The Human Side of Service Engineering, Vol. 108, 2023, 29–37
https://doi.org/10.54941/ahfe1003104

Preparing Data Science Projects – Between Economic Aspects and Requirements Analysis

Claudia Dukino and Damian Kutzias

Fraunhofer Institute for Industrial Engineering IAO, Stuttgart, BW 70569, Germany

ABSTRACT

With the increasing availability of data in enterprises of all industry sectors, new date-based ideas arise, including artificial intelligence (AI) solutions. For those enterprises which have no experience in the implementation of such projects, data science process models can assist in structuring them. We have observed that the majority of the available models do not involve people, their activities, and the associated processes in detail. A possible reason for this is, that many of these models were created with a focus on the data processing and not necessarily to introduce ongoing data-based applications. To close this gap, this paper analyses the aspects which should be included in project preparation, especially requirements analysis, and which methods and tools are adequate to support these steps. These considerations are even more important for AI projects, since it is not necessarily clear from the beginning to which extent the required information is contained in the available data and whether the data is sufficient for the project goals. In addition, it should also be checked strategically whether the idea fits the company's goals and thus offers added value for the company. During the requirements analysis, affected users, their activities and processes are specifically focussed. During these steps, some conceptual information such as formalised current and target processes can be documented, which in turn can help when the implementation is done and the solution is brought to operation.

Keywords: Data science, Process model, Data analytics, Project management, Methodology, Project team, Economics, Industrial science, Transformation process, Key activities, Artificial intelligence, Processes

INTRODUCTION

Project ideas for new data-based solutions can come from both sides, the workforce, and the management. The difficulty of both approaches is to bring the other party along in the process and to understand what their motives are for the project idea. According to a survey by IDG Research Service, 32 percent of German companies still see the lack of acceptance by employees as the most critical factor for the failure of data-based projects (Reder, 2021). That's why it's important, regardless of whether top-down or bottom-up projects are launched, to involve people and not let the project start with the data. What does this mean for the employees in the specialist departments and for the management level in the company? A detailed analysis of the technical, organisational and human requirements should be carried out to successfully

implement the project (Hacker, 2018; Strohm, 1997). A structured approach, which also provides methods and tools, can support this.

STRUCTURED PROCEDURE FOR PROJECT PREPARATION

A structured and detailed analysis in the beginning is an important success factor for utilising the results and bringing solutions into practice in the end of the project. This is even more important for data-based projects, as the effort required is often significantly higher than for classic IT projects. One reason for this is that it is rarely clear from the outset to which extent the existing data contains the relevant information and whether it is sufficient for the project idea.

The importance of recording the requirements for the project is shown, for example, by the findings of the BMBF project "SmartAIwork" (Tombeil et al. 2021). In this project, attempts were made to pilot the introduction of AI applications (artificial intelligence) in processing at three companies from the trades industry and public administration. It became clear that it is not sufficient to simply record standard requirements for a project, but to deal with them in detail and in an integrated manner:

- Needs from user perspective
- Affected processes
- Key activities and
- Legal conditions.

It became clear that the objectives and economic viability, as well as the overall requirements from the analyses at the end of the stocktaking, indicate whether the project idea can go into conception or not.

Based on these findings, it was our concern to investigate whether there are models or methods that can support the selection and implementation process. We researched and analysed freely available data science process models. The respective process steps of the models were analysed, especially the use of methods and tools (Kutzias et al. 2023). The results of the comparative analysis of existing data science process models showed that technical aspects such as data preparation, understanding and exploration as well as model selection and construction are usually explicitly addressed and provided with concrete tool recommendations. However, aspects that are necessary especially for the preparation of a data-based projects are rarely focused. Except for the goals and economic efficiency, none of the analysed models deals with these aspects in detail. CRISP-DM (Cross-Industry Standard Process for Data Mining) (Chapman et al. 2000), TDSP (Team Data Science Process) (Microsoft, 2020) and DASC-PM (Data Science Process Model) (Schulz et al. 2020) even with concrete tool recommendations. The situation is similar when it comes to recording the requirements for the project. Here, the CRISP-DM, DASC-PM, and EDDA (Engineering Data-Driven Applications) (Hesenius et al. 2019) provide concrete assistance.

Requirements that arise from the needs of the stakeholders are only addressed in the CRISP-DM, but without tools and only considered as a marginal issue in the KDD (Knowledge Discovery in Databases) (Fayyad et al. 1996).

The need to deal with the affected processes and the associated key activities is not considered in any model, except for CRISP-DM as a marginal phenomenon. The situation is different for legal issues, where four of the seven models (CRISP-DM, ILG (The lightweight IBM Cloud Garage Method for data science) (Kienzler, 2019), EDDA, DASC-PM) state that it is important to deal with this fact, but not with concrete recommendations for implementation.

This shows that the existing models in the field of requirements elicitation at the beginning of a data-based project have gaps which should be closed for the project manager. To fill these gaps, the project preparation and its steps are discussed below. Each step is briefly described and enriched with examples of methods and tools. The sequence of steps is not mandatory, it only serves as an exemplary implementation. It may make sense to work on steps simultaneously, in a different order or not at all if they are not relevant for the project.

OBJECTIVE AND ECONOMY

The first step is to define the goals of the project. If the project idea comes from the specialist department, it is advisable to check whether there is an overarching AI strategy for the company to align the project with it if necessary and to prevent premature termination. Another point to consider is whether the project can be commercially viable, e.g. explicitly by providing a new service or business model, or implicitly by optimising an internal process. To answer this question, it is advisable to look at the requirements of the project for making a final assessment. This step can be supported, for example, by a business model canvas that can be filled in throughout the entire preparation phase.

Business Model Canvas

With the help of the Business Model Canvas, which was developed by Alexander Osterwalder, it is possible to visualise and structure new ideas in order to discuss them with all participants (Osterwalder and Pigneur, 2011). Numerous free templates are available for this purpose.

NEEDS FROM USER PERSPECTIVE

Especially in data-based projects, user acceptance is a not unimportant component (Reder, 2021). For this reason, it may be necessary to address users' needs early on and integrate them into the process. In this way, they can communicate their needs and fears, and these can be considered when developing the requirements and aligning the goals. The technology acceptance model or the persona method (see below) can support the development of requirements.

Technology Acceptance Model (TAM)

With the help of TAM (Davis, 1989) the technology acceptance can be examined from the view of the participants, for the pending project. The origin of TAM lies in the psychological theory of reasoned action (TRA), which

attempts to explain behaviour (Fishbein and Ajzen, 1975). This was used to examine two major factors in the first simple model of TAM, perceived ease of use and usefulness related to the new application and the associated dependent variable of the user behavioural intention (King and He, 2006).

Persona Method

The persona method was developed by A. Cooper for user-centred human-computer interaction (Cooper, 1999). It aims to define future user models for a fictitious characteristic person of this target group (Pruitt and Grudin, 2003; Schneidewind et al. 2012). This can help designers and developers to align (data-based) project with the needs of the users (Holzinger et al. 2022).

PROCESSES INVOLVED

Weske defines that a business process consists of activities that are carried out in a coordinated manner in an organisational and technical environment. With the purpose of achieving a common goal (Weske, 2019). The process may have been specifically created and implemented or it may have grown over time. By using data or AI, a (partially) manual process can become a supported or even automated process in the future.

In order to analyse where optimisation potential exists in processes by identifying process steps that do not add value or have the potential to automate or support activities through the use of data, the process should be captured in its as-is state. The process mapping and makigami methods described below can be used for this purpose.

If the target process is clear, it should be modelled, especially with regard to the expected human-technology interaction. Process modelling languages can serve this purpose.

Process Mapping

According to the authors, process mapping is an instrument for structured process mapping involving affected employees and experts. It makes it possible to visualise interrelationships within the process and including the existing interfaces within and outside the workflow, thus revealing weak points or work steps that do not add value. It also provides rules for redesigning processes. The analysis is often done through workshops and brown papers (Hunt, 1996; Hofmann, 2020).

Makigami

Sonntag and Alexander describe the Makigami method as a way of making company processes transparent to identify value-adding and non-value-adding activities that are necessary in the process and to derive optimisation potential from them. For this purpose, time sequences (throughput times, action times, idle times, etc.), the number of interfaces (problems) between departments and process owners as well as the exchanged documents and data carriers are recorded. The survey can be carried out very easily on a sheet of paper, according to predefined rules (Sonntag and Alexander, 2015).

Process Modelling Languages

Probably the most widespread graphical process modelling languages are BPMN (Business Process Model and Notation) and UML (Unified Modelling Language), which have been standardised and specified by the Object Management Group (OMG).

The main objective of BPMN is to use simple notation to make a business process comprehensible to all people involved. This includes first drafts, up to technical developments or the later administration and monitoring of processes and their responsibilities (Object Management Group 2013).

The UML specification by Cook et al. (Cook et al. 2015) aims to support system architects, software engineers and software developers in the analysis, design and implementation of software-based systems. It also supports project managers in modelling business and similar processes. For this purpose, rules for semantics and syntax were defined, which must be adhered to. The specification of human-readable notation elements to represent the individual UML modelling concepts and rules can, in combination, lead to a variety of different diagram types such as activity diagrams, component diagrams, class diagrams, etc., which correspond to the different aspects of the modelled systems and processes.

For the use case of human-technology interaction to automate or support sub-processes, it is relevant to show who will take over which task in the future and when.

For the use of these tools, paper, various office tools or specially designed process modelling tools, which are available on the market in large numbers, can be used.

KEY ACTIVITIES

When considering data-driven projects, it is important to find out which activities can be automated or supported. The distinction between routine and non-routine activities can help here. Routine activities describe tasks that can be performed by the computer in a limited and well-defined set of cognitive and manual activities, following explicit rules (Autor et al. 2003). Non-routine activities require decisions for complex and possibly novel problems and communication activities, in which the computer can assist, but usually does not take over the whole task (Autor et al. 2003; Autor and Dorn, 2013; Frey and Osborne, 2013). It is important to analyse the activities in the project context for these characteristics to identify the potential for automatability. The recording of all activities in the process can be done using shadowing methods and the analysis and classification of activities into routine and non-routine using a matrix for cognition and interaction requirements, both of which are briefly presented below.

Shadowing Method

This method supports the recording of previous activities in the process concerned by observing users performing their activities without disturbing or influencing them (McDonald, 2005). This helps to understand the process better and helps to identify routine and non-routine tasks. However, the

method should consider that, depending on the task, data protection may be an issue.

Matrix on Cognition and Interaction Requirements

Tombeil et al. describe how the question of the automatability of key activities can be investigated by means of the matrix on cognitive and interactional requirements. In doing so, the proportion of cognitive and interactive demands on the activity is investigated. For visualisation purposes, the two requirements are compared on two axes. The lower left quadrant represents the routine activities, where both requirements are classified as low. All other quadrants are formally classified as non-routine activities for the time being (Tombeil et al. 2020). Hacker divides the dimension of cognitive requirements into knowledge work and innovation work. Algorithmic thinking can be seen as knowledge work, which can be replaced by computer-based algorithms. A simple example is the generation of automatic texts from digital information for weather or sports reports. Innovation work is becoming increasingly important but remains with humans for the time being (Hacker, 2016). The interaction requirements of the users result from the relationship structures between service providers and service recipients (Böhle and Weihrich, 2020). For example, the interaction requirements are low in archive management and high in telephone customer service.

LEGAL

When using AI or data in products and services, law, regulation, and ethics are serious issues that the company must deal with. This means, for example, that legal requirements must be fulfilled within the framework of the EU General Data Protection Regulation (EU-GDPR) and the traceability of decisions or the creation of trustworthy AI must be ensured (Rodrigues, 2020; Vocelka, 2023). We refrain from recommending a method at this point, as the regulations and obligations are very extensive and also depend on the legislation in the respective country. Furthermore, the current legal situation must be reviewed and considered individually at any time to consider possible changes. Together with the complexity of the domain already mentioned, we advise against non-experts becoming active here and recommend outsourcing relevant issues to internal or external legal experts.

REQUIREMENTS

In the final point of the requirements, the results from the preceding analyses are brought together and, if relevant for the project, supplemented by further requirements such as usability (ISO/FDIS 9241-210:2019) or explainable AI (Haque et al. 2023). Afterwards, extensive information is available to decide for or against the project implementation.

CONCLUSION

In this paper, the early phase of project preparation was discussed and analysed to address frequent gaps in existing data science process models. The

topics of affected processes, key activities, actor needs, and legal aspects were analysed, which can yield important requirements of AI or data-based projects. These were enriched with examples of possible methods and tools to give a first impression of how to approach these topics. By means of this structured approach of the early phase of a project, (project) managers are enabled to make thoroughly informed decisions about project implementation. In addition, the analyses of the discussed topics provide a good basis for the further course of the project, as the results can be utilised in later phases of the project.

REFERENCES

Autor, D. H./Levy, F./Murnane, R. J. (2003). The Skill Content of Recent Technological Change: An Empirical Exploration. The Quarterly Journal of Economics 118 (4), 1279–1333. https://doi.org/10.1162/003355303322552801.

Autor, David H./Dorn, David (2013). The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market. American Economic Review 103 (5), 1553–1597. https://doi.org/10.1257/aer.103.5.1553.

Böhle, Fritz/Weihrich, Margit (2020). Das Konzept der Interaktionsarbeit. Zeitschrift für Arbeitswissenschaft 74 (1), 9–22. https://doi.org/10.1007/s41449-020-00190-2.

Chapman, Pete/Clinton, Julian/Kerber, Randy/Khabaza, Thomas/Reinartz, Thomas/Shearer, Colin/Wirth, Rüdiger (2000). CRISP-DM 1.0. Step-by-step data mining guide.

Cook, Steve/Bock, Conrad/Rivett, Pete/Rutt, Tom/Seidewitz, Ed/Selic, Bran/Tolbert, Doug (2015). OMG Unified Modeling Language TM (OMG UML). Available online at https://www.omg.org/spec/UML/2.5 (accessed 2/6/2023).

Cooper, Alan (1999). The Inmates are Running the Asylum. In: Udo Arend (Ed.). Software-Ergonomie '99. Design von Informationswelten ; [gemeinsame Fachtagung des German Chapter of the ACM, der Gesellschaft für Informatik (GI) und der SAP AG, Walldorf vom 8. bis 11. März 1999 in Walldorf/Baden. Stuttgart/Leipzig, Teubner, 17.

Davis, Fred D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. MIS Quarterly 13 (3), 319–340. https://doi.org/10.2307/249008.

Fayyad, Usama/Piatetsky-Shapiro, Gregory/Smyth, Padhraic (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of the ACM, Vol. 39, No. 11.

Fishbein, Martin/Ajzen, Icek (1975). Belief, attitude, intention and behavior. An introduction to theory and research. Reading, Mass., Addison-Wesley.

Frey, Carl Benedikt/Osborne, Michael A. (2013). The Future of Employment: How Susceptible Are Jobs to Computerisation?

Hacker, Winfried (2016). Vernetzte künstliche Intelligenz / Internet der Dinge am deregulierten Arbeitsmarkt: psychische Arbeitsanforderungen. Journal Psychologie des Alltagshandelns.

Hacker, Winfried (2018). Menschengerechtes Arbeiten in der digitalisierten Welt. Eine Wissenschaftliche Handreichung. Zürich, vdf Hochschulverlag AG an der ETH Zürich.

Haque, A. BahalulK. M./Islam, A. K. M. Najmul/Mikalef, Patrick (2023). Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. Technological Forecasting and Social Change 186. https://doi.org/10.1016/j.techfore.2022.122120.

Hesenius, Marc/Schwenzfeier, Nils/Meyer, Ole/Koop, Wilhelm/Gruhn, Volker (2019). Towards a Software Engineering Process for Developing Data-Driven Applications. In: 2019 IEEE/ACM 7th International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering (RAISE), 2019 IEEE/ACM 7th International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering (RAISE), Montreal, QC, Canada. IEEE, 35–41.

Hofmann, Martin (2020). Prozessoptimierung Als Ganzheitlicher Ansatz. Mit Konkreten Praxisbeispielen Für Effiziente Arbeitsabläufe. Wiesbaden, Springer Fachmedien Wiesbaden GmbH.

Holzinger, Andreas/Kargl, Michaela/Kipperer, Bettina/Regitnig, Peter/Plass, Markus/Muller, Heimo (2022). Personas for Artificial Intelligence (AI) an Open Source Toolbox. IEEE Access 10, 23732–23747. https://doi.org/10.1109/ACCESS.2022.3154776.

Hunt, V. Daniel (1996). Process mapping. How to reengineer your business process. New York, Wiley.

ISO/FDIS 9241-210:2019. Ergonomie der Mensch-System-Interaktion, 2019. Berlin.

Kienzler, Romeo (2019). The lightweight IBM Cloud Garage Method for data science. A process model to map individual technology components to the reference architecture. Available online at https://developer.ibm.com/articles/the-lightweight-ibm-cloud-garage-method-for-data-science/.

Kutzias, Damian/Dukino, Claudia/Kötter, Falko/Kett, Holger (2023). Comparative Analysis of Process Models for Data Science Projects. In: Ana Paula Rocha/Luc Steels/Jaap den van Herik (Eds.). Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART 2023). Setúbal, SciTePress - Science and Technology Publications Lda, 1052–1062.

McDonald, Seonaidh (2005). Studying actions in context: a qualitative shadowing method for organizational research. Qualitative Research 5 (4), 455–473. https://doi.org/10.1177/1468794105056923.

Microsoft (2020). Team Data Science Process Documentation. Available online at https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle.

Object Management Group (Hrsg.) (2013). Business Process Model and Notation (BPMN), Version 2.0. Available online at https://www.omg.org/spec/BPMN/2.0.2/.

Osterwalder, Alexander/Pigneur, Yves (2011). Business Model Generation. Ein Handbuch für Visionäre, Spielveränderer und Herausforderer. Frankfurt/New York, Campus Verlag.

Pruitt, John/Grudin, Jonathan (2003). Personas: practice and theory. In: Jonathan Arnowitz (Ed.). Proceedings of the 2003 conference on Designing for user experiences. New York, NY, ACM, 1–15.

Reder, Bernd (2021). Machine Learning 2021.

Rodrigues, Rowena (2020). Legal and human rights issues of AI: Gaps, challenges and vulnerabilities. Journal of Responsible Technology 4, 100005. https://doi.org/10.1016/j.jrt.2020.100005.

Schneidewind, Lydia/Horold, Stephan/Mayas, Cindy/Kromker, Heidi/Falke, Sascha/Pucklitsch, Tony (2012). How personas support requirements engineering. In: Tiziana Catarci (Ed.). 2012 First International Workshop on Usability and Accessibility Focused Requirements Engineering (UsARE). 4 June 2012, Zurich, Switzerland; [part of the 34th International Conference on Software Engineering (ICSE), 2012 First International Workshop on Usability and Accessibility Focused Requirements Engineering (UsARE), Zurich, Switzerland, 04.06.2012 - 04.06.2012. Piscataway, NJ, IEEE, 1–5.

Schulz, Michael/Neuhaus, Uwe/Kaufmann, Jens/Badura, Daniel/Kerzel, Ulrich/Welter, Felix/Prothmann, Maik/Kühnel, Stephan/Passlick, Jens/Rissler, Raphael/Badewitz, Wolfgang/Dann, David/Gröschel, Alexander/Kloker, Simon/Alekozai, Emal M./Felderer, Michael/Lanquillon, Carsten/Brauner, Dorothee/Gölzer, Philipp/Binder, Harald/Rhode, Heiko/Gehrke, Nick (2020). DASC-PM v1.0 - Ein Vorgehensmodell für Data-Science-Projekte.

Sonntag/Alexander (2015). PROMIDIS Handlungsleitfaden - Instrument Makigami. Available online at https://www.inf.uni-hamburg.de/de/inst/ab/itmc/research/completed/promidis/instrumente/makigami.

Strohm, Oliver (1997). Die ganzheitliche MTO-Analyse: Konzept und Vorgehen. In: Oliver Strohm/Eberhard Ulich (Eds.). Unternehmen arbeitspsychologisch bewerten. Ein Mehr-Ebenen-Ansatz unter besonderer Berücksichtigung von Mensch, Technik und Organisation. Zürich, vdf Hochschulverl. an der ETH Zürich, 21–37.

Tombeil, Anne-Sophie/Dukino, Claudia/Zaiser, Helmut/Ganz, Walter (2021). KI-Ambition als Treiber für die Realisierung von Digitalisierung: Wann ist weniger mehr? Stuttgart, Fraunhofer Verlag.

Tombeil, Anne-Sophie/Kremer, David/Neuhüttler, Jens/Dukino, Claudia/Ganz, Walter (2020). Potenziale von Künstlicher Intelligenz in der Dienstleistungsarbeit. In: Manfred Bruhn/Karsten Hadwich (Eds.). Automatisierung und Personalisierung von Dienstleistungen. Methoden - Potenziale - Einsatzfelder. Springer Gabler, Wiesbaden, 135–154.

Vocelka, Alexander (2023). AI Governance for a Prosperous Future. In: René. Schmidpeter/Reinhard Altenburger (Eds.). Responsible Artificial Intelligence. Challenges for Sustainable Management. Cham, Springer International Publishing; Imprint Springer, 17–90.

Weske, Mathias (2019). Business process management. Concepts, languages, architectures. Berlin/Heidelberg, Springer.