Validation protocol

Table of Contents

System implementation

The proposed solutions were implemented within a prototype system for experimentation. This prototype consists of two subsystems: (1) a cloud server for maintenance ontologies storage and (2) a HoloLens-based AR application. Figure 1 presents the languages and platforms utilised to code each subsystem. The cloud server storage uses the graphical database Neo4j [1] to store maintenance ontologies and Cypher [2] and neosemantics [3] to support data transfer through OWL and RDFS. Besides, the server incorporates a web-based application coded in EJS [4] for maintenance experts to input maintenance data, which has already been described in [ref]. And also a service provider built in NodeJS [5] to transfer ontology data (e.g. classes, individuals, etc.) and related files (e.g. obj, png, etc.) using HTTP requests and JSON objects to the AR application. The HoloLens-based AR application has been coded and deployed using Unity Game Engine [6] and Visual Studio [7]. The programmable content and pattern-matching algorithm have been coded using C# [8]. The interaction through HoloLens has been enable with MixedRealityToolkit [9]. Besides, Vuforia [10] has been used to enable registration and tracking capabilities in the AR application and has been coded to use the same JSON-based API to transfer from the cloud server necessary ontology-related files like Vuforia's model targets.

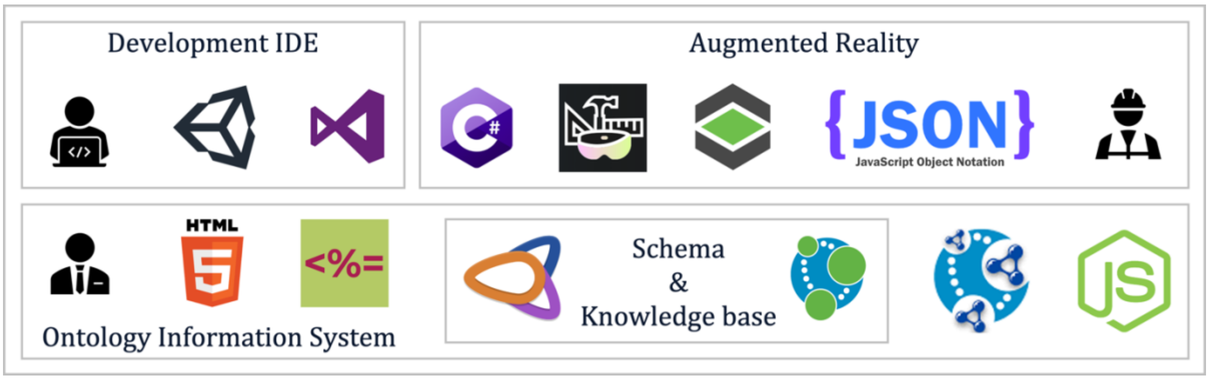

Figure 1. Description of the automatic authoring proposal's implementation as software system.

Experiment design

This paper proposes a real-time, ontology-based, pattern-matching technique for automatic adaptive authoring in multiple maintenance operations. Previous sections explained the methods utilised to separate authoring's content creation and information management processes and to automate the former. Hence, this research's validation should aim to evaluate produced content adaptiveness to multiple maintenance operations.

In academia, usual approaches to evaluate content adaptiveness to a maintenance operation are comparisons of efficiency (**time** and **errors**) [11,12] and **usability** [13,14] effects of different AR and non-AR solutions. In this research, it is also necessary to evaluate such effects on multiple maintenance operations. For this reason, validation methods should compare this research's proposal against alternative authoring and non-AR solutions in different maintenance operations.

The authors identified two already-published, alternative authoring solutions available for experimentation. One is called ARAUM [15] and focuses on off-line context-aware authoring for repair operations. The other one called SMAARRC [16] describes rule-based authoring for remote diagnosis. In order to validate the proposed authoring method (PMAU) against these two, this research considers the following hypotheses:

- Completion errors do not vary significantly among authoring and no-AR solutions for each maintenance operation.
- Completion time decreases with authoring solutions compared to no-AR solutions for each maintenance operation.
- Completion time does not vary significantly among authoring solutions for each maintenance operation.
- Content usability does not vary significantly among authoring solutions for each maintenance operation.

For the abovementioned measures to be appropriate for evaluating these hypotheses, the following assumptions must hold true:

- Time and errors can be a direct representation of efficiency if a consistent quality is assumed at the experimented maintenance operations. In order to ensure so, this validation assumes pre-determined operations whose quality does not depend on the tester's performance.
- Usability of augmented content can affect maintenance efficiency if content is not compatible with maintenance environment or manual operations. Hence, it is necessary to evaluate testers' perceived usability to evaluate maintenance operations' quality.

The authors employed two different research methods to evaluate these hypotheses' validity according to the quantitative and qualitative measures described above. These are stopwatch time and errors studies and usability surveys, and they are described in the following subsections.

Stopwatch time and errors studies

Stopwatch time and errors studies aim to analyse the proposed authoring **solution** (PMAU) effect over maintenance efficiency on different **operations** compared to alternative solutions (ARAUM, SMAARRC, NOAR). It is assumed that AR-improved semantic understanding of real-world objects increases efficiency of maintenance tasks [17]. In such scenarios, it can be said that efficiency solely depends on time for similar levels of effectiveness (quality).

Time can be described by the number of seconds required by a tester to find, understand and complete a maintenance **task**. Quality, also understood as **errors**, can be defined as the number of tasks completed by a tester that deviate in form or result of what was pre-determined. Besides, semantic understanding is assumed to affect efficiency through the authoring **solution** utilised and the **step** of a maintenance **operation** being experimented.

Based on previous definitions, it can be said that if errors (quality) are invariable, then the effect of authoring solutions through semantic understanding over maintenance efficiency can be evaluated based on its effect on completion time. Such evaluation should be made over different maintenance operations to demonstrate the validity of this research contributions. If the assumptions above are correct, then it is reasonable to expect the following results:

- Errors do not vary with the use of different solutions for each maintenance operation.
- Time is reduced with the use of authoring solutions compared to non-AR solutions for each maintenance operation.
- Time does not vary significantly between authoring solutions for the same maintenance operation.

The study described above considers one response variables (**time** and **errors**), one control variables to test assumptions (**step**), and two independent factor variables (**solution** and **operation**). Table 1 defines these variables. Besides, each factor variable can have different levels, which are defined in Table 2.

Table 1. Description of response, control and factor variables for stopwatch studies.

Variable	Type	Definition
Time	Response	Time taken by a tester to identify, understand and complete a maintenance task
Errors	Response	Tasks completed with form or result deviations from its pre-defined target
Step	Control	Specific assignment to be undertaken by a tester as part of a maintenance operation
Solution	Factor	Authoring solution employed to generate augmented content support to conduct maintenance tasks
Operation	Factor	Nature of tasks being conducted which belong to a specific step in the maintenance process

Table 2. Description of factor levels for stopwatch studies.

Factor	Level	Description
Solution	PMAU	Use of this research proposal to generate AR support
	ARAUM	Use of an ad-hoc authoring solution for maintenance repair
	SMAARRC	Use of an ad-hoc authoring solution for maintenance remote diagnosis
	NOAR	Use of non-AR solutions to support maintenance operations
Operation	Repair	Maintenance tasks aiming to return equipment to its working conditions
	Diagnosis	Maintenance tasks aiming to identify the cause of an equipment's failure

These experiments aim to test the proposed authoring solution against other ad-hoc and non-AR authoring solutions in two different maintenance operations. In order to simplify the evaluation process, the tasks experimented at the ad-hoc authoring solutions researches [15,16] will be re-utilised for these experiments. These cases of study comprising different maintenance steps and equipment are presented in Cases of study.

Each experimental study, one per operation, consisted of a tester conducting the operation's steps with two different authoring solutions. Besides, results from previous researches for non-AR support will be re-utilised to use them as baseline comparators. Therefore, testers will be grouped in six different groups according to the abovementioned procedure and factors. Table 3 defines these groups.

Table 3. Description of experimental groups according to factors' levels.

	PMAU	ARAUM	SMAARRC	NOAR
Repair	A	B		C
Diagnosis	B		A	D

The reason to re-use testers on two different maintenance operations is for them to be able to compare the usability of two different authoring solutions. This is comparison is necessary because testers are assumed to have none or very little previous experience in maintenance or AR. Besides, experimental maintenance steps (Cases of study) can be considered sufficiently different for not expecting carry-over effects between experiments.

Usability surveys

Usability surveys aims to evaluate the perceived validity of the proposed authoring solution to enhance semantic understanding compared to alternative authoring methods. Usability refers to the ability of the authoring solution to deliver information appropriately to the user regarding the maintenance operation to be conducted. Besides, it is a feature perceived by users and so subject to opinion. Therefore, it is necessary to use qualitative criteria for its evaluation. Based on similar research [13,14,18], the criteria utilised in these surveys is that presented by Nielsen in his 1993 book "Usability Engineering" [19]. These usability criterions aim to evaluate different aspects of the authoring solution regarding its formats and its impact on maintenance operations. Table 4 presents these criterions and their related AR aspects.

Table 4. Description of criterions and aspects for usability evaluation.

Criterion	Aspect	Scale
Ease-to-learn	Start, Finish, Intuitiveness	Likert 1-5
Ease-to-use	Gestures, Text, Buttons, Images, Models, Holograms, Animations	Likert 1-5
Accuracy	Overlay, Shaking, Occlusion, Visualisation, Latency	Likert 1-5
Effectiveness	Efficiency, Confidence	Likert 1-5
Satisfaction	Design, Feeling, Overall	Likert 1-5

Each criterion includes a separate survey section with several statements for each aspect regarding the authoring solutions tested in experiments. Testers were asked to determine their agreement with these statements in a Likert Scale (1-5). The results collected serve to evaluate the authoring solution's usability compared to other authoring approaches. Besides, operational quality is also evaluated in terms of efficiency and confidence improvements. There are some assumptions to consider regarding these surveys:

- Errors are not evaluated in qualitative terms as they may be dependent on user expertise, which can vary for potential users of this solution.
- It is assumed that the quality is of consistent level for the stopwatch time studies if the results of the questionnaire provide a similar result to the experiments.

The protocol to collect and analyse experimental and survey data is described in Experimental protocol. The following section presents the experimental cases of study along and testing sample.

Cases of study

The cases of study comprise two maintenance operations (repair and remote diagnosis) to be experimented in two complex-engineering assets. These cases of study were already presented and discussed in the two publications [15,16] regarding the experimental alternative authoring solutions. In order to accommodate these cases of study to ontology-based information systems, the mapping procedure from Cullot et al. [20] was used. Figure 2 presents an overview of both cases of study, including equipment, resulted ontologies for PMAU application and views of alternative authoring solutions. The resultant ontologies produced to replicate the databases from previous researches can be consulted at 10.17862/cranfield.rd.12213380.

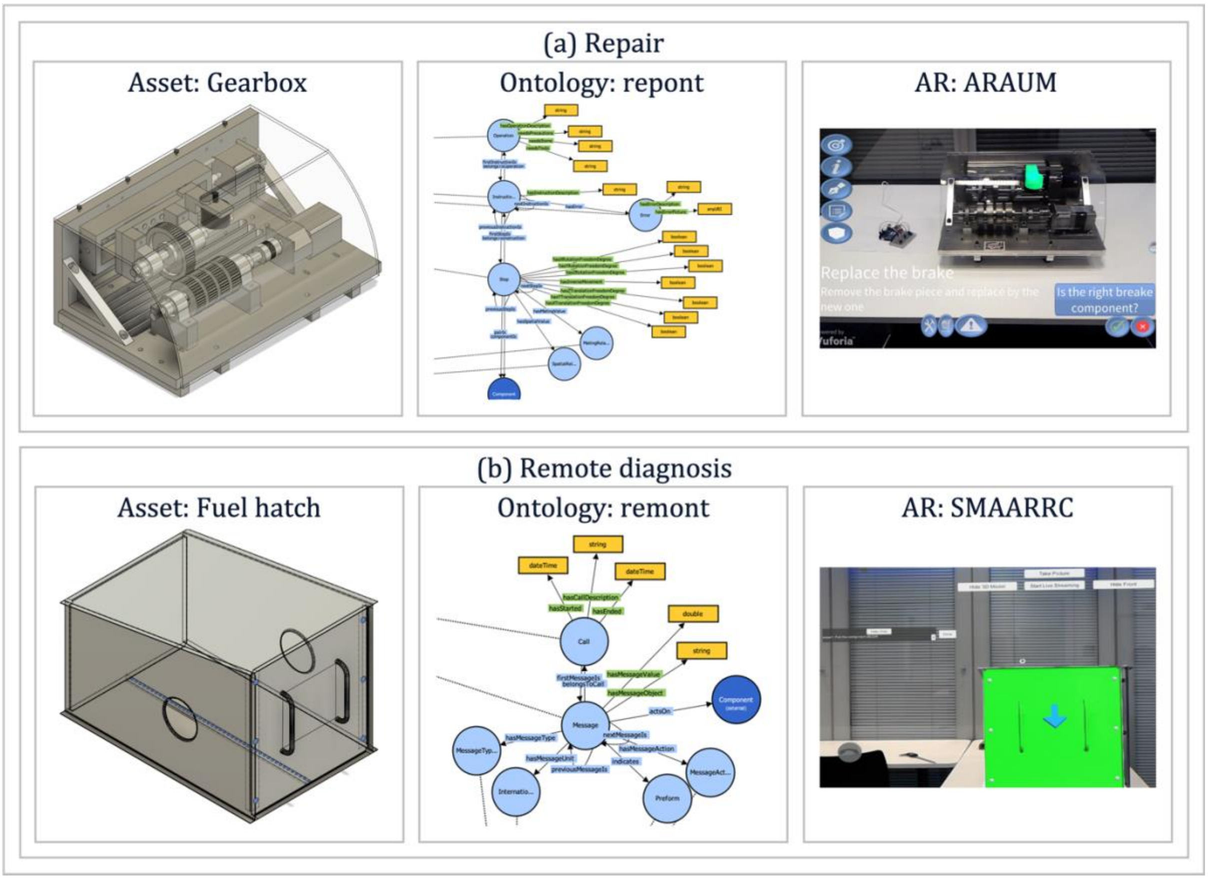

Figure 2. Description of experimental cases of study for repair [15] and remote diagnosis [16] operations including asset, ontology and AR applications.

Maintenance repair

The first case study is the same one described by Erkoyuncu et al. [15]. It represents a repair operation in complex engineering assets for the Defence Industry. These are focused mainly in mechanical, electric and hydraulic systems and assembly and replacement procedures. The case-study equipment is a laboratory prototype of a gearbox for studying gear-wheels degradation that represent real-life conditions of asset-repair scenarios. The experiments described in [15] focus on a specific repair operation composed of several assembly, disassembly and replacement steps involving mechanical components. These experiments aimed to analyse the effect of an ad-hoc tablet-based authoring solution called ARAUM, which aimed to simplify the generation of animations. The experimental repair scenario conducts an operation to replace a gearbox's component (brake wheel) that has been worn away. Figure 3 describes this repair operation’s steps using PMAU content.

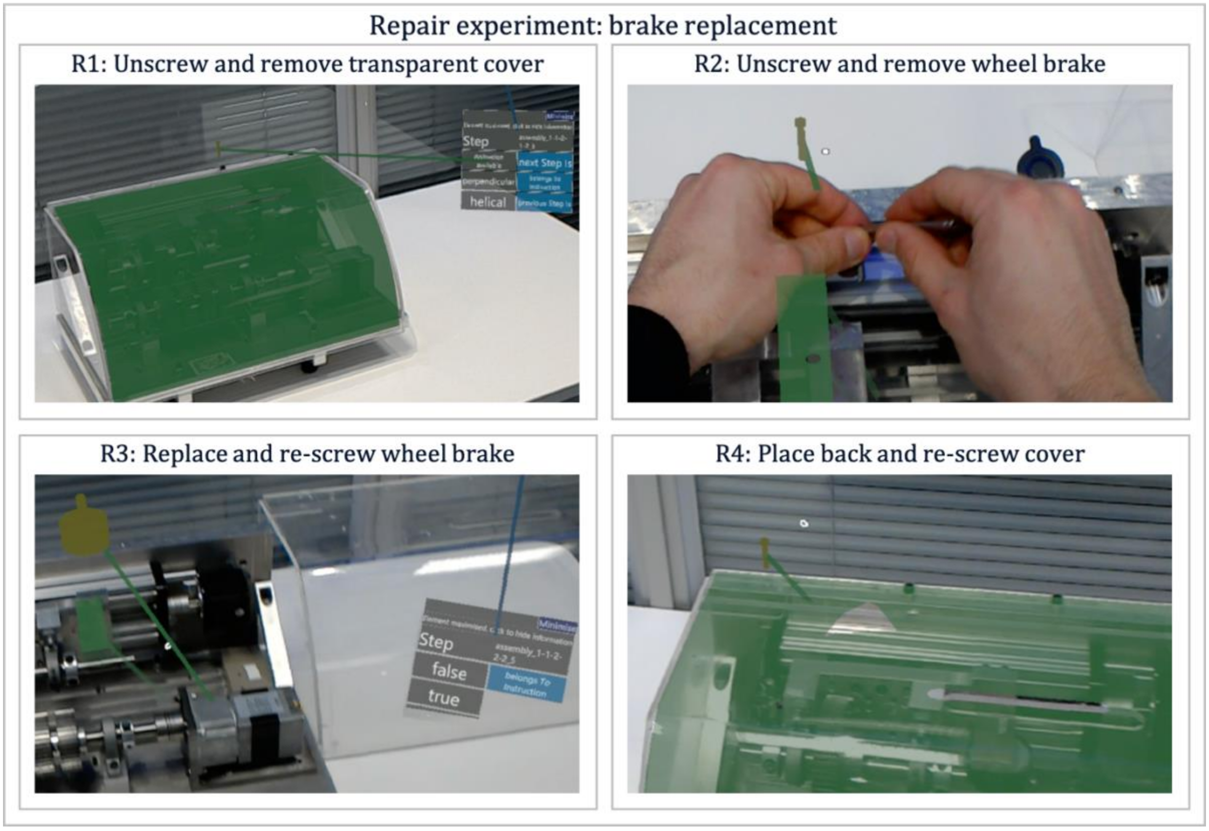
Repair experiment: brake replacement
R1: Unscrew and remove transparent cover
Step
next Step is
perpendicular
helical
R2: Unscrew and remove wheel brake
R3: Replace and re-screw wheel brake
Step
false
true
R4: Place back and re-screw cover

Figure 3. Description of repair experiment using PMAU content.

Maintenance remote diagnosis

The second case study is the same one presented by Fernández del Amo et al. [16]. It describes a remote maintenance diagnosis operation for complex engineering assets in the Aerospace Industry. The focus of these operations is purely in mechanical systems. In this case study, AR aims to develop effective communication-support tools for enhancing remote diagnosis in ‘decision-to-fly’ scenarios. The case-study equipment is an aircraft’s fuel hatch prototype with unidentified imperfections that are the diagnosis target. The experiments described in [16] focus on a diagnosis operation that comprises inspection, measurement and repair of mechanical components. These experiments aimed to analyse the effect of an ad-hoc

HoloLens-based authoring solution called SMAARRC, which aimed to simplify the understanding of complex messages. The experimental diagnosis scenario conducts an operation to identify several defects that the fuel hatch has and resolve them if necessary. Figure 4 describes this remote diagnosis operation's steps using PMAU content.

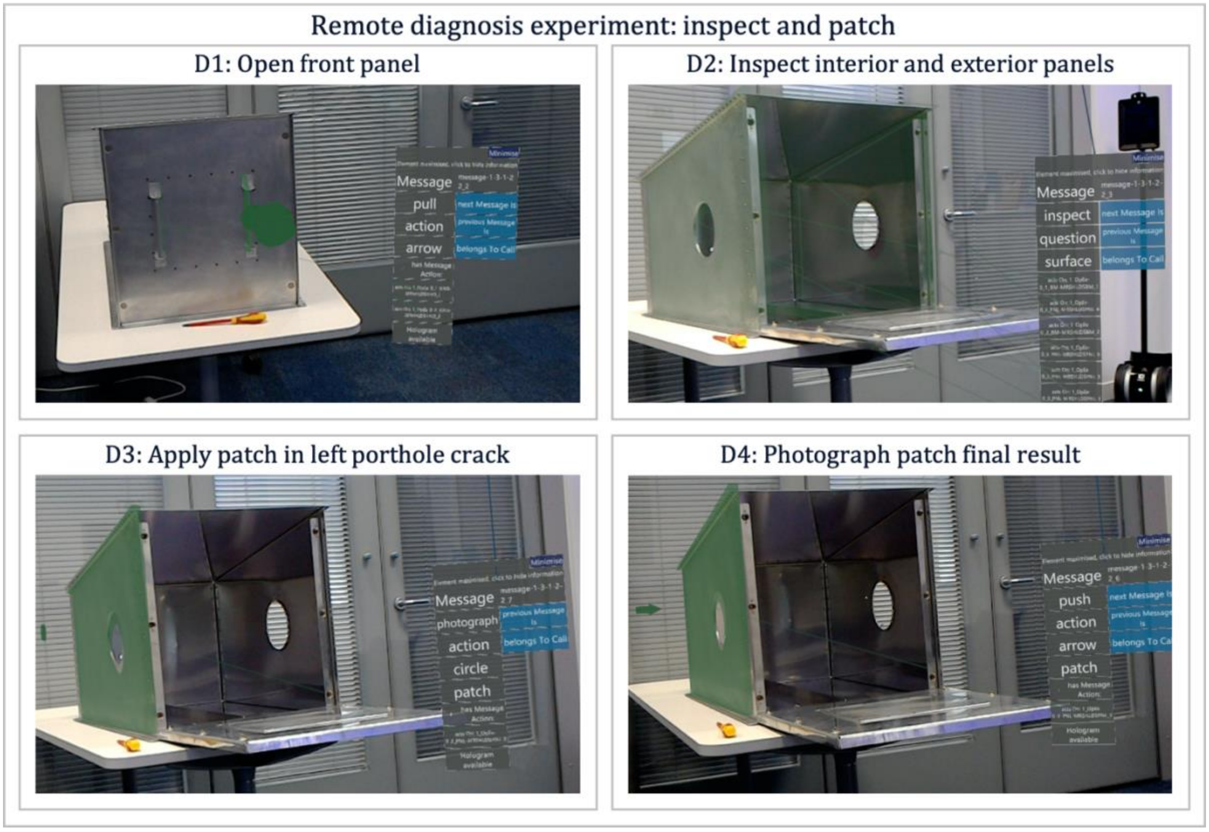
Remote diagnosis experiment: inspect and patch
D1: Open front panel
Message
pull
action
arrow
D2: Inspect interior and exterior panels
Message
inspect
question
surface
D3: Apply patch in left porthole crack
Message
photograph
action
circle
patch
D4: Photograph patch final result
Message
push
action
arrow
patch

Figure 4. Description of remote diagnosis experiment using PMAU content.

Experimental population sample

A total of 30 MSc students (24 males and 6 females) participated as testers in laboratory experiments. Their ages range from 22 to 29 years and they are all enroled in engineering-related MSc degrees. Although they have some basic knowledge in AR and maintenance due to their courses, they have no previous hands-on experience in any of them. So, they were given a short training on AR devices right before experimentation to avoid the presence of any learning curves. Testers were randomly allocated to one of the two groups (A (15) or B (15)) to avoid "carry-over" effects between maintenance procedures while using two different authoring solutions. Besides, the results from previous researches [15,16] were re-used for NOAR solution's groups (C and D).

Experimental protocol

The protocol comprises the steps to collect and analyse experimental and survey data for validating this research proposal against its expected contributions. It implements this validation methods in the case study contexts described above. The following list summarises this protocol:

1. **Data collection** (30 testers per experiment):

a. AR-maintenance introduction: to briefly describe testers the purpose of experiments as well as the use of AR solutions in maintenance operations.

b. Stopwatch time and errors experiments: to capture quantitative data on the effect on efficiency of different authoring solutions for diverse maintenance operations.

c. Usability surveys: to capture qualitative data on tester's opinions regarding usability of the authoring solution proposed compared to other alternatives used within experiments.

2. **Data analysis** (45 testers per experiment):

a. Errors effect study: to ensure the validity on the assumption that quality is kept among experiments. Results should reflect that there are no significant differences on the errors made by testers using different solutions in maintenance operations. Basic statistics, one-way ANOVA tests and graphical analysis will be used for this matter.

b. Time effect study: to analyse the correlation between the response variable (time) and considered factors (solution and operation). Results should reflect that the proposed authoring solution (PMAU) does not present significant differences on time compared to alternative authoring solutions (ARAUM and SMAARRC) in different maintenance operations. They should also reflect that these are significantly different to NOAR solutions. Experiments are set independently for each maintenance operation, and so the factors to consider in the analyses (Step and Solution). Due to the number of factors (2 -

step and solution), a two-way ANOVA analysis will be used to tests these hypotheses for each experiment. Moreover, additional post hoc (Tukey HSD) test comparisons will be used to evaluate interactions between factors' levels.

c. **Usability study:** to quantitatively evaluate testers' opinions on the proposal's content usability. Results should reflect that usability does not compromise the effectiveness of the supported maintenance operation. Due to the quantitative nature of these results, basic statistics and graphical analysis will be used for this matter.

This experimental protocol aims to validate this research's proposal against its expected contributions. For this validation to be coherent, there are few assumptions to consider:

- In order to keep consistency with previous researches [15,16] the experiments were conducted in a laboratory environment in order to keep constant other factors (e.g. ergonomics or lighting conditions) that may affect the results. This enabled to reutilise results from previous research regarding the testing of NOAR solutions for the case study operations.
- Additional effects studied in previous researches are not considered in this protocol. The aim is to prove that the new authoring method achieves similar times to alternatives, so the contributions achieved with those should also be applicable to this new authoring method.
- Experimental sample size for the abovementioned statistical tests can be estimated "a priori". Such estimation can be done using a F test for the most requiring analytical test (two-way ANOVA). With 12 factor groups (solution and step factor levels), a variance of 0.25 (partial eta squared), a type-I error of 0.1 (alpha) and a power of 0.9 (1 – beta), the resultant sample size is 51 people. That is quite close to the 45-sample size achieved: 30 testers from this research experiments and additional 15 testers results obtained from previous researches [15,16]. Besides, these numbers are bigger compared to similar researches that achieved sample sizes of 30 testers [12–14].

- As described above, testers are MSc students with none or very little experience in AR or maintenance. Although this ensures a baseline for measuring maintenance efficiency, further experiments should be required to corroborate laboratory results in real-life working conditions.

This protocol's results are discussed in the following section.

Results

This research's validation aimed to corroborate the hypotheses listed in Experiment design using the experimental protocol described in Experimental protocol. Its results are analysed and discussed in the following subsections to evaluate this research's hypothesis validity. The complete results datasets and analysis can be consulted at 10.17862/cranfield.rd.12213380.

Errors effect study

Stopwatch experiments consisted of testers completing two maintenance operations: repair and remote diagnosis. Errors are defined as the number of tasks within steps completed with form or result deviations from their pre-defined targets. Testers made use of authoring (PMAU, ARAUM and SMAARRC) and NOAR solutions to support their selves with augmented information while completing operations' steps. Hence, if the information utilised was the same although in different content formats, then errors should not differ among solutions for each experiment.

Figure 5 and Table 5 present errors results per tester and on average grouped by operation and solution factors. A conservative estimate on errors for novice testes can be taken at 0.5 errors per step (2 per experiment). So, total number of errors can be considered low with 38% of testers making 1 error and only 3% of testers making more than one. On average, average errors grouped per operation and solution range from 0.267 to 0.6. In repair, average errors with PMAU (0.467) are the lowest, while ARAUM (0.6) is higher than NOAR (0.533). In diagnosis, PMAU and NOAR are equivalent (0.333), while SMAARRC (0.267) are the lowest.

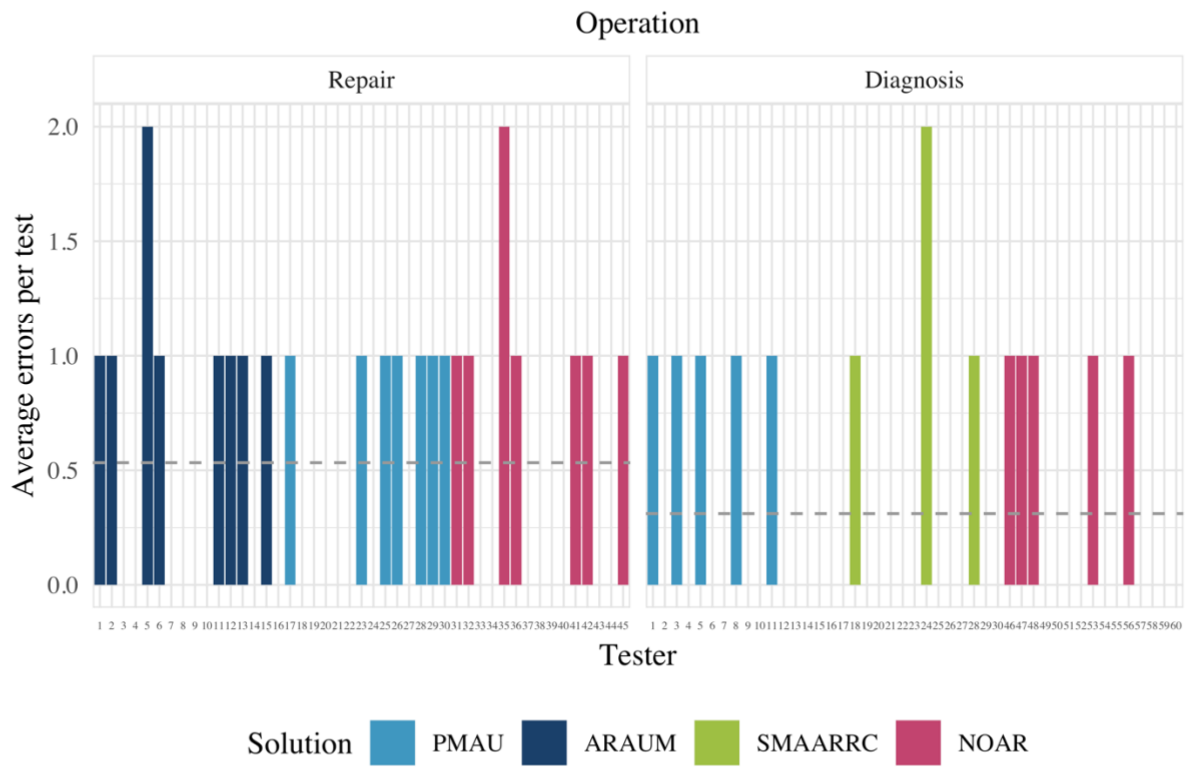

Figure 5. Average errors per experiment classified by operation and solution.

Table 5. Mean and std. deviations on completion errors per operation and solution groups.

Operation	Solution	Tester	Mean	Std. deviation
Repair	PMAU	15	0.467	0.516
	ARAUM	15	0.600	0.632
	NOAR	15	0.533	0.639
Diagnosis	PMAU	15	0.333	0.488
	SMAARRC	15	0.267	0.594
	NOAR	15	0.333	0.488

Further analyses (Table 6) can identify significance of factors on errors results. One-way ANOVA tests made on errors over solutions for each operation indicate that the solution factor is not significant (p-value < 0.05), with a p-value of 0.831 in repair and 0.923 in diagnosis. Besides, t-test results on errors over operations for all solutions suggest that the operation factor is close to be significant, with a p-value of 0.059.

Table 6. Statistical tests on errors results per solution and operation factors.

Factor: Solution:Repair		**Test:** One-Way ANOVA				
Effect	**Df**	**Sum Sq**	**Mean Sq**	**F Value**	**Pr (>F)**	**Significant (95% ci)**
Solution	2	00.133	0.067	0.186	0.831	No
Residuals	42	15.067	0.359			
Factor: Solution:Diagnosis		**Test:** One-Way ANOVA				
Effect	**Df**	**Sum Sq**	**Mean Sq**	**F Value**	**Pr (>F)**	**Significant (95% ci)**
Solution	2	00.044	0.022	0.080	0.923	No
Residuals	42	11.600	0.276			
Factor: Operation		**Test:** t-test				
			T	**Df**	**p-value**	**Significant (95% ci)**
			1.908	86.483	0.059	No

According to previous discussions, the following considerations can be considered valid:

- Number of errors per tester can be considered low with an average of 0.422 errors per test.
- There is significant variance on errors results per solution and per operation.

Therefore, the validation's errors hypothesis can be considered true and so, task completion time can be understood as a direct measure of efficiency. The following subsection analyses the results on experimental completion times.

Time effect study

Time measures the number of seconds taken by a tester to identify, understand and complete an maintenance step (Figure 3 and Figure 4). Because authoring and NOAR solutions did not demonstrate a significant effect on errors, time can be considered a direct representation of maintenance efficiency. Hence, time can be evaluated as the main effect of AR content support on maintenance operations through semantic understanding.

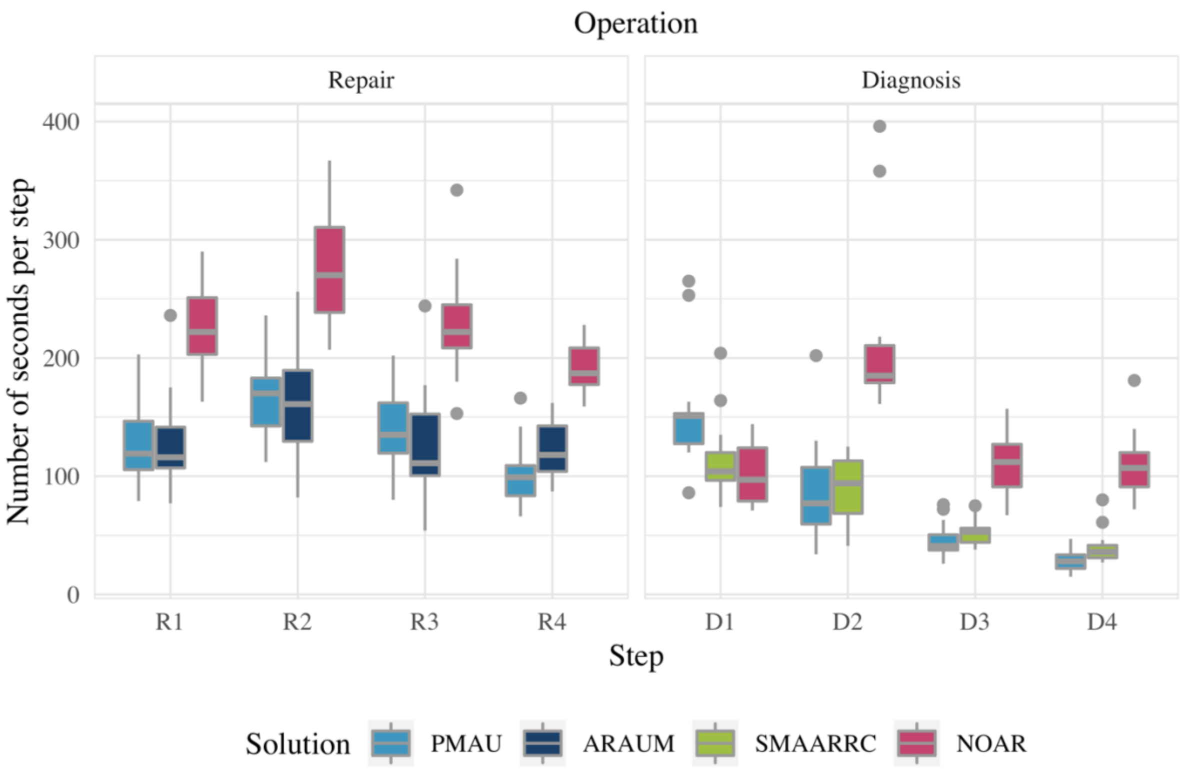

Figure 6. Box and whiskers plot on completion times per step, and solution and operation factors.

Figure 6 presents average time results per step and grouped by operation and solution factors. It displays a difference in completion times per step for each maintenance operation. Besides, it also shows a difference between AR and non-AR solutions, but not between different authoring solutions. A relevant case is D1, it can be seen that in this case the effect of AR solutions is not significant. This case is similar to the findings presented by Fernández del Amo et al. [16], where the kind of step had an effect on AR impact. Table 7 presents means and std. deviations for completion times grouped by solution and operation. These range from 134 to 231 seconds in repair and from 74 to 134 seconds in diagnosis. These numbers show a difference between repair and diagnosis operation. Thus, indicating that the assumption for separate experiment analyses was valid. In repair, means show a considerable difference (42%) in completion times among NOAR and AR (PMAU and ARAUM) solutions. In diagnosis, means also show a substantial difference (43%) in completion times between NOAR and AR (PMAU and SMAARRC), although there is also a smaller difference (~5%) between SMAARRC and PMAU.

Table 7. Means and std. deviations on completion time per operation and solution factors.

Operation	Solution	Testers	Mean	Std. deviation
Repair	PMAU	60	134.48	39.37
	ARAUM	60	134.52	42.43
	NOAR	60	230.82	48.69
Diagnosis	PMAU	60	78.82	57.76
	SMAARRC	60	73.95	37.55
	NOAR	60	133.78	61.03

Further analyses can identify the significance of comparisons discussed above. Table 8 and Table 9 presents the two-way ANOVA tests conducted to analyse time variance according to step and solution factors for each operation. According to repair results (Table 8), it can be said with a confidence interval of 95% (p-value < 0.05) that both factors (Step and Solution) have a significant effect on completion times, but not their interaction. Hence, it can be said that for repair operations, the support AR provides does not depend on the type of step being conducted. For diagnosis experiments (Table 9), ANOVA results also indicate a significant effect of step and solution factors as well as their interaction. These confirm the results presented in [16], where AR support was found more effective for higher complexities of steps being conducted.

Table 8. Two-way ANOVA test results on completion time for step and solution factors in Repair operation.

Operation: Repair						
Factor	**Df**	**Sum Sq**	**Mean**	**F value**	**Pr (>F)**	**Significant (95% ci)**
Step	3	086309	028770	020.12	3.42e-11	Yes
Solution	2	371076	185538	129.79	2.00e-16	Yes
Step:Solution	6	011061	001843	001.29	2.65e-01	No
Residuals	168	240168	001430	-----	-----	-----

ANOVA tests results help to corroborate the first time-related hypothesis presented in Experiment design. Based on these, it can be said that task completion times are dependent on the solution being used. Moreover, completion times group means (Table 7) indicate that these times decrease with the use of authoring (PMAU, ARAUM and SMAARRC) compared to NOAR solutions for each maintenance operation.

Table 9. Two-way ANOVA test results on completion time for step and solution factors in Diagnosis operation.

Operation: Diagnosis						
Factor	**Df**	**Sum Sq**	**Mean**	**F value**	**Pr (>F)**	**Significant (95% ci)**
Step	3	176247	58749	53.08	2.00e-16	Yes
Solution	2	132501	66250	59.86	2.00e-16	Yes
Step:Solution	6	137561	22927	20.71	2.00e-16	Yes
Residuals	168	185940	00117	-----	-----	-----

Besides, post hoc comparisons results from Tukey HSD tests (Table 10 and Table 11) can help to compare differences between factors groups on time means for each maintenance operation. Although ANOVA results suggest that the solution factor is a significant effect, solutions' time means differences between authoring solutions are low compared to the difference with non-AR solutions. Moreover, post-hoc comparisons for repair and diagnosis operations show that the mean differences for same-step groups of PMAU and alternative authoring solutions (ARAUM and SMAARRC) are not significantly different (p-values < 0.05). Hence, it can be said that the main effect is driven by the difference between AR and NOAR solutions rather than in-between AR solutions. Thus, suggesting the validity of the time-related second hypothesis (Experiment design), which enounced that completion times do not vary significantly among authoring solutions (PMAU and ARAUM, and PMAU and SMAARRC) for each maintenance operation.

Table 10. Significance (p-value) results on post hoc comparisons (Tukey HSD) in Repair operation.

Operation: Repair | **Legend:** R = Step, P = PMAU, A = ARAUM, N = NOAR

	R1:P	**R2:P**	**R3:P**	**R4:P**	**R1:A**	**R2:A**	**R3:A**	**R4:A**	**R1:N**	**R2:N**	**R3:N**	**R4:N**
R1:P	---	0.2085	0.9987	0.7285	1.0000	0.5576	1.0000	1.0000	0.0000	0.0000	0.0000	0.0004
R2:P	0.2085	---	0.8018	0.0003	0.2456	1.0000	0.1498	0.0842	0.0020	0.0000	0.0006	0.7620
R3:P	0.9987	0.8018	---	0.1586	0.9994	0.9828	0.9953	0.9776	0.0000	0.0000	0.0000	0.0140
R4:P	0.7285	0.0003	0.1586	---	0.6770	0.0031	0.8152	0.9160	0.0000	0.0000	0.0000	0.0000
R1:A	1.0000	0.2456	0.9994	0.6770	---	0.6128	1.0000	1.0000	0.0000	0.0000	0.0000	0.0006
R2:A	0.5576	1.0000	0.9828	0.0031	0.6128	---	0.4548	0.3088	0.0002	0.0000	0.0000	0.3740
R3:A	1.0000	0.1498	0.9953	0.8152	1.0000	0.4548	---	1.0000	0.0000	0.0000	0.0000	0.0002
R4:A	1.0000	0.0842	0.9776	0.9160	1.0000	0.3088	1.0000	---	0.0000	0.0000	0.0000	0.0001
R1:N	0.0000	0.0020	0.0000	0.0000	0.0000	0.0002	0.0000	0.0000	---	0.0179	1.0000	0.4349
R2:N	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0179	---	0.0469	0.0000
R3:N	0.0000	0.0006	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0469	---	0.2481
R4:N	0.0004	0.7620	0.0140	0.0000	0.0006	0.3740	0.0002	0.0001	0.4349	0.0000	0.2481	---

Table 11. Significance (p-value) results on post hoc comparisons (Tukey HSD) in Diagnosis operation.

Operation: Diagnosis | **Legend:** D = Step, P = PMAU, S = SMAARRC, N = NOAR

	D1:P	**D2:P**	**D3:P**	**D4:P**	**D1:S**	**D2:S**	**D3:S**	**D4:S**	**D1:A**	**D2:A**	**D3:A**	**D4:A**
D1:P	---	0.0000	0.0000	0.0000	0.0674	0.0000	0.0000	0.0000	0.0028	0.0001	0.0418	0.0184
D2:P	0.0000	---	0.0284	0.0001	0.6144	1.0000	0.1455	0.0059	0.9913	0.0000	0.7271	0.8727
D3:P	0.0000	0.0284	---	0.9644	0.0001	0.0207	1.0000	1.0000	0.0004	0.0000	0.0000	0.0000
D4:P	0.0000	0.0001	0.9644	---	0.0000	0.0001	0.7091	0.9988	0.0001	0.0000	0.0000	0.0000
D1:S	0.0674	0.6144	0.0001	0.0000	---	0.6833	0.0001	0.0001	0.9984	0.0000	1.0000	1.0000
D2:S	0.0000	1.0000	0.0207	0.0001	0.6833	---	0.1140	0.0041	0.9960	0.0000	0.7880	0.9119
D3:S	0.0000	0.1455	1.0000	0.7091	0.0001	0.1140	---	0.9961	0.0041	0.0000	0.0002	0.0005
D4:S	0.0000	0.0059	1.0000	0.9988	0.0001	0.0041	0.9961	---	0.0001	0.0000	0.0000	0.0000
D1:A	0.0028	0.9913	0.0004	0.0001	0.9984	0.9960	0.0041	0.0001	---	0.0000	0.9997	1.0000
D2:A	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	---	0.0000	0.0000
D3:A	0.0418	0.7271	0.0000	0.0000	1.0000	0.7880	0.0002	0.0000	0.9997	0.0000	---	1.0000
D4:A	0.0184	0.8727	0.0000	0.0000	1.0000	0.9119	0.0005	0.0000	1.0000	0.0000	1.0000	---

Overall, previous discussions support the validity of the following considerations regarding the effect on completion time of authoring and NOAR solutions:

- For repair operations, completion times for PMAU and ARAUM authoring solutions are 42% faster than NOAR solutions.
- For diagnosis operations, completion times for PMAU and SMAARRC are respectively 41% and 45% faster than NOAR solutions.
- Differences in completion times between authoring and NOAR solutions can be considered significant for each maintenance operation.
- Differences in completion times among authoring solutions in each maintenance operation cannot be considered significant for each operation's step.
- Effect of authoring solutions is dependent on steps conducted for diagnosis operations but not for repair operations.

These results support the validity of this research's hypotheses regarding the positive effect on efficiency of the proposed authoring solution for multiple maintenance operations. Such effect is assumed to be achieved by the proposed authoring's ability to automatically produce content that is adaptive for enhancing semantic understanding of maintenance operations. A relevant method to further evaluate content adaptiveness is measuring its usability. Hence, the following subsection analyses the usability surveys that testers completed after experiments.

Usability study

Usability can be described as a qualitative measure of the degree to which augmented content achieves user's semantic understanding of maintenance operations. Based on Nielsen's criteria [19], Table 4 presented a set of criterions for evaluating usability along with content's aspects against to which criterions can be assessed.

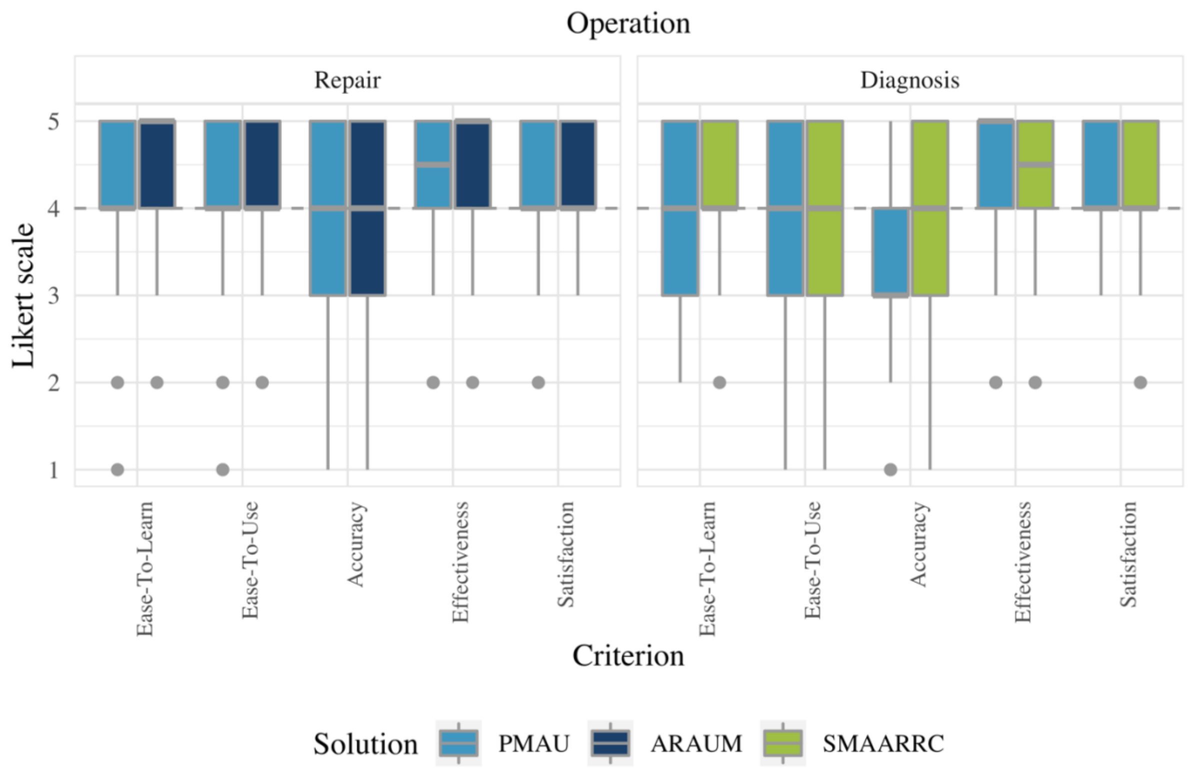

Figure 7. Box and whiskers plot on usability criterions' responses per solution and operation.

Figure 7 illustrates a box and whiskers plot to summarise experts' responses for each usability criterion according to two operation and solution experimental factors. It shows that criterions are not considerably different among PMAU and ad-hoc authoring solutions (ARAUM and SMAARRC) in repair and diagnosis operations usability. Most criterions scored above 4 in a Likert Scale out of 5, with higher variabilities in diagnosis scenarios. Table 12 presents the means and std. deviations for testers' responses usability criterions grouped by operation and solution. In absolute numbers, group means range from 3.9 to 4.1 in a Likert Scale (1-5) with the exception of PMAU's accuracy in diagnosis, which goes down to 3.4. Percentual differences between PMAU and ad-hoc authoring solutions means range in between -1% and 12%. In repair, ARAUM is considered more usable (5%-8%) regarding all criterions except for Satisfaction. In diagnosis, SMAARRC and PMAU have similar considerations for all criterions but for accuracy, where SMAARRC is considered 12% more accurate. Overall, these numbers suggest that PMAU's content achieves similar usability than that from other authoring solutions. The only exception is PMAU's accuracy in diagnosis operation. A reason for this might be related to an event occurred during experiments that was connected to the HoloLens camera behaviour: tracking was being lost when testers were asked to get closer for inspecting the equipment.

Table 12. Means and std. deviations of usability criterions' responses per solution and operation.

Operation	Criterion	Solution	Testers	Mean	Std. deviation
Repair	Ease-To-Learn	PMAU	15	4.045	1.011
		ARAUM	15	4.400	0.780
	Ease-To-Use	PMAU	15	3.940	0.974
		ARAUM	15	4.218	0.841
	Accuracy	PMAU	15	3.920	0.955
		ARAUM	15	4.067	0.977
	Effectiveness	PMAU	15	4.244	0.928
		ARAUM	15	4.367	0.785
	Satisfaction	PMAU	15	4.244	0.743
		ARAUM	15	4.222	0.704
Diagnosis	Ease-To-Learn	PMAU	15	3.978	0.892
		SMAARRC	15	4.182	0.756
	Ease-To-Use	PMAU	15	3.843	1.010
		SMAARRC	15	3.793	1.108
	Accuracy	PMAU	15	3.413	1.001
		SMAARRC	15	3.893	1.085
	Effectiveness	PMAU	15	4.333	0.874
		SMAARRC	15	4.300	0.867
	Satisfaction	PMAU	15	4.133	0.694
		SMAARRC	15	4.111	0.804

The total number of testers responses per criterion's aspect (60) provided sufficient data to analyse each of them separately. Independent box and whiskers plots (Figures Figure 8-Figure 12) for each criterion showing response averages per aspect can provide additional insights regarding further improvements on PMAU's usability.

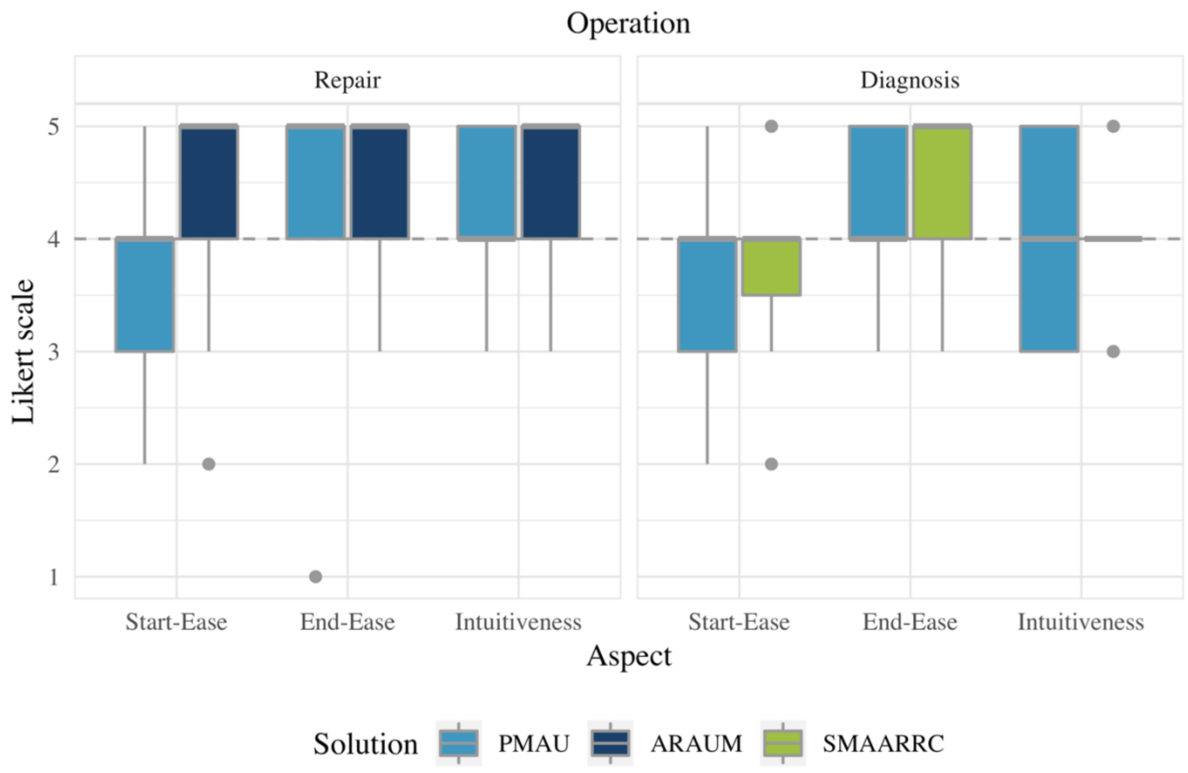

Figure 8. Box and whiskers plot on Ease-To-Learn aspects' responses per solution and operation.

Figure 8 displays average testers' responses regarding Ease-To-Learn aspects compared by authoring solutions and operations. These results suggest that PMAU's content was slightly more difficult to learn compared to other authoring solutions. ARAUM (tablet-based) had almost no differences between ease-to-use at start and at finish, while SMAARRC's had a slightly smaller difference between start and finish compared to PMAU. In terms of intuitiveness, only ARAUM's results indicate a better performance.

Figure 9 presents average testers' responses on Ease-To-Use aspects. These do not show interesting differences between authoring solutions in terms of content formats. Tablet-based solutions (ARAUM) showed better responses for text and buttons, while SMAARRC showed the worst results for 3D models.

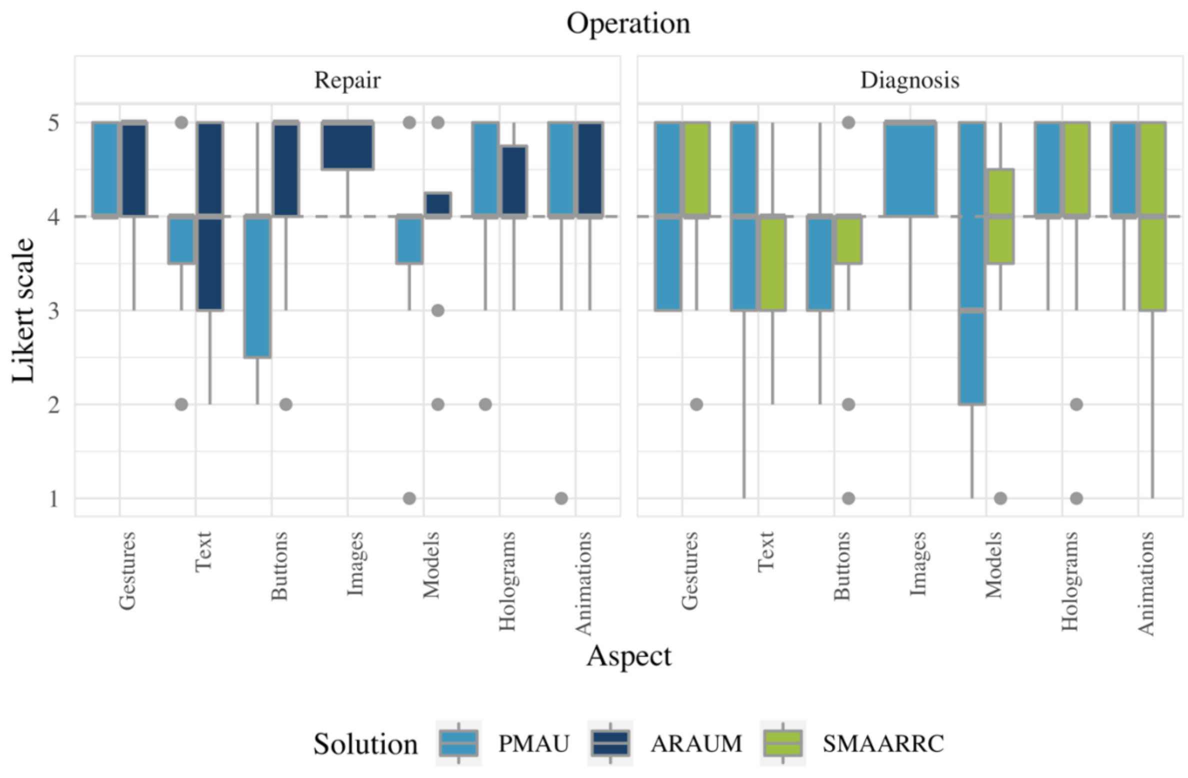

Figure 9. Box and whiskers plot on Ease-To-Use aspects' responses per solution and operation.

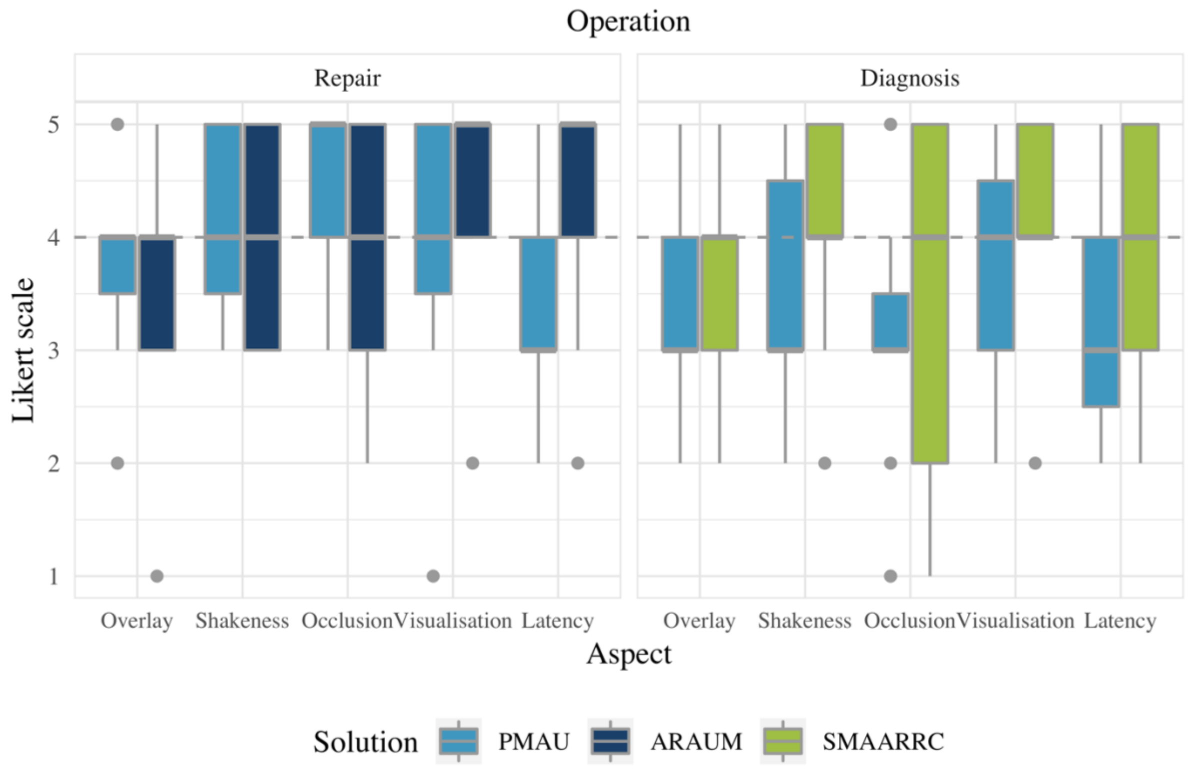

Figure 10. Box and whiskers plot on Accuracy aspects' responses per solution and operation.

Figure 10 describes testers' responses regarding Accuracy aspects of authoring solutions. These indicate that PMAU had a slightly worse performance in terms of latency. That could be explained due to the real-time PMAU's requirements regarding content generation. For other aspects, responses are quite similar for all three authoring solutions except for occlusion, where SMAARRC received a great variability on its responses.

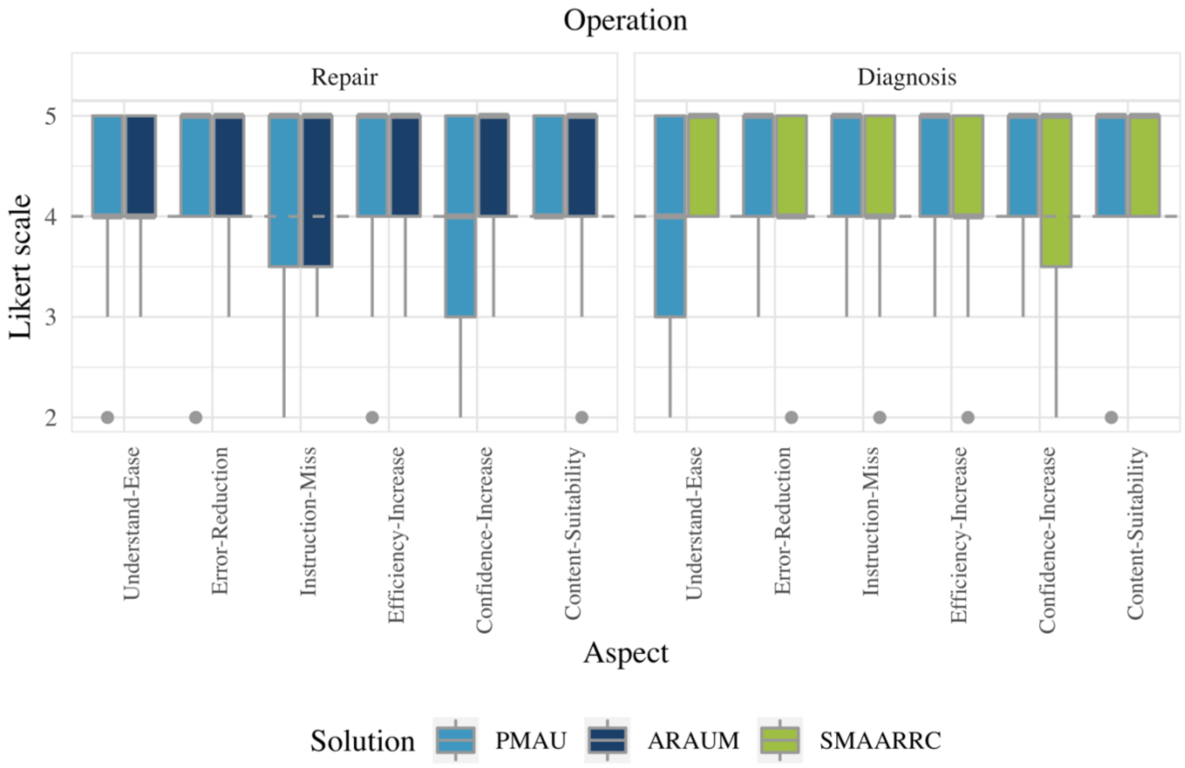

Figure 11. Box and whiskers plot on Effectiveness aspects' responses per solution and operation.

Figure 11 presents average testers' responses regarding Effectiveness aspects. These results indicate that all authoring solutions were considered similarly in terms of their abilities to reduce errors, missed instructions and improve efficiency and confidence. One exception is PMAU's variability in ease-to-understand for diagnosis operations. Few testers noted during experiments that ontological naming conventions were sometimes difficult to understand. Thus, it seems important to adapt ontological wording for improved usability.

Finally, Figure 12 summarises testers' responses regarding Satisfaction aspects compared by solutions and operations. Satisfaction results were reasonably higher for PMAU compared to other authoring solutions. A reason for this can be the potential improvements testers identified about PMAU's ontological approach. Some of them noted the ability of PMAU's approach to track user's performance through accurate content monitoring. Because content is generated in real-time, content visualisation times can be easily tracked to further analyse content usage times and so, content adaptation effectiveness. Although it may require additional user-tracking techniques to ensure accurate measures.

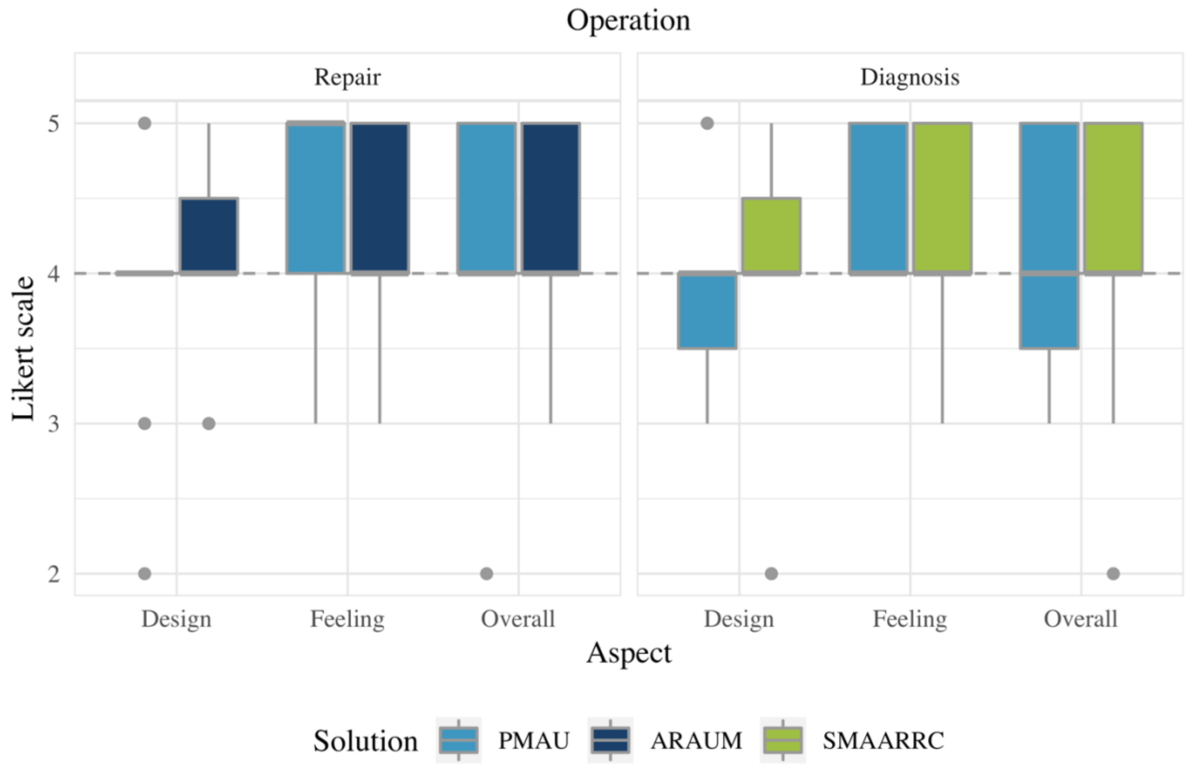

Figure 12. Box and whiskers plot on Satisfaction aspects' responses per solution and operation.

Overall, testers survey results did not suggest a significant difference on content usability among authoring solutions. PMAU scored relatively lower in accuracy and text understanding, which are areas for further improvements. Moreover, PMAU's ability to track user's performance through content monitoring was also perceived as a good solution to further adapt content according to user's expertise. Hence, it can be said that these results indicate validity of the last research's hypothesis regarding insignificant content usability variance among authoring solutions for each maintenance operation.

Discussion

Previous analyses aimed to evaluate this research's hypotheses described in Experiment design, which intended to demonstrate the validity this research's contributions.

The first validation hypothesis stated that "completion errors do not vary significantly among authoring and no-AR solutions for maintenance operations". Errors effect study analysed the correlation of operation and solution factors with experimental errors. It showed that number of errors per tester could be considered low, with an average of 0.422 errors. Results of ANOVA and t-tests did not indicate a significant variance on error results per solution and per operation. Therefore, the first hypothesis can be considered valid within the context of the experiments conducted. Thus, pondering completion times as a direct measure of maintenance efficiency. Nevertheless, number of errors were counted per test and not per test's step. So, it could not be studied the effect on completion errors of each experimental step. Though this may be an interesting element to evaluate, it was out of this study's scope because steps were predetermined to ensure maintenance quality's consistency among experiments. Future studies could investigate such effect by experimenting with the proposed authoring solution in real-life maintenance operations.

The second validation hypothesis assumed that "completion time decreases with authoring solutions compared to no-AR solutions for each maintenance operation". Instead, the third one stated that "completion time does not vary significantly between authoring solutions for each maintenance operation". Time effect study analysed the effect on time of solution and operation factors grouped by case of study. In repair operations [15], completion times for PMAU and ARAUM authoring solutions were found 42% faster than NOAR solutions. In diagnosis operations [16], completion times for PMAU and SMAARRC are respectively 41% and 45% faster than NOAR solutions. Besides, two-way ANOVA results indicated a significant difference on time between authoring and NOAR solutions but not among authoring solutions (PMAU, ARAUM and SMAARRC). Also, the effect of authoring solutions in diagnosis operations was found dependent on the maintenance step. Thus, confirming the results presented by Fernández del Amo et al. [16], which analysed that correlation. The authors considered that analysis out of this research's scope because it aimed at proving the similarity between the effects of different authoring solutions. Nevertheless, future works can investigate the relation between augmented content usability and maintenance complexity to further improve content adaptiveness and relevant discard rules for content formats pairing. Overall, these results prove valid the second and third hypothesis in the contexts of this research's laboratory experiments. And so, it can be said that the proposed authoring's content can achieve similar effects on maintenance efficiency than other operation-specific authoring solutions.

The final validation hypothesis indicated that "content usability does not vary significantly among authoring solutions for each maintenance operation". Thus, aiming to evaluate whether the automatically generated content was usable from a tester's perspective for gaining semantic understanding of maintenance operations. Usability study evaluated usability criteria [19] according to different augmented content aspects. These results did not show significant differences on testers' responses about content usability between authoring solutions. Thus, confirming the assumptions of the abovementioned hypothesis. Nevertheless, PMAU scored relatively lower in accuracy and text understanding. But testers also noted its improved ability to track user's performance through more accurate content monitoring and further adapt content according to user's expertise. These are areas where future works can focus their efforts to achieve better effects of AR solutions in maintenance operations.

The analyses results discussed above aimed to validate this research's contributions for their ability to automatically create adaptive content for multiple maintenance operations. Although this validation's hypotheses can be considered proven in the context of this research's cases of study, the following paragraphs consider some relevant aspects to discuss.

The first contribution described a method to declare programmable formats that semantically describe their data, user and environmental requirements for producing augmented content. Its aim is to create templates with certain augmentation behaviours that can later be matched with maintenance datasets. These formats and their behaviours comprise different combinations of visualisation and interaction modes. Thus, enabling AR developers to create content for all kinds of maintenance tasks, scenarios and expertise levels. Most content formats implemented in this research replicate those presented in [15] for repair and in [16] for remote diagnosis operations. This research also developed some more generic formats to ensure augmentation of all individual properties of datatype and object types. Moreover, there exist two on-going researches where this research's authoring proposal is being utilised to develop new formats for thermographic assessment [ref] and diagnosis reporting [ref] operations. The reason to implement formats from different researches was to demonstrate that the proposed authoring method can create content for multiple maintenance operations. This can be further corroborated in future works by using adaptive formats already researched such as those from Chang [21] and Wang [13] for assembly. Future works can also develop more adaptive formats for less researched operations such as monitoring. Besides, future works can also focus on more basic research for improving content adaptiveness. These can investigate the relation among visualisation and interaction methods (e.g. animations) with human performance (e.g. sight). Thus, designing more accurate descriptors for content formats to enable automatic adaptation to user (e.g. expertise) and environmental (e.g. light) conditions.

The second contribution proposed a real-time, ontology-based, pattern-matching algorithm to pair content formats with ontology individuals for automatically creating, adapting and locating augmented content. The algorithm comprises different assignation and discard rules to match individual's properties with formats' data, user and environment facets. Although these rules have proven sufficient to match ontology individuals with specific content formats for repair and diagnosis applications, they still depend on formats' declarations made by AR developers. Future works can research more advanced methods to declare content formats and rules to pattern-match them. These may include techniques like natural language processing, environment and user tracking (e.g. light conditions or user attention) and so forth.

The proposed algorithm also parts content creation and information management authoring processes to automate the first one. For this reason, the authors considered that authoring efficiency experiments were out of this research's scope. Nevertheless, maintenance experts still need to perform information management processes. In this research, the web-based ontology reporting tool presented in [ref] has been used for this purpose. Future works can further improve this process. They can analyse the effect of ontology wording and its impact on AR semantic understanding and design tools for declaring user-adaptive ontologies. Besides, they can also further evaluate the impact of different authoring solutions in AR deployment costs. Thus, easing the implementation of AR technologies in maintenance organisations.

A relevant feature of the proposed algorithm relates to its ability for generating augmented content in real-time. This allows not only to enable AR applications such as remote collaborative diagnosis [16] but also to perform tracking of content being used (**Error! Reference source not found.**). This research's system implementation enabled reporting capabilities to trace individuals augmented and their content creation dates, although its benefits have not been explored. Future works can further study this ontology-based content-tracking feature and its impact on content adaptiveness as well as maintenance performance evaluation. Moreover, they can also improve its accuracy with more advanced user-tracking techniques like eye-tracking and other biometric technologies.

Another relevant algorithm's feature involves its use of ontologies. Unlike other ontology-based authoring techniques [11,13,22,23], this algorithm does not use ontologies for inferencing purposes but to enable information standardisation for semantic analysis. Thus, allowing individuals to be augmented adaptively through different formats according to user and environment facets. Hence, it is feasible to consider that this algorithm could also be used with other information management methods (e.g. SQL or graphical databases), if those were to meet these standardisation requirements. Future works could further investigate on those requirements for extending this algorithm's applicability to other information management methods. Besides, this algorithm aims to augment existing ontology individuals but does not consider the creation of new instances. Future works can extend this research by including new algorithm features as well as content formats to enable for ontology individuals to be created using AR applications. Thus, enabling AR technologies not only to transfer but also to capture knowledge. Moreover, ontology individual's instantiation may require defining links to existing individuals. The number of individuals to select from in new instantiations can be high and so, their content can overload the augmented scene. Therefore, future works should also investigate recommendation techniques for improved information filtering on AR-based knowledge capture applications. These can consider the relation between recommender systems and the algorithm's content-tracing capabilities to enable maintenance context-aware recommendations.

This research's contributions aim at automatically producing adaptive content for multiple maintenance operations. Although the described experiments proved so for repair and remote diagnosis tasks, these were conducted in laboratory setups. The reason to do so was to maintain consistency with previous researches. Nevertheless, future works described above can further corroborate this research results with experiments in real-life conditions, including evaluation of other relevant factors in AR usability like ergonomics. Besides, the proposed contributions focus on AR-maintenance applications and so, they include some assumptions regarding the use of certain AR methods like tracking and registration. Therefore, future works can study the applicability of these contributions to other AR fields of application such as medicine, tourism and so forth. Thus, aiming to develop a framework for automatic authoring in AR that could ease its implementation in commercial and industrial markets.

References

[1] Zhu Z, Zhou X, Shao K. A novel approach based on Neo4j for multi-constrained flexible job shop scheduling problem. Comput Ind Eng 2019. doi:10.1016/j.cie.2019.03.022.

[2] Panzarino O. Learning Cypher. 1st ed. Packt Publishing; 2014.

[3] Barrasa J. neosemantics 2019.

[4] Eernisse M. EJS 2015.

[5] Surhone LM, Tennoe MT, Henssonow SF. Node.Js. Beau Bassin, MUS: Betascript Publishing; 2010.

[6] Unity Technologies. Unity Game Engine. Https://UnityCom 2019. https://unity.com.

[7] Microsoft Corporation. Visual Studio. Https://VisualstudioMicrosoftCom 2019.

[8] Hejlsberg A. The C# programming language. Addison-Wesley; 2011.

[9] Microsoft Corporation. Mixed Reality Toolkit Documentation. Microsoft Corp n.d. https://microsoft.github.io/MixedRealityToolkit-Unity/Documentation/GettingStartedWithTheMRTK.html (accessed April 29, 2020).

[10] PTC Corporation. Vuforia SDK. PTC Corp n.d. https://library.vuforia.com/getting-started/overview.html (accessed April 29, 2020).

[11] Flotyński J, Walczak K. Conceptual knowledge-based modeling of interactive 3D content. Vis Comput 2015;31:1287–306. doi:10.1007/s00371-014-1011-9.

[12] Longo F, Nicoletti L, Padovano A. Ubiquitous knowledge empowers the Smart Factory: The impacts of a Service-oriented Digital Twin on enterprises' performance. Annu Rev Control 2019;47:221–36. doi:10.1016/j.arcontrol.2019.01.001.

[13] Wang X, Ong SK, Nee AYC. Multi-modal augmented-reality assembly guidance based on bare-hand interface. Adv Eng Informatics 2016;30:406–21. doi:10.1016/j.aei.2016.05.004.

[14] Gimeno J, Morillo P, Orduña JM, Fernández M. A new AR authoring tool using depth maps for industrial procedures. Comput Ind 2013;64:1263–71. doi:10.1016/j.compind.2013.06.012.

[15] Erkoyuncu JA, Fernández del Amo I, Dalle Mura M, Roy R, Dini G. Improving efficiency

of industrial maintenance with context aware adaptive authoring in augmented reality. CIRP Ann - Manuf Technol 2017. doi:10.1016/j.cirp.2017.04.006.

[16] Fernández del Amo I, Erkoyuncu J, Vrabič R, Frayssinet R, Vazquez Reynel C, Roy R. Structured authoring for AR-based communication to enhance efficiency in remote diagnosis for complex equipment. Adv Eng Informatics 2019;45:15. doi:10.1016/j.aei.2020.101096.

[17] Azuma RT. The Most Important Challenge Facing Augmented Reality. Presence Teleoperators Virtual Environ 2016;25:234–8. doi:10.1162/PRES_a_00264.

[18] Chi HL, Chen YC, Kang SC, Hsieh SH. Development of user interface for tele-operated cranes. Adv Eng Informatics 2012;26:641–52. doi:10.1016/j.aei.2012.05.001.

[19] Nielsen J. Usability Engineering. 1st ed. San Francisco, California, USA: Morgan Kaufmann; 1993.

[20] Cullot N, Ghawi R, Yétongnon K. DB2OWL : A Tool for Automatic Database-to-Ontology Mapping. Proc Fifteenth Ital Symp Adv Database Syst SEBD 2007:1–4.

[21] Chang MML, Nee AYC, Ong SK. Interactive AR-assisted product disassembly sequence planning (ARDIS). Int J Prod Res 2020;0:1–16. doi:10.1080/00207543.2020.1730462.

[22] Walczak K, Flotyński J. Inference-based creation of synthetic 3D content with ontologies. Multimed Tools Appl 2019;78:12607–38. doi:10.1007/s11042-018-6788-5.

[23] Zhu J, Ong SK, Nee AYC. A context-aware augmented reality assisted maintenance system. Int J Comput Integr Manuf 2015;28:213–25. doi:10.1080/0951192X.2013.874589.