Quality of Data Measurements in the Big Data Era

Quality of data measurements in the Big Data era - Lessons learned from MIDAS project

Introduction

In recent years, digitalisation of traditional manual processes with a tendency towards a sensorised world and person-generated information streams has led to a massive availability and exponential generation of heterogeneous data in most areas of life. This has been facilitated by the cost reduction and capability improvements of Information and Communications Technology (ICT) for storage, processing and transmission.

The key technologies which make it possible to ingest, store and process Big Data (BD), under the original 3V-s (i.e. Volume, Velocity and Variety) definition, have been developed into a mature state, bringing forward a once hyper-hyped topic into a reality. Starting from the available BD many authors have discussed the benefits and methodological approaches for extracting value from it by enabling rich Data-Driven Decision Making (D^3M) [1], [2], compared to traditional knowledge-based or low precision indicators-based D^3M. But most authors report on the need to measure the uncertainty of the captured data in order to make reliable decisions based on BD. Therefore, the veracity of the captured BD needs to be guaranteed in order to extract Value from such data. Veracity is where the Quality of Data (QoD) comes into play, to measure and control the uncertainty and provide an indicator to decision makers on how reliable the data is for decision-making.

In this paper, we report on the QoD challenges, approaches, and experience gained in the MIDAS project [3], whose aim is data-enabled policy making in healthcare. The MIDAS Project (Meaningful Integration of Data Analytics and Services) aims to map, acquire, manage, model, process and exploit existing heterogeneous health care data and other governmental data along with external open data to enable the creation of evidence-based actionable information and drive policy improvements in the European health sector (implementing four pilots in different EU countries with the participation of the corresponding health department and public health provider). Due to characteristics of the project the following reporting is focussed on QoD on provided datasets ingestion and processing, and not in the uncertainty measurement on the data acquisition from empirical world.

Within the following material, we will elaborate on the following topics:

- Data quality dimensions to be better understood with respect to QoD context, data quality indicators to provide decision makers with reliability information and methods for evaluating QoD.
- Challenges identified, and approaches followed to assure QoD in the context of a healthcare BD project, the MIDAS project.

Data quality dimensions

The traditional context of science and technology includes well-structured and validated procedures designed for data acquisition and data quality management [4]. However, this is not the case for the BD context, where many existing data sources are reused for new use cases, and new data sources may be included as they become available. The impressive proliferation of data sources and the exponential growth in data volumes that characterize BD makes it hard to assess the quality of the available information. Additionally, data quality is usually limited to syntactical aspects such as missing data and for checking metadata constraints (e.g. data types or ranges). Considering this heterogeneous and dynamic context, and that BD building system behaviour reproduces computational models from data, then analysing the different dimensions of data quality becomes crucial.

Many authors and organisations have described different definitions of dimensions for data quality assessment [1], [5], [6], to reference a few of them. As an example of this discrepancy, DAMA UK WG [5] defined them as: completeness, uniqueness, timeliness, validity, accuracy and consistency; while the Canadian Institute for Health Information [6] has defined them as: accuracy, timeliness, comparability, usability and relevance. Many of these discrepancies are related to naming or grouping dimensions, and most authors agree that depending on the specific application some dimensions are relevant. Interesting research has been carried to evaluate which dimensions are most considered in different application fields (e.g. public health information systems [7] or electronic health record data reuse [8]). A reference work has analysed different data quality dimension proposals for synonyms and inter-relationships between dimensions and presented a richer categorization of the data quality dimensions, following a grouping of data quality dimension concepts into clusters based on their similarity [9]. They propose the dimensions described below [9], which are used as the reference standard through this paper (Figure 1). We have adopted these dimensions as they are a result of a well-driven review and analysis of different state of the art quality dimension

proposals, grouping similar dimension concepts with the objective to obtain an inclusive definition of data quality dimensions [9]:

- **Accuracy**, correctness, validity, and precision focus on adherence to a given reality of interest.
- **Completeness**, pertinence, and relevance refer to the capability of representing all and only the relevant aspects of the reality of interest.
- **Redundancy**, minimality, compactness, and conciseness refer to the capability of representing the aspects of the reality of interest with the minimal use of informative resources.
- **Readability**, comprehensibility, clarity, and simplicity refer to ease of understanding and fruition of information by users.
- **Accessibility** and availability are related to the ability of the user to access information from his or her culture, physical status/functions, and technologies available.
- **Consistency**, cohesion, and coherence refers to the capability of the information to comply without contradictions to all properties of the reality of interest, as specified in terms of integrity constraints, data edits, business rules, and other formalisms.
- **Usefulness**, related to the advantage the user gains from the use of information.
- **Trust**, including believability, reliability, and reputation, catching how much information derives from an authoritative source. The trust cluster also encompasses issues related to security.

Figure 1. Data quality dimensions used as reference

QoD indicators and Methods

The development of QoD indicators is key to aid decision makers to judge the reliability of the source data over which processing and decisions are being made. At the same time, data quality indicators guide data engineers on the data preparation task (e.g. performing further cleansing) and data scientists on developing the analytics and visualisations (e.g. discarding non-reliable data sources for analytics). Therefore, it is important to define, contrast and validate the QoD indicators with stakeholders involved in the chain from data to decision making.

According to the DAMA UK WG a common approach for the assessment of data quality would follow the steps described below [5]. (Depicted in Figure 2):

- Select the data to be assessed
- Assess which data quality dimensions to use as well as their weighting

- Define the thresholds for good and bad quality data regarding each data quality dimension
- Apply the assessment
- Review the results to determine whether the data quality is acceptable or not
- When appropriate, perform corrective actions
- Perform a follow up monitoring by periodically repeating the procedure

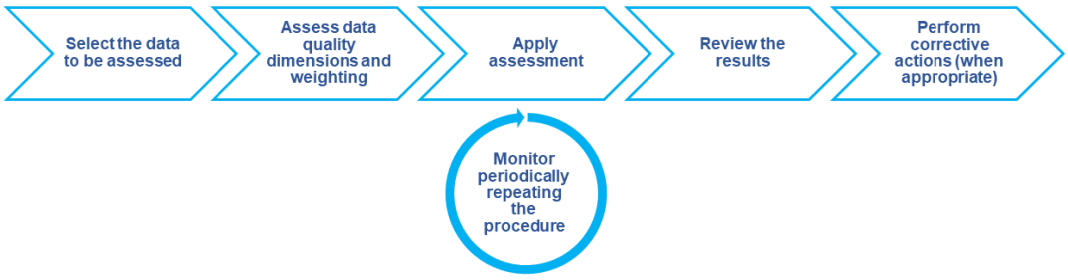

Figure 2. Common steps of data quality assessment

Briefly, the steps listed above describe a methodology to define and validate application specific data quality indicators, and to achieve reliable data for decision making guided by such indicators.

Following this methodology, we developed and presented a web tool for tabular data quality assessment and improvement in the context of health data (TAQIH) [10]. TAQIH enables and supports users to carry out exploratory data analysis (EDA) on tabular health data and to assess and improve its quality. The application menu layout is sequentially arranged as the conventional EDA pipeline helping to follow a consistent analysis process. First, it provides interfaces to understand the dataset, to gain an understanding of the content, structure and distribution. Then, it provides data visualization and improvement utilities for the data quality dimensions of completeness, accuracy, redundancy and readability. More detail on how different quality dimensions are covered by TAQIH is provided in the *Data quality measurement in the BD Context* section.

For the MIDAS project, the data to be assessed were mainly (patient de-identified) health provider's data exports (demographics, prescriptions, diagnosis, hospital entry and discharge information, and questionnaire data) and governmental open data (e.g. social indicators, air and water quality) in tabular format (mainly csv files). Starting from the available data, completeness, accuracy, redundancy and readability dimensions have been identified as dimensions to be assessed using the TAQIH tool's missing

values, correlations and outliers features. Objective quality indicators were defined for the completeness and redundancy dimensions, based on feature and sample missing values and number of highly correlated feature pairs consecutively. Weights were initialised to a default value and experimentally adjusted per each dataset.

In the following section, we will describe the challenges identified while developing the introduced quality assessment tool, and challenges identified while evolving the tool to support more sophisticated BD scenarios.

Challenges

Access to people with knowledge over data and permeating this information through the project

Access to data owners to be able to understand the data and exploit it is an essential item. The correct communication among people with knowledge over data, includes developers and stakeholders who aid to identify, describe and visualize the selected variables in an effective way. When working in a multinational project involving diverse research topics like MIDAS, one problem for the researchers is that the practical meaning of the data and trends (with known cause) is beyond the researchers' scope of knowledge.

MIDAS project has established a data ingestion methodology with the aim to upload to the data repository pre-processed and high-quality data (Figure 3).

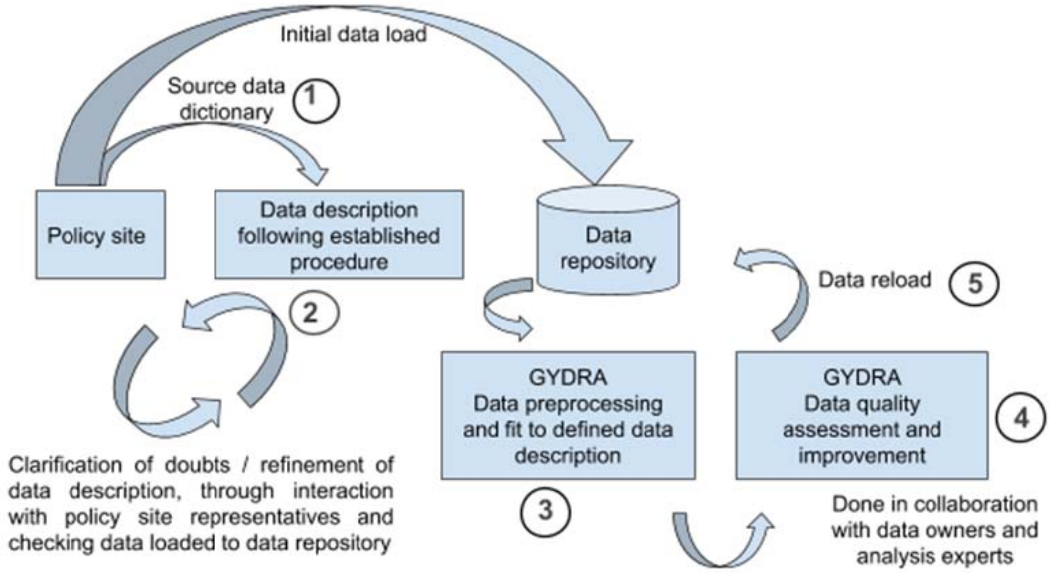

Figure 3. Data ingestion methodology diagram

The first step in the data ingestion methodology is that policy site responsible representatives share an initial data dictionary and the source dataset for an initial data load (1). In order to have this information described in a standard way for all datasets (to aid data analysts and data visualisation experts work), a document has been created describing the procedure to be followed to describe the datasets and has been applied to each of the datasets (2). In this phase, it is necessary to work together with the policy site representatives and check the initially uploaded data from the data repositories with the aim of clarifying any queries with the dataset. At this point we have also analysed each of the data sets to add some initial quality metrics to the dataset description.

Once the data is uploaded to the repository and the dataset description is made, data pre-processing is carried out using the data preparation tool [10] (renamed to GYDRA - *Get Your Data Ready for Analysis*) (3). The main objective of this step is to improve the data quality and to fit it to the defined data description. The next step is to carry out a data quality assessment and to improve it using the tool (4). This step is done in collaboration with data owners and analysis experts. Finally, this pre-processed and high-quality data is reloaded to the data repository (5).

The introduced dataset description file is created and updated in parallel to the dataset preparation following the described ingestion methodology. The dataset description document is used to capture key knowledge on the dataset, following a defined structure and template (i.e. general description including context, structure and observed issues; description of lower level structures and particular level issues; and variable level explanation of the content - including privacy, format, coding and pre-processing information). Having this document as the main interaction point, knowledge over data can be refined by experts on source data (or by people with access to them) and permeate the enriched information among developers and stakeholders (who may also request to further detail some aspects of the document).

Data quality measurement in the BD Context

Assessment methods of many quality dimensions are dependent on pre-existing knowledge of data source. Moreover, assessment of some dimensions involves a level of subjectivity (e.g. trust dimensions involves judgement of data source reputation), and in many cases only a partial interpretation of quality dimensions can be assessed objectively (e.g. accuracy dimensions can be targeted by outlier analysis, but a feature with no outlier might be representing an incorrect reality).

Therefore, the needs of prior information about the data, and the subjective assessment of (part of) the quality dimensions, limits the direct applicability in an automatic manner of the quality assessment. We consider that instead of looking for fully automatic tools for data quality assessment, in many cases either interactive tools or tools to facilitate data exploration are the most appropriate approach.

In the data preparation tool presented at [10], we provide web-based interfaces to understand the dataset in order to gain a better understanding of the content, structure and distribution, to allow the user better judge subjective quality dimension. A missing values section deals with the completeness dimension of data quality. The correlations section presents the correlations among variables, helping to identify possible redundancies among variables or incoherent data, related to the redundancy and accuracy dimensions of data quality. The outlier section identifies outliers in the variable and instances axes which is also related to accuracy, redundancy, readability and trust dimensions in data quality.

The introduced tool's sections provide an exploratory and interactive means for judging different quality dimensions, but no objective means to evaluate the QoD of a given dataset. To overcome this, the quality section summarizes the current state of data quality through QoD indicators of the dataset for the dimensions automatically assessed by the tool (i.e. completeness and redundancy). It permits creating a quantitative report about the data quality of the dataset so that objective decisions can be made depending on the results, such as discarding the dataset or performing additional improvement procedures. The quality section allows for customized weighting of the data quality dimensions for the final estimation of data quality scoring, as well as having the possibility to set a quality threshold for dataset acceptance in line with assessment steps suggested by the DAMA UK group [5].

Despite the proposed tool's approach of dealing with subjective judging of data quality dimensions and (partial) automation of the objective quality indicators, we were missing prior knowledge of data sources to provide a complete context for quality evaluation. Consequently, we have extended our approach to take advantage of the project adopted Isaacus metadata model approach [11], which is used to describe a dataset as well as individual variables in a computer interpretable format starting from a dataset description document. Integrating basic description information (i.e. data types, ranges and units) from computer interpretable metadata allows us to automatically assess syntactic aspects, starting from the data expert's prior knowledge.

Moreover, Isaacus metadata model approach includes some QoD specific elements (e.g. default missing value, factors affecting the quality of the variable, changes that happened in the variable generation) and study level administrative information (e.g. confidentiality, update methodology) that could help evaluating quality dimensions more objectively. But it is not mandatory to fill some of these elements and many are filled as free text, hindering the automatization feasibility.

Summarising, we believe that the correct approach should be focussed on developing and including quantifiable elements of targeted data quality dimensions within a metadata model (e.g. specialising the actual Isaacus metadata model) and providing metadata-automated data quality indicators together with the currently provided syntactic and data extracted ones.

Moving from traditional data preparation tools to large datasets

New challenges appear when moving from traditional datasets which could be loaded and would fit at once into computer memory to data volumes considered in the BD context (i.e. large datasets not fitting in a computer memory and expected to be growing). As a representative example, in one of the MIDAS project pilot sites we had a 17GB prescription dataset (a csv file) that was not possible to load at once into a development PC memory (Intel i5 - 8GB RAM).

When it comes to data preparation and QoD assessment, traditional python-based or R-based do not directly handle dataset that do not fit into a computer's memory. A temporary solution could have been to make use of a more powerful workstation with larger amount of RAM (considering that loading a csv file into memory with its structure and data types takes more space than file size), but this option was discarded as we expected to receive new larger datasets and to combine existing dataset for further processing.

Additionally, many traditional general statistics or quality assessment algorithms need to keep global variables for their computation, which for example for cardinality calculation might require to grow as much as the data source size. This makes existing data quality algorithms not directly applicable for distributed parallel computing.

Besides, we have identified two more issues when moving QoD assessment to large datasets, which are the visualisations used to allow the users to explore the data to evaluate its quality and that data preparation tasks cannot be run synchronously anymore. Traditional visualisations (e.g. missing values or outliers) mainly work by plotting all the instances of the dataset, which requires pulling all instances of the

dataset, and having the user's client applications to manage all the data to visualise and to respond to users' interactions. This is not feasible anymore and having the user wait until a data cleansing task over a large dataset that might require hours or more is not realistic. As a reference, using our non-BD version of the data preparation tool (built using Django Python web framework, Pandas, Numpy and Scikit-learn Python packages, and HTML5 web interfaces) running with a Desktop PC (Intel i5 - 8GB RAM) was fairly interactable (few seconds) for datasets smaller than a hundred megabytes, but working without a good interaction (response taking up to few minutes) for datasets of few hundred megabytes and not working (browser not being able to handle the amount of data for visualisation) for datasets of one gigabyte or bigger.

To overcome the presented data volume challenge, we have opted for using algorithms which provide approximations and to evolve the tool presented in [10] into an asynchronous processing framework (using Celery Distributed Task Queue library with the RabbitMQ message broker solution for asynchronous communication, devoting previous Django web framework-based solution to visualisation and preparation task definition, and configuring remote processing workers for the data preparation tasks). For those algorithms which have distributable or parallelized versions, BD computing infrastructures have been used, while for those requiring adaptations, state-of-art proposals have been implemented following BD computing approaches were possible (using Apache Spark), and per-chunk processing (taking advantage of Pandas per-chunk data processing feature) where more fine-grain control of shared global variables is required.

For the BD QoD indicators visualisation issues, approximations requiring a limited and controlled amount of data to be displayed have been implemented. The computation and generation of the visualisation is done in the asynchronous remote computing machines to reduce processing load and smoothen the user experience on the client side. This way, data-intensive visualisations are loaded from previously created files, improving the time required to render them.

In parallel to the implementation of the algorithm approximations, a pool of different datasets fitting in memory are being tested comparing the traditional implementations with the BD implementation to validate the results obtained.

Information set re-loads, streaming data ingestion

Initially BD applications and parallel distributed processing tools were focussed on the rapid processing of rather static large datasets. Nowadays, it is common that real life

BD applications involve dataset updates at different velocities, in some cases they can be continuous by either streaming data or live API calls, or bulk data loads to upload updated data export for certain period. Examples of continuous data updates can be an IoT device sending new data every minute, and an example of an uploaded data export could be a certain clinical dataset export that is updated every six months.

A data updating scenario opens new challenges to data preparation and specifically to QoD assessment. Each data upload, be continuous or periodical, involving stream processing or batch processing, requires data quality to be assessed to guarantee its veracity for a successful D^3M. In contrast to static large datasets quality assessment, manual assessment of updating datasets becomes impractical. In this context, the automation of the assessment becomes a must. This need is also highlighted in a data preparation products comparison report [12], analysing main commercial tools (e.g. Trifacta, Unifi or Datameer), as the need to formalise, share and collaborate on data preparation recipes, to avoid replicating the same work.

To tackle this challenge, we have developed a data transformation pipeline definition functionality for our data preparation tool [10]. This functionality implements visual definition of transformation pipelines to facilitate non-technical people their definition. Next, we have defined a pipeline export format to enable the reusability and easy deployment pipelines. Currently, we can apply such pipelines to periodically updated datasets running through batch processing. We are exploring how to apply them in stream processing scenarios where the steps where QoD is assessed can vary. For this task, we are testing the use of Apache Kafka and Apache Spark Structured Streaming feature, as our current solution uses Apache Spark (despite other alternatives as Apache Flink or Apache Storm where considered).

We are aware that automation of QoD improvement processes in the form of data handling, storage, entry and processing technologies can also have negative effects. Automation can be a good solution for dealing with data updates, while it can create a different set of data quality issues due to uncovered data sources' specifics. So, it is important to keep in mind and apply the last action of Assessment of Data Quality Steps (Figure 2), "Perform a follow up monitoring by periodically repeating the procedure".

QoD issues detected when developing algorithms and processing data

Despite the efforts placed solving QoD issues during data preparation phase, there are usually still issues left which cannot be noticed before the data is applied in the real analytics.

One challenge in data pre-processing is the case in which multiple data sources share one or more attributes, which need to be used combined, but are have a different representation. The inconsistency, such as different abbreviation of a value of a categorical variable, can be inconspicuous when going through dozens of data tables in a database. By using dataset description and metadata, this type of inconsistencies could be identified and solved easier. In the MIDAS project, an example of this issue was happening where different health data tables contained location information but had different coding schemas on some of them (even if most category values seemed similar). Despite efforts are being made towards unified EHR systems, many times harmonisation tasks are not complete and this is reflected on exports (data and metadata) shared with research or data exploitation projects, which requires to go back from analysis to data preparation and to update the metadata, even if a well-defined requirement gathering and architecture is designed. This is usually motivated by the previously introduced limited access to people with knowledge of the source data, knowledge over different data tables being distributed among different people, and expert people not being aware of their data issues (especially those that arise when combining different datasets).

Another issue detected during analytics development was the lack of necessary information to solve a research problem. In the MIDAS project this was caused by having different planned research data tables delivered progressively or having initially data available only for a limited period. Open data was explored to find more information, and expertise was derived from different departments, which provided decisive supplement to current datasets. Appendix tables were created based on these external data sources to present the linkage between the current datasets and the expected information. These efforts enhanced the usefulness of the data and achieved completeness when crucial information was absent.

Using Isaacus metadata approach we could easily export the defined variables with their additional information, such as data types, and deliver them to data-scientists developing different algorithms for data analysis. The exported metadata information was then used for choosing the algorithm parameters based on their data types. Actual datasets for different MIDAS pilots were stored in HIVE data warehouse that lies on top of distributed HDFS data. The selection of HIVE and HDFS distributed storage technologies was motivated by MIDAS pilots core data being large retrospective data exports, and to enable better performing distributed processing analytics. HIVE was

selected given the structured query features it provides. During the HIVE data extraction, based on the Isaacus metadata, certain discrepancies were discovered mostly due to inconsistency between the data types loaded in HIVE and data types defined in metadata. To minimise this type of issues, we have extended our data preparation tool [10] with an alignment tool and a data preparation sync functionality. The alignment tool allows to make sure that the metadata description provided by people with knowledge on source data, meet data preparation tool inferred variable names and types. Once alignment is achieved, the data preparation sync functionality automates and assures the coherent data and metadata deployment for analytics.

Some MIDAS pilot datasets had missing variable values which hindered the correct analytics development. To palliate this issue, missing value imputation was carried out using different methods, taking advantage of available variable values. In some cases, it was necessary to create new variables, combining two or more existing variables. This helped in boosting the QoD indicators of readability and usefulness for each of MIDAS pilots, as well as enhancing the data uniformity needed for each data analytics model.

Redundancy QoD dimension needs to be carefully assessed, especially when creating new data pools from heterogeneous sources for a given data analysis model. This is achieved by choosing specific variables and tables from the dataset and reducing the total number of data tables. Variables with a high rate of missing values are discarded. The number of duplicated observations is also reduced by carefully tailoring data pools to get the best quality data needed for model input.

The data preparation sync functionality has been developed to easily deploy data for analytics, upon a data preparation or quality improvements task identified during the development of analytics models.

Conclusions

The development of BD technologies in recent years has enabled the timely ingestion, storage and processing of heterogeneous large dataset responding to Volume, Velocity and Variety dimensions of BD definition. But, in order to achieve reliable Value from the processing of BD, and to enable a reliable data driven decision making, it is key to ensure the Veracity of the decision involved data. Veracity is where the QoD comes into play, to measure and control the uncertainty and provide a veracity indicator to decision makers.

In this paper, we first study the QoD context (dimensions and indicators) and then we report on the QoD faced challenges and adopted approached during the execution of a healthcare BD project, the MIDAS project, whose aim is data-enabled policy making in healthcare. We believe that the lessons learned and shared in this paper could be useful guidelines for the veracity assurance of BD projects and for further development of data preparation and QoD assessment tools.

Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 727721 (MIDAS).

References

[1] L. Mari and D. Petri, "The metrological culture in the context of big data: managing data-driven decision confidence," *IEEE Instrum. Meas. Mag.*, vol. 20, no. 5, pp. 4–20, Oct. 2017.

[2] F. Mari and P. Masini, "Big data at work: the practitioners' point of view," *IEEE Instrum. Meas. Mag.*, vol. 20, no. 5, pp. 13–20, Oct. 2017.

[3] Midas Consortium, "MIDAS – Meaningful Integration of Data Analytics and Services," 2019. [Online]. Available: http://www.midasproject.eu/. [Accessed: 10-Jun-2019].

[4] J. McNaull, J. C. Augusto, M. Mulvenna, and P. McCullagh, "Data and Information Quality Issues in Ambient Assisted Living Systems," *J Data Inf. Qual.*, vol. 4, no. 1, pp. 4:1–4:15, Oct. 2012.

[5] The Dama UK Working Group, "The Six Primary Dimensions For Data Quality assessment," Oct-2013. [Online]. Available: https://www.dqglobal.com/wp-content/uploads/2013/11/DAMA-UK-DQ-Dimensions-White-Paper-R37.pdf. [Accessed: 08-Mar-2018].

[6] Canadian Institute for Health Information, "The CIHI Data Quality Framework," Ottawa, ON, Canada, 2009.

[7] H. Chen, D. Hailey, N. Wang, and P. Yu, "A Review of Data Quality Assessment Methods for Public Health Information Systems," *Int. J. Environ. Res. Public. Health*, vol. 11, no. 5, pp. 5170–5207, May 2014.

[8] N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research," *J. Am. Med. Inform. Assoc. JAMIA*, vol. 20, no. 1, pp. 144–151, Jan. 2013.

[9] C. Batini and M. Scannapieco, "Data Quality Dimensions," in *Data and Information Quality*, Cham: Springer International Publishing, 2016, pp. 21–51.

[10] R. Álvarez Sánchez, A. Beristain Iraola, G. Epelde Unanue, and P. Carlin, "TAQIH, a tool for tabular data quality assessment and improvement in the context of health data," *Comput. Methods Programs Biomed.*, Dec. 2018.

[11] THL, "National Metadata Descriptions - THL," *The National Institute for Health and Welfare (THL), Finland*, 2019. [Online]. Available: http://thl.fi/en/web/thlfi-en/research-and-expertwork/projects-and-programmes/national-metadata-descriptions. [Accessed: 11-Jun-2019].

[12] Ovum, "Ovum Decision Matrix: Selecting a Self-Service Data Prep Solution, 2018–19," 2018.

Author BIOs

Gorka Epelde (corresponding author)

Gorka studied computer science at the University of Mondragon and received his B.Tech degree in 2003. In 2014, Gorka obtained his Computer Science PhD from the University of the Basque Country. He is a Project Leader and Senior Researcher in Vicomtech. His fields of interest include interoperability architectures, data engineering, as well as the human computer interaction and the advanced visualisation of data.

Andoni Beristain

Andoni studied Computer Engineering at the Basque Country University where he obtained his PhD in Computer Science in 2009. Since 2010 he is part of the eHealth & Biomedical Applications department at Vicomtech. He has worked as a researcher in various FP6, FP7 and H2020 projects and has coordinated several regional and national projects as well as supported EU projects coordination.

Roberto Álvarez

Roberto received the B.Tech. degree in Computer Science in 2006 from the Complutense University of Madrid. He is a researcher in Vicomtech with a considerable experience in communication protocols and interoperability. His expertise and research interests include interoperability architectures, cloud architectures, big data and distributed architectures, data harmonization and data analytics.

Mónica Arrúe

In 2016 Mónica completed her degree in Biomedical Engineering at the University of Navarra. She has been working as a Research Assistant in Vicomtech at the department of e-Health and Biomedical Applications since then, specifically in the line of Big Data and Personalized Medicine. She is currently completing a Master's Degree in Data Science at the Universitat Oberta de Catalunya (UOC).

Iker Ezkerra

Iker is BIOEF’s responsible for MIDAS Basque platform deployment and Technical-Production Director at NorayBio. He's an expert in the development of IT products for managing and exploiting biosciences data (including various FP7 and H2020 projects). He holds a Master in Big Data and a postgraduate degree in agile methodologies.

Oihana Belar

Oihana is the Quality manager at Basque Biobank. She holds a PhD in Cell Biology by University of Basque Country (2011), Master's degree in Neoplasic Diseases at the University of Basque Country (2008). She actively participates in different projects of the Spanish Biobank network as well as in different European projects related to biobank's protocols quality, identification of new biomarkers, and Big Data.

Roberto Bilbao

Roberto holds a PhD on Gene Therapy by Navarra University (1999), and Master's degree in leadership management for Science at the Pompeu Fabra University.

He set up and is the director of the Basque Biobank. He's also the coordinator of the R&D program of the National Platform of Biobanks. He has been principal investigator of several national funded projects and participated in European research projects.

Gorana Nikolic

She graduated in 2010 at the School of Electrical Engineering in Belgrade, at the department of Software Engineering, where she completed her MsC studies in 2013. After working in the software industry (software engineer and technical lead), she enrolled in a PhD program at KU Leuven, and she is currently in the final year. Her research interests include applied machine learning and privacy preserving data mining.

Xi Shi

She worked in a consulting company as a researcher whose main area is stochastic models for pension system and healthcare system. She received a Master's degree in Science in 2017 from KU Leuven and she is currently working on a PhD in Engineering Science at KU Leuven. Her research interests include data mining, statistics, and machine learning. She is currently working as a researcher in H2020 project.

Bart De Moor

He received a doctoral degree in applied sciences, in 1988, from the KU Leuven. He is a full professor at the KU Leuven. His research interests include numerical linear algebra, system identification, advanced process control, data mining, and bio-informatics. He is the (co-)author of several books and papers, some of which have been awarded. He received the Leybold-Heraeus Prize in 1986, the Leslie Fox Prize in 1989, the Guillemin-Cauer Best Paper Award, of the IEEE Transactions on Circuits and Systems,

in 1990, the biannual Siemens prize in 1994, and became a Laureate of the Belgian Royal Academy of Sciences in 1992.

Maurice Mulvenna

He gained a MPhil. in Information Systems in 1997, and a PhD in Computer Science in 2007, both from Ulster University. He is Professor and Chair of Computer Science at Ulster University. His research areas include data analytics, artificial intelligence, digital wellbeing, innovation and assistive technologies.