Overcoming Limited Dataset Availability when Working with Industrial Organisations

Torben Jess, Philip Woodall, Duncan McFarlane
Department of Engineering, University of Cambridge
Cambridge, United Kingdom
Email: tj282@cam.ac.uk

***Abstract*—Increasing data security and privacy requirements combined with the need for additional data management research leads to a conflict for industrial companies. In order to solve their industrial data management problems companies need to share some of their data, but their internal confidentiality rules sometimes hamper this sharing process. Existing techniques for sharing data without releasing company secrets often loose some of the problems/characteristics within the data. This paper therefore presents a qualitative process to overcome this problem of industrial data sharing while still enabling external researchers to develop relevant solutions to organizational problems. It is based on initial trials with two industrial case studies and showed some promising results.**

Keywords—data privacy, security; indsutrial data management; data sharing; working with industry

I. INTRODUCTION

Industrial companies are continuously increasing their internal data security and confidentiality policies. This results in various rules and time-consuming processes for release of data. Many organisations do not release any data for fear of inadvertently revealing business secrets [1], [2]. The case of AOL, where search requests from various users were exposed by accident although AOL thought that they annonymised the data [3], is an example of an unwanted release companies strive to avoid.

Industrial companies, such as manufacturers, have a significant challenge to release data to academic researchers in an ad hoc manner. Often they do not have mature and frequently used processes to sanitise and then release data. Other fields such as healthcare frequently release data to medical researchers. In contrast to these fields, industry has a multitude of different types of secrets. Hence, when they are faced with a request to share data, it is much more difficult for them to release it confidently. In the same way AOL underestimated the secrecy of data that can be associated with the actual content of the search queries, companies are afraid to not consider a certain type of data secret during the release. In addition to just supplier or customer names manufacturers may also have secret production formulas and recipes, secret materials, secret product design plans, secret inventory levels etc.

At the same time they are relying more and more on consultants, external service providers, and academic researchers to help solve their industrial data management problems. This leads to a significant problem: how can industrial data management problems be researched and solutions be developed without the external party having any access (or very limited access) to actual industrial data? While some generic problems might potentially be solved without having access, once the problems get more complex and specific, access to the underlying data becomes necessary. A lot of industries have this problem but it especially true for industries such as defence, pharmaceuticals, or banking for example.

This paper therefore aims to address the question of how to provide convincing solutions to important industrial data management problems without having access to actual datasets from an organization. Existing techniques often work on anonymising the data, but they are not always able to retain the original problems, like data quality [1], [2] for example.

We present a process consisting of various steps that was developed for working with industrial companies when we could not get access to the underlying data. The process suggests a continuous interaction with the industrial company to identify and analyse their data management problems. The characteristics of the problem are extracted and a combination of public data sets and automatic data generation is then used to recreate an environment with very similar characteristics.

We extracted this process using two case studies. In the first case study we aimed to develop a data quality tool for a large industrial company to address one important data quality problem. For this case our data requirement was for a dataset exhibiting various and actual data quality problems existing within the company—so that we could establish which one was important, and then develop solutions to this problem. In the second case study our goal was to develop a data valuation approach, where datasets could be recommended to the user as additional information to improve their decision-making. We required a realistic industrial data management environment (consisting of table structure, column headings, data types and data volume) for testing and further development. In both cases the company was not able to share any of their data, making it difficult to solve their specific industrial data problems and verify the applicability of the developed tool or approach.

This paper describes the steps we took to overcome this data unavailability and to what extent we were able to continue and complete our research based on not having the data we needed. Using our approach we created mock-up datasets of

industrial data management environments. In the data quality case it enabled us develop a tool for the company, which they are now investigating how to use it within the business. For the data valuation method we showed that the identified mock-up environment enabled some initial testing leading up to a first publication of some results. However, for the second case study the final result is still to be determined because we have not finished all of the research.

This paper is further divided in the following sections. Section 2 presents the research background. Section 3 describes the process of overcoming the limited data availability. Followed by section 4, which presents the case studies leading to the development of this process. Section 5 presents the lessons learnt within an initial case studies before the final conclusion and potential future work in section 6.

II. Research Background

A. Industrial Data Security

Companies are increasingly worried about data security. Which is based on facts such as the number of data security incidents increasing by 48% in 2014 [4] for example. Having a security breach for certain kinds of data can be very expensive for companies [4], [5]. Sony for example spent 15 Million Dollars after a lot of their internal information got hacked in November 2014 [6] (not even counting the potentially even larger reputational damage). This shows the impact that lost data can have on companies. This risk for industrial companies combined with increasing governmental standards on data security standards [4] has lead companies to implement strict data security rules. These often include strict regulations for export of datasets outside of the company or even between divisions within the company.

B. Increasing Data requirements and the need for increasing external advice for industrial data management

Industrial companies have increasing data management challenges like the upcoming trend of big data as a competitive advantage [7], an ever increasing amount of data [8] and data quality problems [9], for example. This challenge combined with historically grown information systems makes them rely increasingly on external service providers like tool developers such as parts of IBM or Oracle for example, but also external advisors or researchers to help them address these problems.

C. Problem for Indsutrial Data Management Research

Combining these two challenges is a problem for many companies. They have to share information about their data in order to work with external companies or advisors but at the same time follow strict data security procedures and regulations. To overcome this challenge existing research has mainly tried to find ways to change the data in a way that it looses all its secret information, but can still be used for research; there has been much research in the healthcare domain where datasets containing patient details need to be anonymised before being released.

Woodall et al. [2] showed that existing techniques have various limitations with regard to data sharing. They either cannot transfer the full information about data quality problems or do not fully obfuscate the data. There exists a trade-off between privacy and utility of the annonymised dataset [10]. They are especially not able to ensure this data obstruction to a degree of mathematical certainty [10], [3]. Leaving companies with an additional risk of future methods or intelligent data observers uncovering their data secrets. In the previously mentioned AOL example they were certain that the original user could not be traced from their search queries. Therefore companies still require a data release process for anonymised datasets. Even if a good approach for data anonymisation were to be found, the data would still have to go through the data releases process of a company. This release often takes several months before researchers can have access to data.

The issue is therefore still not entirely solved and is a huge problem for companies with very strict data sharing rules, which allow no sharing of data even if it has been obstructed. Working towards closing this gap in the current research is the goal of this paper.

D. Alternative methods for analyzing problems in a research environment

Other research areas have identified different approaches to analyse problems in a research environment, while keeping the original dimensions (or characteristics) of this problem [11] like the development of airplanes and ships in fluid dynamics [12],[13]. When a new airplane is developed it needs to be tested how these airplane's wings will react to certain types of wind. While some of this testing is done with computer simulations some aspects are tested in practice. However, because it is too expensive to build a complete airplane, they only build a much smaller representation of the airplane for a wind tunnel. In order to ensure that the tests in the wind tunnel are still representative to how the actual airplane would react they try to make sure that certain key measurements like the ratio of width to length for example (so called dimensions) of the model airplane are similar to the actual airplane [13].

III. A Process for Overcoming Limitted Data Availability

Our approach adapts the philosophy of the dimensional analysis approach discussed in section II.D to the industrial data management domain and shows a clear process to capture the characteristics of an industrial data management problem and using these characteristics as basis for the research. Using the two case studies (see section 4) and observing best practises during these case studies we were able to find a set of common steps between both cases and combine it to a common procedure. It is based on the following eight steps:

1. **Identify contacts:** The key contacts help in providing the additional information for following steps. They need to have a broad understanding of the company to provide the access for understanding the industrial data management problems or provide links to the right contacts in the organization. For data management in industry these could often be IT or Research divisions. But dependent on the problem and the organisation the criteria for the key contacts can vary.

2. **Problem domain:** Using the contact(s) identified in step 1 the general questions about the problem domain need to be understood and clarified—what are the operational goals that the company wants to achieve and how does their data problem affect these goals?

3. **Problem characteristics:** The key problem characteristics vary based on the industrial data management problem. These characteristics can be column headings, specific data types, data profile, error types, number of data occurrences and so on. They need to be sufficient to accurately translate the underlying problem into a mock-up dataset. These characteristics are the basis for the target environment and the mock data generation in the following sections. An exemplary overview about the potential characteristics can be found in Table I. Selecting which criteria are required for accurate tool development is a major challenge. Table I gives a set of criteria to consider to whether or not a certain characteristic is relevant. If these questions can be generally answered with yes, then identifying this characteristic is important for the experiment.

4. **Data environment:** The ideal representative environment is defined based on the understanding of the domain of the data, and its key data characteristics (see Table I).

5. **Mock-up:** Based on the target environment the researcher can now develop a mock-up data environment to "simulate" the problem existing in the industrial company they are working with.

6. **Validate:** After generating the data environment it should be shared with the key contacts, so that they can evaluate whether it is a suitable and correct representation of the real problem.

7. **Solve:** In this step the researcher solves the actual problem utilising the data environment.

8. **Feedback:** This step applies the developed method or tool in the industrial company. The key contacts and their contacts in the company will test the solution under the guidance of the researcher. They can then provide feedback for the identified tool.

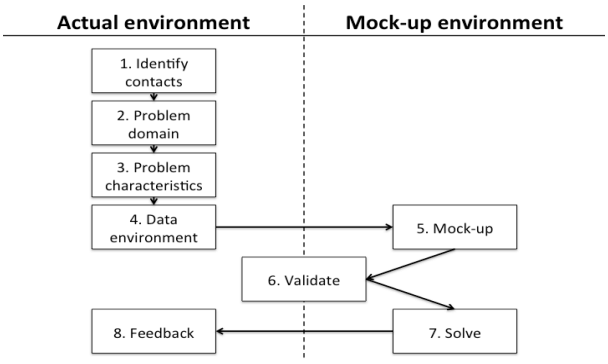

Fig. 1. Flow chart showing the steps taken between the actual industry data environment and the mock-up data environment

In order to generate good input from the industrial partners it is necessary to have enough time to collect the required information from them. The process (outlined in Figure 2) is therefore very interactive and requires various discussions with industrial contacts. It is not necessary linear and might require iteration between steps. This iterative nature of our process should especially be considered in step 8. While being the key step of the actual work on the industrial data management problem it is important to remain focused on the industrial company and regularly ask for feedback. This more flexible and iterative development of solutions helps in saving time during the development process.

Examples of data characteristics

TABLE I. CHARACTERISTICS

#	Data characteristics		
	Characteristic	***Description***	***Example selection criteria***
1	Data structure	Describing the different components of a data table (e.g. number of columns and their column name) and how they are linked with each other (e.g. primary keys and foreign keys)	Is the problem due to or affected by the connection between various datasets or tables?
2	Data types	Describing the kind of data within each column (e.g. Text, Integer, Address, Date)	Does the problem exist in different kinds of data? Would different kinds of data show different kind of problems?
3	Data format	Describing the detailed format of a column; like the format of dates (dd/mm/yyyy etc.)	Are different data formats a reason for the problem?
4	Data profile	Describing the profile of the dataset(s) and/or table(s) like the number and type of data occurrences (e.g. number of NULL values or typos in a specific column)	IS the problem due to or described by different distribution of different data types or occurrences?
5	Data volume	Describing the volume of data for the different tables and the whole table structure (e.g. number of rows for a certain table)	Is the problem in the scalability towards larger datasets?
6	User behaviour	Describing typical interactions of the user with the data and decisions the user makes based on data (e.g. data presented to certain users)	Is the user interaction with the data part of the problem or a potential solution?

Fig. 2. Examples of key data characateristics

IV. CASE STUDIES

To extract the process in different environments we identified two cases in which we needed to extract information about data management problems from an industrial company.

We documented our steps for two case studies (see Table II) and identified common threats between those two cases to generate a more generic process for future use in industrial companies.

A. Case Study A: Data Quality Tool

As part of our research we developed a generic tool to address a particular data quality problem, which was important for the organisation we worked with. We had been told that data quality issues were a major issue in various data management operations within the company. However due to strict privacy constrictions the company was not able to share any kind of data with us. We tried various attempts to get the

Case studies

TABLE II. SPECIFIC STEPS CARRIED OUT IN EACH CASE STUDY

Process steps	**Case study**	
	Case study A: Data quality tool	***Case study B: Data valuation***
Step 1: Identify key contacts	We searched and found key contacts in the company. They worked in the research and development unit of the organization, which is distinct from the operational part of the business.	Based on previous interactions we identified a key contract in the company's research division. The contact had a mix of IT knowledge and a good domain understanding due to previous internal research experience in the area.
Step 2: Understand the problem domain	Within the initial meetings we selected procurement to be the main problem domain. During a set of additional interviews we developed a better understanding of their specific processes and problems.	We conducted an initial literature review to identify areas most promising for our type of data valuation technique. Using 3 discussions and brainstorming sessions with 2-5 key contacts we were then able to identify areas (or domains) in which the application provides the most immediate benefit.
Step 3: Identify critical problem characteristics	In the domain of procurement we identified the data type, formats and profile as the main characteristics using the selection criteria for these characteristics. We set up meetings with various users in this domain who may experience data quality problems and conducted interviews with them. To get further details about these characteristics we presented a list of typical data quality problems. The list was based on previously identified problems in the data quality literature. We asked the users to state whether they experience the problem or not and to identify more specific instances of this type of problem. We selected the data quality problem that appeared to be the most pervasive and important for the company to solve (i.e. would have a large positive impact on the company if the problem was solved)	Within the domain we then identified data characteristics for table structure, data types, data format, data volume, and user behaviour. This way a clearer understanding of the actual data and its use was developed. We then developed sheets containing the list of various information required for the generation of this data. We started with initially identifying the table structure in various discussions. Based on this table structure we then identified the additional information such as types, format or volume using iterative discussions with our key contacts.
Step 4: Identify target data environment	Using the characteristics from step 3, we identified that various tables of data containing similar industrial data (having similar data types, formats and profiles) and containing the data quality problem, was the ideal dataset to use for the development of our data quality tool.	Using this information we were able to define what type of tables, their content and the structure between tables was needed to mock-up a realistic environment.
Step 5: Mockup data environment	We searched for publically available online data that contains these data quality problems in a similar type of data. Once a suitable data source had been found on a website, we manually extracted small amounts of the data from the website to produce our first dataset, which would contain enough data to develop simple scenarios with and would allow us to start building the initial part of the solutions. We later wrote a script to automatically extract a large portion of the data from the data source (website) for larger scale experiments as an iterative step from the following steps we took.	An automatic data generator developed by us was used to generate this dataset. It used a combination of public datasets and random data generation as input for the data mock-up. We used a list of parts, their numbers and a list of peoples name as input for the data generation for example. The data generator ensured a generation of a representative amount of volume and in the development we ensured that the data was consistent among different tables (e.g. matching primary and foreign keys, consistency with regard to part numbers to part names, etc.). In the process we continuously checked with the key contacts due to new questions about the characteristics arising in the generation process.
Step 6: Validate data environment	We presented the scenario (multiple times) to the key contacts in the organization to confirm that the data was representative of the real problem data.	We showed the generated test data to our key contacts and then included their feedback. This feedback process took various iterations and was done on a very detailed level within the data to avoid any mistakes and develop an accurate representation of the company's data structure.
Step 7: Address problem in mockup environment	We wrote the first versions of the solution and continuously tested it on our mockup dataset.	Within the mock-up environment we then conducted our set of experiments using the identified user behaviour and the mock-up dataset. We used to tool to identify the value of specific datasets, which should be recommended to certain datasets.
Step 8: Feedback problem solving method	Once the solution worked well enough we presented this to the key contacts in the organization (as a feedback loop with step 7). We then iteratively extended the solution to cope with the various instances of the DQ problem in the dataset while continually reviewing, with the key contacts, that the issue was still important and our dataset was representative of the real problem.	We were not able to apply this approach towards the same dataset within the company yet. However we were able to get qualitative feedback from the company verifying or not verifying our answer, which we used to adjust our approach.

Fig. 3. Table describing the experience of the two case studies when applying the process for sharing industrial data management problems

data through their data releases mechanism but never received a representative dataset. This made the targeted development of tools for this company very difficult. We had no clear knowledge about the specific data quality problems, their magnitude and the type of data. Developing techniques for solving this was a great challenge. This had a high risk of doing unnecessary work with little industrial impact. In order to conduct our research we therefore needed to find a way to solve these data quality problems without having access to the actual dataset. We documented our steps and were able to develop our solution based on the dataset we extracted using these steps. We returned it to the company for additional testing. A detailed case execution description can be found in Table II.

B. Case Study B: Data Valuation

As part of our research with an industrial company we needed to solve some of their data overload problems. They needed to identify which datasets are worth being presented to the user and which datasets should not be presented. Doing the development of a tool for this problem requires representative information systems environment to test it. It also required a detailed understanding of the underlying data. However, the company was not able to share the data with us. This made the actual research very difficult. Without an overview about the datasets it is hard to identify which datasets were actually valuable to the users. We followed the steps outline in table I and documented these. We had a series of interviews with the company to extract this information from them. We were then able to develop our solution and conduct a series of experiments based on this extracted information.

V. LESSONS LEARNT

During our case study with this process we made the following qualitative observations:

1) Disadvantages of the process:

- It was serendipitous for step 6 that we found a suitable publically available data source, which contained data that exhibited all the data characteristics relevant to the problem we wanted to address. We found the structure of our process certainly does not guarantee that this will always be the case.
- The initial steps 1,2 and 3 can be very time consuming. Getting them right is key for the following steps. However, this can require various interactions with different departments in the organisation.
- Due to the limited data availability there was a risk of choosing the wrong focus in the process selection and definition. One may be tempted to select the problem that public data is available for. However, this may not be an important problem for the company.
- When interacting with the company the process heavily relies on having the right and representative contacts. We were never certain that the problem was important to the operational part of the organization and the data was representative because our key contacts work in the research part of the organization. Therefore, we were relying on their knowledge of the operational part of the business being up-to-date. We tried to address this issue by continually reviewing and checking with key contacts on monthly call. There was still the risk that our key contacts perceived some issues differently then how they were actually existent in the company.
- Following the process for this specific case was a high workload on key contacts. We had 7 one-hour phone calls and one 2 hour personal meeting to develop a detailed understanding of the data environment for the second case. Due to the large number of interactions and the different type of questions the key contacts also spend additional time finding this information within their company in between the meetings. This time availability is not always achievable dependent on the key contact, might require additional resources and can take a long time where no or little research can be done on the problem.
- The actual process of generating data based on the company's specification can be very time consuming. Especially if it needs to ensure that the data is consistent between various tables. There are currently no good tools to help with this process. The development of a data generator took roughly 3-4 weeks of development for the second case.
- Due to the high amount of manual effort the data generation process is prone to errors and requires various quality checks. We had to take 3 iterations in the adjustment of the data generator for the second case.
- There was no clear link for knowing (and justifying to the company) what the actual impact of solving the problem would be in the company (because the data was not based on original data sources).
- Not all business logic can be covered with this approach, which makes the mock-up dataset potentially less accurate then the dataset within the company

2) Advantages of the process

- The data obtained enabled the solution to be developed, and appears at this stage that it is relevant and useful to the company because it is now being integrated into the company's operations (although we are still waiting to complete this activity).
- The separation of public data selection from internal data sources enabled a more targeted search for datasets and therefore enabled it to work more specifically on one specific rather then having to deal with the complexity of all problems within a company dataset. We had a lot of control on these complexities. Sometimes the actual dataset would contain characteristics that were not relevant for the specific analysis we want to conduct (e.g. data quality might not be essential for initial data value analysis or NULL values need additional coding to

be addressed). They might be more of a distraction than added realistic characteristics because additional non-problem-focused code needs to take care of this issue. This is not always beneficial in a research environment. They therefore did not have to be addressed in the solution developed for the problem.

- We got deeper insights into the workings of the company by having an iterative interview process in comparison to having just handed a dataset, because we understood the data users tasks and the business process underneath the data, which helped in the prioritization of the development.
- The process helped in setting up an environment, which we used for some initial testing.

For the first case the result was that we were able to develop a data quality tool that was successfully implemented within the company. For the second case the process enabled us to generate a dataset that we could use for a series of experiments within the data evaluation problem. However, the final development still needs to be further verified within the company by actual experiment other than the current qualitative feedback (see Table I step 8.)

Given the complexity and the amount of detail required as part of the target dataset for our analysis the process was still relatively efficient, but an actual access to the underlying system would have saved this time.

VI. Conclusion and Future Work

In this paper we developed a process to enable researchers and other advisers to work on specific industrial data management challenges when access to the industrial data is not possible due to companies' data sharing rules. We presented two specific cases and the general process was developed from both of these cases. The process proved to work well within those industrial case studies and was able to help in the development of a tool for an industrial company. However various improvements are necessary for this process to make it more efficient, effective and reliable. A set of public datasets and readily easy usable data generator techniques (which enable data generation including data quality errors and consistency among various databases) could help to make it more applicable and easier to use. Additional techniques to improve key contact identification and interviewing will also help. Further input from existing interview techniques might be a potential approach for this area. With regard to evaluation this process needs to be tested in a larger sets of case studies. This will enable further improvement of this process and also a more quantitative evaluation.

References

[1] M. Oberhofer, P. Woodall, and A. Borek, "Solution Architectures for Generating Synthetic Data while Retaining Data Quality Problems," in *In Proceedings of the International Conference on Information Quality (ICIQ)*, Little Rock, Arkansas, USA, 2013.

[2] P. Woodall, M. Oberhofer, and A. Borek, "A Preliminary Study on Methods for Retaining Data Quality Problems in Automatically Generated Test Data," in *In Proceedings of the Americas Conference on Information Systems (AMCIS)*, Seattle, WA, USA, 2012.

[3] M. Zimmer, "'But the data is already public': on the ethics of research in Facebook," *Ethics Inf. Technol.*, vol. 12, no. 4, pp. 313–325, Jun. 2010.

[4] Price Waterhouse Coopers (PwC), "Global State of Information Security Survey 2015," Sep. 2014.

[5] Verizon Enterprise Solutions, "2014 Data Breach Investigations Report," MC15912 04/14, Apr. 2014.

[6] S. Frizell, "Sony Is Spending $15 Million to Deal With the Big Hack," *Time*, 04-Feb-2015.

[7] T. H. Davenport and J. G. Harris, *Competing on Analytics: The New Science of Winning*, 1st ed. Harvard Business School Press, 2007.

[8] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, Report, May 2011.

[9] O. Kwon, N. Lee, and B. Shin, "Data quality management, data usage experience and acquisition intention of big data analytics," *Int. J. Inf. Manag.*, vol. 34, no. 3, pp. 387–394, 2014.

[10] T. Li and N. Li, "On the tradeoff between privacy and utility in data publishing," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, Paris, France, 2009, pp. 517–526.

[11] R. C. Pankhurst, *Dimensional Analysis and Scale Factors*. Chapman and Hall, 1964.

[12] Y. S. Shin and L. D. Santiago, "Surface ship shock modeling and simulation: Two-dimensional analysis," *Shock Vib.*, vol. 5, no. 2, pp. 129–137, Jan. 1998.

[13] P. R. Spalart and S. R. Allmaras, "One-equation turbulence model for aerodynamic flows," *Rech. Aerosp.*, no. 1, pp. 5–21, Jan. 1992.