Adaptive Production Management for Small-Lot Enterprise

Daria Kazanskaia[1(✉)], Yaroslav Shepilov[1], and Bjorn Madsen[2]

[1] SEC "Smart Solutions", 1201-17 Moskovskoye Shosse, Samara, Russia
{kazanskaya,shepilov}@smartsolutions-123.ru
[2] Multi-Agent Technology Ltd., 3 Ashbourne Close, London, UK
bm@multiagenttechnology.com

Abstract. Currently the methodology of eliminating the negative effects of the issues in ramp-up stage mostly involves the increase in investment and updating the design data. In the paper the authors consider an approach that can be applied on every level of ramp-up production: from suppliers to shopfloor operators. The architecture of the system is described and the first implementation results are given.

Keywords: Adaptive planning · Small-lot production · Ramp-up production · Multi-agent technologies · Production management

1 Introduction

The ramp-up stage is typical for the modern enterprise since new products are developed and introduced frequently to keep up with the market needs. The key challenge for management at this stage is to cope with disruptive events, whilst having to increase production volume at short notice. To put this into the context, production usually operates with period-based plans (usually, monthly or, in the best case, daily).

However, this perfect plan rarely fits the reality. In fact, the range of unexpected factors can influence its execution:

1. Suppliers failures (including non-conformities, not delivered parts and delayed deliveries);
2. Overestimated production rate;
3. Unpredictable time in decision-making;
4. Urgent additional orders.

Since the plan is not revised after it is issued to the production the lack of adjustments result in a growing backlog for products. For consecutive periods (week, month, year) the effect is accumulative. The main task for management is therefore to increase the productivity to eliminate the backlog systematically.

Despite understanding this, contemporary systems for production planning still tend to use traditional methods [1] that cannot reflect the environment that is changed almost every moment.

V. Mařík et al. (Eds.): HoloMAS 2015, LNAI 9266, pp. 157–168, 2015.
DOI: 10.1007/978-3-319-22867-9_14

The attempt to cover the typical issues of ramp-up production together with supply chain was taken in Adaptive Ramp-Up Management (ARUM) project by the FP7 of the European Commission. The approach considered in this project was described in the several papers [2,3] and is based on the combination of using the multi-agent planning to deal with the unexpected changes in the scheduler, ontology to gather and store information about the domain, intelligent service-bus to provide the interaction between the different modules.

In the first section of the paper we will describe the current production process of one of the industrial partners of the project (Iacobucci Holding Ferentino, IHF). In the second section the main production issues are highlighted. The third section describes the architecture of the ARUM system. In the fourth section we describe how the system addresses the main challenges, while chapter five provide the results of the experiments.

2 Production Process

The case study considered in the ARUM project covers production (including testing, warehouse and management), interfaces to development, finance, procurement and logistic of IHF. At the center of this study is the production area that is divided into a number of production lines for specific product types:

1. Coffee machines (CM) and espresso machines (EM), which are the most popular products. The assembly lines for these two products are interchangeable including the operators, who can apply the same skillset.
2. Trash compactors (TC), which is an expensive long-term durable product. The current demand for TC is on less than coffee machines, but the ordering profile is more volatile.
3. Induction heating units (IHU) – commonly known as ovens – represent a recently introduced product, which currently is experiencing growth in demand (ramp-up). Ovens are produced at a relatively slow rate with potential for increase in throughput through the ARUM system.

There are eight functions involved in IHF's production:

1. **Production engineering,** which provides the specifications for production and suppliers, such as assembly instructions, technical drawings of parts etc.
2. **Procurement**, which is responsible for supplier contracts and ordering of the parts, required for production the production line.
3. **Customer Service**, which manages the customer contact and maintains the overview of planned and forecasted orders.
4. **Production planning**, which constructs the production schedule to which everyone else is working (from procurement to dispatch of quality certified products). Production planning interfaces with customer service to assure that customers are kept informed about progress.

5. **Warehouse incoming inspection**, which is responsible for receiving and inspecting supplies and to indicating if any parts are delivered short, missing, broken or otherwise non-conform.
6. **Warehouse pick & packing**, which picks the assembly-kits that are consumed by the production line.
7. **Production**, which assures the assembly according to certified processes.
8. **Quality Assurances** (QA), which test all products before dispatching to the customers. This final QA interacts with the product developments quality management department, which is involved in the investigation of any non-conformity from the certified process, and feeds back into product development.

Information about orders, bill-of-materials and inventory is stored and processed in an AS400-database, which was developed in-house. All other information is managed in office documents (PDF, Excel).

The ARUM system influences the order-to-delivery process, whereby it is essential for the reader to understand the sequence of activities where the ARUM system can contribute to improvement of productivity during ramp-up. The process is illustrated below (Fig. 1):

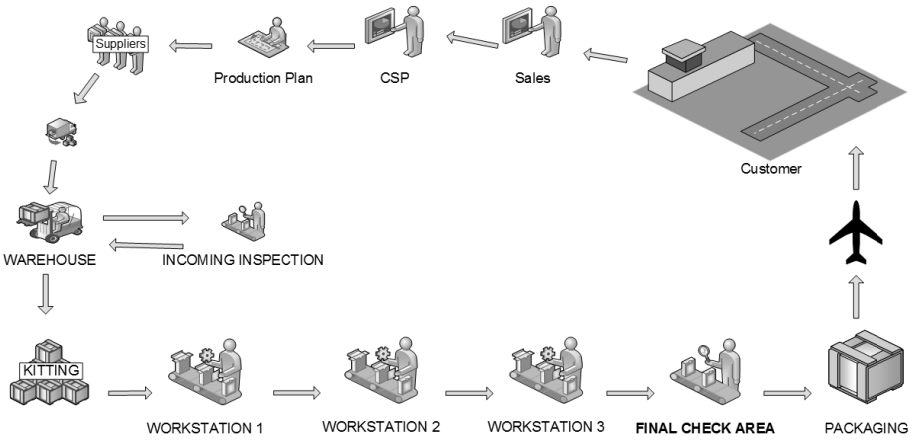

Fig. 1. Order-to-delivery process in IHF

The processing of a new order starts when customer service receives a customer order with a requested date of delivery. Orders vary in quantity, frequency and content/choice of products. As a guideline, orders are given 90 days in advance to ensure receipt of parts from suppliers and shipping to the customer. Some customers provide long-term orders to establish a periodic (re-)delivery of the products during the year, whilst others come with notices as short as 45 days. A key in prioritization criteria is whether the order is for a brand-new aircraft or as replacement for older ones, as delay of delivery to brand-new aircrafts delays the delivery of the aircraft as a whole, and therefore is unacceptable. In either case customer service is the key contact point

for interpretation of the commercial priorities and evaluation of the consequence of any changes given by account managers at short notice.

The ability to fulfill the order on time is verified with other departments (checking capacity, production capability, required supplies etc.) which finally is approved by CEO and released into AS400 database as confirmed demand which need to be incorporated into the production plans. The production is coordinated at the highest level with reference to a master plan, which uses the data from the AS400. Based on the master plan, the production planning department provides the detailed production plan which clarify which products are to be produced by the end of the month at the level of serial numbers including any units made to stock. The detailed plan is then aggregated into work orders, which reflect the number of products that a single assembly line can produce per week. Each work order is then assigned to the assembly lines according to the types of product types, which the line is certified for. The release of the work order to all departments is a used as a trigger for authorization of staffing, purchase of parts by procurement, kitting for the production line by the warehouse, etc.

The material flow is logical: Procured materials are received, inspected, stored, picked to kits, consumed sequentially on the assembly line, tested, labeled for traceability, packed on pallets and shipped to the customer.

The production planning department provides a weekly report on the progress, which is tied into the regular management meetings, though daily trouble-shooting is done directly between the departments

3 Main Challenges

From the analysis of IHF processes and interviews with employees, a wide range of issues were identified which inhibit the ramp-up process from being efficient. From planning through production to delivery of the goods to the customer, the following disruptive events are of key importance (following the order-to-deliver process):

1. Sales and customer service:
   (a) Extra demand: An example is a sudden request of a major airline for the delivery of about 100 TC within four months (at a planned capacity of about 140 units per year). When the a large amount of products is to be produced in short time, two issues have to be solved:
      (i) Resource reallocation. The demand for one type of products may require moving the operators from the lines that assemble other products. If the resources are still not enough, additional capacity can be gained by involving office personnel that has the required certification or staff from EASA 45 line that mostly operates the maintenance of the units supervised by EASA authorities.
      (ii) Keeping the delivery dates. It is obvious that the delivery dates for the other orders must be kept as much as possible. However, if there is no chance to prevent the violation of the due date, the orders should be planned to reduce the penalty. New delivery times should be communicated and negotiated with the customers.

(b) Contract problems (supplier & customers): Prices are based on annual quantities which allow the supplier to operate effectively, but volumes ordered by procurement are not divisible in batch-sizes that are viable for the supplier to deliver. This causes shortage or over delivery.
(c) Updates to orders: Change in delivery dates or required amounts, cancellation of orders trigger the changes that will result in completely new delivery schedule for the current period. That results in the problems with the supplies and affects the delivery dates of other products.
(d) Updates to forecast: When forecast is wrong, there is request from customer to provide additional number of products. Usually the company can handle small amounts (2-3 products); however, these additional orders should be approved by warehouse, procurement and production.

2. Production planning: Production planning has to deal with the daily updates from production and sales. All the data is collected manually, usually in talks and phone calls. Then the plan has to be manually updated in Excel sheets.
3. Procurement:
(a) Delivery delays. Though the orders to the suppliers are communicated year in advance, the suppliers have issues on their side that result in violation of agreed supply dates.
(b) Quality flaws. The parts received from the supplier may be different from the required design because of production flows or inconsistent design data given to the supplier. This results in insufficient stock.
4. Incoming inspection: Materials do not reflect drawings and instructions provided by design engineering. Incoming inspection is a potential bottleneck, since there is no way to learn that the material is delivered/not delivered or if it is conform until if passes the incoming inspection. Therefore, any major issues that require involving the incoming inspection personnel may result in delay in delivery the materials to the production.
5. Warehouse: The warehouse employees discover that there is a lack of certain part only when they start preparing the assembly kits.
6. Production line:
(a) Production capability: in the case of the IHF primarily supply problems are to be expected. Nevertheless, the very cost- as well as quality-effective technology may kindle the demand faster than currently planned. Typical ramp-up problem: incoming inspection has checked the part against the drawing (usually used in hardcopy on-site, available in electronic form in the shared folder) while they were changed or updated (electronic form), so they and the part are not correct anymore. This must create the task that the stock that was inspected under the old inspection instructions are re-inspected under the new instructions.
(b) Defects: Instruction on assembly line is different from physical materials. The defect may result in that a certain part of stock becomes unavailable. Moreover, sometimes the decision regarding the defect resolution requires the coordination of several departments.

(c) Incomplete assembly kits. Sometimes the assembly kits arrive in the production lacking several parts that means that certain subassemblies cannot be completed. If the required supplies will not arrive before the subassembly starts, the management will need to solve this problem.

4 ARUM System

4.1 System Architecture

The architecture of the ARUM system is designed within the context of ramp-up systems for manufacturing, where there are conflicts between the need for control and rigor and the reality of rapid changes. Ramp-up systems often require end-to-end integration from strategies, systems to tools (i.e. at all control and optimization levels). Further, vertical integration is required from MES down to shop floor and horizontal integration from engineering to production system planning to steady state production processes.

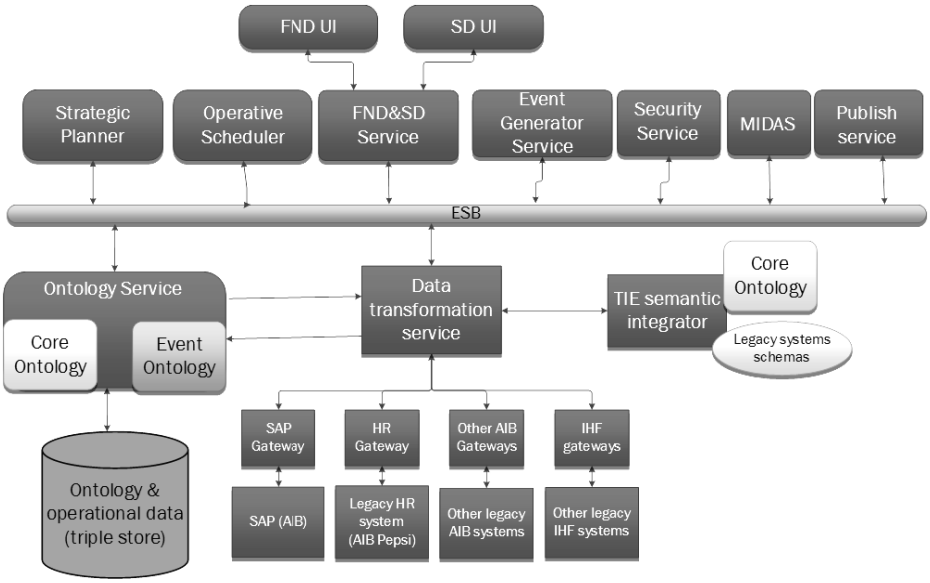

Fig. 2. ARUM system architecture (image courtesy: Cesar Marin, Vadim Chepegin)

The address the above challenges and industrial requirements the ARUM architecture integrates the key features of service-oriented-architecture, holonic multi-agent systems and legacy systems and links them via an enterprise service bus (ESB), providing communication, monitoring, interoperability and aggregation of information across existing legacy systems at all production levels to support real-time automatic negotiation, planning, planning and optimization within and across factories. The envisaged technologies of distributed multi agent system within a holonic architecture

is expected to help in integrating legacy systems, information aggregation from high level systems (MES, ERP, etc.) to factory floor automation (e.g. metal cut and assembly systems). Based on such requirements reflected in the system architecture, the main functionalities of ARUM system include planning, planning, production management and manufacturing process supported by actual information delivered from a variety of sources such as legacy systems, sensors, and user inputs.

Based on the user and domain specific requirements as well as on the results of previous research and implementation [4,5] the logical view of ARUM system not only identifies the key components and services that make up the ARUM system but also demonstrates the interrelations between them, as depicted in Figure 2.

This architecture was considered in more details in [6]. In the paper we will highlight only the key elements of the architecture to support the case of IHF that include:

- Gateways to extract data from the legacy systems.
- Ontology to describe the domain to the system.
- Multi-agent adaptive scheduler to create the plans for the production.
- User interface for production and planning managers to create schedule and ensure its execution.
- User interface for the shopfloor operators to receive the tasks according to the schedule, report their completion and discovered problems.

4.2 Method of Planning

The core of the system is the world of agents consisting of multiple agents that interact with each other by exchanging messages notifying on the certain events the agent is subscribed to.

The message exchange is implemented with the special mechanism called message whiteboard. The whiteboard itself is a high-level agent that coordinates the tasks between the agent that requires the resources (i.e. demand agents) and the agents that can provide the resources (i.e. resource agents). The agents can indicate their demands and resources by making the matching rules. Each agent sends its rule with the demands and resources to the whiteboard. According to these rules, the whiteboard selects the most optimal matches.

With this mechanism, the schedule is not created from the scratch every time, but adjusted according to the events in real time. There adjustments are the result of conflicts, negotiations and compromises between the agents.

The current version of the system implements multi-threading planning mechanism with the ability of the agents to process their messages in parallel in different CPU threads. After an agent processes its message, the thread becomes free and it is ready to receive another agent for operation (or the same agent with a new message). Processing of different messages can take different time, however, it is does not result in additional delay since the agents work asynchronously and occupy different processor threads.

There are four types of agents in the implemented multithread planning system that comply with the demand-resource classification described in [1]:

The *employee agent* (resource) represents an employee that can perform a certain type (or types) of jobs, has specific skills, can use the equipment and is ready to perform any relevant job.

The *equipment agent* (resource) represents a unit of equipment that has the specific model that can be used by the employee to perform specific type of jobs.

The *workshop agent* (resource) is looking for jobs and services from other workshop agents to perform them in their facilities.

The *job agent* (demand) is representing a technological operation that is looking for its allocation in the schedule according to the given criteria (employee, equipment). To satisfy the requirements of the job agent, the employee agent must be able to perform the job of this type, have required skills and be able to work on the specific equipment model required for the job.

The system calculates the schedule in the real-time mode, when the agents go through the cycle of initialization, interaction and achieving the results:

1. *Start.* At this stage the world of agents is created. In the world there exist and interact the instances of applications implementing the required agents functions including the basic agent interaction mechanisms.
2. *Data load.* The loading of the initial and updated data for the system operation;
3. *Creation of agents.* At the first stage the world of agents sends the creation and activation messages to all agents that were created ("wake-up" message);
4. *Agents initialization.* The agents define their goals, priorities, criteria according to the data they received from the agents world. Each agent decides to which updates from the specific agents it will be subscribed. At this stage the threads are started and the parallel operation of the agents can be started;
5. *Agents validation.* Agents specify how accurate their goals, priorities and criteria are defined;
6. *Agents operation.* The agents start operating according to their instructions to achieve their goal in parallel asynchronous mode;
7. *Achieving the compromise.* The agent finds the best solution by negotiating with other agents or on its own, after that the agent operation is stopped;
8. *Saving the results.* Solution achieved by the agents is saved;
9. *Receiving the events.* Notification on the events from the real world is received by the world of agents;
10. *Cycle repeated.* The data is uploaded or updated according to the received event (stage 2) and the cycle is repeated.

After the completion of the cycle, the agents transmit to the pending state when they do not perform any actions until they receive a specific message from other agents.

In this process, the job agent is the most active one: it reacts to the allocation request from the employee, can be initiated by the agent of the related job or just take part in the conflicts resolution. In order to be allocated to the specific slot in the schedule, the agent must satisfy all criteria. An employee agent must be relevant to the given parameters to satisfy the demands of the job agent. This can result in long interactions between the agents, that are avoided by using the message whiteboard described above. A job agent leaves the required demands in the rules while the employee agent leaves the resources that it can provide in the rules. Then the whiteboard agent analyses the rules and informs the agent on the matches found.

Negotiations take considerably long time since there is a huge number of job agents that want to be allocated to the best slot in the schedule, while at the same time many employee agents of and equipment can match many jobs. The number of agents considered during the allocation can be decreased by several criteria, e.g., priority, availability and response time.

The main planning process is done during the stage of achieving the compromise. The agent finds the best allocation option by the negotiations with other agents or by its own means. Then the agent activity is stopped and checked of the event planning accuracy starts. This check consists of correct event processing and the schedule consistency checks. The event processing check is required to ensure that all changes triggered by the planning were effected (for example, the fired employee has no operations in his schedule or a new order is completely scheduled). Only after the check for plan accuracy, the schedule is stored and available to the user.

The mandatory condition of agents operation is the existence of the world of agents. The world is considered as active if at least one agent is active. During its operation, the world runs the parallel operation of the agents by running the CPU threads. All CPU threads can be run simultaneously and work in parallel. For example, if the CPU has eight cores, a maximum of eight agents can process their messages at the same time. After the message processing is completed, the thread is disengaged and will be occupied by the agent that the scheduler chooses to activate. The threads can be free during a certain time, but a thread can be occupied only by one agent at a time.

5 Application of ARUM System

As it was shown in the above chapters, ARUM system covers most of the aspects of factory operation due to its structure and architecture. To provide the reader a clear picture, let us go through the problems that were highlighted above and describe how the system addresses them.

In the case of extra demand received by the customer service the solution is provided by the coordinated operation of strategic planner and operational scheduler. The strategic planner allows the managers to investigate different possible solutions (extra lines, change in workshop layout, etc.) and select the best one in terms of profit. Operational scheduler ensures that the resources will be allocated in the most efficient way (to cut the costs and keep the deadlines) within the set-up provided by the managers.

Contract problems with supplier and customers are solved by the operational scheduler considers not only the production process, but also the inventory profile of the required stock. Therefore, the management can put the orders for supplies according to the needs of production, which helps to eliminate shortage and over-stocking.

Any change in the orders or forecast reflected in one of the legacy systems (either Excel or AS400) is immediately processed by the operational scheduler that updates the current plan for production.

The issues of the production planning are resolved by the automatic updates to the plan done by the operational scheduler will cut the time for communication between the production planning and other departments. Instead of updating the numerous tables, the planning manager can focus on providing the required KPIs values by adjusting the planning properties.

If the required stock was not provided by the supplier in time or in insufficient quantity, the operational scheduler will indicate the problem and will reallocate the resources correspondingly.

The operational scheduler can provide the actual order priorities to the incoming inspection, therefore, the staff will know what parts should be processed first. The incoming inspection operators can be scheduled as production ones, while the two departments and their schedulers can communicate via p2p network. Moreover, when the problem is discovered in the warehouse, the operational scheduler can reallocate the resources.

For the production line, the operator tablets with the installed operator UI ensure that all staff members have up-to-date engineering data that is updated automatically when the new product is assigned to the line. The time for line refurbishment is cut making production more flexible. The operator UI also helps to report the problem without any paperwork. The report can be later received by the managers and be an input in the process of problem resolution while the scheduler reallocates the resources to prevent idle time. In case of incomplete assembly kits the operational scheduler can allocate the operations of the current batch until the materials in the kit allow it. Then the operations from the next batch will be allocated to prevent idle time.

Furthermore, the ARUM system can provide the support in applying the lean principles by highlighting the bottlenecks and reacting to the events and the information received. The system also reduces the time required for communication between the departments and amount of the corresponding paperwork by providing the user interfaces by all roles relevant for the process.

6 Results

In the paper we will investigate the influence of the ARUM system on the production process of IHF in the following set of experiments:

1. The basic case. Describes IHF performance based on the data provided for year 2013.
2. The perfect case. We assume that all orders are known in advance and plan them in the most efficient manner.
3. The realistic case. The orders are received according to the 2013 data. They are planned in the efficient manner.

Considering the perfect case as an ultimate example, we will use its KPIs values to measure the other two cases.

Let us consider in more details the measures presented in the table. The productivity is calculated as following:

$$N = \frac{Q_{out}}{Q_{in}} = \frac{Q_{output}}{Q_{pt} + Q_{empl}},$$

where Q_{output} is the units output in euro, Q_{pt} is the input for part in euro, Q_{empl} is the input for employees in euro.

The delays are calculated as following:

$$D = T_{actual} - T_{contract},$$

where T_{actual} is the actual date of delivery, $T_{contract}$ is contract delivery date.

The utilization is calculated as following:

$$U = \frac{\sum_{i,k} j_{i,k}}{N_r \cdot (t_2 - t_1)},$$

where $j_{i,k}$ is a duration of the specific job, N_r is the number of the resources, t_1, t_2 are the start and the end of the considered time interval.

The assumptions on the resources based on the skill matrices and data provided by IHF representatives were considered in the evaluation presented in Table 1.

Table 1. Data *used in the tests*

Product	Number of lines	Operators per line	Total items per 2013	Order production, man-hours	Operator cost per hour, €	Unit parts cost, €
CM	4	2	768	20	6	679,5
TC	1	4	200	35,68	6	4737,8
IHU	1	2	12	61,07	6	5615

During the experiments, for each of three cases described above, the schedule for the period of one year was calculated. The results are given in Table 2.

Table 2. The *results of the experiments*

Scenario	Productivity	Delays, day			Utilization, %
		Min	Max	Average	
The basic case	1,45	0	434	32,5	99
The perfect case	1,47	0	287	0	70
The realistic case	1,47	0	363	0	67

The slight increase in productivity in the perfect and realistic case is achieved by reducing the penalties to be paid for the delayed orders. This indicator can be increased by taking the extra orders (in comparison with 2013 data).

Resource utilization is reduced in perfect and realistic cases by more efficient planning. That means that new orders can be taken to achieve the full workload. However, the company may would like to maintain the same customers demand, but reduce or reallocate the resources instead. Another possibility for the efficient use of capacity is taking the outsource orders.

Again, the efficient planning resulted in reducing the delays in order delivery thus reducing the penalties to be paid to the customers.

7 Conclusion

The results of the experiment shows that the improved coordination in planning can lead to reducing the delays in order delivery and free capacity. That means that despite the potential impact of the disruptive events, the company can take extra orders or eliminate the backlogs from the previous years. Therefore, the application of the ARUM system provides the possibility to increase the company profit with the same number of resources

Moreover, the experiments have proved that coordination with customers plays significant role, since the performance of the company depends not only on production, but also on the dates when the orders for supplies were placed. This opens the wide field for the further experiments and investigation.

Acknowledgment. The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007- 2013 under grant agreement n° 314056.

References

1. Skobelev, P.: Multi-agent systems for real time adaptive resource management. In: Industrial Agents: Emerging Applications of Software Agents in Industry (Invited Chapter). Elsevier (2014) (in publishing)
2. Leitão, P., Barbosa, J., Vrba, P., Skobelev, P., Tsarev, A., Kazanskaia, D.: Multi-agent system approach for the strategic planning in ramp-up production of small lots. In: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC 2013), October 13-16, 2013, Manchester, UK, pp. 4743–4748 (2013)
3. Tsarev, A., Kazanskaia, D., Skobelev, P., Kozhevnikov, S., Larukhin, V., Shepilov, Y.: Knowledge-driven adaptive production management based on real-time user feedback and ontology updates. In: Proceedings of the EEE International Conference on Systems, Man, and Cybernetics (IEEE SMC 2013), October 13-16, 2013, Manchester, UK, pp. 4755–4759 (2013)
4. De Bra, P., Aroyo, L., Chepegin, V.: The next big thing: Adaptive web-based systems. Journal of Digital Information **5**(1) (2006)
5. Chepegin, V., Perales, F., de la Maza, S.: CREATE Software architecture (2012). https://itea3.org/project/workpackage/document/download/862/10020-CREATE-WP-2-D21Architecture.pdf
6. Marin, C., Moench, L., Leitao, P., Vrba, P., Kazanskaia, D., Chepegin, V., Liu, L., Mehandjiev, N.: A conceptual architecture based on intelligent services for manufacturing support systems. In: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC 2013), October 13-16, 2013, Manchester, UK, pp. 4749–4754 (2013)