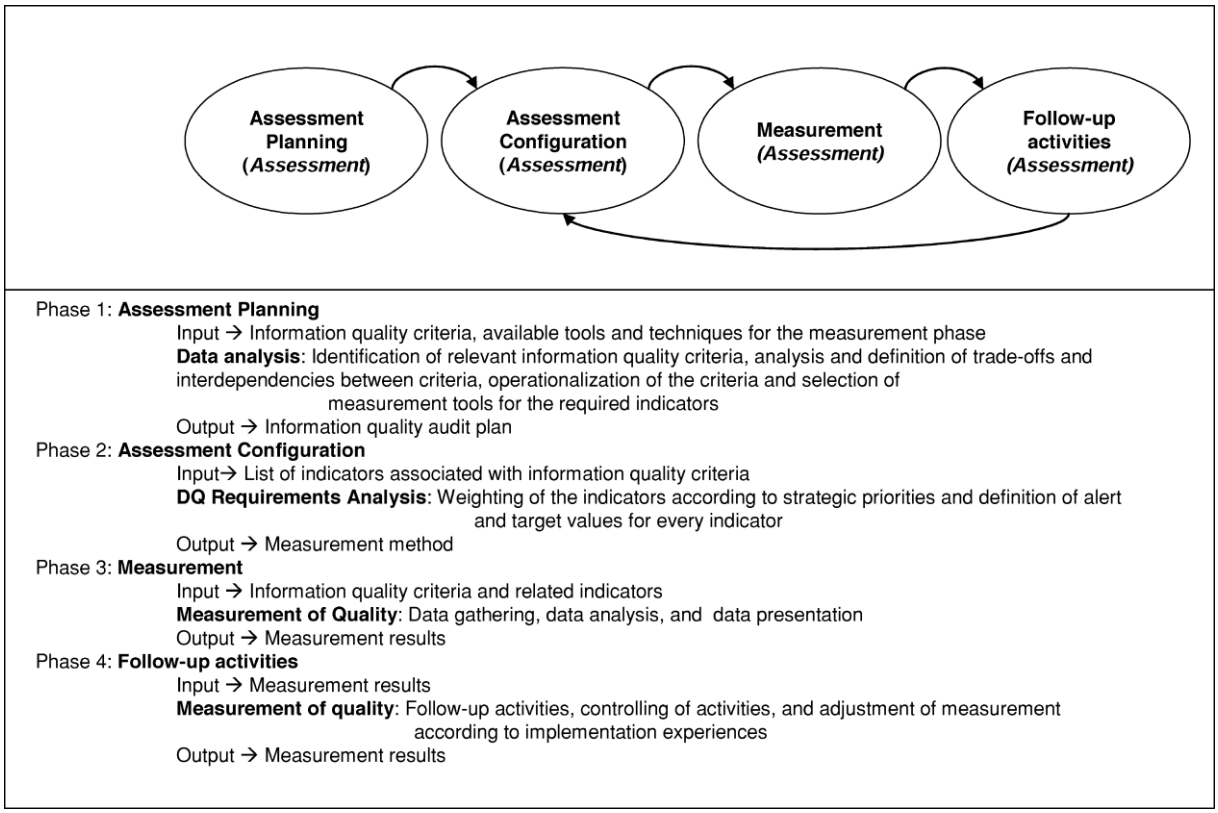
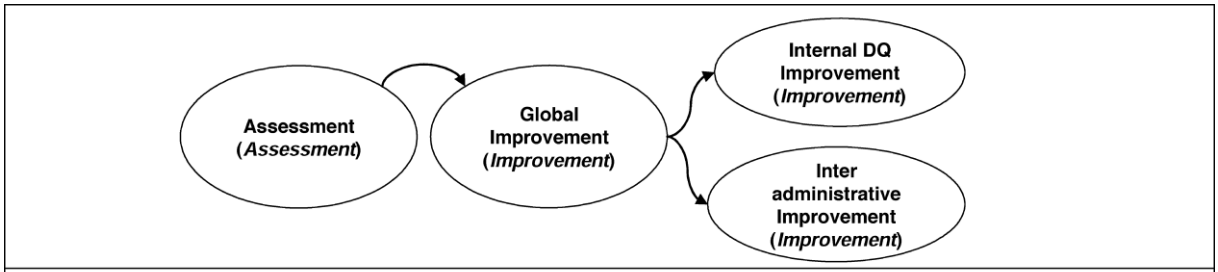


Figure 12: Information Quality Measurement (IQM) [28]



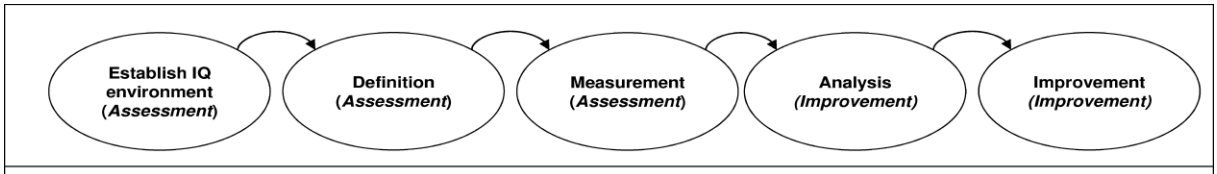
Phase 1: **Assessment**
- Input → Information quality criteria and requirements
- **DQ Requirements analysis** – Isolate from a general process analysis relevant qualities for address data
- **Identification of critical areas**: Find critical areas by using statistical techniques and by analyzing sample database
- **Measurement of quality**: Measure data quality dimensions and communicate results of assessment to single agencies
- Output → Information quality assessment results

Phase 2: **Global improvement**
- Input→ Information quality assessment results
- **Identification of the causes of errors**
- **Design of data improvement solutions**: Perform record linkage among relevant national databases
- **Assignment of data responsibilities**: Establish a national data owner for specific fields
- **Process control and Process Redesign**: Design improvement solutions on processes, the results of the global assessment to decide specific interventions on processes
- **Selection of strategies and techniques**: Make or buy, and adapt, tools for most relevant techniques
- **Improvement management**: DQ activities to deliver to agencies
- Output → Improvement method

Phase 3: **Internal DQ improvement (for each agency, autonomous initiative)**
- Input → Improvement directives from global improvement
- **Process Redesign:** Standardize acquisition format and standardize internal exchange format using XML
- **Improvement Monitoring**: Perform specific local assessments, design improvement solutions on data and processes in critical areas
- Output → Improvement solutions

Phase 4: **Inter administrative improvement**
- Input → Improvement directives from general improvement
- **Process Redesign**: Standardize interadministrative flows format using XML, redesign exchange flows, using a public and subscribe event driven architecture
- Output → Improvement solutions

Figure 13: Italian National Bureau of Census (ISTAT) [31][41]



Phase 1: **Establish IQ Environment**
- Input → Business context
- **Data Analysis**: Evaluation of the enterprise context
- **DQ requirements analysis**
- Output → Dimensions of PIQ, time series of business activities

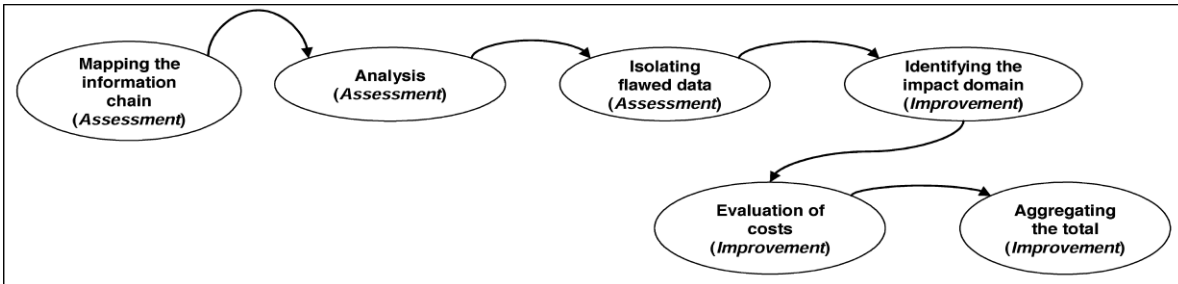
Phase 2: **Definition**
- Input→ Characteristics of the PIQ
- **Process Modelling**: Modelling the enterprise objects
- **Identification of critical areas**: Identification of the PIQ dimensions of Enterprise Activity
- Output → PIQ specifications

Phase 3: **Measurement**
- Input → PIQ specifications and dimensions
- **Measurement of quality**: Identification of Performance measures of PIQ for each activity, selection of a small, manageable pilot project, Definition of complementary metrics and of a final questionnaire to collect data about PIQ, and measures validation
- Output → Measurement results

Phase 4: **Analysis**
- Input → Measurement results
- **Identification of causes of errors**: Analysis of poor information quality dimensions, gap analysis and evaluation of effective performance with respect to the enterprise goal
- Output → Root causes for poor quality dimensions

Phase 5: **Improvement**
- Input → Root causes for poor quality dimensions
- **Assignment of process and data responsibilities**: Application of information resource management (IRM) to improve the PIQ performance in enterprise activities.
- **Process control and redesign**: Application of self-organization management (SOM) to reengineer information flow and workflow with infrastructure
- **Improvement Monitoring**: Development of enterprise objective management (EOM) to realign the value and characteristics of EG
- Output → Long-term improvement of PIQ

Figure 14: Activity-based Measuring and Evaluating of Product Information Quality (AMEQ) [83]



Phase 1: **Mapping the information chain**.
Input → Enterprise sources
Process Modelling: Identification of steps through which the input data is converted into usable information
Output → Information chain: a map composed of the processing stages connected by directed information channels. Data flows: definition of what data are being used within the system and at which information channel source and target points the data pass through.

Phase 2: **Analysis**
Input→ Information chain maps and data flows
DQ requirement Analysis: Employees and customers interviews
Output → IQ problems

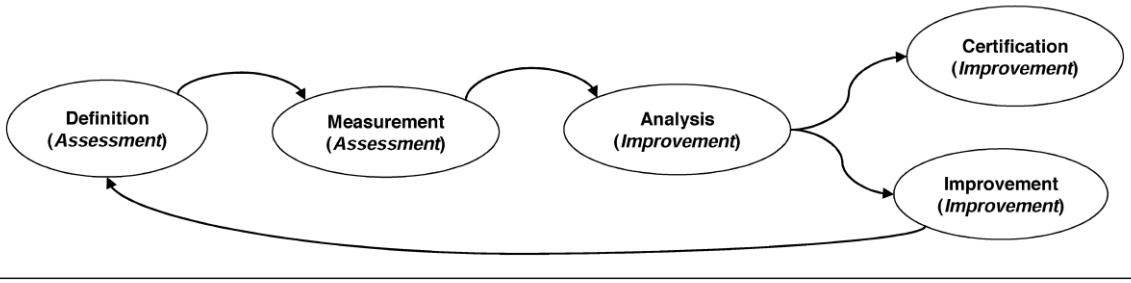
Phase 3: **Isolating flawed data**
Input → Interviews
Process Modelling: Annotation of the information chain
Output → Identification of activities attributable to data flaws

Phase 4: **Identifying the impact domain**
Input → Identification of activities attributable to data flaws
Selection of strategies and techniques: Identification of operational, tactical, and strategic impact
Output → Classification of activities related to poor data quality

Phase 5: **Evaluation of costs**
Input → Classification of activities related to poor data quality
Evaluation of costs: Association of a cost to each operational, tactical, and strategic impact
Output → Data quality scorecard matrix

Phase 6: **Aggregating the total**
Input → Data quality scorecard matrix
Evaluation of costs: Aggregation of costs
Output → Poor quality cost

Figure 15: Cost-Effect of Low Data Quality (COLDQ) [56]



Phase 1: **Definition** .
Input → Data
DQ Requirements Analysis: Definition of the quality criteria and data quality target
Output → List of the required quality

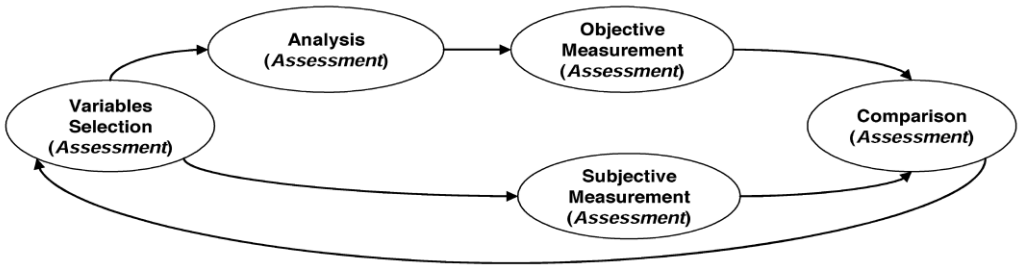
Phase 2: **Measurement**
Input→ IQ dimensions
Measurement of Quality: Definition and application of the IQ metrics
Output → IQ problems

Phase 3: **Analysis**
Input → IQ problems
Identification of the causes of errors
Output → Actions for improving data quality

Phase 4: **Improvement**
Input → IQ metrics
Selection of strategies and techniques: Identification of key areas for improvement
Output → IQ improvement techniques

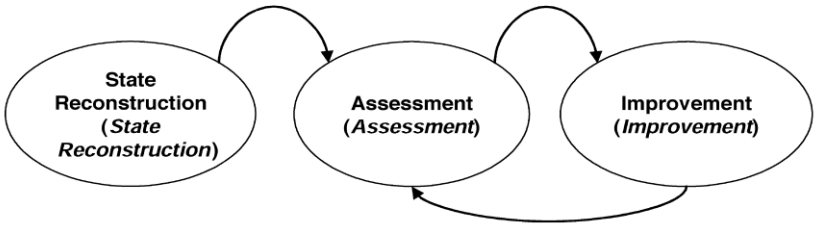
Phase 5: **Certification**
Input → IQ metrics
Creation of the certificate
Output → IQ certification

Figure 16: Data Quality in Cooperative Information Systems (DaQuinCIS) [73]



Phase 1: **Variables selection**
Input → financial variables and the financial context
Data analysis: identification and description of the most relevant financial registry variables, classification of the variables along three classes: qualitative/categorical, quantitative/numerical, and dates
Output → classification and groups of primary variables
Phase 2: **Analysis**
Input→ classification and groups of primary variables, data quality analysis techniques, business rules, process analysis, observation
DQ Requirements Analysis: Analysis of data quality dimensions: syntactic and semantic accuracy, internal and external consistency, completeness, currency, timeliness, and uniqueness
Identification of critical areas: Application of data quality rules
Output → Identification of errors
Phase 3: **Objective measurement**
Input → Identification of errors
Measurement of quality: Definition of indices for the evaluation and quantification of the global data quality level
Output → Quantitative objective assessment
Phase 4: **Qualitative subjective measurement**
Input → Business and data quality expertise, measurement, and classification and groups of primary variables
Measurement of quality: Business expert assessment, Financial operator assessment, Data quality expert assessment, Comparison of subjective assessments
Output→ Qualitative subjective assessment
Phase 5: **Comparison**
Input→ Quantitative objective dimensions, qualitative subjective dimensions
Measurement of quality: Calculation of discrepancies
Output → Assessment and improvement

Figure 17: Quality Assessment of Financial Data (QAFD) [25]



Phase 1: **State Reconstruction**
Input → Internal DBs and information flows, external sources, Organizational structure and rules, processes
State Reconstruction: Identification of role of organizational units in data usage and information flow, identification of role of organizational units in processes, identification of role of external sources in organizational processes, description of processes in terms of their macro-processes, services and norms
Output → Description of the organization along its data, information flows, and processes
Phase 2: **Assessment**
Input→ Description of the organization along its data, information flows, and processes; DQ dimensions, Data activities and techniques, budget
Data Analysis
DQ Requirements Analysis
Measurement of quality
Output → DQ values
Phase 3: **Improvement**
Input → DQ values
Identification of the causes of errors
Selection of strategies and techniques
Evaluation of costs
Output → Costs and benefits, Optimal improvement process, measured/improved DBs and information flows, Controlled/Reengineered Processes

Figure 18: Complete Data Quality (CDQ) [7]