INTERNATIONAL DESIGN CONFERENCE - DESIGN 2014
Dubrovnik - Croatia, May 19 - 22, 2014.

TOWARDS IDENTIFYING PATTERN IN ENGINEERING DOCUMENTS TO AID PROJECT PLANNING

L. Shi, J. A. Gopsill, C. Snider, S. Jones, L. Newnes and S. Culley

Keywords: engineering document pattern, project planning, pattern identification, knowledge reuse, visualisation

1. Introduction

Engineering projects consist of complicated processes and involve large quantity of collaborations, communications and distributed resources. The increasing of complexity of engineering projects causes uncertainty for project execution, and creates challenges for project design and management. Complexity and uncertainty affect project planning in a variety of ways including the identification of goals and objectives, selection of organisational form, selection of project inputs, control of time, cost and quality [Baccarini 1996]. As project planning has an immediate impact on project success, eliminating the complexity and uncertainty effects is necessary for improving the success rate. It has been revealed that the reuse of knowledgeable information from previously completed projects is able to reduce the uncertainty for current project planning [Markus 2001], [Disterer 2002]. However, identifying and extracting suitable knowledge from previous projects for reuse purposes are difficult, as (i) the structure of knowledge can be complicated, and (ii) the knowledge may have various representations according to different points of view [Chandrasegaran et al. 2013].

This research facilitates knowledge reuse in the context of aiding project design by defining knowledge and activity representation of engineering projects, and proposing methods of identifying the representation from engineering documents.

The remainder of this paper is organised as follows. Section 2 reviews the challenges of project planning. Section 3 describes the approaches proposed. Section 4 includes an experimental study using industrial data. Section 5 concludes this work.

2. Related Work

Recent research shows that simplicity should be a primary principle in project design and planning [Vidal et al. 2011], [Giezen 2012]. The characteristics of a project such as project structure, timeline, cost and risk, are always taken into account by the project design process, and the characteristics also have a significant impact on the decision making process in project planning. However, most projects have numerous characteristics, and each of them may have different importance weightings for project planning purposes, which can cause difficulties for project planning, (i) considering excessive number of characteristics is time-consuming and expensive, and (ii) considering characteristics with low significance weight may exaggerate their minor influences, and therefore prevent from identifying the project objectives. On this basis, project planning should not treat all characteristics equally, and its decision making process needs to focus on mainly the characteristics above a certain level of significance, which would reduce the difficulty of planning, and therefore improve the effectiveness and efficiency of project design and management [Berkun 2005], [Chapman and Ward 2007].

Dvir et al. [2003] indicate the processes of project planning, such as deciding timelines, setting specific tasks, allocating resources, managing costs, are considered to be unlikely to perform precisely at the initial planning stage, as there is insufficient evidence to support decision making at this stage. Meanwhile, a project is usually treated as a unique and individual case, and the evidence and profiles from previously completed projects are therefore not really considered in the planning process [Engwall 2003]. Furthermore, numerous factors, such as changes of interests, purposes and constraints, have impacts on project execution, thus the applicability and rationality of the initial project plan can be significantly affected [Turner and Cochrane 1993], [Turner and Müller 2003].

The knowledge of an engineering project has two aspects, (i) the knowledge as content, i.e., the methodology, rationale and other useful information that the project was applied and followed, and (ii) the knowledge as the activity, i.e., the way that the project was conducted and proceeded. Recent research highlights that reusing such knowledge of previously completed projects is an effective approach to facilitating project design and management [Baxter et al. 2008], [Ettlie and Kubarek 2008].

In the context of project planning, the knowledge of previous projects can contain significant project characteristics and evidence, therefore they are useful to improve the applicability and rationality of project plan. The main reasons are, (i) the project characteristics such as project structure, timeline, cost and risk, can provide essential indications to the current project decision makers, and (ii) the evidences such as solutions of particular problems encountered, can be used to support the decision making process, therefore the process is based on facts rather than estimations.

In practice, the definition and classification of project-related knowledge can vary dependent on the intended uses. Moreover, the domain of knowledge can also vary dependent on these different project characteristics. Therefore, to identify suitable knowledge for large collection of projects can be complicated [Xie et al. 2011], [Carey et al. 2013]. To solve that, Sawyer et al. [2005] suggest to use a type of shallow knowledge to represent projects. The shallow knowledge is generated from project related documents by using corpus-based statistical approaches, and it is easier to be identified than domain specific knowledge, thus shallow knowledge is a suitable knowledge representation for large scale project-related knowledge in practice.

3. Engineering Document Patterns

An Engineering Document Pattern (EDP) is defined as a type of shallow knowledge extracted from engineering documents by analysing their content using data mining and semantic content analysis techniques. The use of EDP aims to help decision makers understand the initial requirements and potential risks of a current project. In this research, the EDP of a project will cover three aspects: (i) time related; (ii) people related, and (iii) file-related.

3.1 Time-related EDP Identification

An engineering document can contain lots of time-related information that is reusable for future project design and control, e.g., setting project milestone, deciding the execution time of a task, etc. In this research, time-related information contained by engineering documents is treated as time-related EDP, and it is used to identify project time-line, page time-stamp, percentage of date occurrence and cumulative frequency of within page dates. The detailed explanation of the identification process is shown as follows.

3.1.1 Project Time-line

In practice, most engineering documents contain explicit start/end dates, e.g., the In-Service reports analysed include a cover page that displays the dates explicitly, so that the time-line of a project can be extracted directly from its engineering documents. For an engineering document without explicit start/end date, there are two solutions, (i) analysing each page content to identify the most likely start/end date, and (ii) entering start/end date manually.

3.1.2 Page Time-stamp

The pages of an engineering document are likely to have been produced at different times. If a page contains multiple dates, the first occurrence of the date will be considered as the page time-stamp, and the time-stamp must be in the range of project time-line. The time-stamp of a page indicates the occurred time of its related task, therefore different tasks of the project can be linked to the project time-line based on their time-stamps.

3.1.3 Percentage of Date Occurrence

Assuming a given report contains multiple dates d_n, i.e., $\{d_1, d_2, ..., d_n\} \in r$, then the percentage of date occurrence (pdo) of d_i regarding r is defined as:

$$pdo(d_i) = \frac{occu(d_i)}{\sum_{j=1}^{n} occu(d_j)} \tag{1}$$

It measures the importance of a date by considering its occurrence. It is argued that a higher value means the date has a higher importance.

3.1.4 Cumulative Frequency of Within Page Dates (WPD)

Assuming a page p_x of the given report r contains multiple dates $d_{x,m}$, i.e., $\{d_{x,1}, d_{x,2}, ..., d_{x,m}\} \in p_x$, then the cumulative frequency of WPD regarding p_x is defined as:

$$cum\_ferq(p_x) = \frac{\sum_{i=1}^{m} occu(d_{x,i})}{\sum_{j=1}^{n} occu(d_j)} \tag{2}$$

It measures the project activity of page related task by using the cumulative frequency of the within page dates. A higher value is considered to mean that the page related task has a higher project activity.

3.2 People-related EDP Identification

The people-related information of an engineering document includes the participants' names, job titles, contacts and expertise. It is potentially reusable for future projects, e.g., personnel selection, human resource allocation, etc. In this research, people-related information contained by engineering documents is treated as people-related EDP, and it is used to identify percentage of people occurrence and cumulative frequency of within page people. The detailed explanation of the identification process is shown as follows.

3.2.1 Percentage of People Occurrence

Assuming a given report r contains multiple people names pn_n, i.e., $\{pn_1, pn_2, ..., pn_n\} \in r$, then the percentage of people occurrence (ppo) of pn_i regarding r is defined as:

$$ppo(pn_i) = \frac{occu(pn_i)}{\sum_{j=1}^{n} occu(pn_j)} \tag{3}$$

Similar to percentage of date occurrence, this measures the importance of people by considering the occurrence of his/her name. A higher value means the person has a higher impact on the project and possibly a higher importance.

3.2.2 Cumulative Frequency of Within Page People (WPP)

Assuming a page p_x of the given report r contains multiple people names $pn_{x,m}$, i.e., $\{pn_{x,1}, pn_{x,2}, ..., pn_{x,m}\} \in p_x$, then the cumulative frequency of WPP regarding p_x is defined as:

$$cum\_ferq(p_x) = \frac{\sum_{i=1}^{m} occu(pn_{x,i})}{\sum_{j=1}^{n} occu(pn_j)} \tag{4}$$

Similar to cumulative frequency of WPD, this measures the activity of the page related people by considering the cumulative frequency of their names within the page. A higher value means the activity of people within the page related task is higher.

3.3 File-related EDP Identification

Each page of engineering document has a file type, e.g., text, image, correspondence, repair instruction, etc., thus an engineering document can have various file types, and each file type can have different quantities. For a project, its document-contained file types reflect the evolution of the project, and the quantity of file types can measure the project activity. In this research, the file-related information contained by engineering document is treated as file-related EDP, and it is used to identify file type distribution and generated page distribution. The detailed explanation of file type categories and the identification process are shown as follows.

3.3.1 File Type Category

The categories applied in the file-related EDP identification process include file type A, B and C (see Table 1).

Table 1. File type categories

File type A	File type B	File type C
Correspondence	Daily Repair Request	Outgoing
Text	Damage Report	Incoming
Image	Request for Information	Internal
	Information Supply	
	Technical Disposition	
	Repair Instruction	
	Repair Design Approval Sheet	

The file categories are used to label the engineering document pages. For each page, the file-related EDP identification process analyses the page content, then selects a label from each category based on the analysis result, and assigns the selected labels to the page. For example, the labels of a page can be "Correspondence + Request of Information + Incoming".

3.3.2 File Type Distribution

Assuming a given report r contains multiple file types f_n, i.e., $\{f_1, f_2, ..., f_n\} \in r$, then the file type distribution value of f_i regarding r is defined as:

$$ftd(f_i) = \frac{|f_i|}{\sum_{j=1}^{n} |f_j|} \tag{5}$$

It measures the percentage of a file type, and it is used to represent the distribution of file types under each category.

3.3.3 Generated Page Distribution

Generated page distribution is another method to measure the project activity by considering the number of document pages with the same time-stamp. A greater value means the number of generated pages on the time-stamp indicated date is higher, which also indicates the project activity on that date is higher.

3.4 Summary

Engineering Document Pattern (EDP) identification aims to identify the patterns or profiles and extract shallow knowledge from combinations or series of engineering documents. The proposed methods of EDP identification uses the time-related, people-related and file type related information from the documents to identify the project time-line, page time-stamp, percentage of date/people occurrence, and cumulative frequency of WPD/WPP. The terminologies and related descriptions are summarized in Table 2.

Table 2. Terminologies and descriptions of EDP identifying process

Method	Description
Project Time-line	Representing the project life-cycle
Page Time-stamp	Indicating the created date of document page
Percentage of Date Occurrence	Identifying the critical date of a project Measuring the importance of different dates
Cumulative Frequency of WPD	Measuring the time-related project activity
Percentage of People Occurrence	Identifying the critical people of a project Measuring the importance of different people
Cumulative Frequency of WPP	Measuring the people-related project activity
File Type Distribution	Representing the project evolution
Generated Page Distribution	Measuring the file-related project activity

4. Experimental Study

To evaluate the methods, an experimental study is introduced in this section. The experimental study uses a collection of Engineering In-Service repair design reports to test the proposed EDP identification methods. In this sections, two repair cases, i.e., a standard one (project A) and a complicated one (project B), are used to demonstrate the EDP identification results.

4.1 EDP-based Project Activity

Project activity is the key indicator of measuring a project performance, and a high project activity on a date means the level of project input or output on that date is high. As mentioned before, cumulative frequency of WPD is used to identify the project activity. The visualisations of cumulative frequency of WPD with project time-line regarding project A and B are shown in Figure 1 and 2 respectively.

In practice, the processes of both projects include the requirement setting stage, information request stage, and problem solving/evaluation stage. According to Figure 1 and 2, the visualisations of both projects contain the same number of peaks that correspond to the processes of the projects. As project

A is simpler than project B, the visualisation of the former is clearly smoother than the latter. In other words, the change of project activity in project B is more frequent than project A, as project B could have complicated process or difficult execution, and each of them can affect the change of project activity, hence project A is considered more stable than project B, which matches the facts.

In general, the average project activity is inversely proportional to the length of project time-line. The project time-line of project B is longer than project A, thus the average activity of the former is most likely lower than the latter. As shown in Figure 1 and 2, the blue horizontal line in two figures represents the average project activity of each project, and the average activity of project A is clearly higher than project B.

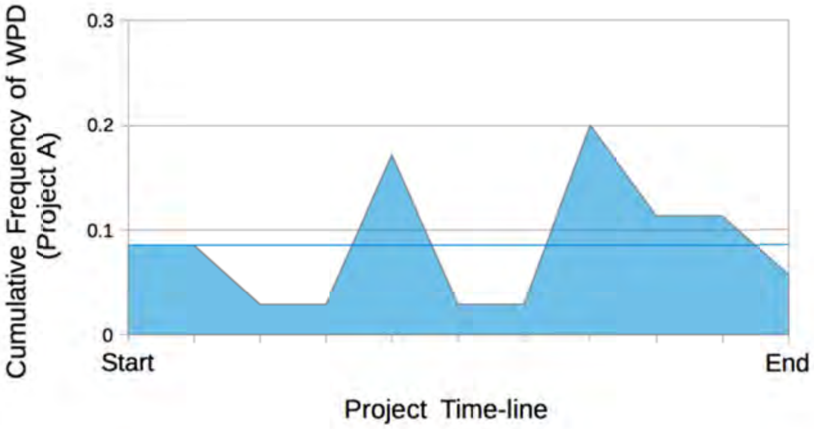

Figure 1. Project A: Cumulative frequency of WPD with project time-line

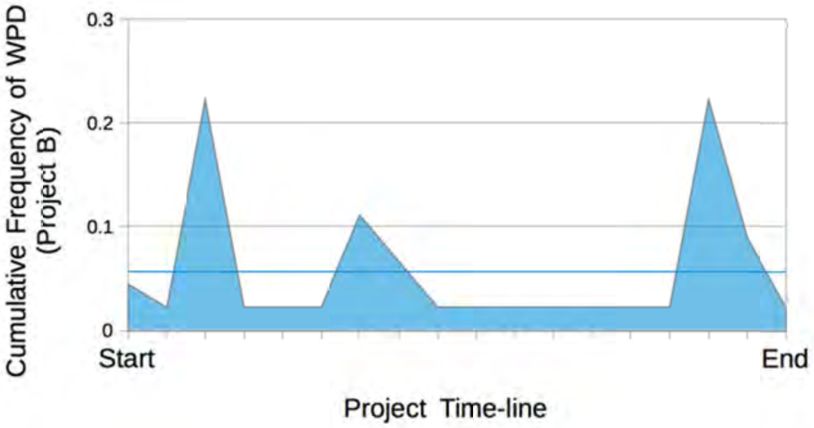

Figure 2. Project B: Cumulative frequency of WPD with project time-line

4.2 EDP-based People Activity

People activity is another measurement of project performance, and a higher people activity within a date means the task on that date requires more human resource. Cumulative frequency of WPP is used to identify the people activities. The visualisation of cumulative frequency of WPP with project time-line regarding project A and B are shown in Figure 3 and 4 respectively.

In general, the people activity is proportional to the project activity regardless the project is simple or complicated. The visualisations regarding project activity and people activity of project A have considerable similarity (see Figure 1 and 3). Although project B has different characteristics from project A, the visualisations regarding its project activity and people activity of project B are also similar (see Figure 2 and 4).

Similar to the average project activity, the average people activity is inversely proportional to the length of project time-line. As shown in Figure 3 and 4, the green horizontal line in two figures represents the average people activity of each project, and the average activity of project A is clearly higher than project B.

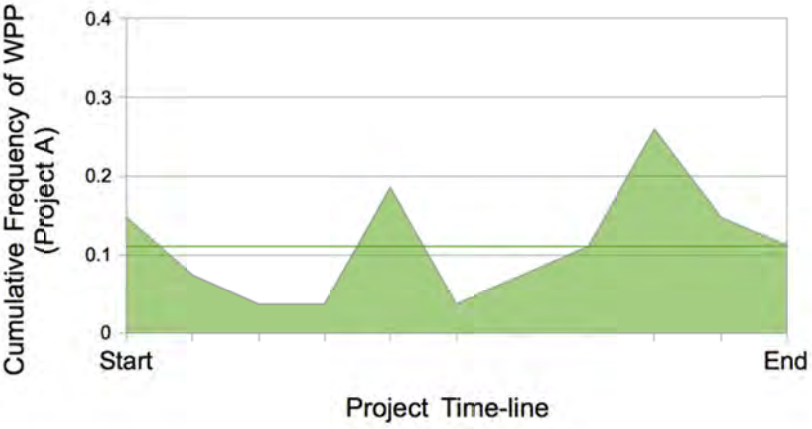

Figure 3. Project A: Cumulative frequency of WPP with project time-line

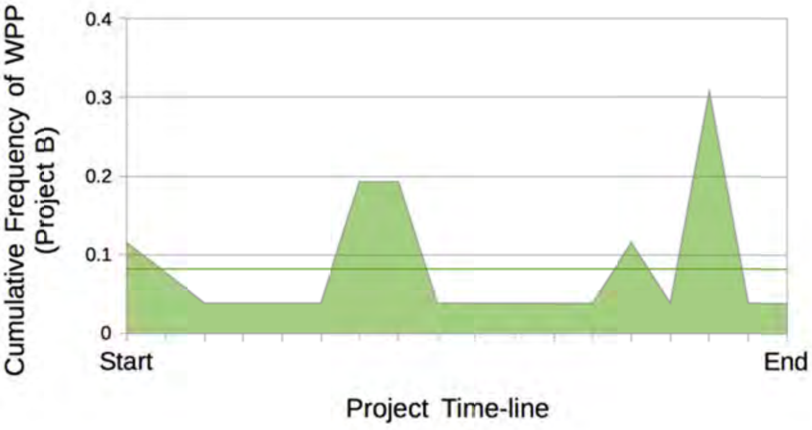

Figure 4. Project B: Cumulative frequency of WPP with project time-line

4.3 EDP-based Project Evolution

The file type of an engineering document page is determined by the page related task. For example, pages about general information requests are classed as correspondence, pages about detailed damage information supply are classed as image files. The file-related EDPs with project time-line can represent the evolution of project. File type distribution is used to identify the project evolution. The visualisations of file type distribution (file type A) with project time-line regarding project A and B are shown in Figure 5, 7 and Figure 6, 8 respectively.

As shown in Figure 5 and 6, correspondence is the first appeared file type in both projects, and the following file types are image and text. The sequence of file types shown in the visualisations reflects the actual project evolution. For example, correspondence reflects discussions of problems at the early stage; image reflects demonstrations of the detailed damage information after the discussions, text then reflects descriptions of the formal solution and summaries of the project. The visualisations also match the characteristics of project A and B, e.g., as the standard one, the visualisation of project A has a relatively simple structure.

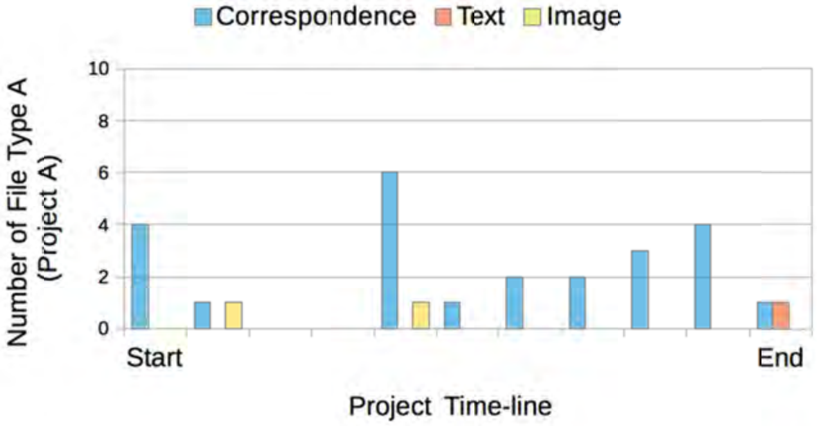

Figure 5. Project evolution of project A: Number of file type A with project time-line

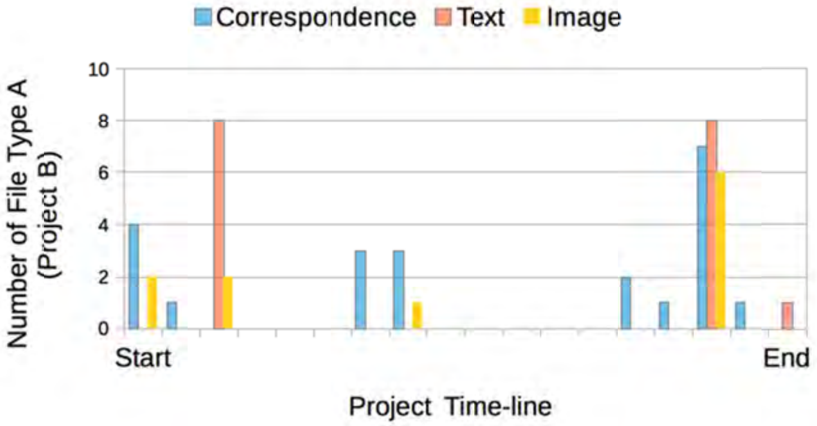

Figure 6. Project evolution of project B: Number of file type A with project time-line

As shown in Figure 7 and 8, the cumulative file changes of project A is less than project B. The reason is that the former had more specific project goal, clearer problem description and fewer requirements than the latter.

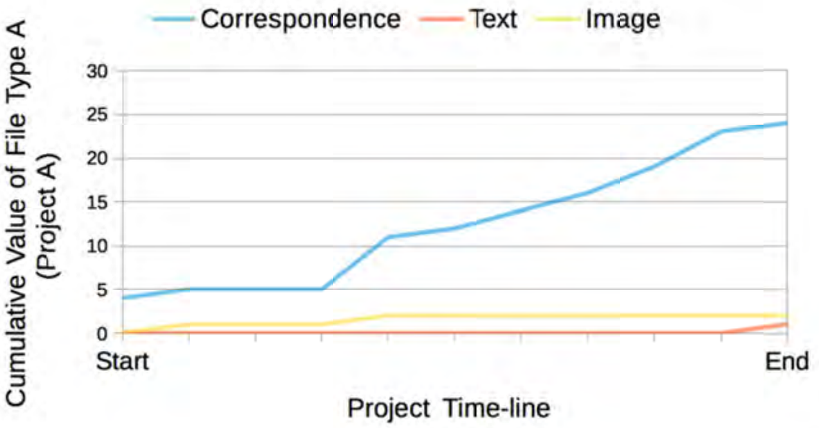

Figure 7. Project evolution of project A: Cumulative value of file type A with project time-line

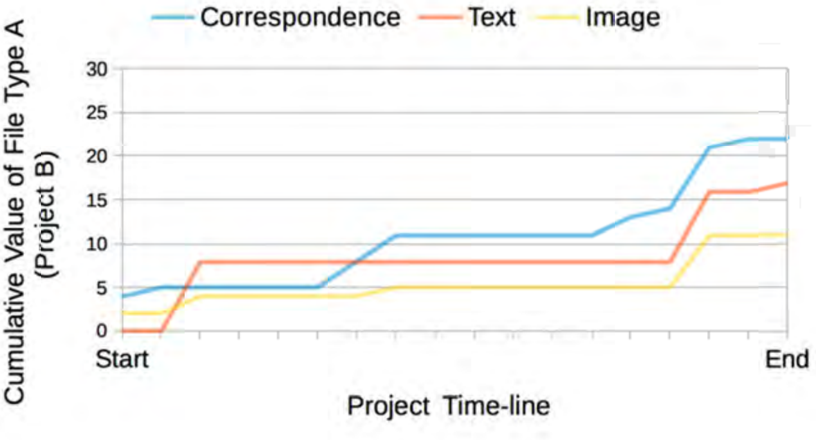

Figure 8. Project evolution of project B: Cumulative value of file type A with project time-line

5. Conclusions

To facilitate the project design, reusing the knowledge from previously completed projects is essential. The paper introduces the definition of the engineering document pattern (EDP), which is the shallow knowledge contained by the engineering documents of previous projects. Furthermore, the methods of EDP identification are also introduced. The methods use the time, people and file type related information to identify and extract related EDPs from engineering documents. The EDPs are then used to measure the project activity, people activity and project evolution. The experimental study of the methods is based on industrial data, and the result shows the EDPs and their visualisations can accurately represent the actual characteristics of the projects. This is illustrated in the paper by two quite different profiles of a simple and complex case. It is clear that this profiling and generation of project patterns has considerable potential to assess when difficulties or problems may be occurring. Further work is being undertaken to increase the level of analysis and will include extending the covered aspects of EDP identification and testing the proposed methods using large-scale data from different industry sectors.

Acknowledgements

The research reported in this paper is funded by Engineering and Physical Sciences Research Council (EP/K014196/1). The authors would like to thank the industrial collaborators and their engineers for their input and support on this project.

References

Baccarini, D., "The Concept of Project Complexity - a Review", International Journal of Project Management 14(4), 1996, pp. 201-204.

Baxter, D., Gao, J., Case, K., Harding, J., Young, B., Cochrane, S., Dani, S., "A Framework to Integrate Design Knowledge Reuse and Requirements Management in Engineering Design", Robotics and Computer-Integrated Manufacturing 24(4), 2008, pp. 585-593.

Berkun, S., "The Art of Project Management", O'Reilly Media Inc, 2005.

Carey, E., Culley, S., Weber, F., "Establishing Key Elements for Handling in-Service Information and Knowledge", In: ICED 13 - 19th International Conference on Engineering Design. Seoul, Korea, 2013, pp. 11-20.

Chandrasegaran, S. K., Ramani, K., Sriram, R. D., Horváth, I., Bernard, A., Harik, R. F., Gao, W., "The Evolution, Challenges, and Future of Knowledge Representation in Product Design Systems", Computer-Aided Design 45(2), 2013, pp. 204-228.

Chapman, C., Ward, S., "Managing Project Risk and Uncertainty: A Constructively Simple Approach to Decision Making", John Wiley & Sons, 2007.

Disterer, G., "Management of Project Knowledge and Experiences", Journal of Knowledge Management 6(5), 2002, pp. 512-520.

Dvir, D., Raz, T., Shenhar, A. J., "An Empirical Analysis of the Relationship between Project Planning and Project Success", International Journal of Project Management 21(2), 2003, pp. 89-95.

Engwall, M., "No Project Is an Island: Linking Projects to History and Context", Research Policy 32(5), 2003, pp. 789-808.

Ettlie, J. E., Kubarek, M., "Design Reuse in Manufacturing and Services", Journal of Product Innovation Management 25(5), 2008, pp. 457-472.

Giezen, M., "Keeping It Simple? A Case Study into the Advantages and Disadvantages of Reducing Complexity in Mega Project Planning", International Journal of Project Management 30(7), 2012, pp. 781-790.

Markus, M. L., "Toward a Theory of Knowledge Reuse: Types of Knowledge Reuse Situations and Factors in Reuse Success", Journal of management information systems 18(1), 2001, pp. 57-94.

Sawyer, P., Rayson, P., Cosh, K., "Shallow Knowledge as an Aid to Deep Understanding in Early Phase Requirements Engineering", Software Engineering, IEEE Transactions on 31(11), 2005, pp. 969-981.

Turner, J. R., Cochrane, R. A., "Goals-and-Methods Matrix: Coping with Projects with Ill Defined Goals and/or Methods of Achieving Them", International Journal of Project Management 11(2), 1993, pp. 93-102.

Turner, J. R., Müller, R., "On the Nature of the Project as a Temporary Organization", International Journal of Project Management 21(1), 2003, pp. 1-8.

Vidal, L.A., Marle, F., Bocquet, J.C., "Measuring Project Complexity Using the Analytic Hierarchy Process", International Journal of Project Management 29(6), 2011, pp. 718-727.

Xie, Y., Culley, S. J., Weber, F., "Applying Context to Organize Unstructured Information in Aerospace Industry", In: ICED 11 - 18th International Conference on Engineering Design. Lyngby/Copenhagen, Denmark, 2011, pp. 424-435.

Dr Lei Shi
University of Bath, Department of Mechanical Engineering
Bath, BA2 7AY, United Kingdom
Email: L.Shi@bath.ac.uk