DATA FOR ENGINEERING DESIGN: MAPS AND GAPS

Chiarello, Filippo (1,4);
Coli, Elena (2,4);
Giordano, Vito (2,4);
Fantoni, Gualtiero (3,4);
Bonaccorsi, Andrea (1,4)

1: Department of Energy, Systems, Territory and Construction Engineering, University of Pisa;
2: Department of Information Engineering, University of Pisa;
3: Department of Civil and Industrial Engineering, University of Pisa;
4: B4DS Lab - Business Engin

ABSTRACT

Data, information and knowledge are strongly involved in Engineering Design (ED) process. Despite the crucial role played by data in the design process, there is a lack of studies about how different data are used and generated by the various phases of the ED process. This study is a first attempt to fill this gap by mapping which data types are involved in the different ED phases from a research perspective.

In order to achieve this objective, we used a methodology based on Text Mining. Firstly, we retrieve a corpus of scientific papers related to ED; then, we build two lexicons to recognize ED phases and data types; finally, we collect these entities within ED papers and map the relations between them.

The methodology application allows the building of a network graph for visualizing the relations among data lexicon and ED lexicon. Then, we investigate the specific relations among data types and ED phases by building a heatmap to investigate data types from 3 different perspective.

The insight coming from our analysis shows that ED studies have a great potential in the usage of many data sources, but also that there exist some gaps to be solved in order to reach a more effective data usage in the context of ED.

Keywords: Semantic data processing, Knowledge management, Information management, Data Types, Text Mining

Contact:
Coli, Elena
University of Pisa
Department of Information Engineering
Italy
elena.coli@phd.unipi.it

1 INTRODUCTION

During the Engineering Design (ED) process, engineers make intensive use of data, information and knowledge (Chaudhari et al., 2020). According to the model introduced by Ackoff (1989), data are the building block of the knowledge pyramid and are defined as properties of observables phenomena. From data it is possible to move to information (i.e., the description inferred from data), and from information to knowledge (i.e., know-how that comes from learning). Without proper data it is hard to collect design knowledge, and with the widespread of information given by digitalisation, this is not a trivial task.

Despite the crucial role played by data in the design process, there is a lack of studies about how different data are used and generated by the various phases of the ED process. This study is a first attempt to fill this gap by identifying and mapping which data types are involved in the different ED phases from a research perspective. The resulting map can be used by ED researchers and practitioners to understand potential data gaps and opportunities. It is in fact relevant to understand the relation between data and ED, aiming at helping ED researchers to focus on the impact that advances in Artificial Intelligence and Machine Learning are having on the field of ED.

In order to achieve this objective, we used a methodology based on Text Mining (TM) and applied it to a corpus of scientific papers in the ED field. The method is similar to the one described by Chiarello et al. (2021), where the authors analyse how different techniques for data analysis are used in ED. In the present paper we add the building block of data, in order to offer a complete picture. Scientific publications are chosen as input because these are reliable and updated data sources for Engineering Design, as demonstrated by Chiarello et al. (2019). Text Mining is chosen as research method to have the possibility to mine a large quantity of documents (i.e., 17,104) in an efficient way. Furthermore, such an approach has proven to be more replicable with respect to expert driven approaches, as discussed by Chiarello et al. (2019, 2020, 2021) and Fantoni et al. (2021). For the matter of replicability, code and data used to carry on the present analysis are shared in an open folder on GitHub[1].

The paper is structured as follows: firstly, in Section 2 we explore the Engineering Design process and describe the relevant data involved. In Section 3 we show the methodology to recognize the ED phases and the data types in a corpus of scientific papers. Section 4 presents the main findings of the application of the methodology and explores the intersection between ED and data types from different points of view. Section 5 concludes, showing how the emerging data related gaps in the ED literature are driven by the lack of proper interfaces inside the ED process, between ED practitioners and researchers, and between ED and other disciplines.

2 THE ROLE OF DATA IN ENGINEERING DESIGN

Engineering Design is a complex and iterative process, the success of which strongly depends on existing design experience in the domain area of the design team (Zhao et al., 2020). The role of cumulated design expertise is crucial, given that the bulk of engineering designs is related to modifications of previously proven designs to accommodate new requirements (Kim et al., 2005).

Designers must keep track of a vast amount of data while performing system analysis and synthesis (Krupa, 2019). This is a serious challenge that arises from the early stages of conceptual design, since designers must be aware and access their expertise and build up a synthesis. To this, we must add the huge amount of external data available to solve design problems. This process is time-consuming and laborious: designers must browse (mentally and manually) a large number of databases, web documents and scientific papers in order to access the relevant data (Liu et al., 2020). A possible solution to these problems is open data for design, but a complete and operational framework in this context is still far to be achieved by scholars or practitioners (Parraguez & Mayer, 2017).

Deutsch (2015) made a great effort to summarize the benefits derived from systematic data usage in the design process: data could not only help in design and planning, but also in mitigating the subjectivity of decision-making process. At the same time, Deutsch (2015) points out that even if architects, engineers and contractors have been working with data for ages, many data sources are brand new and new players may be unfamiliar, even to those already established in the industry. One of the main issues to be faced by humans in the design of complex engineering systems is in fact information overload. When data exceeds a certain size threshold, a designer may become overwhelmed

---

[1] https://bit.ly/33rW6Y8

and no longer be able to recognize the underlying meaning of the design problem at hand, which results in premature or poor results (Bang & Selva, 2016).

There is a clear need for a methodology to identify and map the data that are relevant in the various stages of ED, in order to unveil how data is used in ED process. This need becomes even more relevant in routine design, that makes use of existing standards, guides, procedures and design cookbooks to define solutions for given functional requirements. For this reason, the value of the proposed methodology is mostly focused on routine design activities, where ED procedures are known, and design variables are stable. In this context, it is desirable to synergize the designers' creativity and the capacity of exploring large amount of data minimizing manual intervention to most value-added steps (Bi, 2019).

3 METHODOLOGY

This section describes the methodological steps undertaken to map the data sources of ED using Text Mining on scientific articles. The methodology is shown in Figure 1 and is subdivided in three macro-tasks: (1) selecting and retrieving the scientific papers related to ED; (2) building the lexicons to recognize ED phases and data types by using Named Entity Recognition (NER) techniques; (3) applying NER for collecting the listed entities in the lexicons within the ED scientific papers and mapping the relationships between these entities. In the following Subsections (3.1-3.3) we describe in detail each step of the methodology.
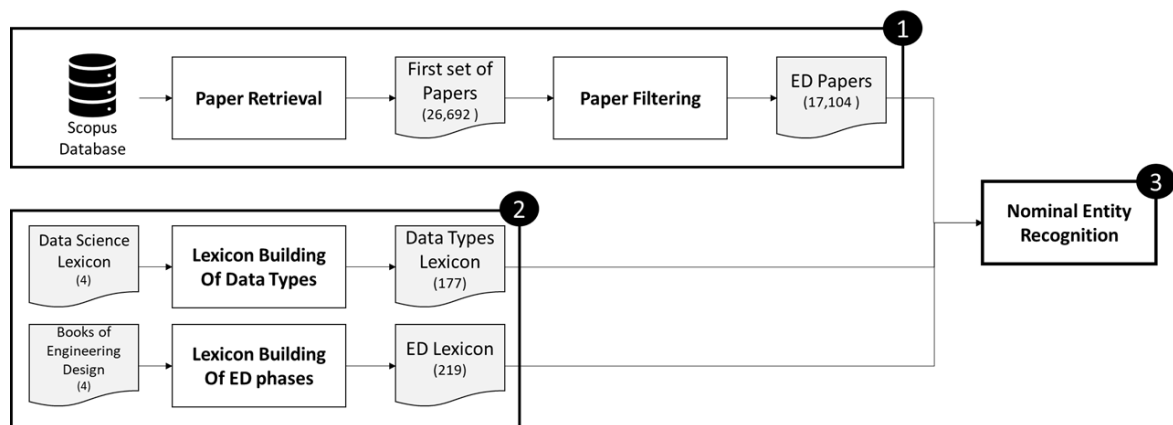

Figure 1. Flowchart of the methodology.

3.1 Paper Retrieval

The goal of this stage is collecting scientific articles covering the Engineering Design field. We collected all the scientific papers (both articles in journals and conference papers) published in a 20-year time window (2001-2020) and containing the string "engineering design" in the title, abstract or keywords. The papers were retrieved from Scopus, a well-known database of abstracts and citations launched in 2004 by Elsevier. This approach, executed in November 2020, led to the collection of 26,692 scientific papers. From this starting set, we selected only papers belonging to journals and conferences in the fields of Engineering, Computer Science, Business, Management and Accounting, Decision Sciences and Multidisciplinary (according to the All Science Journal Classification, ASJC), obtaining 17,104 documents.

3.2 Lexicon Building

The collected ED papers in Section 3.1 are analysed to automatically extract the ED phases and the data types using Nominal Entity Recognition (NER) techniques. NER is the process of detecting lexical units in a word sequence that refers to a predefined entity, thus determining what kind of entity the unit is referring to, such as food names, technologies, users and so on (Chiarello et al., 2018).

The approach to NER used in this article is based on gazette or lexicon (a structured set of expressions belonging to a specific semantic domain), a well-known terminological-driven NER. This method aims to map mentions of entities within texts to terminological resources (e.g., Wikipedia) (Nadeau et al., 2007). We start from two lists of relevant entities: (1) ED phases lexicon and (2) data types lexicon.

It is important to stress here the concept that the lexicons added a strong value to the corpus, since they enabled not only the filtering of relevant papers, but also the identification of the main use of data in the different ED phases. For this reason, and for reproducibility purposes, we made the lexicons and the dataset of scientific articles available on GitHub, as explained in Section 1.

First of all, we need to build an **Engineering Design lexicon** in which each expression of the lexicon is associated to a specific ED phase. For this reason, we followed a hybrid top-down and bottom-up approach based on ED state-of-the-art. Firstly, we selected a small number of authoritative textbooks in the field of ED and synthesised a shared view of the main phases of the process. The books taken into account during this step are: (1) "Engineering Design: a systematic approach" (Pahl et al., 2007), (2) "Engineering Design" (Dieter & Schmidt, 2009), (3) "Product design and development" (Ulrich, 2003) and (4) "Engineering Design process" (Haik et al., 2018). Since there is not a shared terminology for the ED process, we based our framework on 4 vertical phases that were independently identified by two of the four authors and that are shared also by the other two. They are numbered in order of time unfolding, from Problem definition to Conceptual, Embodiment, and Detailed design, respectively. Furthermore, we added 6 horizontal or transversal activities (i.e., set of activities that have an impact on the whole ED process) to the framework. The first column of Table 1 (see Subsection 4.1) shows the 10 identified ED phases.

Following this methodology, we obtained an ED lexicon of 333 words associated to 10 ED phases.

In order to build the **data types lexicon**, we referred to data mentioned in Data Science studies, since it has many sub-fields (e.g., Machine Learning, statistics, big data, etc.) and allows a detailed exploration of different kinds of data. For this reason, we merged 4 publicly available lexicons: (1) the Machine Learning vocabulary developed by Google, (2) the Glossary of common Machine Learning, Statistics and Data Science terms by Analytics Vidhya, (3) the Data Science Glossary by Kaggle and (4) the Outline on Machine Learning by Wikipedia. We then used a word2vec approach based on Glove (Pennington et al., 2014) to find words and n-grams (contiguous sequences of n words) similar to those contained within the two lexicons. For each entry of the two lexicons, we extracted the top-30 similar (in terms of cosine similarity) words contained in the corpus of scientific articles of ED, and manually selected the relevant ones. Moreover, we manually associated each word of the lexicon to one of 27 different classes. These classes were designed following 4 different classification criteria, in order to explore data types from different perspectives. The first and second columns of Table 2 show the 4 classification criteria and the 27 data types classes, respectively (see Subsection 4.1).

Following this methodology, we obtained a data types lexicon of 201 words associated to 27 data type classes.

3.3 Nominal Entity Recognition

The approach to NER used in this article is based on gazette approach that maps entities of the previously defined lexicons within texts. We start from the two lists of relevant entities developed in Subsection 3.2: (1) ED phases lexicon and (2) data types lexicon. Then the occurrences of all these entries are identified in paper corpus, using the trie based search (Chu & Culbert, 1997).

The lexicons were used in order to filter the papers, i.e., only the papers containing at least one term for each lexicon (in the title, abstract or keywords) were selected. Out of our original references, 6,337 were selected for the analysis.

We examine the corpus in two steps. First, we focus separately on the phases of the ED process and on data types. Second, we identify the intersections between the two dimensions of ED phases and data types by building co-occurrence matrices from the lexicons.

4 RESULTS

We present the key findings in two main subsections: (1) the Engineering Design phases and data types, results of NER (Subsection 4.1) and (2) the ED map of data types (Subsection 4.2).

4.1 Engineering Design phases and data types

The application of the methodology leads to the identification of: (1) 219 unique entries of ED lexicon that are referred to the Engineering Design phases, grouped in 10 macro phases, and: (2) 177 unique entries of data types lexicon, structured in 27 classes, following 4 different classification criteria.

Table 1 shows a breakdown of the number and the percentage of papers that deal with the various phases and activities of the ED process. The papers analysed are those containing at least one entry of the data types lexicon. Since every paper is associated with one (or more) ED phases and uses one (or more) data types, the table shows a distribution of the actual usage of data in the different ED phases.

Table 1. Number of papers that contain at least one entry in the data types lexicon, by phase and activity of the Engineering Design process

	Engineering design phases	**Number of papers**	**% of total number of papers**
Vertical ED Phases	1 - Problem Definition	632	9.97 %
	2 - Conceptual Design	918	14.49 %
	3 - Embodiment Design	642	10.13 %
	4 - Detailed Design	878	13.86 %
Horizontal ED Phases	Design Education	2904	45.83 %
	Design for X	108	1.7 %
	Innovation & IP Management	1561	24.63 %
	Project Management	1529	24.13 %
	Prototyping	811	12.8 %
	Sustainability	311	4.91 %

From Table 1, it is evident that the 4 vertical phases of ED are almost equally distributed among data sources (about 10 % each). This is an evidence of the fact that the whole process of ED is data driven, and that a large literature (6,337 scientific papers) talking about these data sources exists.

Moving to horizontal ED phases, it is evident that a large fraction of scientific papers under analysis mention Design Education-related data sources (45.83 % of papers). This was an expected result. In fact, many scientific works in ED use surveys and interviews to analyse the human (student)-behaviour in Engineering Design or to assess the learning outcomes of the classroom from an Engineering Design perspective (Bartholomew et al. 2017). The extent to which the use of data in Design Education translates into an intense use of data in the practice of design is an open question, which points to some weaknesses, to be discussed below.

Second, the next horizontal ED phase which make intensive use of data is Innovation & IP Management. This result is led by the fact that a large stream of literature uses data coming from patent, a large and open repository of information about designs and innovations (Moehrle & Caferoglu, 2019). Very close in terms of percentage of papers there is Project Management, where data are produced by the team interaction phases. This is an interesting result, considering the COVID-19 breakdown. Since the pandemic is already shaping technological changes and the way people work (Melluso et al, 2020), the go-remote process is increasing the amount of data that are generated during the interaction between workers. This is another opportunity to collect data on how people perform design activities.

In general, by looking at Table 1, we are led to conclude that the most intensive use of data is not made in the central pipeline of design, from problem definition to detailed design, but in infrastructural and transversal activities, such as Project Management and Innovation Management, as well as in the remote preparation of students and practitioners.

Table 2 shows a summary of data types and their distribution in the context of ED scientific papers. In particular, it shows 27 data classes identified on the basis of 4 different logical criteria, in order to observe Engineering Design data from several points of view.

First, not surprisingly the business functions that exchange the larger amount of data with ED processes are Design (22.44 %), Intellectual Property (19.66 %) and Planning and Control (15.21 %), all data-intensive activities by nature. On the contrary, much less use of data is found in a business function such as Production (6.82 %), that in theory should be prepared to the high levels of data exchange and processing associated to the Industry 4.0 revolution. Even lower use of data is found in other business functions such as Accounting, Marketing and Sales, and Purchasing. Since these activities do make intensive use of data, we interpret these findings as suggesting a distinction between internal and structured data (e.g., CRM, or vendor rating systems) and the whole range of structured and unstructured data that are made available to firms by advanced data-driven methods. Going further

in the interpretation of these findings, it can be said that the inter-functional collaboration between design and market-oriented business functions, such as purchasing (upstream), and marketing (downstream) is still in a poorly developed stage.

Table 2. Number of papers that cite data types, by classification criteria

Classification criteria	Data types classes	Number of papers	% of total number of papers
Business Function	Design	1422	22.44 %
	Intellectual Property	1246	19.66 %
	Planning and Control	964	15.21 %
	Production	432	6.82 %
	Accounting	14	0.22 %
	Inspection	8	0.13 %
	Marketing and Sales	8	0.13 %
	Purchasing	1	0.02 %
Data Format	Image and Video	737	11.63 %
	Textual	434	6.85 %
	Audio	92	1.45 %
	Numerical	11	0.17 %
Data Element	Structure	648	10.23 %
	Feature	411	6.49 %
	Format	37	0.58 %
Sources	Human Interactions	1469	23.18 %
	Simulation	1064	16.79 %
	Web	448	7.07 %
	Machines	205	3.23 %
	Empirical Evidence	134	2.11 %
	Software	23	0.36 %
	Literature	18	0.28 %
	Historical	12	0.19 %
	Digital World	11	0.17 %
	Machine Learning	9	0.14 %
	Open Sources	3	0.05 %

Second, it is rather surprising to observe that Image and Video type of data outperform Textual data. It remains to be seen whether this result is driven by the importance of Design Education, in which these kinds of data are regularly used in the classroom and in the field (e.g., recording of ethnographic observations).

Third, Human interaction dominates the ranking of data sources, followed by Simulation. At the bottom of this ranking, we find an interesting entry: "Open Sources". Despite the clear advantage linked to open data, our findings suggest that they are poorly adopted in ED (0.05 %). This evidence, linked with the large amount of data coming from Education, is another evidence of the fact that there exists a lack of open industrial data in the context of ED. This creates limitations in terms of the phenomena that can be studied and in terms of the overall reproducibility of research in ED. Researchers should strive to identify other data sources to study ED principles and tools. We acknowledge that this process requires trust-building and a deep connection to industrial settings that are traditionally hesitant to share design data.

4.2 The ED map of Data Types

This subsection discusses the relationship among data types and ED phases. In this analysis we use the classification criteria of Table 2 that hold the best promise in terms of analysis: we excluded the entries with very small number of observations (e.g., Digital World, Machine Learning and Open Sources) and those with limited internal variability (e.g., entries in Data Element).

Figure 2 shows the graph where each node is an entry of our two lexicons; an edge between two nodes exists if there is at least one scientific paper that mentions both nodes in the title, abstracts or keywords; the thickness of the edge depends on the co-occurrences between the nodes (i.e., number of papers in which they co-occur); the size of the nodes depends on the node degree; and the colour of the nodes represents the group that contains the entry, as shown in the legend panel.

The graph allows us to observe the ED literature from a data-based point of view. A detailed investigation of Figure 2 gives three key insights:

1. the Innovation & IP Management (in yellow) and the Project Management (in light blue) are the bridge phases that allow to connect the data types lexicon entries (in teal), such as *simulation, copyright, requirement, procedure* and *sensor* data, and the Design Education phase (in lilac). The central role of these horizontal phases can be seen as an opportunity to bring together ED research and ED practice: we have here the evidence that ED research is largely using Intellectual Property related data to carry out research, and that companies can take great advantage of that. With the proper technological transfer, scholars can disseminate innovative approaches towards companies, and companies can share their data to scholars. ED will struggle taking advantage of the new opportunities brought by Machine Learning if this is not going to happen;

2. the web and survey data play a central role. This can represent a switch in the paradigm: it is well known that surveys are an important tool to mine customers point of view, but this has drawbacks. In the last years, research is moving towards the substitution of questionnaires with more effective data sources in terms of quantity and quality of data, such as social media and blogs. This shift is having an important impact on ED, considering that this data is open and easily collectable;

3. conceptual Design phase (in red) has a few relationships with Innovation & IP Management, Project Management and Design Education, but it is more linked with the data types of lexicon and is peripherical with respect to the layout of the graph. This shows that conceptual design is one of the most internal phases of ED, where stronger relations exist only inside the design team. In this phase the team need to synthetize the problem towards feasible solutions, taking into consideration the needs of the users. It is interesting to notice that this part of the graph is the opposite to the education side. To carry out scientifically sounding experiments on conceptual design, there is the need of ED expertise, and students cannot be easily used for this purpose. From a data perspective, this is the reason why knowledge bases are the closer data type to this phase. This open up to another possible solution to the lack of open data in the context of ED, that is the disclosure of ED knowledge bases. Again, this is not a trivial process given the value of knowledge bases both for companies and universities.
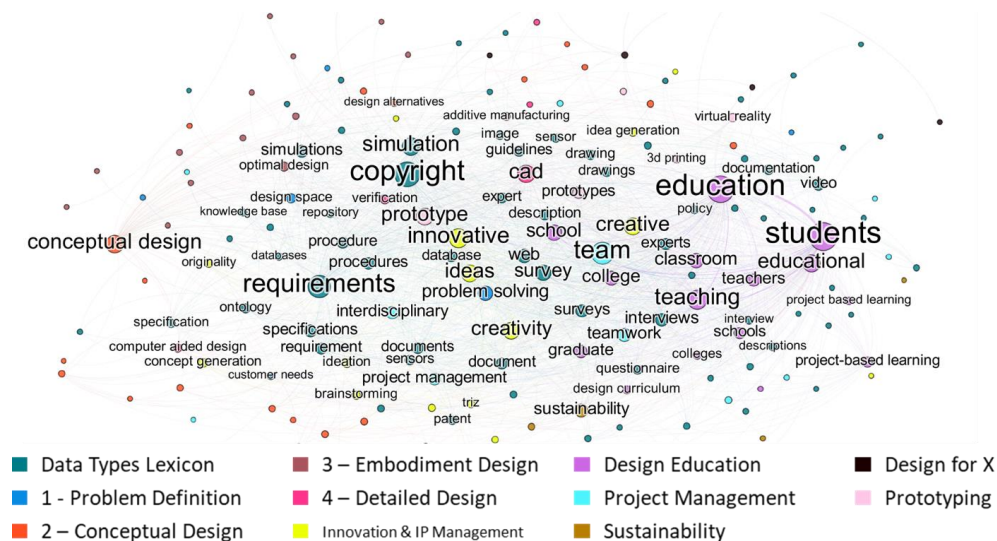

Figure 2. Graph of ED and data types lexicons and their relationships in analysed articles.

Moving to Figure 3, we can see the 3 heatmaps of the co-occurrences among each ED phase and the data types. In particular, 3 heatmaps were built according to the 3 different data types classification criteria taken into consideration (see Subsection 4.1). In particular moving from left to right, the heatmaps shows co-occurrences among ED phases and (i) data provided by different Business Functions (on the left), (ii) data in a certain Data Format (in the centre) and (iii) data provided by different Sources (on the right). Each heatmap gives evidence of the cardinality of an intersection by using colours: large numbers of scientific articles, intersecting an ED phase and a data type, are shown in white, while small numbers are shown in darker purple. In addition, heatmaps make gaps and blind spots explicit, identifying the empty regions of the matrices. In each cell, we also report the number of papers mentioning both types of entities (data types and a given ED phase). The x-axis shows the different data types divided in the 3 areas we want to investigate (Business Function Data, Data Format, Source of Data). The y-axis reports the ED phases that are divided into two main groups: vertical phases and horizontal phases, as explained in Subsection 3.2.
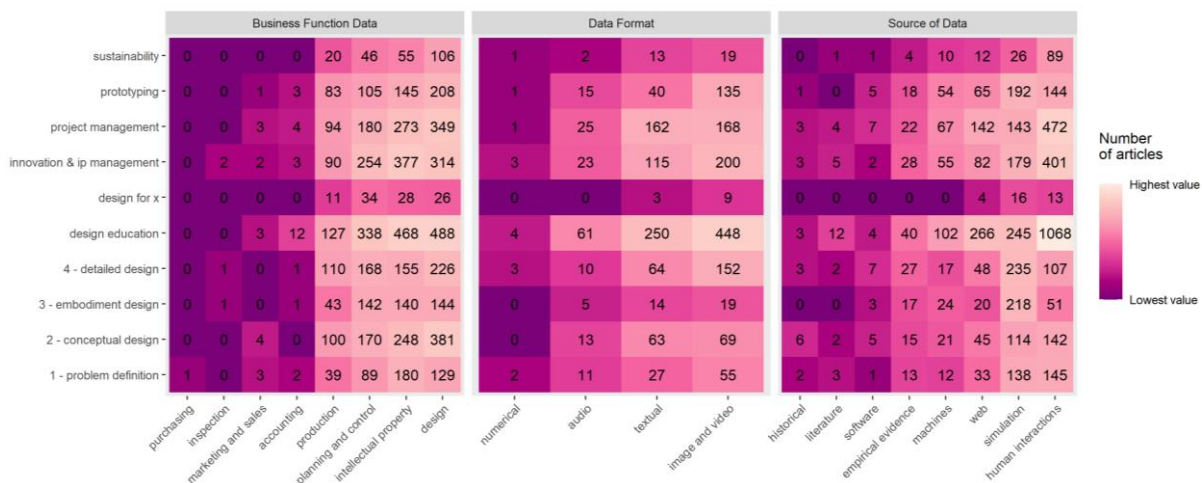

Figure 3. Number of papers in the dataset mentioning (i) data provided by different Business Functions (on the left), (ii) data in a certain Data Format (in the centre) and (iii) data provided by different Sources (on the right), by ED phase.

As expected, the first heatmap of Figure 3 highlights that data coming from the design function are the most used within ED process. Furthermore, the high transversality of data belonging to Intellectual Property, Planning and Control and Production functions confirms the relevance of previously proven designs, coming from patents, operational plans and product databases in each step of ED.

Data coming from Marketing and Sales and Purchasing functions seem to be less used in ED, even if they are relevant especially in the translation of customer requirements into technical specifications. Probably this is due to the fact that the importance of data integration in ED has only recently been recognised (Yu et al., 2018). This is an evidence of the fact that a stronger integration is needed not only among ED research and practice, but also between different functions that deal with ED in order to properly exploit different data sources.

The heatmap on Data Format (the central one) shows what kind of data ED researchers are asked to process in the different ED phases, but also what are the technologies needed to process data (i.e., Optical Character Recognition for images, speech to text software for audios). In particular, we found that image and video are the most used in ED and in particular in Design Education, followed by textual data, confirming the growing need of Text Mining applications in ED in recent years (Zhao et al., 2020). For what concerns blind spots, it is evident a lack of usage of numerical data. This is because rarely data from the design phase comes in a structured format (tidy tables to be analysed), but rather has to be pre-processed in order to become numbers. A stronger focus on structuring ED related unstructured data (i.e., video, images, text and sounds), and sharing the results of the structuring phase, can help ED researchers to the use a larger quantity of data of different type. Here, the interface should be between ED scholars and researchers in other disciplines (e.g., computational linguistics, Virtual Reality experts), since structuring this kind of data require specific computation skills to be integrated with Design Expertise.

Finally, analysing the third heatmap, we can state that many ED phases make great use of web data (e.g., forums and social networks), as already evident from the analysis of Figure 2. Just a few ED phases (such as Conceptual Design) involve historical data, and this is again linked to the fact that historical data on ED are rarely available for research.

All ED phases use data from human interactions (e.g., interviews or experts), while only a few specific phases (in particular Design Education and Project Management) make more massive use of machine-generated data or data coming from software (e.g., Enterprise Resource Planning and Manufacturing Execution System). This may show that ED struggle in becoming digital or 4.0, using data coming from machines to design better product and processes. Anyway, this can be a bias in our data set, since the significant impact of technologies 4.0 on ED in industrial context. For this reason, we explore how simulation (one of the enabling technologies of Industry 4.0 more focused on data) impacts on the different ED phases. Contrary to what noted above with respect to Machine Learning, data from simulations seem to be quite widespread in ED, evidence that simulation technologies are well established in this field. This is an interesting finding, suggesting that the adoption of Industry 4.0 technologies does not start with more ambitious but less understood Machine Learning techniques, but mainly after the well-known practices associated to simulation, which takes place in CAD suites and in Virtual Reality Environments.

With respect to blind spots, all the data sources have low values for sustainability, except for human interactions: these insights confirm the hypothesis that sustainability is a new challenge for most manufacturing companies (Lin et al., 2018) and for this reason there are few historical and empirical data available.

5 CONCLUSION

The present paper maps data sources in the context of ED studies. The map comes from automatic text analysis and shows an efficient and reproducible method for analysing scientific literature. The insight coming from our analysis shows that ED studies have a great potential in the usage of many data sources, but also that there exist some gaps to be solved in order to reach a more effective data usage in the context of ED. These gaps can all be linked to problems of interfaces of ED.

First, interface between ED research and industrial practices. Data Science tools have already been used in ED by companies, but a little evidence of transfer of the outcomes of data-driven design research into practice exist, as well as standard approaches for the successful transfer of research into practice.

Second, interface between ED and other disciplines. ED embraces competences of several scientific sectors and several disciplines of study. A proper use of data coming from ED should bring together multidisciplinary teams, with experts on the data and experts in the approaches to process this data.

Finally, interaction between ED related business functions. Data are created within and across phases of ED and this aspect is not visible in the literature. The reason why the academic papers of the present analysis do not address this issue, is that localized studies are easier from research perspective. Scholars should challenge themself finding a way to bridge different business functions, unveiling the value of ED related data.

REFERENCES

Ackoff, R.L. (1989), "From Data to Wisdom", *Journal of Applied Systems Analysis*, Vol. 16, No. 1, pp. 3-9.

Bang, H. and Selva, D. (2016), "iFEED: Interactive feature extraction for engineering design", *Proceedings of the ASME 2016 International Design Engineering Technical Conference*, Charlotte, August 21–24, 2016. https://doi.org/10.1115/DETC2016-60077

Bartholomew, S.R., Reeve, E.M., Veon, R., Goodridge, W., Lee, V.R. and Nadelson, L. (2017), "Relationships between access to mobile devices, student self-directed learning, and achievement", *Journal of Technology Education*, Vol. 29(1), pp.2-24. https://doi.org/10.21061/jte.v29i1.a.1

Bi, Z. (2019), *Finite Element Analysis Applications*, Academic Press.

Chaudhari, A.M., Bilionis, I., and Panchal, J.H. (2020), "Descriptive models of sequential decisions in engineering design: An experimental study", *Journal of Mechanical Design*, Vol. 142(8): 081704. https://doi.org/10.1115/1.4045605

Chiarello, F., Cimino, A., Fantoni, G. and Dell'Orletta, F. (2018), "Automatic Users Extraction from Patents", *World Patent Information,* Vol. 54, pp. 28–38. https://doi.org/10.1016/j.wpi.2018.07.006

Chiarello, F., Melluso, N., Bonaccorsi, A. and Fantoni, G. (2019), "A text mining based map of engineering design: Topics and their trajectories over time", *Proceedings of the design society: International conference*

on engineering design, Vol. 1, No. 1, pp. 2765-2774. Cambridge University Press. https://doi.org/10.1017/dsi.2019.283

Chiarello, F., Bonaccorsi, A. and Fantoni, G. (2020), "Technical Sentiment Analysis. Measuring Advantages and Drawbacks of New Products Using Social Media", *Computers in Industry*, Vol. 123: 103299. https://doi.org/10.1016/j.compind.2020.103299

Chiarello, F., Belingheri, P., Fantoni, G. (2021), "Data Science for Engineering Design: State of the Art and Future Directions", *Computers in Industry (in press)*

Chu, K.C. and Culbert, D.J. (1997), "Efficient high speed trie search process", U.S. Patent No. 5,640,551. Washington, DC: U.S. Patent and Trademark Office. Available at: https://patents.google.com/patent/US5640551

Deutsch, R. (2015), "Leveraging data Across the Building Lifecycle", *Procedia Engineering*, Vol. 118, pp. 260-267. https://doi.org/10.1016/j.proeng.2015.08.425

Dieter, G.E. and Schmidt, L.C. (2009), *Engineering design*, Boston: McGraw-Hill Higher Education.

Fantoni, G., Coli, E., Chiarello, F., Apreda, R., Dell'Orletta, F. and Pratelli, G. (2021), "Text mining tool for translating terms of contract into technical specifications: Development and application in the railway sector", *Computers in Industry*, Vol. 124C: 103357. https://doi.org/10.1016/j.compind.2020.103357

Haik, Y., Sivaloganathan, S. and Shahin, T.M. (2018), *Engineering design process*, Nelson Education.

Kim, S., Bracewell, R. and Wallace K. (2005), "A framework for design rationale retrieval", in *Proc. of ICED 05, 15th International Conference on Engineering Design,* Melbourne, Australia, 252-253.

Krupa, G.P. (2019), "Application of Agile Model-Based Systems Engineering in aircraft conceptual design", *The Aeronautical Journal*, Vol. 123(1268), pp.1561–1601. https://doi.org/10.1017/AER.2019.53

Lin, J., Rohleder, C. and Nurcan, S. (2018), "Material Management in LCA Integrated PLM" in *Proc. of 2018 IEEE International Conference on Engineering, Technology and Innovation*, ICE/ITMC 2018, Stuttgart, Germany, 17-20 June 2018. https://doi.org/10.1109/ICE.2018.8436374

Liu, Q., Wang, K., Li, Y. and Liu, Y. (2020), "Data-Driven Concept Network for Inspiring Designers' Idea Generation", *Journal of Computing and Information Science in Engineering*, 20(3): 031004. https://doi.org/10.1115/1.4046207

Melluso, N., Bonaccorsi, A., Chiarello, F. and Fantoni, G. (2020), "Rapid detection of fast innovation under the pressure of COVID-19", *PloS one*, 15(12): e0244175. https://doi.org/10.1371/journal.pone.0244175

Moehrle, M.G. and Caferoglu, H. (2019), "Technological speciation as a source for emerging technologies. Using semantic patent analysis for the case of camera technology", *Technological Forecasting and Social Change*, Vol. 146, pp.776-784. https://doi.org/10.1016/j.techfore.2018.07.049

Nadeau D. and Sekine S. (2007), "A survey of named entity recognition and classification", *Lingvisticae Investigationes*, Vol. 30, no. 1, pp. 3–26. https://doi.org/10.1075/li.30.1.03nad

Pahl, G., Beitz, W., Feldhusen, J. and Grote K.H. (2007), *Engineering Design: A Systematic Approach*, ed. K. Wallace and L. Blessing., London: Springer-Verlag. https://doi.org/10.1007/978-1-84628-319-2

Parraguez, P. and Maier, A. (2017), "Data-driven engineering design research: opportunities using open data", in *Proceedings of the 21st International Conference on Engineering Design (ICED 17)*, Vol. 7: Design Theory and Research Methodology, Vancouver, Canada, 21-25.08.2017, pp. 041-050. ISBN: 978-1-904670-95-7

Pennington, J., Socher, R. and Manning, C.D. (2014), "Glove: Global vectors for word representation", *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP),* pp. 1532-1543. https://doi.org/10.3115/v1/D14-1162

Ulrich, K.T. (2003), *Product design and development*, Tata McGraw-Hill Education.

Yu, J., Wang, G., Ming, Z., Yan, Y. and Lan, X. (2018), "Ontology-based unified representation of dynamic simulation models in engineering design", in *Proceedings of the ASME Design Engineering Technical Conference*, Vol. 2A-2018. https://doi.org/10.1115/DETC2018-85536

Zhao, S., Xu, C. and Wang, R. (2020), "Knowledge structure generation and modularization based on binary matrix factorization in engineering design", *Journal of Mechanical Science and Technology*, Vol. 34, pp. 4657–4673. https://doi.org/10.1007/s12206-020-1024-4