Realising an Applied Gaming Eco-system

Grant agreement no.: 644187

D8.1 – RAGE Evaluation Framework and Guidelines

RAGE – WP8 – D8.1

<table>
<thead>
<tr>
<th>Project Number</th>
<th>H2020-ICT-2014-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Due Date</td>
<td>31 January 2016 (M12)</td>
</tr>
<tr>
<td>Actual Date</td>
<td>19-02-2016</td>
</tr>
<tr>
<td>Document Author/s</td>
<td>Christina Steiner, Alexander Nussbaumer, Eric Kluijfhout, Rob Nadolski, Barbara Bazzanella, Samuel Mscarenhas, Pablo Moreno Ger, Mihai Dascalu, Stefan Trausan-Matu, Jana Becker, Li Yuan, Paul Hollins</td>
</tr>
<tr>
<td>Version</td>
<td>2.4</td>
</tr>
<tr>
<td>Dissemination level</td>
<td>PU</td>
</tr>
<tr>
<td>Status</td>
<td>Final version</td>
</tr>
<tr>
<td>Document approved by</td>
<td></td>
</tr>
</tbody>
</table>

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 644187
### Document Version Control

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Change Made (and if appropriate reason for change)</th>
<th>Initials of Commentator(s) or Author(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>01-09-2015</td>
<td>First draft structure</td>
<td>CS</td>
</tr>
<tr>
<td>0.3</td>
<td>16-10-2015</td>
<td>Further drafting and notes on sections, inclusion of contents from ICERI paper</td>
<td>CS</td>
</tr>
<tr>
<td>1.2</td>
<td>13-12-2015</td>
<td>Draft version uploaded to RAGE site</td>
<td>CS, EK, PMG, SM, JB,</td>
</tr>
<tr>
<td>1.4</td>
<td>28-12-2015</td>
<td>Updated draft version</td>
<td>CS, RN, BB, LY, AN</td>
</tr>
<tr>
<td>1.8</td>
<td>08-01-2016</td>
<td>Finalisation of draft for peer review</td>
<td>CS</td>
</tr>
<tr>
<td>2.0</td>
<td>21-01-2016</td>
<td>Updated version incorporating feedback from the peer reviews (WW, KS, JJ)</td>
<td>CS, EK, LY, PH</td>
</tr>
<tr>
<td>2.1</td>
<td>22-01-2016</td>
<td>Final version submitted to the coordinator</td>
<td>CS</td>
</tr>
<tr>
<td>2.2</td>
<td>25-01-2016</td>
<td>Final updates (section 3, header/footer)</td>
<td>CS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Very final version submitted to the coordinator and the EC</td>
<td>CS</td>
</tr>
<tr>
<td>2.4</td>
<td>19-02-2016</td>
<td>Shortened version submitted coordinator</td>
<td>CS, EK</td>
</tr>
</tbody>
</table>

### Document Change Commentator or Author

<table>
<thead>
<tr>
<th>Initials</th>
<th>Name of Author</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td>Christina Steiner</td>
<td>TUGRAZ</td>
</tr>
<tr>
<td>AN</td>
<td>Alexander Nussbaumer</td>
<td>TUGRAZ</td>
</tr>
<tr>
<td>EK</td>
<td>Eric Kluijfhout</td>
<td>OUNL</td>
</tr>
<tr>
<td>RN</td>
<td>Rob Nadolski</td>
<td>OUNL</td>
</tr>
<tr>
<td>BB</td>
<td>Barbara Bazzanella</td>
<td>OKKAM</td>
</tr>
<tr>
<td>SM</td>
<td>Samuel Mscarenhas</td>
<td>INESC ID</td>
</tr>
<tr>
<td>PMG</td>
<td>Pablo Moreno Ger</td>
<td>UCM</td>
</tr>
<tr>
<td>MD</td>
<td>Mihai Dascalu</td>
<td>UPB</td>
</tr>
<tr>
<td>STM</td>
<td>Stefan Trausan-Matu</td>
<td>UPB</td>
</tr>
<tr>
<td>JB</td>
<td>Jana Becker</td>
<td>FTK</td>
</tr>
<tr>
<td>LY</td>
<td>Li Yuan</td>
<td>UOB</td>
</tr>
<tr>
<td>PH</td>
<td>Paul Hollins</td>
<td>UOB</td>
</tr>
<tr>
<td>WW</td>
<td>Wim Westera</td>
<td>OUNL</td>
</tr>
<tr>
<td>KS</td>
<td>Kam Star</td>
<td>PLAYGEN</td>
</tr>
<tr>
<td>UU</td>
<td>Johan Jeuring</td>
<td>UU</td>
</tr>
</tbody>
</table>
TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXECUTIVE SUMMARY</td>
<td>6</td>
</tr>
<tr>
<td>1  INTRODUCTION</td>
<td>7</td>
</tr>
<tr>
<td>2  EVALUATION IN RAGE</td>
<td>9</td>
</tr>
<tr>
<td>2.1 Evaluation Phases</td>
<td>9</td>
</tr>
<tr>
<td>2.2 Evaluation Objects</td>
<td>10</td>
</tr>
<tr>
<td>2.3 Stakeholders and User Groups</td>
<td>10</td>
</tr>
<tr>
<td>2.4 Evaluation Instruments</td>
<td>11</td>
</tr>
<tr>
<td>3  RESEARCH DATA MANAGEMENT AND EVALUATION GUIDELINES</td>
<td>12</td>
</tr>
<tr>
<td>3.1 Roles and Responsibilities in RAGE Evaluation</td>
<td>12</td>
</tr>
<tr>
<td>3.2 Checklist for Data Management in Evaluation Studies</td>
<td>12</td>
</tr>
<tr>
<td>3.3 Tools &amp; Tips</td>
<td>13</td>
</tr>
<tr>
<td>3.3.1 Planning Evaluation – Data Management Planning</td>
<td>13</td>
</tr>
<tr>
<td>3.3.1.1 Planning Data Documentation</td>
<td>13</td>
</tr>
<tr>
<td>3.3.1.2 Recruiting Evaluation Participants</td>
<td>13</td>
</tr>
<tr>
<td>3.3.1.3 RAGE Information Letter and Consent Form</td>
<td>13</td>
</tr>
<tr>
<td>3.3.2 Carrying Out Evaluation – Collecting Data</td>
<td>14</td>
</tr>
<tr>
<td>3.3.2.1 Informed Consent</td>
<td>14</td>
</tr>
<tr>
<td>3.3.2.2 How to Use Online Surveys</td>
<td>14</td>
</tr>
<tr>
<td>3.3.2.3 Audio and Video Recordings</td>
<td>15</td>
</tr>
<tr>
<td>3.3.2.4 Blind Experiments</td>
<td>15</td>
</tr>
<tr>
<td>3.3.3 Working with the Results – Processing, Storing, and Sharing Data</td>
<td>15</td>
</tr>
<tr>
<td>3.3.3.1 Organising Data Files</td>
<td>15</td>
</tr>
<tr>
<td>3.3.3.2 How to Anonymize</td>
<td>15</td>
</tr>
<tr>
<td>4  RAGE EVALUATION MODEL</td>
<td>16</td>
</tr>
<tr>
<td>5  EVALUATION OF GAME DEVELOPMENT &amp; ASSETS</td>
<td>17</td>
</tr>
<tr>
<td>5.1 Evaluation Objectives and Questions</td>
<td>17</td>
</tr>
<tr>
<td>5.2 RAGE Evaluation Approach</td>
<td>18</td>
</tr>
<tr>
<td>5.2.1 Participants</td>
<td>19</td>
</tr>
<tr>
<td>5.2.2 Research Design</td>
<td>20</td>
</tr>
<tr>
<td>5.2.2.1 Evaluations in the Context of Use Cases</td>
<td>20</td>
</tr>
<tr>
<td>5.2.2.2 Training and Evaluation Workshops</td>
<td>20</td>
</tr>
<tr>
<td>5.2.2.3 Asset-specific Evaluation Studies</td>
<td>21</td>
</tr>
<tr>
<td>6  EVALUATION OF ECOSYSTEM SERVICES AND PROCESSES</td>
<td>22</td>
</tr>
<tr>
<td>6.1 Evaluation Objectives and Questions</td>
<td>22</td>
</tr>
<tr>
<td>6.2 RAGE Evaluation Approach</td>
<td>22</td>
</tr>
<tr>
<td>6.2.1 Phased Approach</td>
<td>24</td>
</tr>
<tr>
<td>6.2.2 Participants</td>
<td>24</td>
</tr>
<tr>
<td>6.2.3 Research Design</td>
<td>25</td>
</tr>
<tr>
<td>6.2.3.1 Ecosystem Evaluation Workshops</td>
<td>25</td>
</tr>
<tr>
<td>6.2.3.2 Evaluations of Semantic Annotation Services</td>
<td>26</td>
</tr>
<tr>
<td>6.2.3.3 Evaluation in the Context of Application Scenarios</td>
<td>26</td>
</tr>
<tr>
<td>7  VALIDATION STUDIES IN APPLICATION SCENARIOS</td>
<td>27</td>
</tr>
<tr>
<td>7.1 Evaluation Objectives and Questions</td>
<td>27</td>
</tr>
<tr>
<td>7.2 RAGE Evaluation Approach</td>
<td>28</td>
</tr>
<tr>
<td>7.2.1 Educational Effectiveness of Applied Games</td>
<td>29</td>
</tr>
<tr>
<td>7.2.1.1 Participants</td>
<td>30</td>
</tr>
<tr>
<td>7.2.1.2 Research Design</td>
<td>30</td>
</tr>
<tr>
<td>7.2.1.2.1 Real-world Pilot Studies</td>
<td>30</td>
</tr>
<tr>
<td>7.2.1.2.2 Laboratory Studies</td>
<td>31</td>
</tr>
<tr>
<td>7.2.2 Cost-effectiveness for Game Development</td>
<td>31</td>
</tr>
<tr>
<td>7.2.2.1 Participants</td>
<td>32</td>
</tr>
<tr>
<td>7.2.2.2 Research Design</td>
<td>32</td>
</tr>
<tr>
<td>8  CONCLUSIONS AND NEXT STEPS</td>
<td>33</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>35</td>
</tr>
</tbody>
</table>
ANNEX 1: ADDITIONAL CONSIDERATIONS ON EVALUATION IN RAGE.......................... 43
A1.1 Goals of Evaluation ........................................................................................................ 43
A1.2 Research Target Framework .......................................................................................... 45
A1.3 Person Variables ............................................................................................................ 46
ANNEX 2: ADDITIONAL INFORMATION ON RESEARCH DATA MANAGEMENT........... 48
A2.1 The Evaluation Process and Research Data Lifecycle .................................................. 48
A2.2 Ethics, Privacy and Open Access ................................................................................... 49
A2.2.1 Main Considerations for RAGE Evaluation Studies ................................................ 49
A2.2.2 Securing Open Data Access .................................................................................... 53
A2.2.3 Support and Monitoring Structure .......................................................................... 54
A2.3 Information Letter and Consent Form ........................................................................... 57
A2.4 Textfile Template for describing Data Set ...................................................................... 59
A2.5 National Contacts & Supervisory Authorities on Data Protection & Privacy .......... 59
ANNEX 3: ADDITIONAL INFORMATION ON EVALUATION OF GAME DEVELOPMENT & ASSETS .......................................................................................................................... 61
A3.1 Introduction to RAGE Assets ......................................................................................... 61
A3.2 Relevant Related Work and Evaluation Approaches .................................................... 61
A3.2.1 Game Development and Course Authoring ......................................................... 61
A3.2.2 Layered Evaluation of Adaptive Systems .................................................................. 62
A3.2.3 Open Learner Models .............................................................................................. 63
A3.2.4 Social Believability in Games .................................................................................. 63
A3.3 Elaboration of Evaluation Questions ............................................................................ 64
A3.3.1 General Evaluation Questions on Assets and Asset Application ......................... 64
A3.3.2 Evaluation Questions on Specific Asset Functionality ............................................ 65
A3.4 Instruments .................................................................................................................. 66
ANNEX 4: ADDITIONAL INFORMATION ON EVALUATION OF ECOSYSTEM SERVICES AND PROCESSES ............................................................................................................ 68
A4.1 Introduction to the RAGE Ecosystem ........................................................................... 68
A4.2 Relevant Related Work and Evaluation Approaches .................................................... 68
A4.2.1 Software Repositories and their Evaluation ............................................................... 68
A4.2.2 Digital Libraries and Virtual Research Environments and their Evaluation .......... 69
A4.2.3 Further Relevant Work ............................................................................................ 70
A4.3 Elaboration of Evaluation Questions ............................................................................ 70
A4.4 Instruments .................................................................................................................. 72
A4.5 Evaluation Software Equalia ....................................................................................... 75
ANNEX 5: ADDITIONAL INFORMATION ON VALIDATION STUDIES IN APPLICATION SCENARIOS ......................................................................................................................... 76
A5.1 Introduction to the RAGE Application Scenarios ........................................................... 76
A5.2 Relevant Related Work and Evaluation Approaches .................................................... 76
A5.2.1 Evaluation of Computer Games and Serious Games .............................................. 76
A5.2.2 Cost-Benefit Analysis ............................................................................................... 78
A5.3 Elaboration of Evaluation Questions ............................................................................ 79
A5.4 Instruments .................................................................................................................. 81
A5.4.1 Instruments for Evaluation of Educational Effectiveness ...................................... 81
A5.4.2 Evaluation Asset ....................................................................................................... 84
LIST OF FIGURES

Figure 1: RAGE evaluation phases aligned with the technology development process ........................................ 9
Figure 2: RAGE evaluation model .............................................................................................................. 16
Figure 3: Evaluation levels on game development and assets ...................................................................... 19
Figure 4: Interaction components and evaluation topics in the evaluation of the Ecosystem ................................. 23
Figure 5: Evaluation levels on the Ecosystem ......................................................................................... 24
Figure 6: Evaluation levels on the applied games for the RAGE use cases ................................................... 29
Figure 7: Evaluation dimensions in RAGE ................................................................................................. 46
Figure 8: Steps of the evaluation process .................................................................................................. 48

LIST OF TABLES

Table 1: Overview of evaluation questions and variables on game development and assets and addressed stakeholder groups ........................................................................................................ 17
Table 2: Overview of evaluation questions and variables on the Ecosystem and addressed stakeholder groups ........................................................................................................................................... 22
Table 3: Overview of application scenarios and responsible partners .............................................................. 27
Table 4: Overview of evaluation questions and variables for the applied games in the RAGE use cases and addressed stakeholder groups ........................................................................................................ 28
Table 5: Research data lifecycle stages and how they map to the steps of the evaluation process ...................... 48
EXECUTIVE SUMMARY

This document presents the RAGE evaluation methodology. It provides the framework and accompanying guidelines for the evaluation and validation of the quality and effectiveness of the project outputs. Formative and summative evaluations of the different RAGE technologies and their underlying methodologies – the assets, the Ecosystem, and the applied games – will be carried out on the basis of this common framework.

Evaluation in RAGE addresses applied gaming in a holistic manner, taking into account the interests of different stakeholder groups. This means targeting the effectiveness of the developed applied games, as well as evaluating the process of developing them and the benefit that the RAGE technologies afford this process. The evaluation framework therefore constitutes a multi-perspective and holistic approach to systematic and comprehensive evaluation of applied games technology. Based on the specification of the evaluation objectives and questions for each strand of evaluation work and a review of existing methods, the evaluation approaches for RAGE are elaborated. This includes the identification of the planned types of evaluation studies, the description of their proposed research designs and envisaged instruments. An evaluation model integrating the different evaluation tasks and their perspectives on evaluation is presented.

The RAGE evaluation framework forms the basis for the analysis of the educational and cost effectiveness of the RAGE technologies, and for ensuring that they reflect the needs of their stakeholders. The assets will be evaluated in respect of their software quality and the added value they bring to the game development process. In addition, the relevance and benefit that assets provide through their pedagogical functionality shall be evaluated from an educational viewpoint. The Ecosystem will be assessed in terms of users’ reactions to the system, quality of resources provided, the capability of attracting relevant stakeholders, user contributions, and the added value for users. The evaluation of effectiveness of the games applying the assets and developed for the application scenarios will cover the analysis of the games’ effect on end-users’ perceptions, learning, skills acquisition and knowledge transfer. Unobtrusive data tracking and sensing through RAGE assets will complement more traditional instruments (such as questionnaires), thus providing a wide spectrum of evaluation methods capable of obtaining in-depth insights to players’ experiences. Furthermore, explorative and contextual cost-benefit analyses for game development and educational application of applied games will be undertaken for each of the six use cases.

This document also provides evaluation guidelines, which serve as a manual for implementing the evaluation approaches in the empirical work with users. In particular, these guidelines focus on relevant aspects of research data management to be taken into account when conducting evaluation studies. General procedures to be followed and tools and tips are described on how to implement requirements in terms of ethics, privacy, and open access in collecting, managing and sharing data in evaluation studies.

The RAGE evaluation framework and guidelines are the foundation for planning and organising scientifically sound, iterative, and mixed-method evaluations on the project outcomes and thus, serve as the common reference point for all evaluation studies carried out in the project. The evaluation approaches outlined in this document will be translated into concrete evaluation procedures and materials for the individual studies in the two main evaluation phases. The methods and outcomes of these formative and summative evaluation studies will be reported in detail in the RAGE evaluation reports (D8.3 in M33 and D8.4 in M47). The evaluation framework – in addition to its use in RAGE evaluation – aim at contributing to evaluation methodologies in the context of applied and serious games, on a wider scope, and may be (re-)used in and adapted for other applied gaming projects. In this way, the evaluation approaches, outcomes and experiences of RAGE may serve as best practice for future applied game evaluations.
1 INTRODUCTION

RAGE aims at fostering the adoption of digital game-based learning in game industry and in education. RAGE will provide a collection of reusable, and interoperable software components (‘gaming assets’ or simply ‘assets’) for game development. These assets will provide functionalities to undertake various data analyses, like competence assessment, emotion detection, comprehension measurement, or motivation identification. Another group of assets will enable game intelligence and adaptation, e.g. in terms of competence-based personalisation, natural language processing, motivational adaptation, cognitive interventions, and social gamification. The gaming assets will be provided via an online ‘Ecosystem’, which will also make available a broad range of literature and training material, as well as collaboration tools. The RAGE Ecosystem will therefore serve as a central access point and social affinity space. The RAGE technologies will be applied and tested in asset-based learning games (mobile and desktop implementations). These games will address different types of employability skills in the context of six application scenarios.

To demonstrate and seize the educational and motivational potential of applied games technologies and to foster their uptake meaningful quality and impact measurements are indispensable. Well-designed evaluation demonstrating the educational effect as well as the return on investment of applied games may foster broader adoption by educational institutions and training providers, and support the development of the applied games industry. Although there is a growing body of evidence on the efficacy of games for learning, evaluation is often poorly designed, incomplete, biased, if not entirely absent (e.g. Connolly, Boyle, MacArthur, Hainey, & Boyle, 2013). A critical aspect of research on the effectiveness of educational games is, in fact, how to approach and operationalize the measurement methodologies.

This document sets the overall scope and framework of evaluation work in RAGE. It provides a compendium outlining the overarching evaluation model and guidelines, and the collected approaches and methods of all evaluation tasks. The evaluations in their entirety aim at gathering a holistic understanding of the quality and effect of RAGE technologies for applied game development and application for education and training. This deliverable provides the general reference document for evaluation and is mainly targeted at the WP8 researchers conducting evaluations. In the course of planning the formative and summative evaluation studies the evaluation instruments and procedures will be further elaborated and will be provided in separate guidance and instruction documents to the partners involved in carrying out the individual evaluation tasks.

Aside from a systematic analysis of the games’ effectiveness for learning, which is traditionally done in serious game evaluations (e.g. Connolly et al., 2013), the broader benefit for training providers or educational organisations shall be taken into account in assessing empirical evidences. Focusing purely on the effectiveness of games, though, would not be sufficient for a comprehensive understanding of the added value of RAGE technologies for the game industry. Therefore, the underlying processes of using these new tools and methods for actual game development need to be targeted in evaluation, as well. The RAGE evaluation framework therefore takes the position of a broad evaluation of the whole process of applied gaming, from designing and developing games, using and sharing assets and resources via the RAGE Ecosystem, to applied game experiences, and to cost-benefit analysis of applied games.

This document is structured as follows: In section 2 the evaluation dimensions in RAGE are summarised. Section 3 provides guidelines on how to implement relevant aspects and procedures of research data management in the empirical evaluation work. Section 4 gives an overview of the overarching RAGE evaluation model, which integrates the different evaluation tasks and represents the holistic perspective on evaluation that is pursued in RAGE. Grounding on this, in sections 5, 6, and 7 the RAGE evaluation tasks with their evaluation questions and approaches are discussed in more detail – the evaluation of game development and assets (section 5), the evaluation of the Ecosystem services and processes (section 6), and the
validation studies in the application scenarios (section 7). Section 8 provides conclusions and an outlook to future work.
2 EVALUATION IN RAGE

The tools and methods that RAGE produces and will make available are of interest to the wider serious gaming research and industry communities, as they strive to improve the quality of serious games. To meaningfully evaluate these tools and methods and to ensure that they meet the needs of industrial and educational stakeholders a comprehensive and multi-perspective evaluation approach is required. Annex 1 further elaborates on the overall goals and challenges of evaluation work in RAGE and its research target framework.

2.1 Evaluation Phases

Technology evaluation, in general, is suggested to be oriented along the iterative design and development process of a system (Van Velsen, Van der Geest, Klaasen, & Steehouder, 2008). Evaluation may therefore be considered as a continuous process with a gradually shifting focus that is aligned with the current phase of design or implementation reached (Harvey Higgison, & Gunn, 2000). The three main phases of evaluation consist in requirements analysis, formative evaluation and summative evaluation (e.g. Gena & Weibelzahl 2007; Harvey et al., 2000). The requirements phase is the first in the system design process and is dedicated to identifying what users require or desire from a software system. Formative evaluation constitutes a preliminary evaluation phase usually carried out for a prototype of a system developed, in order to identify issues with the technology and to provide information for further modification and improvement. Summative evaluation refers to the final evaluation phase at the end of the development process. It aims at investigating the quality of the finished software product.

![Figure 1: RAGE evaluation phases aligned with the technology development process](image)

In RAGE this iterative approach of evaluation will be adopted aligned with the timing of the development cycles and releases of the RAGE technologies (see Figure 1). Formative and summative evaluations will be carried out successively but overlapping in time for the different types of technologies and alongside the development in the game projects for the RAGE use cases. A requirements analysis with game studios has been conducted in the context of WP1 (Saveski et al., 2015) early on in the project and is complemented by more specific need elicitations in the context of the development work packages. Very initial formative evaluation data related to game development, for example on perceived usefulness and expected benefits of assets, may be collected already at an early development stage (e.g. Davis & Venkatesh, 2004). Such preprototype evaluations incorporate participatory design ideas (Danielsson & Wiberg, 2006); they provide the possibility of gathering first evidence and validation of the significance of technologies and may provide indications for potential refinements in design and
functionalities, thus complementing to requirements analysis. Actual formative evaluations of the Ecosystem and the assets will be carried out after the first versions and bundles of components are available. The formative evaluation phase on the applied games and their deployment for training in the application scenarios will start at the end of year two with the release of the initial versions of games. Formative evaluation outcomes will provide information of initial technology benefits (usability, appreciation, empirical evidence regarding assumptions) and may provide inspiration for further enhancement of asset use and functionality, applied games, and Ecosystem. Similar to the formative evaluation phase, summative evaluations on assets and Ecosystem in year 3 and 4 will precede those on the final game versions in the final project year. Summative evaluations aim at showing increased benefits or effects compared to formative evaluation results and shall provide a concluding assessment of the contributions to RAGE’s strategic goals.

2.2 Evaluation Objects

Evaluation object(s), also called ‘evaluand(s)’ constitute the entities that are to be evaluated, for example individual learning objects, entire courses, an educational program or component (e.g. Williams, 2006). For the design of an evaluation approach it is of course necessary to clarify the initial object of reference in order to be able to define the appropriate evaluation questions and to choose the appropriate evaluation methods.

In RAGE the main objects of evaluation are given by the different kinds of technologies and underlying methodologies developed within the project:

1) the gaming assets developed in WP2 and WP3
2) the Ecosystem developed in WP6
3) and the applied games developed in WP4 for the application scenarios of WP5

2.3 Stakeholders and User Groups

The stakeholders of an evaluation are the people interested in using the evaluation results. In case of RAGE the immediate stakeholders are the project partners working on and using the RAGE applied game technologies and the underlying methodologies. Results from the evaluation studies have different applications: they are used as feedback for improvements to be made on the game technologies; they serve for giving evidence of the benefit and value of the research and development accomplished in the project; they are used for and reported in scientific publications and presentation. The European Commission as well as the experts invited as project reviewers can also be understood as stakeholders, as they are interested in the achievements and impact of the project. Similarly, the project outcome and evaluation results of RAGE may be of relevance to other project initiatives and researchers in the field of applied games and gamification.

Another perspective is to consider the stakeholders of the RAGE developments, itself i.e. the different groups of people who will be affected by and will be using the technologies developed and provided during the project. These are given by:

- asset developers,
- game developers (including game studio managers as well as game engineers who actually implement games technology)
- training providers (educational providers, intermediary organisations), and
- end users (learners) in application scenarios coming from within or outside the project consortium. These stakeholder groups together represent the industrial, organisational, educational, and learning perspectives and interests on applied games.

The above-mentioned user groups are the main target groups of the evaluations studies conducted in RAGE. The RAGE technologies (i.e. evaluation objects) can be mapped to the relevant user groups addressed as target audience/users by the respective technologies:

- **Gaming assets** will be used by **game developers** for creating applied games.
- The **Ecosystem** shall be used by **asset developers**, to upload and share their own assets and documents or papers, and to access assets of others, as well as other
resources from the digital library and media archive. The Ecosystem will furthermore be used by game developers, to access and use the game assets as well as other kinds of resources. In addition, training providers are potential future users of the digital library and media archive contents provided by the Ecosystem.

- The applied games developed for the six use cases of RAGE will be used by training providers as part of their training measures or courses. The games will be provided to the learners/end users, who will actually play the developed games in these application scenarios.

## 2.4 Evaluation Instruments

A wide range of instruments for evaluation data collection is available (for an overview see e.g. Gena, 2005; Gena & Weibelzahl, 2007; Harvey, 1998; Niegemann, Hessel, Hochscheid-Mauel, Aslanski, Deimann & Kreuzberger, 2012). Among the most popular methods are: interviews (e.g. Sewell, 1999), focus groups (e.g. Steward, 2007), questionnaires, observations (e.g. Bakeman & Gottman, 1986), think alouds (e.g. Someren, Barnard, & Sandberg, 1994), interaction logs, expert reviews, and cost analyses (e.g. Sewell & Marczuk, 2004).

Well-designed evaluations are methodologically eclectic – i.e. incorporating a mix of quantitative and qualitative evaluation methods. Triangulation through a mix of qualitative and quantitative methods (methodological triangulation), and data collection from multiple sources (data triangulation, e.g. collecting feedback from learners as well as from training providers) in general helps to cross-check results and to enhance the rigour and validity of an evaluation.

The evaluation studies in RAGE will use a mix of different evaluation instruments to gather evaluation data. These will be selected and developed in the course of the actual planning of the individual evaluation studies. Instruments will be chosen in line with the individual evaluation task, the evaluation questions defined and the related evaluation variables (envisaged evaluation instruments for each evaluation task are elaborated in Annexes A3.4, A4.4, A5.4). Shared instruments within evaluation tasks and across pilots will be strived for, as possible and appropriate.

The evaluation instruments will be made available in manuals and guidance documents which will also provide further information on the evaluation procedures and the use of instrument and step by step guides for evaluation tasks and studies.
3 RESEARCH DATA MANAGEMENT AND EVALUATION GUIDELINES

RAGE evaluation studies have to conform to principles of ethics, privacy and data security, and open access. Addressing these principles at the design stage may prevent problems later on during execution. In addition, RAGE has been designated a Horizon2020 pilot project on open access to research data, implying that not only RAGE journal articles based on our evaluation studies have to be openly accessible, but also (part of) the underlying data.

In this chapter we summarize ethics-, privacy- and open access requirements relevant in the context of the evaluations, and provide practical guidelines on the design and conduct of evaluation studies within RAGE. The requirements are further elaborated in Annex 2. The listing is definitely not exhaustive and when in doubt consult the RAGE national contact, your institutional Ethics Committee or data protection officer.

3.1 Roles and Responsibilities in RAGE Evaluation

Research data management in RAGE follows a decentralized approach with localized responsibilities and procedures to address institutional, national and EU-guidelines and legislation on ethics, data protection, and open access. The local partners will ensure compliance of local data collection, processing, and storage with institutional and national regulations and procedures on privacy, ethics and data protection. They take over responsibility that the ethics, privacy and open access aspects and procedures described in this chapter and the data management plan (D10.2) are adhered to. TUGraz as WP8 leader will provide assistance in realising this in the context of the evaluation studies, for example by providing templates for the letters of information and consent forms, hosting a secure online survey service for online data surveys, providing brief manuals on data anonymization, providing an interface to a secure open research data repository etc.

3.2 Checklist for Data Management in Evaluation Studies

This section provides a checklist of the most important things to consider related to ethics, privacy, and open access. It shall assist organising and conducting evaluation studies and collaboration between partners and people involved in the research process.

- Carefully plan data collection for your evaluation study – purpose, audiences/subjects, kind of data, instruments, evaluation procedure
- Ensure compliance with, and when necessary notify or gain approval, with institutional and national regulations
- Assign roles and responsibilities in the data management process (see Annex A2.2.3.4 Data Controller and Data Processor)
- Collect personal data only if really needed (see section A2.2.1.2 Provisions when Processing Personal Data)
- In case of collecting personal data make sure to involve the relevant institutional infrastructure and responsibilities to implement, monitor and support the adoption of research ethics, data security, and open access in RAGE (see Annexes A2.2.3 Support and Monitoring Structure and A2.5 National Contacts & Supervisory Authorities on Data Protection & Privacy)
- Inform evaluation participants about the purpose of the evaluation study (information letter; see section 3.3.2.1 and Annex A2.3 Information Letter and Consent Form)
- Collect participants informed, specific and freely given consent to participate in evaluations (consent form; see section 3.3.2.1 and Annex A2.3 Information Letter and...
3.3 Tools & Tips

This section provides tools and tips for research data management along the steps of evaluation and shall serve as guidelines for collecting, managing, preserving, and sharing your data in the context of the RAGE evaluation studies (and beyond).

3.3.1 Planning Evaluation – Data Management Planning

In planning an evaluation study the evaluation methodology is set up and considerations of how the data will be managed during and after the evaluation process are made. This includes making arrangements for a clear and detailed documentation of the data, the planning of recruitment methods for participants, and the preparation of the information letter and consent form. It may be useful to consider these in advance when designing your evaluation research.

3.3.1.1 Planning Data Documentation

An appropriate documentation of the data should already be taken into account when designing an evaluation study. This will facilitate the management, curation, and sharing of data later on. For a good documentation of data the following information should be planned to be documented for each evaluation study (Van den Eynden et al., 2009, 2011):

- context of data collection
- data collection methods
- dataset structure
- data sources
- data validation, checking, proofing, cleaning procedures applied
- modifications to data over time
- information on access and use conditions or data confidentiality

3.3.1.2 Recruiting Evaluation Participants

Recruitment refers to the process in which the evaluator/researcher identifies and invites participants to join the study. The goal will be to include participants who will provide appropriate and rich source of data relevant to the evaluation questions being investigated. For this purpose ‘inclusion’ and ‘exclusion’ criteria should be defined, who will be approached – e.g. age group, geographic location, language, background) (Eide, 2008). Furthermore, recruitment methods/strategies are to be defined – e.g. online, newspaper, or online advertising, referral, word of mouth (e.g. Krusche et al., 2014). Especially when aiming at randomised controlled trials, the definition of recruitment criteria and methods is important. At this stage also the decision and planning on the use of recruitment incentives need to taken. Incentives are bonuses given to participants to boost enrolment and participation in studies. In particular, it needs to be checked with institutional regulations and ethics committee whether (and what kind) of incentives for participation are allowed.
3.3.1.3 RAGE Information Letter and Consent Form

Participants need to be informed about the purpose of the data collection and agree to participate in an evaluation study. For this purpose, an information letter and consent form need to be prepared. As a basis for that, the template included in Annex A2.3 can be used. Annex A2.3 provides an information letter and consent form template with standard text (general information about RAGE, privacy note, use of data). This text needs to be further detailed and elaborated (and translated, if necessary) for each evaluation study, describing the study purpose and procedure.

Alternatively, new information letters and consent forms may be set up by the researchers. These need to include the following information:
- Purpose of the study
- What will be researched
- How will the research be executed and how the data are to be collected.
- What is expected from the participant in participating in the study (duration, tasks)
- Possible advantages, disadvantages and other implications of participating in the study
- How the data will be used and managed
- Underline the voluntary nature of participation and that consent can be withdrawn at any time
- Information about the responsible research institution and contact person
- The data subject’s rights

Information letters and consent forms will be prepared and administered either online, if possible, or physically. Online consent may help to make participants feeling less pressure to take part in a study, so that they are more likely to participate freely (Eynon, Fry, & Schröder, 2008); besides it reduces evaluation load for RAGE researchers in collecting and managing consent forms. Electronic/online consent, in general, may be obtained in low-risk studies, i.e. if the risks to subjects are low and if subjects are aged 18 or older (Bruckman, 2002). Online consent forms should thereby take participants through each sub-element in a step-by-step manner. The decision on online or physical consent collection is taken for each evaluation study in RAGE individually in the course of planning and specifying the procedure.

In any case, local partners need to verify the text of information letter and consent form with national legislation and translate them for the pilot target audiences, as needed.

3.3.2 Carrying Out Evaluation – Collecting Data

In the phase of actually carrying out an evaluation study and engaging with evaluation participants, the collection of participants’ consent, as well as the proper use of data collection instruments need to be ensured.

3.3.2.1 Informed Consent

Before collecting data form evaluation participants their informed consent needs to be obtained. The consent as a legal basis for processing of personal data must be free, informed and specific and will be collected by using the information letter and consent form prepared for the evaluation study.

The main elements of gathering informed consent are the following:
- The participant must have been under no pressure when consenting and must have been given sufficient information in order to be able to make a choice of whether or not to participate
- The decision of the participant on the consent issue must be evidenced. A copy of the information sheet and signed consent form should be given to the signee.

3.3.2.2 How to Use Online Surveys

The use of online surveys should meet the same ethical, privacy and data security requirements as other data processing methods and tools. Also with online survey the subject should be
informed on the purpose of the survey (equivalent of the Information Letter) explicitly agree on participation (recorded acceptance of the Consent Form).

In case subjects are surveyed anonymously, online (free) survey tools like SurveyMonkey may be used. Take care however that your collected data do not unintentionally get ‘personalized’ by having respondents create an online identity that allows their identification. In case personal data are collected, privacy and data security should be guaranteed. In practice this will exclude many cloud tools as their data security does not meet EU privacy standards. Collecting personal data though online surveys will thus require a secure survey infrastructure, possibly to be hosted by TUGraz. One reason to collect personal data in surveys is to allow repetitive (longitudinal) data collection from the same subject. A possible work around is to have the subject manage his/her own identity, i.e. create and login with a self-managed username (and password). In case the identity is linked to a mail address (which may be necessary to notify the subject of follow-ups), the subject should be warned not to use a mail address that allows easy identification.

3.3.2.3 Audio and Video Recordings
Audio and video recordings almost by definition are ‘personal data’. Anonymization almost invariably implies transcription, which may be fairly labour intensive. Therefore carefully consider alternatives before recording. Possible alternatives for data collection and recording may be screen recordings and written records of observations or interviews (using behaviour observation system or checklists, interview guides or questionnaires as a basis for note taking, or by involving assistant evaluators).

3.3.2.4 Blind Experiments
In case of a blind experiment, the use of an Information Letter and Consent Form explaining the true purpose of the research may not be possible. In such a case the subject should a) be informed as soon as possible about the true purpose of the experiment after its completion (Information Letter), and b) be allowed to withdraw as yet (Consent Form). It is good practice to always submit blind experiments to an Ethics Committee!

3.3.3 Working with the Results – Processing, Storing, and Sharing Data

3.3.3.1 Organising Data Files
Having evaluation data stored in a well-organised and consistent structure of file names and folders facilitates finding and keeping track of data files. File names should provide meaningful indicators to the content and status of the file – they may contain Project acronyms, researchers’ initials, file type information, version number, date (Van der Eynde et al., 2011).

A best practice approach to organising files and folder consists in
- Using meaningful but short names
- Using file names for classification of broad file types
- Avoiding spaces and special characters
- Avoiding very long names

3.3.3.2 How to Anonymize

True anonymization
Simply delete all references to the subject’s identity. In case of electronic data, a ‘soft delete’ may be insufficient! The subject should be informed beforehand (preferably through the Consent Form) at what moment the personal data will be deleted, as at that moment the subject’s ‘access right’ to the data terminates.

Pseudonymization
This typically involves the following steps:
1. In the original data set create an identifier for each subject
2. Create an identifier file with the same identifiers
3. Copy the personal details from the original data set to the identifier file

4. Delete the personal data from the original data set (a ‘soft delete’ may be insufficient!)

The pseudonymized data set may now be submitted to an OAR while the identifier file is kept by the data controller. Note that the data subjects in principle retain their access rights as they are still ‘linked’ to their data set, but it should be clear to them (through the Consent Form) that any publications based on their data after pseudonymization cannot be retracted.
4 RAGE EVALUATION MODEL

Three main evaluation tasks are carried out in the context of WP8 and are covered by the RAGE evaluation framework: evaluation of game development and assets (T8.3, see section 5), evaluation of the Ecosystem (T8.4, section 6), and evaluation of the applied games in the application scenarios (T8.5, see section 7). In each evaluation task research questions, which can be linked to different evaluation levels will be addressed. This levelled approach has been inspired by Kirkpatrick’s four-level model of training evaluation (Kirkpatrick, 1976; Kirkpatrick & Kirkpatrick, 2009). The evaluation levels of all evaluation strands can be integrated to form one combined and overarching evaluation model, as illustrated in Figure 2. This model represents the holistic evaluation approach pursued in RAGE, aiming at gaining a comprehensive understanding of the quality and added value of the applied gaming methodologies and technologies developed in the project for the different target stakeholder groups.

![Figure 2: RAGE evaluation model](image)

The evaluation model forms the basis for analysing the effectiveness of the RAGE technologies, and for ensuring that they reflect the needs of industrial and educational stakeholders. The considerations and approaches on each of the evaluation tasks and levels are described in more detail in the subsequent sections.
5 EVALUATION OF GAME DEVELOPMENT & ASSETS

5.1 Evaluation Objectives and Questions

The assets developed and made available in RAGE provide reusable software components for supporting the creation of applied games (for an overview of the RAGE assets see Annex A3.1). The objective of T8.3 ‘Evaluation of game development and assets’ is to evaluate the application of RAGE assets (i.e. WP2 and WP3 outcomes) as part of game development (WP4). This concerns the questions on whether and how the game assets are able to support applied game development. The assets provide functionalities for adding pedagogical value to a game, which can be categorised as follows:

- Different kinds of analytics of game-based user data (WP2 assets)
- Storage and reporting of analytics data and results (WP2 assets)
- Adaptation of the game experience to the player’s profile (WP3 assets)
- Improvement of socio-emotional believability (WP3 assets)

A main challenge in the definition of the evaluation methodology for game development and assets consists in the diversity of assets and the functionalities they provide. The goal in the context of the evaluation framework is to define an overall approach for this strand of evaluation that captures all assets. This shall be achieved by identifying generic evaluation questions that are applicable for all of them. In addition, evaluation questions that address specific types of functionality provided by different groups of assets are incorporated.

<table>
<thead>
<tr>
<th>ID</th>
<th>Evaluation Variable</th>
<th>Evaluation Question</th>
<th>Evaluation Object</th>
<th>Stakeholder Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>Usability</td>
<td>Are the assets usable?</td>
<td>All assets</td>
<td>Game developers, training providers</td>
</tr>
<tr>
<td>A2</td>
<td>Usefulness</td>
<td>Do the assets enhance game development?</td>
<td>All assets</td>
<td>Game developers</td>
</tr>
<tr>
<td>A3</td>
<td>Relevance</td>
<td>How relevant are the assets for the RAGE application scenarios?</td>
<td>All assets</td>
<td>Game developers, training providers</td>
</tr>
<tr>
<td>A4</td>
<td>Game engineering</td>
<td>Can assets easily be included in game development?</td>
<td>All assets</td>
<td>Game developers</td>
</tr>
<tr>
<td>A5</td>
<td>Benefit</td>
<td>Which benefits do the assets bring for applied games?</td>
<td>All assets</td>
<td>Game developers, training providers</td>
</tr>
<tr>
<td></td>
<td>General evaluation questions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A6</td>
<td>Validity of player modelling</td>
<td>Do the assets successfully infer user characteristics?</td>
<td>WP2 analytics assets</td>
<td>Players</td>
</tr>
<tr>
<td>A7</td>
<td>Perception of dashboard</td>
<td>How do stakeholders value the externalisation and presentation of game-based user data?</td>
<td>WP2 dashboard asset(s)</td>
<td>Players, training providers, game developers</td>
</tr>
<tr>
<td>A8</td>
<td>Appropriateness of adaptation decision making</td>
<td>Are adaptation decisions of assets appropriate and meaningful?</td>
<td>WP3 adaptation assets</td>
<td>Players, training providers, game developers</td>
</tr>
<tr>
<td>A9</td>
<td>Social believability</td>
<td>Are game experience and non-player characters believable?</td>
<td>WP3 assets for enhancing social believability</td>
<td>Players, training providers, game developers</td>
</tr>
</tbody>
</table>
Table 1 provides an overview of the evaluation questions and variables targeted in the evaluation studies on game development and assets. A more detailed elaboration of the evaluation questions is provided in Annex A3.3 Elaboration of Evaluation Questions. For each evaluation variable the respective evaluation objects as well as main stakeholder groups for evaluation are indicated in the table. Especially in the context of formative evaluations these questions may be complemented by additional, detailed questions on individual assets in order to gather specific information for further improvement and development.

5.2 RAGE Evaluation Approach

Evaluating game development and assets aims at showing that the provided game assets are useful means supporting game developers in the creation of games for education and training and, in the end, that they render the serious game market more attractive. Relevant related work on this evaluation task is summarised in Annex 3.2. While for a meaningful evaluation of the Ecosystem (its software repository) the availability of an appropriate set of assets and other resources provided via the Ecosystem is a prerequisite, the evaluation of game development and assets (in the context of T8.3) may be carried out independently of the Ecosystem.

Quality assurance mechanisms on all individual assets (e.g. in the sense of proper design and documentation), as well as testing individual assets, are part of the work on asset development in WP2 and WP3. In the context of WP8 evaluations on a significant subset of assets and targeting relevant evaluation and research questions related to them will be carried out. Apart from these evaluations, the validation studies in the context of the application scenarios (compare section 7) will cover implicitly the evaluation of all assets integrated in an applied game with respect to its contribution to the cost effectiveness for game development as well as to educational effectiveness. An evaluation study in the context of T8.3 will address a small group of assets (e.g. assets from one partner), which constitute meaningful combinations of functionality while – as far as possible – still allowing to trace back identified effects to individual assets. The assets addressed within an evaluation study will be defined in accordance with the evaluation questions and in consultation with responsible development partners, as well as drawing from dependencies between assets, as outlined in the RAGE asset catalogue.

From the overview of evaluation questions in section 5.1 it can be seen that evaluation of game development and assets covers several levels. These levels are illustrated in Figure 3. The first level of evaluation relates to the reaction of users to assets – in terms of perceptions of usability and usefulness – as covered by the general evaluation questions A1 and A2 in Table 1. While data on this level may be gathered to a large extent by standard and customised survey instruments, evaluation at higher levels requires more extensive data collection and intensive dialogue with game developers and other stakeholders. These higher levels focus on the analysis of the actual impact of the technology use on game development. Evaluation level 2 relates to the relevance of an asset for the game development projects in the specific use cases (A3). Level 3 refers to the perception of technical integration from a game engineering perspective (A4). Level 4 addresses the benefit in terms of the perceived value of the pedagogical functionality provided and the added value for the game development process (A5). Comparisons of formative and summative evaluation results (on A1-A5) will be done for assets, if possible. Systematic comparative analysis between RAGE assets are not intended, due to the diversity of function provided by the different assets. More reasonable appears an analysis whether or to what extent RAGE assets are better in comparison to existing commercial assets providing similar functionality (if available); this shall be considered in the context of evaluation level 4. This level also includes asset-specific questions (A6-A9), which are connected to the actual learning game experience enhanced by the respective assets and aim at validating the core pedagogical function and value of those assets (as a basis for a broader evaluation of applied games).

The collection of evaluation data will to a large extent be contextualised by the game development for the application scenarios. The actual use of an asset for the game projects in the use cases is a positive indicator for the significance of this asset for game development.
Conversely, however, not using an asset in the context of the application scenarios might just mean that it does not fit the concrete objectives and design of the respective use case and game, but does not necessarily mean that stakeholders have a negative opinion of it, in general, and that they do not intend to use this asset in future game development projects. It appears therefore reasonable to address the evaluation questions independent of the assets’ application in the use case games. This would entail collecting feedback not only from those game developers and training providers making use of a given asset, but also from those people who do not currently apply it. In addition to evaluations in the context of the use cases therefore also task-based evaluation workshops and specific evaluation studies that may be independent of a particular use case game will be conducted. These will allow examining in more detail specific evaluation questions on groups of assets.

![Figure 3: Evaluation levels on game development and assets](image)

### 5.2.1 Participants

Quite naturally, the main target audience to be addressed as participants in the evaluations of game development and assets are **game developers**. Only participants with expertise and experience in game creation will be able to comprehensively provide feedback about the quality and value of the assets. Game industry partners from the project consortium will therefore be involved as participants. This will also allow repeated measurements from formative to summative evaluation studies and thus, the identification of changes in game developers’ assessment of the game assets. The sample of game developers is intended to be extended in the course of the project by involving also game developers outside of the project and from external (serious) game companies. External game developers will be contacted and recruited via the RAGE partner network, similar as done in the context of the initial requirements analysis carried out by WP1 (cf. Ljubicic Saveski et al., 2015).

Apart from game developers, in a narrower sense, **asset developers** from within the project may be recruited, especially in the formative round of evaluations. Although not directly coming from a game industry background, they have an understanding of game technologies and assets for game creation and may be able to assess and comment on assets of others.

In addition to game developers, **training providers** (i.e. WP5 partners and/or members/trainers of their associated institutions) will be able to provide valuable feedback on some of the evaluation questions, like the assets’ benefit in terms of pedagogical function, the relevance of assets for their particular use case, their usability (if an asset is also used by training providers, e.g. for entering content), as well as asset specific questions (like social believability of game characters). Evaluation data on these questions may be collected without reference to high technical knowledge/detail and game development expertise. Training providers will provide feedback on the assets by taking the perspective of the application of games for training and the significance of assets’ functionality for educational practice. Similar to the sample of game developers, the sample of training providers involved in evaluations shall be gradually extended to external training and educational organisations, as possible.
To investigate some asset-specific evaluation questions, data from actual interaction of learners with games (with integrated assets) are required. The involvement of end users will be done in alignment with the real-world pilots from the application scenarios of WP5, but may in exceptional cases also be achieved by recruiting users from different appropriate populations.

5.2.2 Research Design

Below an overview of the different types of evaluations that are envisaged in the evaluation of game development and assets is given. The evaluations will be planned and carried out by WP8 partners in consultation with asset developers. A description of evaluation instruments envisaged to be applied in these evaluation studies is given in Annex 3.4.

5.2.2.1 Evaluations in the Context of Use Cases

Evaluation studies in the context of the RAGE use cases mean that evaluation data collection is embedded in the game creation process carried out for the different application scenarios. The game projects for the application scenarios entail developers’ concrete experience of and with assets as part of their game design and development work. It will therefore be possible to gather in-depth feedback from game developers based on real-world game development scenarios and tasks. The subset of assets to be evaluated in the context of a use case naturally results from the assets used by game developers in the games. These evaluations will be carried out in an iterative manner aligned with the two main evaluation phases of the applied games.

This kind of evaluation will be used for investigating aspects of general reaction (A1-A2 – usability, usefulness) and relevance (A3) of assets (since this variable explicitly links to the application scenarios), as well as benefit (A5). In particular, this strand of evaluations shall also gather data on the game engineering perspective of integrating assets with game development procedures (A4). Feedback on game engineering and the benefit for game development may be elicited by requesting game developers to compare the game development integrating RAGE assets with their usual game development process, without the respective asset(s) as a benchmark. This needs to take into account that such traditional or previously applied game development approach may include the use of other reusable assets/software. It also needs to take into account that RAGE assets mean a new development approach for game development companies, and that doing things differently might require some change of processes, cost time, etc. Therefore, a comparison on the differences in the respective development approaches could be part of this evaluation. This also links up to planned the cost-benefit analyses of applied games projects for the application scenarios (see section 7.2.2). Assessment on relevance and benefit will additionally be gathered from training providers.

5.2.2.2 Training and Evaluation Workshops

Especially organized workshops presenting/demonstrating assets to game developers and providing training on their application and hands on experience will be used as a basis for collecting evaluation data on game development and assets. Evaluation participants will mainly be game developers not already using the assets in question in their own game projects. The evaluation workshops are thus considered as a complement to the evaluations in the context of the use cases and may also serve as dissemination instrument to promote asset application. They will allow a more complete coverage of assets and a larger data set of feedback. The training and evaluation workshops will feature much more controlled conditions compared to the evaluations in the context of the use cases. Asset use, however, will necessarily be much more artificial and results may not be straightforwardly generalizable to real-world game development.

Such training and evaluation workshops will be carried out for assets forming meaningful combinations of functionality (for example assets for uncovering the current motivational state of a learner and for adapting the game experience to it). Workshops will be organised and conducted by the partners who have developed the assets (if the asset developer is not a WP8 partner, this will be done in close collaboration between asset developers and evaluation partners). The workshops will be in a face-to-face format, but may potentially also be conducted
at distance in form of webinars or virtual meetings (with trade-offs in terms of evaluation instruments that can be used and awareness and controllability of the evaluation setting).

The general procedure of such evaluation workshops will be as follows: After the presentation and demonstration of the assets, participants will try out and work with the targeted assets in a task-based manner. The tasks will consist of small game creation and asset application exercises specifically prepared for the evaluation study. The tasks will be the same for all participants in a given evaluation study and will largely be independent of the concrete use cases of WP5. While working on the tasks, asset experts will provide assistance in case of questions or technical issues. After working on the tasks (in case of an on-site workshop the duration will be the same for all participants), evaluation instruments are administered and assessments and feedback from game developers will be elicited. In this manner, the general reaction to assets (A1-A3) and their benefits for game development and in terms of pedagogical value may be investigated (A5). Depending on the assets addressed, it may also be possible to gather expert feedback on some asset-specific evaluation questions (A6-A9).

Early or pre-prototype evaluations may do without demonstration and hands-on session and may be carried out based on the asset descriptions only (potentially supported by available demos), to gather preliminary feedback on usefulness (A2), relevance (A3), and (expected) benefit (A5) from game developers as well as training providers (cf. Davis & Venkatesh, 2004).

5.2.2.3 Asset-specific Evaluation Studies

Asset-specific evaluation studies will provide an opportunity for investigating asset-specific evaluation questions (A6-A9) that more closely focus on game experience. These evaluations will be carried out for selected assets and for evaluation and research questions of particular interest and relevance in asset and game development. The planning of these studies is done based on a careful analysis and balancing of scientific interests and practical feasibility. The studies will be pursued/pushed by the asset developers and owners and will be carried out in WP8, in close collaboration between evaluation partners and asset developers. Asset-specific evaluation studies aim at examining assets based on concrete game experiences in the context of applied games, with the aim of demonstrating the functionality of assets and validating the pedagogical functionality they provide. Aside from data based on actual end-user experiences, expert judgements from training providers or game developers may also be useful to gather.

The asset-specific questions presented in Table 1 relate to the validation of the actual added value of asset usage on the game and resulting gaming/learning experience. For their evaluation the assets need to be integrated in a game, to make the functionality/effect of the asset tangible. This may be achieved in the context of the applied games for the use cases, but also by the integration in a different or example game (or game part) for demo and evaluation purposes. An evaluation of the perception of the dashboard assets (A7) may even be done in a more isolated manner (without full game integration) based on mock-up data.

Evaluations of this kind are designed in order to clearly trace back outcomes to individual assets. This is in contrast to the evaluation studies in the context of the application scenarios (see section 7), where a game uses a range of different assets, which means that any effects identified may result from the combination of different assets and probably cannot be traced back to individual assets. Asset-specific evaluation studies represent those with the most controlled conditions (compared to the evaluations in the context of use cases and training and evaluation workshops outlined above). For systematic analysis, comparative/control group study designs will be used, where appropriate – for instance for investigating the appropriateness of adaptation decision making (A8) in comparison to a game without adaptation or for investigating the social believability (A9) of game characters implemented with and without the use of assets.
6 EVALUATION OF ECOSYSTEM SERVICES AND PROCESSES

6.1 Evaluation Objectives and Questions

The RAGE Ecosystem provides a central access point to comprehensive functionality and resources for stakeholders in an applied gaming context – software repository, digital library and media archive, training, community/collaboration tools and support to upload and manage content objects or software (for an overview of the Ecosystem see Annex A4.1). The work in WP8 Task 8.3 ‘Evaluation of Ecosystem services and processes’ consists in conducting evaluation studies on the Ecosystem, its services and processes, to investigate their quality and impact in the field of applied gaming. Since the involvement of stakeholders is considered critical for the success of the Ecosystem, users’ attitudes towards, perceived benefit, and actual usage of the software and resource repositories and social system provided by the Ecosystem shall be analysed in detail and with different user groups. Beside users opinion towards the system, successes should also be evaluated in terms of the Ecosystems capability of actually attracting relevant stakeholder groups.

Table 2 gives an overview of the evaluation questions and variables for the evaluation of the Ecosystem. Further description of the evaluation questions can be found in Annex A4.3 Elaboration of Evaluation Questions. The set of evaluation questions is open-ended and may be complemented by specific questions arising in the course of the iterative Ecosystem development (e.g. acceptability of a certain format for bibliography management), to collect focused feedback relevant for further refinement and development of Ecosystem functionality.

<table>
<thead>
<tr>
<th>ID</th>
<th>Evaluation Variable</th>
<th>Evaluation Question</th>
<th>Stakeholder Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>Usability</td>
<td>Is the Ecosystem usable for the targeted user groups?</td>
<td>Game developers, asset developers, training providers</td>
</tr>
<tr>
<td>E2</td>
<td>User acceptance</td>
<td>Do users find the Ecosystem acceptable and intend to use it?</td>
<td>Game developers, asset developers, training providers</td>
</tr>
<tr>
<td>E3</td>
<td>Resource quality</td>
<td>Do the resources provided serve the needs of the users?</td>
<td>Game developers, asset developers, training providers</td>
</tr>
<tr>
<td>E4</td>
<td>Performance</td>
<td>Does the Ecosystem allow to effectively and efficiently search, manage, and use resources?</td>
<td>Game developers, asset developers, training providers</td>
</tr>
<tr>
<td>E5</td>
<td>User community involvement and support</td>
<td>Does the Ecosystem attract and support targeted user communities?</td>
<td>Game developers, asset developers, training providers</td>
</tr>
<tr>
<td>E6</td>
<td>User contribution</td>
<td>Do the stakeholders contribute resources to the Ecosystem?</td>
<td>Game developers, asset developers, training providers</td>
</tr>
<tr>
<td>E7</td>
<td>Added value</td>
<td>What added value do users experience from the Ecosystem?</td>
<td>Game developers, asset developers, training providers</td>
</tr>
</tbody>
</table>

6.2 RAGE Evaluation Approach

The evaluation of the RAGE Ecosystem is operationalized based on the consideration of the relevant interaction components involved – users (game and asset developers, training providers), system (the Ecosystem with its different repositories and tools), and content (the
different kinds of resources made available – assets, papers, training material). This interaction triangle has been used previously in an evaluation model and related evaluations in the field of digital libraries and VRE (Fuhr et al., 2007; Steiner et al., 2014; Tsakonas, Kapidakis, & Papatheodorou, 2004). This evaluation model identifies evaluation variables related to the interaction axes between users, system, and repository content: usability (system-user axis), usefulness (content-user axis), and performance (system-content axis). For a more general overview of relevant related work and evaluation approaches, please refer to Annex A4.2.

For the purpose of RAGE the CULTURA evaluation model (Steiner et al., 2013, 2014) is adopted and further elaborated to form the conceptual basis for the Ecosystem evaluation. The evaluation questions presented in Table 2 reflect the topics of this existing evaluation model and can be mapped to the interaction axes (see Figure 4 for an illustration). Usability (E1) and user acceptance (E2) are located on the axis between system and user (in line with Steiner et al., 2013, 2014). Resource quality (E3) matches with the concept of usefulness from the original Triptych model (Tsakonas et al., 2004) on the content-user axis, thus referring to the usefulness of the resources/content for users. Performance (E4), which is operationalized for the Ecosystem evaluation in terms of response time, relevance and accuracy of system operations on contents (i.e. mainly the semantic services), is located on the system-content axis and thus, also correspond to the original digital library evaluation model. User community involvement and support (E5), user contribution (E6) and added value (E7) address the wider impact and social dimension of the Ecosystem and constitute new topics of evaluation. These evaluation questions involve the interlinkage of all three components of the interaction triangle and focus very much on the actual interaction of users with both the Ecosystem and the content/resources provided. These evaluation questions are arranged on a circle surrounding the interaction triangle (see Figure 4).

When associated with levels of evaluation, the evaluation questions on the Ecosystem can be mapped to five levels (see Figure 5). Level 1 refers to users’ reaction in form of usability and user acceptance perceptions. Level 2 addresses the resource quality in terms of the usefulness of the asset collection and content objects provided. Level 3 addresses performance of system operations on the resources (the evaluation of which actually requires the availability of a relevant set of resources). Level 4 relates to social dimension, i.e. the evaluation of the involvement and support of user communities. On level 5 the evaluation topics on user contributions and the experienced added value can be located. Since the Ecosystem addresses several stakeholder groups as target users, participants from each of these groups should be involved in investigating the five evaluation levels.

A meaningful evaluation of most of the Ecosystem’s components is only possible with an appropriate collection of software assets, and digital library and media contents available.
becomes also clear when considering the interaction triangle between system, content, and users (see Figure 4). The availability of relevant resource collections is therefore a prerequisite for a comprehensive and significant Ecosystem evaluation, in order to enable users to actually test the functionality and interact with the system. The feedback gathered in the evaluation studies will therefore be tied to a certain extent to the resource collections provided. Only the results obtained for evaluation topics that address purely the interaction between system and user (i.e. evaluation questions E1 and E2) should be largely resource-agnostic.

The collection of evaluation data related to evaluation levels 1-3 (as shown in Figure 5) may be carried out largely independent of the concrete use case scenarios of RAGE and can potentially be covered by special evaluation workshops and studies. Evaluation of user involvement and contributions, as well as an analysis of the perceived added value should be embedded in more realistic or real-life settings of creating and using applied games resources. Such usage settings may refer to the game projects for the RAGE application scenarios, but may also address the wider use of the Ecosystem by project partners and external stakeholders.

![Figure 5: Evaluation levels on the Ecosystem](image)

**6.2.1 Phased Approach**

The evaluation approach for the Ecosystem will follow the iterative approach of the Ecosystem development. This means, in each iteration of the Ecosystem an emphasis will be put on the evaluation of the development focus and available system functionality from that phase. With the availability of the first deployment of the Ecosystem in M12 the basic functionality of the Ecosystem will be provided, allowing the collection of assets and knowledge resources (compare D6.1). This means, the first deployment of the Ecosystem will not come with any resources, but shall subsequently be populated by users with resources. As a result, the evaluation of the first Ecosystem version should naturally address the process of ingesting content objects to the digital library and media archive and, respectively, assets to the software repository. This means, the reaction to (i.e. usability aspects and acceptance; E1, E2) and performance (E4) of features for uploading, importing and managing resources may be assessed, as well as the process and types of user contributions (E6).

Once the Ecosystem is equipped with resources, also the features for exploring and searching the repository and archive (E1, E2, E4), as well as resource quality (E3) can be evaluated, and findings can be fed back into Ecosystem development for further enhancement of user experience. With the integration of semantic annotation services and collaboration tools at a later stage it shall be evaluated how these features further support the search and use of resources (E1, E2, E4). Actual involvement of user communities (E5), as well as user contributions (E6) and added value (E7) will be evaluated in more detail on a longer term and with the availability of a more complete and populated Ecosystem version.

**6.2.2 Participants**

The RAGE Ecosystem mainly targets game developers, asset developers, as well a training providers as potential users. The Ecosystem will create a social space allowing all user groups
to connect and collaborate with other users. **Training providers** are supposed to primarily use the digital library and media archive to search and access information on applied gaming and the application of games in education. **Game developers** are assumed to make more comprehensive usage of the resources provided by the Ecosystem – apart from taking advantage from the offered media and training resources, they will make use of the software repository to retrieve and use assets for their ongoing game development projects. **Asset developers** are primarily assumed to make available their assets to the software repository and documentations or papers on these assets to the digital library. Upload and management of content objects is envisaged to be done by game developers and training providers, as well. The user groups of game developers and asset developers potentially overlaps, i.e. game developers are likely to also create and share their own assets.

Aligned with the phase-centred agile development approach of the Ecosystem the evaluation studies will iteratively involve these user groups and communities and extend the respective samples. While the initial evaluation of the Ecosystem providing basic functionality will address primarily asset developers within the project (for populating the Ecosystem with resources), with the evolution of the Ecosystem and the collections of resources made available the sample of evaluation participants will be gradually extended to involve also game developers and training providers from the use cases. Eventually, external members of all user groups are expected to be part of the Ecosystem user communities and should be involved in evaluation attempts.

### 6.2.3 Research Design

To empirically investigate the evaluation questions presented in **Table 2** different types of user-centred evaluation studies will be carried out and are outlined below. An overview of the evaluation instruments envisaged to be used in these studies is given in Annex 4.4. The results obtained from these shall provide useful input for further development and help establishing a deep understanding of the quality, benefit, and (potential) impact of the RAGE Ecosystem.

#### 6.2.3.1 Ecosystem Evaluation Workshops

Ecosystem evaluation workshops are events especially organised for the purpose of collecting evaluation data on the evaluation questions presented in **Table 2**, and additional specific questions arising from the iterative development. Evaluation data collection will thereby occur in a more or less ‘artificial’ setting using a task based evaluation approach. The procedure of the workshop will usually include the presentation and demonstration of the Ecosystem and its functionality, with subsequent hands-on experience, after which participants will provide feedback on their opinion and perception of the system. Evaluation workshops will thereby likely be user group specific, i.e. targeting either game developers (with a more comprehensive coverage of system functionality), or training provider, or asset developers (focusing on features mainly of interest for these stakeholder groups).

This type of evaluation study may be covered in the scope of on-site half/full-day workshops. Users’ interaction with the system will happen under controlled conditions and will be of similar, but limited duration/extent for all participants. The focus will be on the evaluation of levels 1-3 shown in **Figure 5** (evaluation questions E1-E4). Direct feedback from users while using the system (observations, comments during the workshop) may be captured and there is the possibility of immediate response or support in case of any questions/technical issues arising.

Participants’ experience of the system will be guided by pre-defined tasks to work on – in in this way it shall be ensured that the functionalities and tools provided by the Ecosystem and addressed in a particular evaluation are used. A task-based evaluation approach presents virtual scenarios to work from with some concrete predefined tasks; these will be independent of the games developed for the RAGE use cases. Users’ experience in working on the task serves as a basis for collecting evaluation data. The tasks will be defined in line with the system functionalities and evaluation questions addressed in an evaluation. An example would be information retrieval tasks on finding papers or assets on a particular topic. The tasks will be derived based on the definition of personas, i.e. identifying and describing archetypical users, their background and their intent of using the Ecosystem (doing what, searching for...).
To enable a longer-term interaction with the system (probably without or with only initial tasks as a basis for usage) and thus, to target more closely evaluation levels 4 and 5 (E5-E7), workshops would be suitable that realise an initial face-to-face or webinar-like part introducing the Ecosystem, with a subsequent phase in which participants can use the system on their own in a self-directed manner. Such kind of workshop can also be seen as a dissemination instrument (WP9) aiming at fostering the adoption and use of the Ecosystem. Conversely, presentations and demo sessions on the Ecosystem at conferences may be potentially be used as opportunities for collecting evaluation feedback in terms of a very general level overall assessment (e.g. very short online survey). In addition, the general information and training events carried out by WP6 may be taken as an opportunity to gather user feedback on the Ecosystem. Since the training events of WP6 also aim at prompting the further use of the Ecosystem, thus providing further opportunity for the evaluation of user community involvement (E5), user contributions (E6), and added value (E7) at a later stage.

6.2.3.2 Evaluations of Semantic Annotation Services

Specific validation scenarios are envisioned for the semantic annotation services. Initially, these evaluations may be carried out independently of the Ecosystem and related resources, but using well-defined sets of articles available at UPB. This kind of evaluations will focus on the performance of these services (E4). Later, the co-authorship network of papers from the RAGE Ecosystem will be built and may be used as a comparative baseline for the Social Networks Analysis approach applied on the semantic graph of similarities between papers.

A systematic evaluation of the added value that the annotation services provide for Ecosystem users is planned by implementing a comparative approach: Initial task-based evaluations for searching and browsing the Ecosystem will be carried out without the semantic annotation services available. When the semantic services are more deeply embedded into the Ecosystem, another evaluation will be carried out applying the same procedure and tasks. By comparing task performance (e.g. time required for retrieving information) from both evaluations it will be possible to identify whether users are able to find information more easily and reliably with the availability of semantic services.

6.2.3.3 Evaluation in the Context of Application Scenarios

As far as possible, evaluations shall also be carried out in the context of concrete usage scenarios. This means evaluation data collection that is embedded in the stakeholders’ use of the Ecosystem as part of their ongoing game development or application (RAGE use cases and beyond). Users interact and experience the Ecosystem as part of their tasks on applied game creation and use in a real-world setting; i.e. system interaction is not controlled, but is experienced under realistic circumstances. This kind of evaluation would allow gathering most in-depth and practice-oriented evaluation data, since it will be based on exploring and searching the Ecosystem motivated by real needs for resources and assets. This would enable data gathering on all evaluation levels.

Carrying out this kind of evaluation in the context of the RAGE application scenarios of WP5 is complicated by the fact that the decisions on which assets are used in the context of which use case and game are taken in WP4 well before the assets are provided as resources via the Ecosystem. Besides, RAGE partners are envisaged to support populating the digital library and media archive with resources, while the level of Ecosystem usage in terms of searching these archives as part of the creation and application of games will likely be limited. The extent to which an evaluation in the context of the RAGE application scenarios will be possible, and whether other application scenarios (beyond those of WP5) can be found at a later stage of the project, is up to further investigation. There is a chance that through dissemination and training events of WP6 and WP9, and evaluation workshops on the Ecosystem, the real-life use of the system may be initiated, if stakeholders continue to use the system in the context of their everyday tasks and activities. This would, provide good opportunities for further evaluation data collection in the context of application scenarios.
7 VALIDATION STUDIES IN APPLICATION SCENARIOS

The game assets developed in the project are applied in the context of game projects for six use case scenarios (represented by WP5). For all of these use cases applied games are being developed, for application in very concrete training settings (a more detailed introduction to the application scenarios is given in Annex 5.1). An overview of the RAGE use cases and allocated responsible partners is given in Table 3. The validation studies are planned and set up by WP8 partners in consultation with WP5 partners. WP5 partners are responsible for planning and arrangement of locations and recruitment and involvement of user groups for the evaluations in the scope of their pilots. WP5 partners may also be requested to take over the arrangement and execution of data collection in the validation studies, under the guidance of WP8. WP8 partners will be in charge of ensuring scientific warranty and will lead and administer data collection, analysis and reporting. WP4 partners will be involved as evaluation participants for data collection related to the development of the applied games.

<table>
<thead>
<tr>
<th>Application Scenario</th>
<th>Description</th>
<th>Use case partner (WP5)</th>
<th>Game developer (WP4)</th>
<th>WP8 partner for validation studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Professional communication skills (T5.2)</td>
<td>Training on professional communication skills in the vocational education of ICT in the Netherlands</td>
<td>SPL</td>
<td>Playgen</td>
<td>OUNL</td>
</tr>
<tr>
<td>Digital skills/group working (T5.3)</td>
<td>Developing team collaboration skills in the context of a game design course in UK</td>
<td>HCUK</td>
<td>Nurogames</td>
<td>UOB</td>
</tr>
<tr>
<td>Entrepreneurial skills (T5.4)</td>
<td>Training for Arts students and graduates in UK on creative skills that entrepreneurs require</td>
<td>HCUK</td>
<td>Gameware</td>
<td>UOB</td>
</tr>
<tr>
<td>Employability skills in sport (T5.5)</td>
<td>Complementing placement service courses for Italian students by training on soft skills based on a sport metaphor</td>
<td>OKKAM</td>
<td>Playgen</td>
<td>OKKAM</td>
</tr>
<tr>
<td>Police interview skills (T5.6)</td>
<td>Training for Portuguese police professionals on specific communication skills to interview victims and offenders</td>
<td>EPJ/MJ</td>
<td>Gameware</td>
<td>INESC-ID</td>
</tr>
<tr>
<td>Job search skills (T5.7)</td>
<td>Training on job search skills (preparing CV, job interview) for students and unemployed in France</td>
<td>RANDST</td>
<td>BIP</td>
<td>TUGRAZ</td>
</tr>
</tbody>
</table>

7.1 Evaluation Objectives and Questions

The work in task T8.5 ‘Validation studies in application scenarios’ consists in conducting evaluation studies on the applied games developed in WP4 for the use cases of WP5. While T8.3 focuses on the evaluation of the process of game development and the application of the assets therein, the evaluation and validation studies in T8.5 focus on the quality and benefit of the outcomes and results of this game creation process. The effectiveness of asset-based applied games developed for WP5 will thereby be addressed from two different viewpoints – the educational effectiveness of applied games for training, which can be broken down into more concrete evaluation questions (see below), and the effectiveness of applied games for game developers in terms of an explorative cost-benefit analysis.

Table 4 gives an overview of the evaluation questions and variables and relevant stakeholder groups. A more detailed description of each of these questions is presented in Annex 5.3.
These basic evaluation questions are relevant for all application scenarios. They may be complemented by additional use case-specific questions defined in the context of the organisation and planning of the individual pilots and evaluation studies. The evaluation questions in Table 4 address the effectiveness of the applied games, as a whole and will therefore implicitly cover the assets integrated in each of the games.

**Table 4:** Overview of evaluation questions and variables for the applied games in the RAGE use cases and addressed stakeholder groups

<table>
<thead>
<tr>
<th>ID</th>
<th>Evaluation Variable</th>
<th>Evaluation Question</th>
<th>Stakeholder Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>Usability</td>
<td>Are users able to interact easily with the applied games?</td>
<td>End-users, training providers</td>
</tr>
<tr>
<td>G2</td>
<td>Game experience</td>
<td>How do end users experience the use of the applied games?</td>
<td>End-users</td>
</tr>
<tr>
<td>G3</td>
<td>Learning effectiveness</td>
<td>Do the applied games effectively support learning?</td>
<td>End-users</td>
</tr>
<tr>
<td>G4</td>
<td>Transfer effect</td>
<td>Do the applied games support transfer of acquired knowledge/skills to the performance context?</td>
<td>End-users</td>
</tr>
<tr>
<td>G5</td>
<td>Pedagogical costs and benefits</td>
<td>What are the benefits and what are the costs/disadvantages of applying the applied games for training?</td>
<td>Training providers</td>
</tr>
<tr>
<td>G6</td>
<td>Costs and benefit for game development</td>
<td>How cost effective is the application and integration of RAGE applied game technologies and methodologies for game development?</td>
<td>Game developers</td>
</tr>
</tbody>
</table>

Evaluation questions G1-G5 target the educational effectiveness of applied games for learners and training providers. Their investigation shall provide a substantial picture on the significance of using applied games for training, and to eventually raise educational and industrial interest in applied games. Evaluation question G6 addresses the (cost) effectiveness of the applied games from the perspective of game development. While in T8.3 (see section 5) game development is evaluated in a more isolated and asset-specific manner, here a more holistic approach of analysing the estimated costs and perceived or prospected benefits of using the RAGE technologies for concrete game projects shall be taken. This shall serve a better and more comprehensive understanding of asset use in concrete game projects and the practical significance and exploitation potential in the applied games market.

### 7.2 RAGE Evaluation Approach

The evaluation in the context of the application scenarios constitutes an approach of holistically investigating the effectiveness of the applied game technologies in RAGE with their different relevant stakeholder groups. Figure 6 gives an overview of the different evaluation levels for the application scenarios and applied games. These will be taken up and elaborated in more detail in sections 7.2.1 (for evaluation of educational effectiveness from a learning and teaching point) and 7.2.2 (for evaluation of cost-benefit analysis from a game development perspective) outlining the generic evaluation approaches and research designs the two perspectives of applied games effectiveness. Relevant related work and evaluation approaches are summarised in Annex A5.2.

Evaluation and validation studies will be carried out separately for each application scenario. A common underlying evaluation approach and methodology shall align evaluation studies in the individual pilots, to enable comparison of results between the use cases. Nevertheless, the instruments and procedure in the individual validation studies need to account for the specific conditions in the different application scenarios. Standard instruments for aspects of usability and game experience may be employed in all application scenarios and enable a comparison of the effect of the different applied games. The measurement of learning effectiveness and
transfer effect will necessarily be tied to the individual use cases and depend on the learning objectives in question and available learning performance measures. The very concrete designs, instruments, and procedures for all validation studies will be specified in alignment with the requirements and preconditions in each application scenario and the related applied game; this will be done based on continuous communication with use case partners in the phase of scenario arrangement and pilot planning of WP5.

![Application Scenarios - Applied Games](image)

**Figure 6**: Evaluation levels on the applied games for the RAGE use cases

### 7.2.1 Educational Effectiveness of Applied Games

The educational effectiveness of games is considered in RAGE with reference to different aspects and goes beyond the pure learning effect. Apart from the direct impact of a game in terms of knowledge or skill acquisition, also subjective reactions by users as well as a broader consideration of benefits for an educational organisation are relevant aspects of effectiveness from an educational viewpoint and represent evidence of the potential benefits of games for learning in a wider sense. Educational effectiveness therefore incorporates the impact of the games for learners as well as for training providers and is targeted by evaluation questions G1-G5 (cf. Table 4). This evaluation approach takes up the idea of an integral evaluation of training programmes in terms of a four level process, as suggested by Kirkpatrick (Kirkpatrick, 1976; Kirkpatrick & Kirkpatrick, 2009). The evaluation levels are illustrated in Figure 6. On the first level, evaluation will investigate the degree to which learners and training providers react favourably to the applied games. This reaction level entails two facets, perceived software quality operationalized by usability (G1), as well as game experience and enjoyment (G2) including variables like motivation, flow, or presence. Level 2 relates to the intended learning objectives and outcomes of the scenario in question (G3). On this level, evaluation will investigate whether and to what degree learners acquire the targeted knowledge and skills by interacting with the serious game. This might also involve evaluation data collection from training providers, e.g. in terms of their perception of end users’ learning. Level 3 addresses the question whether learners are able to apply/transfer the knowledge and competences acquired during gaming in other/new settings (G4). Evaluation at this level is very challenging and ways for actually capturing at least partial evidence of transfer still need to be further explored and specified with the evolvement of the individual application scenarios arrangements. The investigation of level 2 and 3 needs to be closely aligned with the learning and performance context of each use case, by translating the educational objectives into operational and measurable skills. In order to obtain sound evidence of the effect of serious games for learning and transfer (evaluation level 2 and 3) the extent of prior knowledge will be taken into account and comparative-based research designs will be used. At level 4, evaluation of the educational effectiveness addresses the organisational or institutional perspective and related second-order effects, in terms of the pedagogical added value of the games for training providers and/or educational institutions and the disadvantages or costs for introducing and using this type of learning technology (G5). This is slightly different to the ‘results’ level of the Kirkpatrick model and includes subjective perception and reaction on the games’ pedagogical effectiveness. Although this level addresses costs and benefits, this does not necessarily refer to a monetary calculation, but may consist of a qualitative weighting of added value and drawbacks.
7.2.1.1 Participants

Training providers participating in the evaluations on the educational effectiveness of games will mainly be represented by WP5 partners and, respectively, members and trainers from the educational/training institution associated with those partners. Opportunities of extending this sample to involve also external training providers at a later project stage will be sought. The WP5 partners constitute the link to the end-users/learners of the applied games and will provide access to evaluation participants by arranging and organising their pilots. In addition to these end-users coming directly from the RAGE use cases, possibilities of involving additional relevant end-users by deploying the applied games to a wider audience will be explored in the summative evaluation stage. Thereby, of course, language dependencies (i.e. language of games) need to be taken into account.

7.2.1.2 Research Design

A two-tier approach of evaluating the educational effectiveness dimension is proposed in order to realise a systematic investigation of evaluation questions and effects on and for learning. This will help to gain a deep understanding of the educational effectiveness of the applied games, with the aim of demonstrating their significance for educational practice. In terms of the concrete research designs this means that a mix of real-life pilot studies and laboratory studies will be deployed. The rationale for the two study types is outlined below. Envisaged evaluation instruments are discussed in Annex A5.4.1.

Each evaluation study will address a meaningful collection of evaluation questions and variables. While the pilot studies shall cover a more complete range of the evaluation levels, laboratory studies will be carried out for selected evaluation variables and specific evaluation questions of scientific and/or practical relevance. In total, the evaluation studies carried out for an application scenario should cover all evaluation levels.

7.2.1.2.1 Real-world Pilot Studies

The general approach for conducting validation studies on the applied games is in the context of the real-world pilots carried out in the context of WP5 for the RAGE use cases. This is the preferred type of evaluation study, since the testing of the applied games in the field is considered essential. The pilot studies will serve investigating all levels of evaluation questions (G1-G5) related to applied games’ educational effectiveness of educational practice. The embedding and application in real learning and training scenarios will enable to gather evidence of the practical significance of project results through application in real-world training setting.

Different kinds of comparative analyses will be made. Pre-post measurements on relevant evaluation variables on user experience (G2, e.g. motivation) and learning (G3) before and after (or at the end of) gaming (or after one vs. several gaming sessions) will be used to analyse the direct effect of using games for training. Furthermore, a longitudinal analysis will be applied by comparing the formative and summative evaluation results on educational effectiveness, which will be obtained in the two rounds of WP5 pilots. The comparison between initial and final applied games will allow to prove the evolution and improvement of effects/benefits between the two game versions, the costs and benefits for deploying them, and added value of assets integrated only in the final version of the games.

To measure the added value and effectiveness of an applied game, the comparison with a ‘training-as-usual control condition’ is considered most suitable (Van der Kooij et al., 2015). This will be done while keeping in mind fundamental issues in comparative media studies, like for instance the fact that a game would never beat an excellent teacher lecturing. Where available and possible comparative studies will be carried out for a systematic analysis. Comparisons with relevant baseline measurements of learning performance from previous or parallel trainings/courses without games shall be used as benchmark data to systematically compare and analyse the benefit of the games for learning (G3) and transfer (G4).
7.2.1.2.2 Laboratory Studies

In addition to the real-world pilot studies carried out in each of the use cases, laboratory studies will be carried out for specific selected evaluation questions that turn out to be of specific interest (e.g. to follow-up on pilot study results). While the pilot studies aim at maximising coverage of evaluation levels, laboratory studies will focus on specific research hypotheses related to evaluation questions G1-G4 or additional research questions evolving in the individual game projects and use cases. Students from the application scenarios will be targeted as evaluation participants. Participants recruited from different populations or application scenarios might be an option, this would however complicate comparability with results from the actual target population of the applied game.

Laboratory studies will enable more sophisticated and methodologically sound comparative designs and systematic comparisons through randomised controlled trials (i.e. randomly assigning participants to experimental conditions). An example would be the comparison of different versions of a game representing different experimental conditions, for instance with vs. without the use of assets or using different kinds of assets, applied game vs. placebo/entertainment game. Besides, in the scope of these studies the application of evaluation instruments that require a specific laboratory setup or highly controlled settings, like systematic observations or eyetracking, will be possible.

Overall, laboratory studies will allow the evaluation of the games’ effectiveness under more controlled conditions. They permit controlling or excluding outside influences, to focus on the effect on the evaluation variables to be assessed. On the other hand, the controlled setting is not identical to the real-life setting and results may not be generalizable to the real-world context (e.g. Van Velsen et al, 2008). The planning and conducting of laboratory studies (realising experimental conditions, controlling variables etc.) involves high effort; this kind of study will therefore be defined and organised by carefully considering the tension between research interests/scientific needs and practical feasibility. The relevant research questions to be investigated in laboratory studies will be identified in cooperation with use case partners, as well as with asset and game developers, to ensure the significance of study results for educational practice or, respectively, for (further) development of the games and integrated assets.

7.2.2 Cost-effectiveness for Game Development

Evaluation level 4 for the application scenarios and applied games (as illustrated in Figure 6) does not only involve the analysis of pedagogical benefits and costs and second-order learning for training providers when using applied games for training, but also incorporates the analysis of costs and benefits from a game industry perspective (i.e. evaluation question G6). This refers to the aspect of the ‘costs’ for applying and integrating the RAGE game technologies and methodologies for game development (e.g. Sewell & Marcuk, 2004) and whether and how these can be balanced by their added value. This assessment shall be achieved by conducting an explorative cost-benefit analysis for each use-case in order to identify the cost-effectiveness of the applied game technologies (assets and Ecosystem) for game developers. In this context, the analysis of the current market of assets for game creation and relevant media and software repositories in the field of applied games and related costs, as well as a gap analysis on the market are relevant, in order to determine the advantage of RAGE assets compared to existing solutions on the market and, respectively, how the technologies developed by RAGE match current market needs. An exploratory analysis of the cost benefit of the applied game assets must acknowledge and explicitly take into account the nature of RAGE as a funded research project with the aim of simulating growth towards industrial leadership. Two main areas of a research project are generally agreed to be the production of knowledge i.e. research, and the dissemination of knowledge i.e. meetings and conferences. In addition, the time consuming process of elaborating, discussing, designing and developing assets and game integration, which embodies the research and innovation action of the project, coupled with a lack of existing examples could obscure a view of the cost structure of applied game ecosystem.

Furthermore, the cost-effectiveness of applied games is influenced by both the development cost of the games and the potential revenue that can be generated from these games. The current cost to develop a high quality game tends to be high. The cost, securing funding and
having an experienced team (e.g. game developers and subject experts) is vital to making a successful applied game. However, there is currently little reliable evidence to evaluate the cost-effectiveness of applied game development. If it is possible, the evaluation will be carried out with game developers to project the cost of creating the same game both with assets and without assets to estimate the real cost benefit of the RAGE applied game development approach. The evaluation of the cost benefit will focus on the “use of assets” to demonstrate the “potential” value to future users.

This strand of evaluation work links up to WP7 on business modelling and validation outcomes. At the moment WP7 is engaged with analysing existing business models from the leisure and non-leisure gaming industry; at a later stage business options and models for RAGE will be elaborated. The insights gained in the evaluation of the cost-effectiveness for game development will give a better understanding of the market potential and exploitability of the gaming technologies and may feed into the definition of those business models.

7.2.2.1 Participants
For cost-benefit analysis of applied games development, the game developers involved in RAGE (i.e. WP4 partners) will be involved as evaluation participants. From these partners evaluation data will be collected based on their experience in the game projects carried out for the use cases, as well as on their usual approach to game development without the use of assets. In addition, external game developers may be involved for a general analysis of game development and costs, as well as for the identification of assets available on the market.

7.2.2.2 Research Design
Cost-effectiveness analysis will be carried out on applied game development for each of the RAGE use cases. The envisaged instruments for data collection and carrying out this analysis are described in Annex A5.4.3. Depending on the comparability of data that can be gathered for each game project through interviews with the game developers and the users, a comparative study of the (estimated) development costs and benefits per game is envisaged in order to understand the cost-effectiveness of developing and using the RAGE assets.

The main goal, in essence, is to gather meaningful data on the game development efficiencies by the application of assets in the use cases. The cost-benefit analysis will include the subjective analysis and, where possible, estimation of monetary costs and benefits. However, it is understandable, pinpointing the costs and benefits in numbers is difficult as a funded research and development project, since the real costs and benefits might be intangible and uncertain in this context. Therefore, an approach of using an effectiveness analysis as a basis for cost-effectiveness analysis, as suggested by Giertz (2010), is envisaged. Effectiveness is thereby understood as the extent to which the game technologies fulfil their intended objective. Appropriate measures of effectiveness will be identified and may consist in subjective assessment through game developers. Overall, some efforts will be made to estimate the costs and benefits for developing and integrating assets in the use cases and the data collected will include a range of indicators, like the time for development, cost of integration of assets, and perceived actual benefits to the organisation and end product (applied game). Assuming that an asset is selected for a game project because it is needed, cost analysis may be approached by subtracting the estimated costs for reprogramming the functionality provided by the asset and the approximate costs of asset integration.

The analysis of costs and benefits incurred will start at a stage where sufficient progress has been made in game development and on the use of RAGE technologies. An initial analysis will be carried out mid-term of the project for the initial version of game applications (D4.3), a final analysis will be conducted towards the end of the project for the final version of game applications (D4.4).
8 CONCLUSIONS AND NEXT STEPS

This document has elaborated the RAGE evaluation methodology, which aims at sound quality and impact measurements. For a full coverage of the usage dimension of the RAGE technologies, from searching and generating applied games resources and assets, to developing applied games, to playing applied games, a holistic evaluation framework has been elaborated (an early version is presented in Steiner et al., 2015). It integrates the perspectives of multiple stakeholders on applied game development and application for education and training into a holistic approach for evaluating applied game technologies. Each type of RAGE technology (with their underlying methodologies and incorporated resources) will be evaluated on different levels with the relevant stakeholders. In addition to setting the scope for the different evaluation tasks of RAGE, by describing the general evaluation approaches, guidelines for carrying out evaluations and handling evaluation data have been outlined. This is to ensure that evaluation data management is in line with the data management plan set out in WP10.

The presented evaluation framework and guidelines build the common ground for all evaluation and validation studies conducted in the different phases of the project. Using the RAGE evaluation model delineated in the present document as a reference point, scientifically sound, iterative, and mixed-method evaluations of project outcomes will be carried out. Upon selecting concrete research questions, individual study designs, procedures and instruments will be carefully selected and developed on the basis of the evaluation approaches set out in this document. The evaluation framework thus guides the design and specification of data collection and analysis methods, as well as actual data gathering, processing, and management.

Upon submission of this deliverable, the next steps will consist in planning and carrying out formative evaluations. In year 2 of the project formative evaluation studies on the basic version of the Ecosystem (as part of task T8.4) and of the first bundles of assets and related game development (as part of T8.3) will be carried out. In addition, the first version pilot validation instruments (milestone MS8 in M22) and detailed study designs will be specified (as part of T8.5). Furthermore, in the upcoming project year the conceptual approach for the evaluation asset will be further elaborated (in T8.3). The asset will be implemented (deliverable D8.2 Evaluation Asset in M22) and used as part of the validation studies in the application scenarios. Formative evaluation studies on the applied games will start at the end of project year 2 and continue during year 3. The formative evaluation results from individual studies will be reported back to the project team and stakeholders as soon as available. Formative outcomes from all evaluation tasks will be reported in the first RAGE evaluation report (D8.3 due in M33). Summative evaluations on assets and game development and on the Ecosystem will be carried out in project years 3 and 4. Summative evaluations of the final versions of applied games will follow in year 4 (with MS16 second version pilot validation instruments due in M37). Results from the summative evaluation phase will be compared with those from formative evaluations, where possible. The summative evaluation outcomes of all evaluation tasks will be presented in detail in the second and final evaluation RAGE evaluation report (D8.4 due in M47).

In addition to the documentation of evaluation studies and their outcomes in the two RAGE evaluation reports, the evaluation results, insights gained and experiences with the applied evaluation methodology will be shared with the research community via scientific publications and presentations. The evaluation data generated, collected, and processed and published in RAGE and published in papers will be made openly accessible as part of the EU open research data pilot, thus making research reproducible and providing the possibility for further use and analysis by other researchers and future projects.

The evaluation outcomes obtained in RAGE shall give evidence of the significance and benefit of gaming technologies and stimulate their industrial and educational uptake. The evaluation model and framework – in addition to their use in RAGE evaluation – aim at making a contribution to advancing evaluation methodologies in the context of applied and serious games on a wider scope, and may be (re-)used in and adapted for other applied gaming projects. The
evaluation approaches, outcomes and experiences of RAGE thus may serve as best practice for future applied game evaluations.
REFERENCES
(including references from Annexes)


RAGE Evaluation Framework and Guidelines


ANNEX 1: ADDITIONAL CONSIDERATIONS ON EVALUATION IN RAGE

A1.1 Goals of Evaluation

Work package 8 in RAGE is engaged with evaluation and shall ensure an alignment of the project from a validation perspective – by providing a shared framework underlying all evaluation and validation work. WP8 is responsible for the evaluation and validation of the applicability of assets for game creation (WP2, WP3, WP4), the evidence from the application scenarios (WP5), and the functioning of the Ecosystem (WP6). Besides, a link up with the work on business modelling in WP7 is intended.

The goal of evaluation in RAGE is to measure the effectiveness of the applied game methodologies and technologies developed and made available in a scientifically and methodologically sound way. This refers to the quality and benefit of the assets and the Ecosystem and the educational effectiveness and impact on user experience of the applied games.

Recapitulating from the Description of Action, WP8 aims at designing and carrying out sound evaluation and validation of the effectiveness of the project outputs. This can be broken down into the following objectives:

a. specification of a comprehensive evaluation methodology
b. researching and implementing an approach for in-game evaluation
c. conducting empirical evaluations

The present deliverable elaborates the methodological framework and guidelines and therefore addresses objective a. This work carried out under task T8.1, which represents the common ground for conducting all empirical evaluation and validation work, by identifying and specifying the evaluation dimensions and principles of the project, the evaluation questions and approaches for the empirical evaluation tasks (T8.3-T8.5), and integrating them into a sound overall evaluation model. The evaluation framework and guidelines are the basis for planning and conducting empirical evaluations of the assets (in T8.3), the Ecosystem (T8.4), and the applied games (T8.5) developed in RAGE (i.e. objective c). The evaluation methodology also forms a reference point for the work on in-game evaluation (objective b) as pursued in T8.2, which in turn will be applied in the empirical evaluations of applied games (T8.5 and objective c).

The evaluation framework aims at explicitly considering and responding to challenges of evaluation that are relevant for evaluation in RAGE:

- The evaluation framework needs to reach a balance between methodological soundness and practical feasibility (in terms of effort, timing etc.); i.e. an evaluation methodology needs to be defined that is manageable and suitable for the actual conditions and constraints in evaluation practice. This is especially relevant given the fact that the evaluation framework is defined well in advance before the technical development is finished. The evaluation framework should therefore provide sufficient flexibility for fine-tuning the evaluation approach and incorporating new evaluation questions arising during the project.

- The significance of evaluation outcomes for the different stakeholder groups needs to be ensured. This means, evaluation outcomes should consist of a combination of scientifically interesting but also practically relevant results. In particular, formative evaluation should be taken as an opportunity to gather rich and valuable information and results that are relevant for developers and can be fed back into the development cycle. To ensure evaluation outcomes that are relevant and timely for stakeholders, it is necessary to iteratively revisit the evaluation questions and to continuously consult with and cooperate with the various project partners.

- It needs to be taken into account that the engagement of larger numbers of stakeholders for evaluations might be challenging, especially when aiming at...
stakeholders coming from specific expert groups or comparably small populations, like game developers and training providers. Opportunities and possible strategies for recruiting participants need to be explored and evaluation instruments should be selected in line with the sample of participants. Small sample sizes will enable the use of evaluation instruments that require higher effort in data collection and analysis. For the evaluations in the context of the WP5 pilots, on the other hand, partly very large numbers of participants are envisaged – and the evaluation procedure and instruments administered need to be tailored to these evaluation settings.

- Evaluating learning effects of applied games and learning technologies, in general, is sometimes difficult. This is especially true when aiming at proving transfer effects. To meaningfully demonstrate that games are beneficial for learning, the actual learning activity and gain in knowledge needs to be investigated (instead of relying purely subjective and retrospective reports) and comparative evaluation designs need to be realised. In addition, learning effects and experience should be considered in a more qualitative way, taking into account also motivational or affective effects, which may be interrelated with actual learning achievements.

- Another challenge to be faced is the evaluation of sustainability of learning effects through applied games. This refers to the sustainability of the learning impact in terms of long term retention of the acquired knowledge and actionability and transferability to different contexts across time. Considerations on sustainability should always take into account the actual application of knowledge and skills.

- In the evaluation work of RAGE, not only the quality and validity of the developed gaming methodologies and technologies should be demonstrated, but also their wider actual or prospective impact on the game market and the use of games for education and training. This shall be achieved by examining the social dimension of the Ecosystem and a deeper analysis of the added value, benefits, and costs of game development and application in the context of the game projects for the six application scenarios in RAGE.

- Another challenge that needs to be explicitly incorporated in the evaluation framework is the need for incorporating appropriate mechanisms of research data management complying with relevant ethical principles and data protection regulations, and corresponding to the goals of open access pursued by the European Commission’s open data pilot in which RAGE participates.

Evaluation work in RAGE furthermore aims at applying the following range of evaluation principles:

- **Holistic evaluation**: The different aspects targeted in evaluation are incorporated into an overarching evaluation model that meaningfully integrates the different evaluation tasks and questions – and also reflects the different strands and dimensions of evaluation work and stakeholders. Evaluation will be comprehensive and holistic, considering all relevant aspects of the applied game technologies and integrating them into an overall picture of the quality and achievements of RAGE. Although aiming at an overall assessment of the RAGE project outcomes, nevertheless asset- and component specific evaluation results shall be obtained as an evidence of their significance and as a basis for further improving the technologies.

- **Iterative evaluation**: Evaluation is considered as an inherent part of the technology development lifecycle and will be carried out in alignment with and accompanying the development process. The feedback loop from evaluation back to development will be ensured.

- **Flexibility**: The evaluation framework shall provide appropriate flexibility to respond to evaluation interests and necessities not explicitly foreseen at the time of the definition of the initial framework. These might consist of new or slightly changing research questions arising during the project, or refinement of evaluation methods based on experiences in empirical evaluations. This flexibility will be achieved by regularly revisiting evaluation questions and approaches, and adapting/extending them in line with the current research and development questions and needs, as needed.

- **Process-oriented evaluation**: Many evaluation approaches in the context of serious games purely focus on the evaluation of the game artefact. The RAGE project goals and outputs, however, require the consideration of the whole process of creating and
delivering applied games in evaluation. This involves the analysis of effects on an organisational level and, in particular also of the game development process, i.e. the evaluation of the RAGE tools used for the creation of games in addition to the analysis of the gaming experience itself.

- **Theory-driven evaluation**: A systematic approach of evaluation is pursued, in terms of differentiating between different, theoretically founded levels of evaluation. In the definition of evaluation topics and variables and their measurement we will build upon existing theoretical approaches, operationalization, and instruments. Furthermore, evaluation of the educational effectiveness of applied games needs to be oriented alongside the psycho-pedagogical theories/models incorporated in the design of applied games.

- **Multi-method evaluation**: A combination of evaluation methods and instruments will be applied in the evaluation studies – in terms of a mix of qualitative and quantitative, objective and subjective data. The use of different data sources and methods for evaluation data collection will enable triangulation of data. Where suitable, consistency of measurement instruments will be aimed at for the reason of comparability of evaluation results across studies and use cases.

- **Continuous evaluation**: Evaluation is considered as a continuous process and will not be restricted to purely retrospective assessment, but will also incorporate continuous evaluation data, like non-invasively collected interaction logs and analytics, as well as explicit user feedback during usage.

- **Evaluation and research**: Evaluation shall cover the analysis of practical significance and real-world impact of technologies, as well as the systematic investigation of specific research questions that are relevant from a research/scientific point of view. As a result, evaluation will not only consist of pure application, analysis, and assessment of the RAGE technologies in terms of traditional user-centred evaluation, but shall also cover carefully designed comparative, experimental studies systematically investigating the effect of experimental conditions on dependent variables.

- **Real-life and laboratory**: Following the principle of evaluation and research outlined above, empirical evaluations will be carried in real-life as well as laboratory settings. Laboratory settings enable investigations under controlled conditions, but probably with limited generalizability. Deployments in real-life settings shall serve the investigation and demonstration of the significance and applicability in industrial/educational practice.

- **Gender-sensitivity**: Although gender aspects are not an explicit research topic in RAGE, evaluation will take into consideration gender as a relevant person variable. With respect to samples involved in evaluation studies, we will aim at gender balance in evaluation participants. The existence of any gender differences in evaluation results will be explicitly taken into account and analysed.

### A1.2 Research Target Framework

Three main evaluation dimensions for RAGE can be identified: 1) the evaluation phases (compare section 2.1), 2) the RAGE technologies (compare section 2.2), including the related resources, learning content and context, as evaluation objects, and 3) the target stakeholder/user groups (compare section 2.3). These dimensions form the conceptual framework and evaluation space in which the methodology for evaluation and validation is defined and individual evaluation studies are planned and carried out (see Figure 7). The central goal is to collect evidence for the effectiveness of serious game technologies in a systematic and methodologically sound way.

Evaluation follows an iterative approach – represented by the evaluation phase dimension – aligned with the technology development phases. The main evaluation phases consist of formative evaluations during and summative evaluations at the end of the development process. Formative evaluations are important because they will allow further modifications and improvements of the technologies. Summative evaluations shall demonstrate the effectiveness of the final research and development outcomes.

RAGE evaluation integrates the perspectives of the different stakeholder groups present in the project (i.e. stakeholder group dimension). Evaluations will address the perspectives and
benefits for the different stakeholder groups addressed: asset developers, game developers, training providers, and end users (the actual learners/gamers). This multi-perspective approach will yield a holistic understanding of the quality and impact of the applied game technologies (i.e. the third dimension in the evaluation space) – from the use of and interaction with the Ecosystem, to the asset-based game development process, to the actual interaction with and impact of the developed applied games.

In the three-dimensional evaluation space illustrated in Figure 7 individual evaluation studies can be represented by building blocks, characterising the triplet of technology, stakeholder group, and evaluation phase addressed in a given study. Evaluations addressing more than one user group or longitudinal studies over both evaluation phases can be represented by larger building blocks. Not all possible combinations of building blocks, however, are meaningful and will be covered by in evaluation, since stakeholder groups map diversely to the RAGE technologies. Evaluations addressing mainly game development are represented in blue in Figure 7, while green building blocks illustrate evaluations focusing more on the point of view of educational application of games.

The evaluation approaches elaborated in alignment with the outlined evaluation space and presented in the remainder of this paper incorporate the idea of starting from general evaluation questions per task and realising systematic research design. Nevertheless, there will be flexibility to accommodate to specific research questions arising during the project. Moreover, flexibility to account for the specific conditions of the individual application scenarios and game projects shall be provided. With regards to evaluation instruments, the framework aims at a mixed-method approach, enabling the integration and triangulation of qualitative and quantitative data from multiple sources (instruments) and perspectives (stakeholder groups).

**A1.3 Person Variables**

Characteristics of evaluation participants that might have a systematic influence on evaluation variables measured in evaluation are important to take into account in evaluation. These include gender, education, professional and cultural background, personality traits, preferred learning styles, game experience, instructional/game design experience, attitude to games etc. to name a few. Data on relevant person variables should be explicitly gathered when conducting evaluations, to allow taking them into account in data analysis. Which person variables are relevant will likely differ for the different stakeholder groups and the different use cases. The person variables to take into account will be identified in the individual evaluation tasks and study and data on these variables will be elicited accordingly.
ANNEX 2: ADDITIONAL INFORMATION ON RESEARCH DATA MANAGEMENT

A2.1 The Evaluation Process and Research Data Lifecycle

Conducting evaluation studies can be described as a step-by-step procedure. The main steps of the evaluation process consist of planning the evaluation, carrying out the evaluation, and finally working with the data collected. These main phases can be broken down into more fine-grained steps, as shown in Figure 8. These steps are usually considered and carried out in the course of an evaluation process (even if not in this exact sequence).

![Figure 8: Steps of the evaluation process](image)

The operations that are performed on a data record or data set throughout its life are described in terms of stages of a so-called research data lifecycle. A range of different (but largely overlapping) data management lifecycle models has been proposed (for an overview see Ball, 2012). The research data lifecycle stages can be mapped to the key evaluation steps (see Table 5 for examples). In particular for the evaluation phase of working with the data and results the research data lifecycle provides a more detailed and systematic approach of relevant aspects for data management.

Table 5: Research data lifecycle stages and how they map to the steps of the evaluation process

<table>
<thead>
<tr>
<th>Steps of Evaluation Process</th>
<th>Stages of Research Data Lifecycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning</td>
<td>Research360 Institutional Research Lifecycle Concept (Jones, 2011)</td>
</tr>
<tr>
<td></td>
<td>Plan and design</td>
</tr>
<tr>
<td>Carrying out</td>
<td>Collect and capture</td>
</tr>
<tr>
<td>Working with the results</td>
<td>Interpret and analyse</td>
</tr>
<tr>
<td></td>
<td>Manage and preserve</td>
</tr>
<tr>
<td></td>
<td>Release and publish</td>
</tr>
<tr>
<td></td>
<td>Discover and reuse</td>
</tr>
</tbody>
</table>
A2.2 Ethics, Privacy and Open Access

RAGE evaluation studies have to conform to principles of ethics, privacy and data security, and open access. Addressing these principles at the design stage may prevent problems later on during execution.

All (research) activities carried out under Horizon 2020 must comply with ethical principles and relevant national, EU and international legislation, for example the ‘Charter of Fundamental Rights of the European Union’ and the ‘European Convention on Human Rights’\(^1\). Ethics not only pertains to the use of research subjects in for example medical research, but also to privacy, data protection and research integrity.

Evaluation as described in this document involves the collection, analysis, interpretation and reporting of data. This is what the ‘EU-directive 95/46/EC on the protection of personal data’\(^2\) terms as ‘data processing’. The directive contains an extensive set of rules regarding data processing to secure privacy and data protection. All EU-member states have adopted and implemented the directive in the form of national legislation, with a corresponding institutional structure to facilitate and monitor its execution. As a result the RAGE partners, when conducting evaluations, have to operate within the boundaries of their national data processing legislation, and in many cases also within the data processing guidelines of their own institution/organization.

Data protection legislation generally applies only in the context of personal data, but not for anonymous data, which cannot be linked to an identifiable individual. General principles of ethics and research integrity, however, should be followed in both cases in the sense of good scientific practice.

All projects receiving Horizon 2020 funding have the obligation to make sure any peer reviewed journal article they publish is openly accessible, free of charge\(^3\). In addition, RAGE has been designated a Horizon2020 pilot project on open access to research data, implying that not only RAGE journal articles based on our evaluation studies have to be openly accessible, but also (part of) the underlying data.

In this chapter we summarize ethics-, privacy- and open access requirements relevant in the context of the evaluations, and provide practical guidelines on the design and conduct of evaluation studies within RAGE. These should be taken into account in planning, organising, and conducting of any evaluation study.

A2.2.1 Main Considerations for RAGE Evaluation Studies

In this section we cover the most relevant ethics-, privacy- and open access issues to consider when designing your evaluation study. The listing is definitely not exhaustive and when in doubt consult the RAGE national contact, your institutional Ethics Committee or data protection officer.

A2.2.1.1 Safety and Privacy

Research subjects’ safety and protection of their privacy are the main ethical issues to consider in designing and conducting the RAGE evaluation studies. Safety will most likely not be an issue in RAGE evaluations, but privacy definitely will. Privacy comes in play as soon as personal data are collected and processed. Personal data in this respect is any data that may directly or indirectly link information to a person, and their processing requires special provisions (see Annex 2.2.1.1).

---

\(^1\) http://ec.europa.eu/programmes/horizon2020/en/h2020-section/ethics


A special category of personal data labelled ‘sensitive data’ may only be processed under more stringent rules. This is data relating to race, religion, political opinion, sexual preference, involvement in criminal proceedings, etc.\textsuperscript{4} In RAGE we are not likely to process such data. If personal data is not required for a given purpose of an evaluation study, the collection of any personal data should be avoided.

**A2.2.1.2 Provisions when Processing Personal Data**

In case the evaluation study requires the collection of personal data, the following provisions need to be made:

**Consent.** The subject should have unambiguously given his consent for participating in the evaluation. This is will be done through a RAGE Consent form (see Annex 1), which needs to be linked to an audit trail like a signature or electronic click-through mechanism. In case the subject is under-aged the parent or guardian will (also) have to give the consent. Also in case of collecting anonymous data from the start, informing evaluation participants about the data collection and gathering their consent should be realised as a basic ethical principle and procedure of good scientific practice.

**Transparency.** The subject should be informed by whom, why and how his personal data is processed. Processing can take place a) on the basis of data that is newly collected for this purpose, or b) on the basis of existing data that was collected for other purposes (e.g. student records). In the first case the data subject is informed at the moment the data is collected. In the second case the data subjects has to be traced and contacted. In both cases the data controller\textsuperscript{5} will secure transparency through the provision of a RAGE Information letter (see Annex 1).

**Necessity or data minimization.** The use of personal and identification data should be minimized, avoiding the use of such data when anonymous data can be used to achieve the objectives. In RAGE we encourage to make use of non-personal data whenever possible.

**Legitimate purpose.** Personal data can only be processed ‘for specified explicit and legitimate purposes’ and may not be processed further in a way incompatible with those purposes (as expressed in the RAGE Information letter). So afterwards deciding to use the data for a different purpose is not allowed, and requires due consideration during the evaluation design phase. Of course when the data is anonymized, there is no problem.

**Proportionality.** Personal data may be processed only insofar as it is adequate, relevant and not excessive in relation to the purposes for which they are collected and/or further processed.

**Access right.** The individual should retain control over his/her data, thus should be able to verify, correct or even delete his/her personal data. Once personal data is really anonymized (see next) this no longer applies, and the procedure and time frame for the subject to exert his access right should therefore be included in the consent form.

**Anonymization.** Once personal data are collected, they need to be anonymized as soon as possible as the data shouldn’t be kept in a form that permits identification of data subjects for longer than is necessary. Anonymization is a complex issue (see Annex A2.2.1.3), especially in combination with open access research data, and therefore the general guideline in RAGE is to prevent the collection of personal data whenever possible.

**Security.** Technical and organizational measures should be implemented to ensure that no data is lost or destroyed, even accidentally, only authorized entities may access the data, and

\textsuperscript{4} The legal definition of what constitutes ‘sensitive data’ differs somewhat between EU member states and should thus be checked for the respective case experiments.

\textsuperscript{5} The RAGE partner responsible for processing the data. Note: this is a legal term with legal implications attached to it.
no processing is performed which departs from that for which the data had been initially collected.

A2.2.1.3 Anonymization of Personal Data
In case you do collect personal data, the data should be anonymized as soon as possible as personal data should not be kept longer than necessary, definitely before it is shared with third parties (including other RAGE partners). Anonymization may be realized in different ways:

**Real anonymization**
This is the procedure to follow when you no longer need to be able to link your data sets to its subjects. This is how it works: a) make a copy of the raw data set including personal data; b) remove the personal data from the copy; c) delete the original raw data set permanently. In case of an audio or video recording the procedure may involve transcription before the original recordings are destroyed (as transcription is a labour-intensive process, be cautious with using audio and video recordings in your evaluation study).
Real anonymization will rarely occur when you take care not to collect personal data unless it is absolutely necessary (in which case you would pseudonymize your data sets, not really anonymize it). In practice however real anonymization happens a lot due to research designs that have not considered the implications of collecting personal data, or where personal data is collected unintentionally (so-called ‘collateral data’, e.g. personal data included in a log file of a game play, where the researcher is only interested in the response times).

**Pseudonymization**
This is the procedure to follow if in future you want to be able re-establish the link between the subject and his/her anonymized data set, for example in longitudinal research:\(^6\) This is how it works: a) make a copy of the raw data set including personal data; b) delete the personal data from the copy; c) create a reference file with the identification data; d) delete the original raw data set permanently. The anonymized data set may be published as open access data, while the identification file remains under custody of the data controller.
As it is still possible to link the data set to the subject through the identification file, it is called ‘pseudonymization’.

**Data masking**
This procedure entails creating a data set without person identifiers by stripping out obvious personal identifiers from a data set. Data masking allows individuals to be tracked as part of a longitudinal studies.
This can be done by:
- partial data removal – removing some personal identifiers, like name and address, but keeping others such as date of birth
- data quarantining – providing data to recipients who have no access to the other data needed for re-identification. This may mean supplying unique personal identifiers or reference numbers, while the identification key to link these to particular individuals is not disclosed.

**Aggregation**
Data aggregation means that data is displayed as totals, such that no data relating to or identifying individual persons are represented. Personal data sets are summarised into anonymous statistics, as often applied in scientific publications. Destroying the original data sets after aggregation, however, conflicts with the principle of open access data pursued in RAGE, as deleting the original data would prevent other researchers to validate the evaluation outcomes.

---

\(^6\) However, also in longitudinal studies it is possible to collect anonymous data by letting the subject manage his/her own identification data.
A2.2.1.4 Safe Data Storage and Processing

Procedures for storing and processing personal data and anonymous data differ. Personal data in this respect is data related to an identifiable natural person and data that may directly or indirectly link information to a person (identification data). A subject’s picture is personal data. So is a student ID, or a subject’s date of birth in a small sample. As explained earlier, personal data should only be collected when absolutely necessary for the evaluation, and when it is collected, it should be anonymized as soon as possible after which the original data set is destroyed. (See ‘How to anonymize’ in section 3.3.3.2.).

Storing personal data

In case personal data needs to be retained, for whatever reason, it should be guarded against unauthorised access, processing and loss. Beware that as long as personal data is not destroyed, ownership remains with the subject who should be guaranteed access right.

Managing personal data locally. Most likely your institutional data management infrastructure will not meet the requirements for protection of personal data as specified by the data seal of approval. Especially securing the required level of access protection is highly demanding. Public cloud repositories definitely will not do.

In case personal data is stored and managed ‘locally’ by the data controller:

- To prevent unauthorised access, use dedicated data carriers like flash drives only. Do not use your local hard drive, which is also used for other purposes. Apply password protection to your data carriers.
- To prevent loss, make a copy of raw data sets on a separate data carrier. Keep both data carriers in separate and safe locations, which can be locked. In case you collected your raw personal data on paper, make a digital copy: keep the paper- and digital copy in separate and safe places.
- When working with sensitive personal data, apply encryption.
- When eventually disposing of personal data, do not use a ‘soft delete’ but reformat or use file shredder software.

Submitting personal data to a Trusted Collaboration Repository. Preferably personal data should be transferred a.s.a.p. to a Trusted Collaboration Repository (TCR) with a ‘data seal of approval’. When intending to transfer to a TCR, consider its preferred file formats in advance! If transfer to a TCR is not feasible, the personal data should be managed locally by the data controller (see above).

Storing anonymous data

As soon as anonymous data are collected, it is good practice to immediately make a backup to prevent unintended data loss, and keep the original and copy in separate and safe locations. In case the data sets are intended to be deposited as open access data in a Trusted Digital Repository (TDR), it is important to consider the preferred file formats of that TDR (see ‘Preferred file formats’ elsewhere).

A2.2.1.5 Processing of Personal Data

Also during processing personal data should be guarded against loss (by backup, see above) and unauthorised access. Therefore, processing of personal data (including anonymizing personal data) should take place either on the dedicated data carrier where the personal data are stored (see 3.2.4) or in the Trusted Collaboration Repository – and not on the local hard drive of the researcher or on the institutional infrastructure.

---

7 http://datasealofapproval.org/en/
8 See the topic ‘Safety and privacy’ for what constitutes sensitive personal data.
9 Such trusted collaboration repositories should not be confused with a) repositories in which data sets are merely deposited, or b) repositories in which open access articles are published. Most repositories with a data seal of approval are of type a) and do not support collaboration.
10 http://datasealofapproval.org/en/
A2.2.1.5 Transfer of Personal Data
In principle the transfer of personal data between institutions in EU-member countries is allowed, provided the receiving institution also meets privacy and data security requirements. For practical reasons therefore it is easiest to anonymize data before it is transferred. In case of pseudonymization the identification file is kept by the original data controller. (Note: The data can be collected on behalf of a data controller by a data processor but the responsible entity is the data controller.)

A2.2.2 Securing Open Data Access
RAGE has been designated a Horizon2020 pilot project on open access to research data, implying that not only RAGE journal articles based on our evaluation studies have to be openly accessible, but also the underlying data sets.

For data citation and reference a unique identification mechanism should be adopted. DataCite, which implements DOI identification, is a possible choice.

A2.2.2.1 Where to Submit
It should be possible for other researchers to identify and locate your data. In practical terms this means that you submit your data set to an Open Access Repository (OAR), which is listed in the Registry of Open Access Repositories (ROAR) or the Directory of Open Access Repositories (OpenDOAR). Repositories listed with these registries are accessed by ‘harvesters’ like Databib and re3data.org (in the process of merging).

The person/organization who/which collected and processed the data (the formal data controller) typically submits the data set (depending whether you have a personal account or an institutional account with the OAR).

Available safe research data deposit and sharing repositories are currently explored to identify possible candidates for RAGE. The Dutch DANS repository[^11] promotes sustained access to digital research data and encourages researchers to archive and reuse data. It is geared towards individual researchers working at Dutch research institutions, but not suitable for international projects with partners with varying backgrounds. The EUDAT data services infrastructure[^12] focuses on research data solutions for research communities and seems to provide what is needed for the purposes of RAGE. This and other options are under investigation (as part of the work on the data management plan in WP10) to identify the most preferable and suitable option.

A2.2.2.2 What to Submit
Open Access Repositories (OARs) typically require you to submit[^13]:
- Your (anonymized) data set
- The questionnaire and/or other research instruments used to collect your data
- A fieldwork report, describing how the data was collected
- A code book, or a description of variables and information regarding:
  - The population
  - Type of data (units of observation/analysis)
  - The sample and the sampling procedure
  - Response and non-response
  - Weighting variables
  - Constructed and/or derived variables
  - Information on anonymizing

[^11]: http://www.dans.knaw.nl/
[^12]: http://eudat.eu/
[^13]: Taken from the DANS-EASY OAR.
- The publications based on the data (if available) or a bibliographical description of such publications.
- For data sets collected under RAGE we will submit an additional text-file (.txt) in which the RAGE project and its relation to the data set is described. See Annex 1 for an example.

It may be useful to consider these in advance when designing your evaluation research.

### A2.2.2.3 Which Metadata to Use

For harvesters to be able to locate your research data, it is important to metadate your data set. This metadata consists at least of the same metadata as the publication(s) based on in. Existing metadata standards will be explored and based on this a consolidated set of metadata for the applied game field will be adapted and defined in the course of the project.

### A2.2.2.4 Preferred File Formats

Open Access Repositories allow a limited number of file formats, typically those that are least dependent on proprietary software and/or are considered ‘future-proof’. It may thus be necessary to convert your original data set to another format. When designing your evaluation research, please check the requirements of your envisaged Open Access Repository. For submitting quantitative data, the RAGE DoA favours the Comma Separated Values (CSV) format.

### A2.2.2.5 Licensing your Data

Open Access Repositories typically allow you to select from a number of licenses to assign to your data set:

1. Open access – Unrestricted access (CC0 Waiver No Rights Reserved – [https://creativecommons.org/about/cc0](https://creativecommons.org/about/cc0))
2. Open access for registered users – Unrestricted access for all registered users
3. Restricted: request permission – Registered users, but only after depositor permission is granted
4. Other access – The data are not available via the repository (they are either accessible in another way or elsewhere)

For RAGE, given the EU-directive to provide open access to research data in Horizon 2020 projects, we aim for license-options 1 or 2.

### A2.2.3 Support and Monitoring Structure

In this section we outline the institutional infrastructure and responsibilities to implement, monitor and support the adoption of research ethics, data security, and open access in RAGE.

#### A2.2.3.1 National Data Authority

Each Member State, following article 28 of the Directive, has established a ‘National supervisory authority’ on the protection of personal data. In case your evaluation research will include personal data, the national supervisory authority has to be notified before data processing is allowed to start. The notification contains at least the following information (EU-directive, art. 19):

- the name and address of the controller and of his representative, if any;
- the purpose or purposes of the processing;

---

14 Taken from the DANS-EASY OAR.
15 See Annex A2.5 for the contact details of the Authorities in the partner-countries.
16 See Annex A2.2.3.5
• a description of the category or categories of data subject and of the data or categories of
data relating to them;
• the recipients or categories of recipient to whom the data might be disclosed;
• proposed transfers of data to third countries;
• a general description of the measures taken to ensure security of processing.

The fact that RAGE will process research data in various countries implies that various national
supervisory authorities may be involved. Therefore we apply a ‘decentralised’ model of national
responsibilities, vested in ‘national data controllers’ (see Annex A2.2.3.5).

A2.2.3.2 Institutional Ethics Committee
Most research institutes, including universities, these days have an Ethics Committee. In case
research will involve the processing of personal data and/or medical or psychological
experiments with live subjects, the Ethics Committee usually has to approve the research
design before it can be executed. In research where no personal data is processed and no risks
for the subjects are involved, approval of the research design by the Ethics Committee is usually
not necessary. However, in most institutions it is common practice to submit all research
designs to the Ethics Committee as additional ethical criteria may also require assessment.
The RAGE Data Controllers should thus check a) whether their institution has an Ethics
Committee, b) what the institutional requirements and procedures regarding research ethics are,
and c) how much time scanning/approval of a research design by the committee typically takes.

A2.2.3.3 Institutional Data Protection Officer
Institutional data protection officers have expert knowledge of data protection law and practices. They carry out data privacy assessments and ensure that appropriate policies are in place to comply with data protection regulations. They provide assistance in analysing the lawfulness of data applications, which may arise in the context of RAGE evaluation studies and are involved in the registration and notification of data applications with the national data protection authority.

A2.2.3.4 Data Controller and Data Processor
Responsibilities for data processing (in its broadest meaning) are legally vested in the roles of ‘data controller’ and ‘data processor’. According to article 2 of the EU-Directive 95/46/EC:

• “controller” shall mean the natural or legal person, public authority, agency or any other body
  which alone or jointly with others determines the purposes and means of the processing of
  personal data

• “processor” shall mean a natural or legal person, public authority, agency or any other body
  which processes personal data on behalf of the controller

Even if an individual is given responsibility for data control or processing in an organisation, they
will be acting on behalf of the organisation, which will be the data controller or processor.
Each data controller in principle has to register with the National Supervisory Authority to be
included in a public register so that the general public has access to the names of all data
controllers and the type of processing they do. Data controllers must adhere to the data
processing rules of the Member State where he or she is established even if the data processed
belongs to an individual residing in another State. In practice, legislation may vary slightly from
one country to the next, depending in the national implementation of the Directive.

A2.2.3.5 National RAGE Data Controllers
In RAGE we opt for a decentralised approach (see DoA), with a ‘national responsible
researcher’ (or in fact, the responsibility vested in his/her organization) assigned the role of data
controller for each of the local pilots. In the RAGE evaluation studies these data controllers operate within the context of their national and institutional legislation, as well as within the framework of this evaluation framework under the guidance of TUGraz.

The data controller thus:

- Is registered with the National Supervisory Authority on the protection of personal data in his/her respective country.
- Submit the design of the evaluation study/ies to their institutional Ethics Committee for either information or approval.
- Notifies the national supervisory authority about the RAGE evaluation study/ies before data processing is allowed to start (see 3.4.1).
- During execution of the RAGE evaluation secures compliance with national and institutional legislation.
- Supports local researchers involved in the evaluation study/ies.

A2.2.3.6 Data Controller Agreement

In those evaluation studies, where personal data are collected about national subjects by the national RAGE partner, the responsibility of the data controller is well defined. In evaluation studies with cross-border data collection or a data controller processing data from subjects in another country, responsibilities may be less clear-cut. In such a case a 'data controller agreement' between the RAGE partners involved may be required, depending on national legislation. In general, though, if data from different countries are needed for the evaluation analysis, the identification data can be kept by the data controller at national level, while the anonymized data are processed abroad.

---

17 See Annex A2.5 for the RAGE data controllers in their respective countries and institutions.
A2.3 Information Letter and Consent Form

Research information letter for participants in RAGE evaluation studies

Your are asked to participate in an evaluation study as part of the RAGE project. RAGE, Realising and Applied Gaming Eco-system, aims to develop, transform and enrich advanced technologies from the leisure games industry into self-contained gaming assets that support game studios at developing applied games easier, faster and more cost-effectively. For further information about RAGE please visit the project website: http://www.rageproject.eu/

This research study is conducted as part of the evaluation and validation of the project, with the overall goal of investigating the quality and benefit of the methodologies and technologies developed in RAGE. By participating in this research you will be making an important contribution to the research on advanced technologies for applied gaming.

If you agree to take part in this study, you will be asked to do the following:

The information collected for this study will only be used for research purposes. Personal data will not be passed on to any other party. The collected data will only be evaluated in a means, which ensures the anonymity of the participant. The results of the survey, including any data, may be published for scientific purpose and will not give any identifiable reference to individual participants.

If at any time during or after the study you wish for your personal data to be deleted from the dataset, you may contact the responsible researchers.

Participation is on a voluntary basis. You may withdraw participation at any moment. If you choose to participate, and wish to discontinue participation at any time during the study, there will no penalty or loss of benefits to which you are otherwise entitled.

If you have any questions about this study, please feel free to contact the responsible researchers:

RAGE senior local research coordinator: <name>, <function>
**Consent form for RAGE evaluation participants**

**EVENT:** <event, activity>

**VENUE:** <venue>

**DATE:** <…/…/…>

This form is to be signed by the person (or parent/guardian if the individual is under <age> years of age or is a vulnerable adult) who has agreed to participate in this evaluation study for the RAGE project.

I, the undersigned, agree to participate in this study. I consent to the use of the data resulting from my participation by the RAGE project.

I have been informed about the purpose of the survey and understand what is expected of me. I am participating voluntarily and understand that I can withdraw from the research, without repercussions, at any time, before it starts or while I am participating.

I confirm that I understand the terms of consent and that I consent to be bound by them.

I am satisfied that the assurances of responsible and strict data governance, given by the S-HELP project, will be upheld. I understand that anonymity of data will be ensured at each research stage in the project.

SIGNED ___________________________________ DATED __________________________

In case the individual is under <age> years of age or is a vulnerable adult:

FULL NAME PARENT/GUARDIAN

ADDRESS

CONTACT TELEPHONE

EMAIL ADDRESS

SIGNED ___________________________________ DATED __________________________

RAGE senior local research coordinator: <name>, <function>

Contact details: <address>, <tel.>, <e-mail>

RAGE validation coordinator: <name>, <function>

Technische Universität Graz, Rechbauerstrasse 12, Graz 8010, Austria

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 644187.

This form is adapted from the Web2Rights 3.10 Model Consent Form.
This data set has been collected in the context of the RAGE project. RAGE, Realising and Applied Gaming Eco-system, aims to develop, transform and enrich advanced technologies from the leisure games industry into self-contained gaming assets that support game studios at developing applied games easier, faster and more cost-effectively. These assets will be available along with a large volume of high-quality knowledge resources through a self-sustainable Ecosystem, which is a social space that connects research, gaming industries, intermediaries, education providers, policy makers and end-users. The games technologies made available by RAGE are applied and tested in applied games for six selected application scenarios. For more information about RAGE please visit the project website: http://www.rageproject.eu/

The data set has been collected as part of the evaluation and validation in the Project, with the overarching goal of investigating the quality and benefit of the methodologies and technologies developed in RAGE. Concretely, the data collected addresses the evaluation task on 「include the name of respective WP8 task, i.e. 'evaluation of game development and assets', 'evaluation of the Ecosystem services and processes’, or 'validation studies in application scenarios'」.

「add any additional relevant information describing the data set and its relation to the RAGE project」

### A2.5 National Contacts & Supervisory Authorities on Data Protection & Privacy

<table>
<thead>
<tr>
<th>Issue / country</th>
<th>National RAGE contact</th>
<th>National Supervisory Authority on data protection and privacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAGE data management overall</td>
<td>Eric Kluijfhou</td>
<td><a href="mailto:eric.kluijfhou@ou.nl">eric.kluijfhou@ou.nl</a></td>
</tr>
<tr>
<td></td>
<td>Rob Nadolski</td>
<td><a href="mailto:rob.nadolski@ou.nl">rob.nadolski@ou.nl</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAGE Evaluation Framework and Guidelines</td>
<td>Christina Steiner</td>
<td><a href="mailto:christina.steiner@tugraz.at">christina.steiner@tugraz.at</a></td>
</tr>
<tr>
<td>Austria</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Country</td>
<td>Name(s)</td>
<td>Email(s)</td>
</tr>
<tr>
<td>---------</td>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td>France</td>
<td>Olivier Lepoivre</td>
<td><a href="mailto:olivier.lepoivre@randstad.fr">olivier.lepoivre@randstad.fr</a></td>
</tr>
<tr>
<td>UK</td>
<td>Li Yuan</td>
<td><a href="mailto:lyuan@bolton.ac.uk">lyuan@bolton.ac.uk</a></td>
</tr>
<tr>
<td>Italy</td>
<td>Barbara Bazzanella</td>
<td><a href="mailto:bazzanella@okkam.it">bazzanella@okkam.it</a></td>
</tr>
<tr>
<td>Spain</td>
<td>Ivan Martinez</td>
<td><a href="mailto:imartinez@fdi.ucm.es">imartinez@fdi.ucm.es</a></td>
</tr>
<tr>
<td>UK</td>
<td>Kam Star</td>
<td><a href="mailto:kam@playgen.com">kam@playgen.com</a></td>
</tr>
<tr>
<td>Germany</td>
<td>Matthias Hemmje, Dominique Heutelbeck, Jana Becker</td>
<td><a href="mailto:mhemmje@ftk.de">mhemmje@ftk.de</a> <a href="mailto:dheutelbeck@ftk.de">dheutelbeck@ftk.de</a> <a href="mailto:jbecker@ftk.de">jbecker@ftk.de</a></td>
</tr>
<tr>
<td>Portugal</td>
<td>Rui Prada</td>
<td><a href="mailto:rui.prada@tecnico.ulisboa.pt">rui.prada@tecnico.ulisboa.pt</a></td>
</tr>
<tr>
<td>Romania</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A3.1 Introduction to RAGE Assets

The assets developed in RAGE and made available via the RAGE ecosystem are ready-to-use components that provide advanced applied game technologies and features. The assets hide the underlying complexity of the corresponding technologies so that game developers can add these advanced features without requiring a deep knowledge of their internal design principles. This is especially relevant in applied games that aim to leverage the latest research advances (e.g. emotion detection, advanced tracking) without an extensive background on the underlying research fields.

The assets created within WP2 ‘User data analytics’ focus on gathering detailed information about how each player is interacting with each applied game. They are designed to facilitate a range of information gathering and analysis techniques, ranging from the collection of interaction data to the non-intrusive detection of the player’s competence or emotional state. These assets allow game developers to gather data and perform different analyses for further actions (like interventions that aim for improved learning and/or improved motivation), hiding the complexities of gathering, transmitting and analysing the data in a scalable and secure way.

One of the main aims of the assets developed in WP3 ‘Strategic and social agency’ is to support the creation of game experiences that are socially believable. This is done through assets that will improve the decision-making, the natural language capabilities and the embodiment of virtual characters. Another important goal of WP3 is to tailor the game experience to the player. This involves assets that provide on-line adaptation to the players’ characteristics (e.g. competence state), as well as assets that facilitate off-line generation of flexible learning experiences.

A3.2 Relevant Related Work and Evaluation Approaches

A3.2.1 Game Development and Course Authoring

This section reviews evaluation approaches that were considered relevant as a basis for the evaluation of game development and assets.

The RAGE gaming assets aim at supporting applied game creation and development; they will provide authoring tools for entering relevant domain data and for including and configuring features of game analytics and intelligence. Although these authoring tools and the game development process in RAGE will be quite different from common content authoring, evaluation approaches applied for conventional e-learning and game authoring software may inspire the evaluation of gaming assets.

Different authoring environments and tools exist to assist the creation of serious games and interactive stories, as well as traditional e-learning. Examples are the StoryTec authoring environment for story-based digital educational games (e.g. Mehm, Göbel, Radke, & Steinmetz, 2009; Mehm, Konert, Göbel, & Steinmetz, 2012), the Imaginary Storyboarding Environment for collecting and structuring content for simulation and serious game scenarios (Lau, Thakker, Karanasios, & Ascolese, 2013), the ACTSim authoring tool for learning simulations (Gaffney, Dagger, & Wade, 2010), the Invention Board Unique Authoring tool enabling children to develop their own games (Ibrahim & Yatim, 2014), and the GRAPPLE Authoring Toolkit (De Bra et al., 2013) as an example system for creating traditional adaptive e-learning courses (for a more extensive overview of e-learning and authoring tools see e.g. Haghshenas, Khademi, & Kabir, 2012). Authoring tools for course or game development address professional instructional or game designers and developers, but may also aim at supporting pedagogical practitioners and content providers (Mehm, Göbel, Radke, & Steinmetz, 2009). These different user groups have different levels of expertise in programming and game authoring, and therefore have also
different needs and expectations towards authoring tools, which need to be taken into account in the evaluation of the quality and the benefits of this kind of software.

The most commonly addressed variable in evaluations of authoring tools is usability of these authoring tools, covering different aspects of a tool’s suitability, effectiveness and efficiency for a given task, ease of use and learnability, as well as user satisfaction (e.g. Dag, Durdu, & Gerdan, 2014). Standardised usability scales or heuristic checklists provide systematic instruments for evaluating usability features (e.g. Prümper, 1993, Ibrahim & Yatim, 2013). The research design oftentimes consists of task-based evaluations, presenting an authoring task to evaluation participants, who carry out the task themselves then or, alternatively, by giving instructions to other persons who operate the system (e.g. Mehm et al., 2009; Gaffney, Dagger, & Wade, 2008). Data collected through standard questionnaires is usually complemented by more in-depth feedback gathered through think-alouds or focus groups (e.g. Mehm et al., 2009; Gaffney et al, 2008). These also allow establishing a better understanding of the authoring process and how users experience and use the software. In this way, more detailed information about the specific benefits for authoring can be captured. At earlier stages of development evaluation is sometimes also performed through cognitive walkthroughs (Gaffney et al., 2010).

### A3.2.2 Layered Evaluation of Adaptive Systems

Adaptive hypermedia and learning systems are designed and developed towards the main objective of tailoring information provided by the system and the system itself to the preferences, goals, knowledge and other characteristics of the individual user (e.g. Brusilovsky, 2001) (i.e. adaptation of content). A well-researched approach for evaluating adaptivity consists in so-called modular or layered evaluation. The main idea behind this evaluation approach is that adaptation should not be treated as a monolithic or singular process, but to break down adaptation into its constituents or components for separate evaluation, which may help to identify problems in the adaptation process (Paramythis, Weibelzahl, & Masthoff, 2010). A number of models has been devised and elaborated by researchers in this field (e.g. Karagiannidis & Sampson, 2000; Paramythis & Weibelzahl, 2005; Paramythis et al., 2010; Totterdell & Boyle, 1990; Weibelzahl & Weber, 2003). The separate evaluation steps or layers distinguished in these approaches refer to information processing steps within the adaptation process or to different components involved in the adaptation process. The various frameworks for layered evaluation that have been proposed differ with respect to their granularity level, i.e. the number of steps/components that are distinguished and taken into account for evaluation (Gena, 2005). Two main and distinct phases of adaptation can be distinguished on a high level and have been modelled for layered evaluation (Karagiannidis & Sampson, 2000; Brusilovsky, Karagiannidis, & Sampson, 2004; Totterdell & Rautenbach, 1990):

a) **User modelling or interaction assessment:** The goal of this phase of adaptation is to draw inferences from different aspects of user-computer interaction to user characteristics. User modelling is usually carried out on the basis of low-level information provided by a system’s monitoring mechanisms (e.g. task completion, answers to test or quizzes, keystrokes). The results of the user modelling process are stored and represented in the user model, which captures information on user characteristics that are considered important for a particular application and that are used for the purpose of adaptation.

When evaluating user modelling, the question is whether user characteristics are successfully inferred by the system and stored in the user model. Comparisons of the opinion of experts monitoring users or of users themselves with the conclusions made by the system may serve this purpose.

b) **Adaptation decision making:** The adaptation process itself constitutes the adaptation decision making, where a specific adaptation is selected and grounded in the results of the first phase, i.e. user modelling. The rationale and logic behind this decision process is captured by adaptation rules determining which constituents of adaptation are selected according to the results of user modelling.

When evaluating this layer the question is whether adaptation decisions are appropriate and meaningful given a certain user model constellation. This question can be
addressed by direct knowledge testing (in case of knowledge-based adaptation) or by querying the user on adaptation quality (e.g. whether adaptation improved the quality of the system given a certain goal).

These phases are nicely reflected in RAGE by the distinction of (a) WP2 assets for user data analytics and (b) WP3 assets for tailoring game experience to the learner.

### A3.2.3 Open Learner Models

Open learner models (OLM) pursue the idea of opening up the learner model in e-learning environments to the user, to support reflection on and awareness of learning, and dynamically update that information for a deeper understanding of the learning and teaching process (Bull & Kay, 2010). Opening the learner model is supposed to add pedagogical value to the teaching in learning process and – by enhancing monitoring and reflection – to support educators and learners in performing their tasks. Visualisations in learning analytics have very similar goals, through the illustration of activity and interaction data and the representation of inferences drawn from that data in terms of learners’ skills, knowledge, affective states (Bull et al., 2013). Dashboards are a popular type of visualisation used in learning analytics; they provide a comprehensive view on a large amount of information (Pardo, 2014; Elias, 2011). The RAGE dashboard and analysis asset developed in WP2 will similarly provide different customised (to different stakeholder groups) visualisations of the data and results gathered by the other analytics assets.

Evaluations of open learner models have addressed questions of how learners and instructors value the externalisation and visualisation of learner data, how they interpret and use it, and what they additionally require (e.g. Lazarinis & Retalis, 2007). In addition, the impact of open learner models on metacognitive and self-regulated learning skills has been investigated (e.g. Long & Aleven, 2011; Mitrovic & Martin, 2007). Applied instruments were surveys, interviews, system assessments or performance tests as well as self-assessments. Evaluations of open learner models and dashboards often focus on subjective perception and short-term effects, while it is more rarely evaluated whether these visualisations contribute to behavioural change or new understanding, which would require higher evaluation effort and longitudinal study designs (Verbert, Duval, Klerks, Govaerts, & Santos, 2013).

### A3.2.4 Social Believability in Games

Social believability in games refers to work on modelling and developing believable social agents and non-player characters in games (e.g. Johansson et al., 2013). The believability of a game is a critical subcomponent of player experience. Togelius, Ynannakakis, Karakovskiy, and Shaker (2012) differentiate between character believability (someone believes that the character itself is real) and player believability (someone believes that the player controlling a character is real). Games incorporating believable elements or rather, feature high social believability, are able to enhance players’ attention, higher levels of immersion, and increased enjoyment (Togelius et al., 2012). In the context of applied games, techniques for higher social believability may impact learning effectiveness of games (Linssen, Theuene, de Groot, 2013).

For evaluating believability of game characters from the end user’s perspective, the use of a Game Agent Matrix has been proposed, which describes the highest level of social interaction that a game character is capable of (Warpefelt, Johansson, & Verhagen, 2013; Warpefelt, 2013). The evaluation methodology itself consists of conducting observations in serious game or virtual world, categorising behaviour based on the matrix, and describing problematic behaviour or situations. This analysis is carried out by experts/researchers, who play the game as a player would do; but is essentially also possible on the basis of recorded ‘real’ game sessions (i.e. assessment by experts as external observers). Other evaluation approaches have focused on subjective assessment of believability by players, which may be done by gathering free responses during gameplay or collecting feedback through questionnaires (Togelius et al., 2012). One difficulty of evaluating believability through self-assessment is tied to the broad nature of the concept, which led to different researchers focusing on isolated aspects. Facing this issue, Gomes, Paiva, Martinho, and Jhala (2013) proposed to break down the concept into
a total of nine dimensions that participants can objectively report about: behavior coherence, change with experience, awareness, behavior understandability, personality, visual impact, predictability, social and emotional expressiveness.

A3.3 Elaboration of Evaluation Questions
Work carried out in the context of evaluation task T8.3 consists in carrying out user-centred evaluations to investigate general evaluation questions on testing and demonstrating the assets' usability and benefit for applied game development and their pedagogical significance. Evaluations shall also cover specific questions emerging from the development of the assets and addressing the specific type of pedagogical function of assets – thus, providing information for further refinement and proving pedagogical value. (Note: Software testing is not in the scope of WP8 and will be carried out as part of asset development in the respective WPs).

The definition of the evaluation questions and approaches on game development and assets, the organisation of the evaluation studies, evaluation data collection and analysis is carried out in continuous collaboration with asset developers (as evaluation stakeholders) and game developers (as main target users of the assets).

The evaluation questions on game development and assets are summarised in Table 1 in section 5.1. In the following subsections the evaluation questions are described in more detail.

A3.3.1 General Evaluation Questions on Assets and Asset Application
A range of evaluation questions and topics is of interest for all assets, regardless of their specific pedagogical functionality.

Usability
A1: Are the assets usable?
This addresses the question whether the RAGE assets are user friendly and focuses on the use of the assets themselves. This question relates to whether working with the asset in game development is smooth and whether the asset is easy to use and learn. It may also include aspects of navigation and complexity of an asset. (Note: The appropriate documentation of assets is part of WP2/WP3-internal quality assurance procedures and is not explicitly addressed in WP8 evaluations.) In the context of this usability evaluation it needs to be taken into account that the target audience for assets are primarily game developers, rather than 'average' end-users. The usability aspect may also concern training providers (e.g. in cases where they have to enter content), which would at the same time imply higher usability demands on an asset.

Usefulness
A2: Do the assets enhance game development?
This question addresses the degree to which game developers believe the use of an asset is useful and will help them in their game development tasks. This evaluation question is motivated based on the well-known Technology Acceptance Model (Davis, 1989), where perceived usefulness is suggested as one main factor influencing users’ decision about how and when to use a new technology. The usefulness that game developers perceive when getting in touch with new game assets can be considered as a critical factor for the significance and adoption of assets in game industry.

Relevance
A3: How relevant are the assets for the RAGE application scenarios?
This evaluation question aims at identifying the significance of an asset for the affordances and requirements of the six use cases in RAGE. While the above evaluation questions address the quality of assets, in general, this question puts asset application in the direct context of the RAGE application scenarios. This shall allow correlating use case-specific conditions or features with the assets applied in the respective games. Since this evaluation question does not purely focus on the game development process itself but rather the requirements coming from each use case and the respective educational context,
in addition to game developers also training providers are considered relevant addressees for investigating this question.

**Game engineering**

**A4: Can assets easily be included in game development?**

This addresses the questions whether – from a technical point of view – the assets can easily be integrated with the usual/traditional procedures and technologies of game development, and whether the integration in the game is easy to perform. While usability (see above) addresses the subjective reaction to an asset itself, the focus here is on an asset's actual application and integration with other technologies that game developers use in their game creation. This essentially means an assessment of whether game developers perceive the process of incorporating assets in game engineering as straightforward or as cumbersome. It includes the efficiency and engineering effort required. It also addresses the question of interoperability between assets – sustainable and flexible game components should be integrated and work together easily and effectively.

**Benefit**

**A5: Which benefits do the assets bring for applied games?**

This evaluation question refers to the concrete advantages or added value of an asset for the game development process and/or for the resulting applied games. It relates to specific design and development problems that an asset may help to address or overcome, but also to the pedagogical functionality and value in/of a game and its significance for educational practice. Advantages and disadvantages compared to other reusable assets on the market, which provide similar functionality, may be analysed in this regard. In this context also the extensibility and combination of assets will be addressed, i.e. the question whether additional functionality can be created from the combination of assets. Some assets have the power to be enhanced when used with other assets; assessments of the actual benefit of multiplying functionality in this way will be gathered from stakeholders. This evaluation question is again relating to perceptions of both stakeholder groups, game developers as well as training providers.

### A3.3.2 Evaluation Questions on Specific Asset Functionality

In addition to the evaluation topics described above, for part of the RAGE assets evaluation questions targeting specific aspects of functionality of interest and may be investigated in individual evaluation studies. These evaluation questions relate to the very concrete pedagogical functionality and their investigation is tied to the application of assets in an applied game and the related game experience. Thus, these questions link up to the evaluation of applied games (which is covered by the evaluation studies in the context of T8.5, see section 7).

Since the validation of these questions, however, is highly asset-focused and an intermediate step before evaluating the broader experience of a game, these evaluation questions are considered as part of the game development and asset evaluation.

**Validity of player modelling**

**A6: Do the assets successfully infer user characteristics?**

This evaluation question aims at investigating whether the inferences drawn by analytics assets of WP2 (category a. from the list in section 5.1) from game-based user data to user characteristics are appropriate. Comparisons of the conclusions made by the assets with external or self-assessments of players or with the opinion of experts monitoring players may serve this purpose. This corresponds to the first phase of a layered evaluation of adaptive systems (see Annex A3.2.2), as proposed by Brusilovsky, Karagiannidis, and Sampson (2004). The relevant assets provide different approaches for predicting/identifying comprehension, skills, emotions etc. and the according behavioural/cognitive/emotional dimensions will be considered in their evaluation.

**Perception of dashboard**

**A7: How do stakeholders value the externalisation and presentation of game-based user data?**

This questions refers to the dashboard visualisations of WP2 (category b. as given in section 5.1), which depict the data gathered by the different analytics assets. Evaluation of this question addresses learners', instructors’, and game developers’ experience, perceived usefulness of
and attitude towards the provided visualisations; it may also include the reflection on effects on metacognition and learning/training activities.

**Appropriateness of adaptation decision making**

*A8: Are adaptation decisions of assets appropriate and meaningful?*

This refers to the question whether the adaptation of the learning experience, as provided by WP3 assets (category c. as given in section 5.1) in an applied game is appropriately tailored given a certain player model constellation. Data on this evaluation question may be gathered from players, as well as game developers and training providers. This corresponds to the second phase of a layered evaluation of adaptive systems (see Annex A3.2.2), as proposed by Brusilovsky et al. (2004).

**Social believability**

*A9: Are game experience and non-player characters believable?*

This evaluation question targets WP3 assets (category d. as given in section 5.1), and whether they are able to enhance believability of the game experience. Believable game characters are of essence for player enjoyment and immersion. Believability can thereby be understood in terms of human-like non-player characters, whose actions and social behaviour make sense (e.g. Johansson, Eleidhari, McCoy, & Verhagen, 2013). Feedback on this evaluation question may be gathered from all stakeholder groups.

**A3.4 Instruments**

The evaluation studies on game development and assets will use a mix of different evaluation methods and instruments to gather evaluation data. Below an overview of candidate instruments is presented. The actual instruments to be used in a concrete study will be selected in line with evaluation questions targeted in the course of planning the study design and procedure.

**Interviews and focus groups**

Gathering in-depth qualitative feedback from stakeholders seems indispensible in this strand of evaluation work. The qualitative data collected in this manner is particularly relevant in formative evaluation, to gather information for further development and refinement of assets. Qualitative feedback is important especially with respect to the aspect of game engineering (A4) and game development and pedagogical benefits and related expectations (A5), but also for a deeper understanding of the reasoning on the relevance of assets (A3). Interviews and focus group discussions appear most suitable for this purpose. These will be carried out in a semi-structured manner, guided by a set of pre-defined questions or discussion topics but providing sufficient flexibility and freedom for elaborating deeper on or touching topics not explicitly foreseen.

**Observations**

For getting a better understanding of the game development process and potential issues in the application of assets, observations (optimally combined with think alouds) appear as a suitable option. Observational data will be very difficult to gather in the context of the game projects for the use cases (section 5.2.2.1), but may be feasible in the context of training and evaluation workshops on assets (section 5.2.2.2). Most likely, the observations will take the form of informal observations through evaluators and anecdotal observational notes on noticeable or critical incidents. Observations may furthermore be suitable in the context of asset-specific evaluation studies. Since these studies will be of more laboratory character, a more systematic observational approach – for instance involving think aloud and based on dyadic tests (i.e. pairing up two participants who collaboratively work on a task, in order to support verbalization of thoughts) – would be possible and may provide highly relevant evaluation data on asset-specific evaluation questions.

**Questionnaires**

Questionnaires will be used for the collection of quantitative assessment as well as short qualitative feedback on general evaluation variables related to reaction (A1-A2), but also to gather numerical relevance judgements (A3) and on asset-specific questions (for example on
the perception of dashboard – A7, or the social believability – A9). In this way comparability of evaluation results between different assets, longitudinal measures (e.g. after short-term and longer-term usage, from formative and summative evaluation), and in case of systematic comparisons (e.g. with vs. without asset use) can be achieved.

Standard instruments and existing questionnaires used in related work will be adopted, where available and suitable. Candidate questionnaires for relevant evaluation variables are as follows:

- **Usability (A1)**
  Time available for collecting usability data on an individual asset will most probably be quite limited in the evaluation of groups of assets and due to the need of gathering data also on other evaluation questions. As a result, short questionnaires providing only a general usability assessment appear particularly suitable.
  The System Usability Scale (SUS, Brooke, 1996) may be used for general usability assessments of assets. It is a freely available usability scale and can be used for global and quick assessment of a system’s usability. SUS consists of 10 statements (e.g. ‘I think that I would like to use this system frequently’) and is answered on a 5-point rating scale ranging from ‘strongly disagree’ to ‘strongly agree’. It provides as a result an overall usability and user satisfaction index ranging from 0 to 100, with higher values indicating higher usability. SUS has proven to be a reliable measure of usability (Tullis & Stetson, 2004) and has also been shown to correlate well with other subjective measures of usability (Brooke, 1996).
  An even shorter usability metric is provided by the Usability Metric for User Experience (UMUX) or the UMUX-LITE (Lewis, Utesch, & Maher, 2013, 2015), which cover only 4 and, respectively, 2 items.
  If more detailed feedback on individual usability aspects is needed, more comprehensive questionnaires like the IsoMetrics (Gediga, Hamborg, & Düntsch, 1999) or ISONORM (Prümper, 1993, 2010) may be applied (for a description of theses instruments see Annex A4.4).

- **Usefulness (A2)**
  To gather feedback on usefulness, subscales available from questionnaires on user acceptance (e.g. Davis, Bagozzi, & Warshaw, 1989; Jung, Loria, Mostaghel, & Saha, 2008) may be adapted.

- **Asset-specific evaluation variables**
  To analyse the social believability (A9) of games and game characters existing questionnaires, for example those used by Afonso and Prada (2008) or Lee and Heeter (2012) may be applied.
  The perception of the dashboard (A7) may be evaluated based on the User experience questionnaire (Rauschenberger, Schrepp, Pérez Cota, & Thomaschewski, 2013; Schrepp, Hinderks, & Thomaschewski, 2014), as used by Celi (2015) for the evaluation of user experience of dashboards.

Aside from existing and standard questionnaires, also surveys specifically created for the purpose of RAGE evaluation may be used, for example to gather ratings on relevance (A3), expert assessment of validity of player modelling (A6) or the appropriateness of adaptation decision making (A8). As far as possible, such home-grown questionnaires will be grounded on existing literature and inspired by relevant available instruments.

**Natural language processing tools**
Natural language processing tools will be used as instruments for automated data analysis. Applying opinion mining / sentiment analysis on qualitative data (interview answers or focus group contributions as well as open answers from questionnaires) will facilitate the analysis of qualitative data and support comparability of results. The results may be complemented by additional human-led analysis and can be cross-correlated with quantitative data (ratings, scores) obtained from questionnaires.
ANNEX 4: ADDITIONAL INFORMATION ON EVALUATION OF ECOSYSTEM SERVICES AND PROCESSES

A4.1 Introduction to the RAGE Ecosystem

To support the development and growth of the applied gaming branch the RAGE Ecosystem will foster the merging of the heterogeneous applied gaming communities by providing an effective knowledge and innovation management service tool. The Ecosystem will serve as an interactive knowledge and content management platform and provide a diverse set of services across the knowledge value chain (Salman, Heutelbeck, & Hemmje, 2015).

As a single entry point for applied gaming it will realise centralized access to a wide range of advanced, usable applied gaming software modules, services and resources (or their metadata) by the arrangement of a well-managed and structured asset repository, digital library, and media archive system. The resulting material in the Ecosystem, particularly the textual resources, will be semantically annotated to support searching and access. The semantic annotation services (ReaderBench) will provide, for example, associations of key concepts of a paper and their relevance, concept maps depicting the strength of the semantic similarity between identified central concepts, automatic category matching on the main categories from the RAGE taxonomy.

Besides this, the Ecosystem will be expanded to arrange workshops and offer training courses on an online training portal, covering training for developers and educators in order to amplify applied game uptake. The aim will be to support the self-sustainable production of assets and documentation, training material, workshops and collaboration activities. In addition the social dimension of the RAGE Ecosystem will be supported by community tools for collaboration, annotation, creativity, matchmaking.

The development of the RAGE Ecosystem follows an iterative approach, in which the user-groups to be addressed and thus, requirements to the development process are gradually extended (see D6.1 Ecosystem software platform with underlying software repository, digital library and media archive). The Ecosystem functionality available in each phase will be evaluated in cooperation with WP6 by involving the user-groups matching the individual iteration.

The main objectives of the RAGE Ecosystem are technology push by allowing its participants to get hold of advanced, usable gaming assets, commercial opportunity by providing access to the associated business cases, alliance formation by providing the opportunity to create bonds with peers, suppliers and customers, publicity by giving the chance to advocate their expertise and demands, trade by the development and publication of their own assets, harmonization and focus while contributing to creating a joint agenda and road-map.

A4.2 Relevant Related Work and Evaluation Approaches

The RAGE Ecosystem constitutes a combination of a digital library, media archive, and software repository. With the multitude of resources and collaboration and supporting tools provided, it can be considered similar to a virtual research environment. These kinds of information systems and the Ecosystem share the same core feature – providing (access to) information and resources through networked systems – and thus, they also share similar factors of system success. All those systems usually also include features allowing or supporting collaboration between users. Evaluation methods used for these kinds of systems provide a useful starting point for framing the evaluation of the Ecosystem.

A4.2.1 Software Repositories and their Evaluation

Software repositories are libraries of reusable software assets and make them available to software developers. Software assets include all reusable artefacts from software development,
like specifications, design, code, test data and documentation (Lloyd, n.d.). A popular example is GitHub, a software development Ecosystem with currently over 10 million users. Users can upload or start a software project and collaborate on its development with other users in the community (Dabbish, Sutart, Tsay, & Herbsleb, 2012). Currently active and past projects are retained and can be used as examples and sources of knowledge.

In the evaluation of software repositories, software technological aspects are oftentimes of key interest, while aspects of user interaction and experience sometimes remain off-stage. Most research therefore focuses on technical details and methods for storing and retrieving repository content, while the evaluation of the effectiveness of a repository is oftentimes rather informal or vague (Lloyd, n.d.). There are, however, approaches that explicitly include user-centred aspects. Mili, Mili, and Mittermeir (1998) have defined three types of repository assessment criteria – technical, managerial, and human criteria – which are measured on a 5-point ordinal scale. Based on these Lloyd (n.d.) has proposed an evaluation framework comprising measures of accuracy, usability, cost, and tractability. Another set of software repository evaluation criteria has been proposed by Newhouse (2005, cited after Kurilovas & Dagiene, 2010), which refer to documentation, technical aspects, and management, which partly address also the evaluation of the software made accessible by a repository. Methods applied for evaluating software repositories include gathering direct feedback from repository users or managers, for example via questionnaires or structured interviews (e.g. Zuccala, Oppenheim, & Dhienas, 2008), or evaluations through experts against pre-defined evaluation criteria (e.g. Maril & Luczk, 2009).

### A4.2.2 Digital Libraries and Virtual Research Environments and their Evaluation

Digital libraries and virtual research environments (VRE) are digital repositories equipped with a variety of additional tools supporting users in the exploration, search and interaction with repository contents, like e.g. cultural artefacts. These systems support conceptualising, visualising, and analysing information, and collaboratively working on it. They usually do not consist of one monolithic technology, but cover a collection of tools assembled in one place in order to assist research tasks and processes. Examples are the CULTURA virtual research environment (Hampson, Lawless, Bailey, Yogev, Zwerdling, & Carmel, 2012) or the TextGrid environment for supporting researchers in the arts and humanities (Neuroth, Lohmeier, & Smith, 2011). These environments incorporate tools for text search and analysis, archiving and reuse, collaboration and annotation. Commonly, they are developed for a particular digital collection and address a specific target audience, like professional researchers, research projects or institutions, or teaching communities. A major deficiency of most VREs is that they are designed with a ‘one-size-fits-all’ approach, thus making it difficult for users of varying experience levels to effectively make use of the content and resources provided.

Evaluation approaches for digital repositories and VREs fall into three main categories: a) user-oriented evaluations addressing users’ requirements, preferences, interaction and satisfaction with a VRE; b) system-oriented evaluations focusing on technological aspects of digital information representation and retrieval (e.g. precision, recall); and c) systematic evaluations covering user-oriented as well as system-oriented evaluation goals (Saracevic, 2000). A whole range of evaluation models or frameworks have been devised and used for evaluation of digital libraries and VRE (e.g. Fuhr et al., 2007; Steiner et al., 2007). A review of user-centred evaluation approaches by Heradio, Fernández-Amorós, Cabrerizo, and Herrera-Viedma (2012) has shown that evaluation of digital libraries is centred around the topics of usability and usefulness. Another review (Xie, 2008) found that most evaluation studies are in fact usability studies; further evaluation criteria identified relate to collection quality, service quality, system performance, and user satisfaction. In terms of methods, there is a trend to blend quantitative and qualitative data for evaluation studies, for a deeper understanding of the quality of a system, using e.g. transaction logs, questionnaires, focus groups and interviews, and observations (Heradio et al., 2012; Xie, 2008)
A4.2.3 Further Relevant Work

Apart from evaluation approaches in the context of software repositories and digital libraries, work on the evaluation of instructional material or learning object repositories may provide useful inspiration for the evaluation of the Ecosystem with respect to the training material provided (as part of evaluation question E3 on resource quality). On the one hand, evaluation criteria for learning object repositories and learning objects in e-learning have been proposed (see Kurilovas & Dagiene, 2009, 2010 for an overview); these go beyond the evaluation of the learning resources themselves and involve also more technical aspects of the repository, interoperability etc. Besides, there are tools for educators in order to help them determining whether instructional materials are aligned with current educational standards (Student Achievement Partners, 2013), from which considerations for an explicit evaluation of the training material may be drawn. In addition, ideas of student evaluation of instruction, as known from university courses (e.g. Thielsch & Stegemöller, 2012), may be adopted for this purpose.

With respect to the community tools included in the Ecosystem, approaches on the evaluation of social collaboration tools and technologies might be relevant (e.g. Noble, Buck, & Yeargain, 2001; O’Dea, Harris-Thompson, Malek, Dominguez, & Crandall, 2007). The goal of existing approaches is to measure the impact of theses tools, e.g. in terms of product quality and team efficiency. While the use of such collaboration metrics goes beyond the evaluation goals for the Ecosystem, approaches and tools used for eliciting users’ perception of collaborative tools (e.g. Brodahl, Hadjerrouit, & Hansen, 2011) may be relevant for the purpose of RAGE evaluation. Besides, work and approaches of social network analysis might be relevant and could be used for a deeper investigation of user communities and interaction between users (e.g. Butts, 2008).

A4.3 Elaboration of Evaluation Questions

The RAGE Ecosystem developed in WP6 is strongly linked to and dependent on the methodologies and technologies of the other work packages. The methodologies for asset development and packaging etc. defined in WP1, the assets developed in WP2 and WP3, and the business models elaborated in WP7 – all of these determine what the Ecosystem will actually provide to users. These links and interdependencies need to be taken into account in evaluation work. On the one hand, it is not possible to meaningfully evaluate the Ecosystem in isolation, i.e. without reference to these and without these different kinds of resources. On the other hand, evaluation needs to take care not to focus purely on the assessment of the methodologies and technologies provided, but primarily on the very tools and functionality provided by the Ecosystem itself, and the impact of their combined usage.

The work in evaluation task T8.4 on the Ecosystem is carried out in coordination and collaboration with Ecosystem developers (WP6), for an alignment of evaluation questions with the research interests and development focus of the Ecosystem and to ensure the feasibility of the evaluation approach. The iterative approach of Ecosystem development will be accompanied by an aligned and phased approach of evaluation.

The evaluation questions relevant for the Ecosystem and/or its individual subsystems or services are summarised in Table 2 in section 6.1. A more detailed description of each of the evaluation questions is are presented below.

Usability

E1: Is the Ecosystem usable for the targeted user groups?
This addresses the question whether the RAGE Ecosystem supports users (from the different stakeholder groups targeted) in their tasks in an applied games context (creating, researching, or using applied games). It relates to the interaction and communication of the user with the Ecosystem in terms of the ease of use, suitability for the tasks in questions, self-descriptiveness and learnability of the system, conformity with user expectations, complexity and controllability etc.

User Acceptance

E2: Do users find the Ecosystem acceptable and intend to use it?
The fact that a system is technically sound and usable does not necessarily mean that users actually accept it and want to use it—especially when facing novel software or prototypes. The consideration of aspects of users’ acceptance and influencing factors on acceptance, which are determinants for users’ behavioural intention to use and to actively contribute to the Ecosystem are therefore considered essential in the evaluation of the RAGE Ecosystem. The Technology Acceptance Model (Davis, 1989; Davis et al., 1989) will be used as a theoretical basis for investigating this evaluation question. The model postulates that perceived usefulness and perceived ease of use (which overlaps with usability) are the main factors determining users’ attitude and acceptance towards a technology, which translates into actual usage intentions. Relevant influencing factors are individual differences (e.g. digital literacy skills), organisational context variables (e.g. accessibility), and interface characteristics (e.g. screen design) (e.g. Thong, Hong, & Tam, 2008).

Resource quality

**E3: Do the resources provided serve the needs of the users?**

This evaluation question refers to the usefulness/utility of the different kinds of resources provided by the Ecosystem and whether they serve the information needs of users (cf. Tsakonas & Papatheodorou, 2006). As a result, this question refers to the utility of the assets provided via the software repository and, respectively, the usefulness of documents, papers, training material provided by the digital library and media archive. As a result, resource quality may be addressed separately for the individual types of resources. For the software repository the main question is on the relevance of the overall collection of assets provided (as opposed to the relevance of individual assets, which is in the scope of asset evaluation, cf. section 5.1), but also on the representation format (e.g. metadata). For digital library and media archive contents, in addition to the usefulness and relevance of the content collection to users’ tasks, relevant attributes determining resource quality are the representation format and level, reliability, and timeliness. With respect to specific training material on applied gaming and assets, the question on resource quality refers to the quality of learning objects and courses made available via the online training portal of the Ecosystem. Thereby, the quality of the training material could be broken down into different kinds of criteria relating to content, organisation and presentation, assessment (e.g. SIMRA, 2014).

Performance

**E4: Does the Ecosystem allow to effectively and efficiently search, manage, and use resources?**

The evaluation of performance addresses features for supporting exploration, search and use of the collections available in the digital library and media archive and in the software repository. This evaluation question refers to users’ perception of and reaction to the background operations that the system performs on the resources. It includes the assessment of response time, i.e. time needed by the system to perform operations (e.g. search query) and present the results to the users. It also includes the relevance of resources retrieved by a query or accuracy of semantic annotations. From a technical perspective, this relates to information retrieval evaluation in terms of precision and recall. We will focus more on the user-centric perspective by asking users about their subjective experience of system performance on resource, e.g. the perceived appropriateness of search results. In this context, special emphasis will be put on the assessment and thorough validation of the semantic annotation functionality embedded in ReaderBench that is employed as an external web service for the Ecosystem. In the Ecosystem, which integrates different kinds of resources, the appropriate interlinkage of these resources (e.g. a publication and the related presentation about a particular software asset, as well as the software component itself, all made available through the Ecosystem) is also an aspect of performance.

User Community Involvement and Support

**E5: Does the Ecosystem attract and support targeted user communities?**

The effectiveness of the Ecosystem is not only determined by subjective reaction on its technical functionality and population with relevant resources; eventually its success needs to be evaluated in terms of the capability of actually attracting relevant stakeholder groups – and its population as social environment. This evaluation question therefore focuses on the social dimension and aims at investigating whether the Ecosystem, when launched, is able to attract and involve the anticipated diversity of communities. This goes beyond intention to use, as
covered by user acceptance (see above), but addresses the actual usage of the Ecosystem – the use of its tools and resources and, in particular, the use as a social space. This includes also the question whether or to what extent the community tools made available by the Ecosystem support collaboration (communication and interaction) within and potentially even between user communities.

The evaluation of this question goes far beyond the technical usability or acceptance of the individual system features involved; it is an important aspect of the impact of the Ecosystem and needs to be evaluated on a longer term.

User contribution

E6: Do the stakeholders contribute resources to the Ecosystem?

This evaluation question follows up on the question on user community involvement and support and addresses the aim of fostering the production, contribution, and sharing of resources (assets, documentations, training materials, papers) via the Ecosystem. Since the overarching objective of the Ecosystem is to create a self-sustainable repository and social space populated with applied gaming resources and actively used by its user communities, stakeholders’ contributions to the software repository and digital library and media archive are essential. It will be investigated whether and in what way users contribute to the Ecosystem and share their resources, knowledge, and software via the system. This will also provide useful information on the significance of different components (software repository, digital library, media archive) of the Ecosystem for game and asset developers and training providers in an applied gaming context, and on the significance of the types of resources. Besides, it will include the consideration of the significance and users’ acceptance of the specific methodologies required for developing and contributing assets to RAGE (as defined in WP1).

Added value

E7: What added value do users experience from the Ecosystem?

The RAGE Ecosystem integrates different types of content objects and software assets from different sources and leverages them through semantic annotation. It also acts as a social space bringing together users from diverse stakeholder groups. These aspects of integrating and interlinking the various kinds of resources (and their interoperability), as well as connecting users and user communities represent the innovative and distinguishing character of the Ecosystem and are assumed to provide added value to its users. The specific benefits experienced when using the Ecosystem will be investigated in the scope of this evaluation question.

A4.4 Instruments

Multi-method evaluations will be carried out, triangulating data from different sources, like explicit user feedback gathered after (e.g. questionnaires, interviews) or during (e.g. think aloud, rating) using the Ecosystem and actual user interaction data.

In parallel to the alignment of evaluation with the iterations of Ecosystem development (as described in section 6.2.1), evaluation will use a phased approach in terms of the methods applied. User requirements and initial design decisions on the Ecosystem have been identified and taken in WP6 based on expert round tables with prospective users. Evaluation will take up these and gather user feedback alongside the evaluation questions on them, which will also serve the validation and further refinement of requirements. In addition, formative evaluations will allow to identify features and aspects that are suitable for a longer term and more in-depth analysis through the use of continuous evaluation methods (by applying the evaluation software ‘Equalia’ – see Annex A4.5 below). These continuous methods, in a subsequent step, will allow gathering explicit and non-invasive evaluation data during Ecosystem usage and complement initial, retrospective evaluations for more comprehensive evaluation results.

Below an overview of the main evaluation instruments that will be applied in the Ecosystem evaluations is given.

Interviews and focus groups
Discussions with individual or groups of users will be used to gather in-depth qualitative feedback on the Ecosystem. These are especially suited for identifying issues and information to be fed into further development and will be carried out in a semi-structured manner using a predefined set of questions to be used as a guide for conversation. This will allow comparability over different evaluations and evaluators, but still leaving flexibility in the course of communication for probing for further detail and discussing aspects that might not have been explicitly foreseen. Interviews or focus groups may be used in all kinds of evaluation studies (depending on evaluation settings and constraints in terms of time) and for gathering data on all evaluation questions.

Observations
The observation of users while interacting with the system, optimally combined with think alouds is suitable for gathering a deeper understanding of the process of using the system and identifying critical issues, in particular relating to evaluation questions E1, E4, and E6. Evaluation data is elicited directly while using the system, which can meaningfully complement retrospective self-reports. Systematic observations (with recording of user actions in standardised form, e.g. using predefined checklists) involve a high evaluation effort and, especially when combined with think alouds, require evaluation settings with individual or pairs of users. Observations are envisaged to be applied as evaluation instruments in a more informal way in the context of evaluation workshops, i.e. with evaluators observing the group of participants while working on their tasks (potentially determining in advance what to observe but being alert for unusual incidents) and providing anecdotal records on important or critical actions or events.

Questionnaires
Standardised questionnaires or questionnaires used in evaluation studies reported in the literature will be applied for gathering comparable and general-level assessment on evaluation variables, in particular those relating the evaluation questions E1-E4. Possible candidate questionnaires are listed below.

- **Usability (E1)**
  For a quick general usability assessment, for instance in the context of evaluations with very limited time for evaluation data collection, the System Usability Scale (Brooke, 1996) may be used (for a description of the instrument see Annex A3.4).

  IsoMetrics (Gediga et al., 1999): The IsoMetrics usability inventory is a more extensive usability form and suitable for both formative and summative purposes. Two versions of IsoMetrics are available: one short version (only closed question items) and a long version (containing for each item an open question, to identify malfunctions or weaknesses of the software). For the use in RAGE the use of the short version is envisaged, which still covers 75 items. Those items operationalize the seven dialog design principles of the international standard ISO 9241-10: i) Suitability for the task, ii) Self Descriptiveness, iii) Controllability, iv) Conformity with user expectations, v) Error Tolerance, vi) Suitability for individualization and vii) Suitability for learning. The questionnaire’s items consist of statements that have to be judged on a rating scale (agreement from predominantly disagree to predominantly agree). Example items are ‘The software forces me to perform tasks that are not related to my actual work’ for the aspect ‘suitability for the task’, ‘The possibilities for navigating within the software are adequate’ for the aspect ‘controllability’, and ‘I needed a long time to learn how to use the software’ for ‘suitability for learning’.

Alternatively, the ISONORM (Prümper, 1993, 2010) may be used, which is also suitable for gathering data on the different usability aspects in line with the ISO 9241-10 standard, similar to the IsoMetrics questionnaire. This instrument is considerably shorter and thus may be more feasible for use in Ecosystem evaluation, since evaluations will generally address more variables than just usability. The ISONORM 9241-10 consists of 35 items (5 items for each ISO principles) to be responded on a seven tier, bi-polar answer scale. An example item (for the aspect ‘conformity with user expectations’) is given below.

The software…
makes more difficult the orientation because of a non-conforming design.

simplifies the orientation because of a conforming design.

- User acceptance (E2)
Existing questionnaires grounding on the Technology Acceptance Model will be used to gather users’ perceptions of ease of use, usefulness, and behavior intention to use (Davis, Bagozzi, & Warshaw, 1989). Candidate questionnaires are those used by Thong et al. (2002) and, respectively, Hong, Thong, Wong, and Tam (2002) in the context of digital library evaluation. The questionnaire items would have to be adapted for the purpose of Ecosystem evaluation by making a reference to the Ecosystem (instead of e-library, as used in the previous work). Both questionnaires use the same items for the subscales on perceived ease of use (4 items), perceived usefulness (4 items), and behaviour intention (2 items). In addition, the two questionnaires cover items on different kinds of variables that may further influence user acceptance, like domain knowledge, or screen design. Some of these may be relevant for the purpose of Ecosystem evaluation, as well; this will be further explored when planning the individual evaluations. Items are answered on a 7-point rating scale from ‘strongly agree’ to ‘strongly disagree’.

- Resource quality (E3)
To evaluate resource quality, existing questionnaires on the Triptych evaluation model for digital libraries (Tsakonas & Papatheodorou, 2006) and the CULTURA evaluation model (Steiner et al., 2014) will be used and adapted for the purpose of Ecosystem evaluation. These questionnaires contain closed questions on the evaluation variables situated on the different interaction axes of the evaluation model.
In addition, for the evaluation of training material, items from existing instruments on instructional material evaluation (e.g. SIMRA, 2014; Thielsch & Stegemöller, 2012; Asmawi & Razak, 2006) may be adopted, as appropriate. The 10 criteria form for evaluation of instructional material suggested by SIMRA, for example, contains criteria ratings on the content, accessibility, and pedagogy of the material.

In addition to existing and standard instruments, questionnaires or subscales may be constructed especially for the purpose of gathering evaluation data about evaluation questions and variables on the Ecosystem, in particular related to added value (E7) and to specific in-depth questions relevant for the iterative development. Where possible and meaningful, also in this case items and ideas from existing instruments shall be incorporated and adapted.

Interaction logs
Aggregated log data on the interaction of users with the Ecosystem will be used to gather evaluation data on the actual use of the Ecosystem and its tools. Correspondingly, this evaluation method is particularly relevant for data collection related to evaluation question E7 on system usage and user contributions. In addition, log data will be used for the validation of explicit user feedback given on evaluation variables related to questions E1-E6. Only if users have made use of a certain feature of the Ecosystem, it will be meaningful to consider their explicit assessment of this feature. The analysis of interaction logs will be part of the third step of the phased approach described in section 6.2.1 and will be carried out by making use of the evaluation software Equalia (sensors – see Annex A4.5) integrated with the Ecosystem.

Judgets
Judgets are small widgets integrated in the system under evaluation, which enable users to provide feedback on system features and tools while using them. Through the integration of the evaluation software Equalia (see Annex A4.5 Evaluation Software Equalia) the use of judgets as evaluation instruments, and thus, the collection of continuous feedback from users will be enabled. These will allow to present questions to users during Ecosystem use, providing the possibility of giving feedback and reporting on issues while using the system, i.e. in the moment when those issues (or benefits) are experienced. Judgets will consist in different questions that are injected at different places and addressing different features of the Ecosystem. Answers provided to them are sent to, stored, and analysed by Equalia.
Expert review

Expert or heuristic reviews carried out by evaluators (WP8 partners not directly involved in the development of the Ecosystem) would be suitable for evaluating early versions of the Ecosystem and even Ecosystem features at a preprototype level. This is especially relevant for evaluating usability (E1) and user acceptance aspects (E2). As a basis for such expert review, predefined and existing checklists will be used, for instance the checklist of usability principles suggested by Nielsen (1994).

A4.5 Evaluation Software Equalia

In the context of the EU-funded cultural heritage project CULTURA (http://www.cultura-strep.eu/), the Web-based evaluation software Equalia (Nussbaumer, Hillemann, Steiner & Albert, 2012) has been developed by TUGraz to support evaluators in designing, carrying out, and analysing evaluations. This service has been integrated with the CULTURA system, a virtual research environment, and was used to evaluate it. Equalia is envisaged to be integrated with the RAGE Ecosystem and to be used as supporting tool for its evaluation.

One of the core features of Equalia is the possibility of defining and applying evaluation models (in an authoring module), i.e. what is to be evaluated in terms of evaluation variables (quality model) and how these are evaluated (survey model). Equalia allows and supports different types of data collection. Questionnaires are a very traditional and popular way for retrospective evaluation data gathering. Questionnaires in Equalia may cover rating scale questions and open (free-format text) questions. The questionnaire items are defined directly in Equalia and related to the evaluation variables, which allows automated data aggregation and analysis after data collection. The questionnaires are also instantiated and presented via Equalia. A different way of collecting evaluation data is realised with so-called judgeths (a neologism combining ‘judgement’ and ‘widget’), which are little widgets integrated in the system to be evaluated, where users can give immediate evaluation feedback. They appear in the system to be evaluated as a visual element that may consist of a question with a rating scale and/or open (free-format text) comment field. Judgeths allow to continuously collect evaluation data from users in an explicit way during system usage. Similar to questions in a questionnaire, judgeths are related to quality aspects. As data collection is realised in bits and pieces, embedded in the actual user experience and querying aspects relevant in the situation, a minimum of disruptiveness is assumed. This concept is already well-known in social media and Web 2.0 applications where users can rate content objects (e.g. Dooms, De Pessemier, & Martens, 2011). Sensors represent an instrument that establishes a continuous and non-invasive evaluation method. Sensors are not visible to the users, but monitor and log their usage behaviour and collect evaluation data in this way. Sensors and their calculation are defined in survey models and are also related to quality aspects. For example, the frequency of using the help function of a system in relation to overall system interaction may be used as a measure for the system’s ease of use or learnability. A similar approach of monitoring user behaviour is known from the Contextual Attention Metadata (CAM) format, which is a structured and semantic way of capturing users’ behaviour (Schmitz et al., 2009).

An important feature of the evaluation system is the generation of automatic reports from the collected evaluation data on the basis of the underlying evaluation model. Reports may be derived for the individual data collection instruments or aggregated overall data available for an evaluation variable.

Equalia is planned to be integrated with the Ecosystem in the scope of its iterative development process and will be used for evaluation of the Ecosystem as part of the phased evaluation approach described in section 6.2.1.
ANNEX 5: ADDITIONAL INFORMATION ON VALIDATION STUDIES IN APPLICATION SCENARIOS

A5.1 Introduction to the RAGE Application Scenarios

The case experiments of WP5 provide application scenarios and settings for testing the RAGE gaming technologies with end-users (learners) and intermediary organisations (training providers). There are six different use cases involving different target groups in different contexts; all of them address the acquisition of employability skills, but each case focuses on partly overlapping but different skill types/areas. The application scenario partners partly mimic the role that intermediary organisations have toward the game industry. They have specified their needs and detailed requirements for the design and development of applied games in their application settings and contribute to specifying and creating game contents. In WP4 the needs of application scenario pilot users have been analysed for effective applied game solutions. The game companies in RAGE design, specify, and implement games and gamification for these pilots working closely together with the training providers. The games will use a variety of approaches on various delivery platforms including mobile and desktop and a mix of monolithic games and mini-games depending on the use case. For the creation of these applied games the RAGE asset technologies are applied and tested in the context of game engines and platforms, in cooperation with and providing feedback and needs analysis to the researchers and developers in WP1, 2, and 3. Each applied game thereby integrates a different, but potentially overlapping set of assets. The resulting applied games shall deliver exemplar applications in a variety of non-leisure domains, which serve as validation and samples for other game developers, and provide a basis for the user-centred evaluation of the effectiveness of applied games and the analysis of benefits and costs of applying the RAGE game technologies.

Each application scenario in RAGE consists of a WP5 partner representing the intermediary organisation and application scenario. They provide access to and involve relevant end users for the pilots and validation studies (Note: In case of the WP5 partner OKKAM the application scenario is located at the University of Trento and users from there will be involved.) With each of these application scenarios a game developer from WP4 is associated, who takes over the game design and development for the respective use case. Furthermore, from WP8 (more concretely, task T8.5) there is a partner associated to each application scenario, leading the work on the evaluation and validation studies. Evaluation work will be led and carried out by WP8 partners in continuous exchange and cooperation with the other partners, especially the WP5 partners as access point to evaluation participants.

The goal of this section is to describe the overall approach for the validation studies across all different application scenarios. This illustrates that a certain level of alignment in the evaluation methodology for all use cases is envisioned. The specification of the very concrete instruments and procedures for the evaluation studies in the individual cases will be based on this general approach, but will of course respond to and incorporate the specific focuses and conditions (learning objectives addressed, specific target audience, learning and performance context etc.) in each of the application scenarios.

A5.2 Relevant Related Work and Evaluation Approaches

A.5.2.1 Evaluation of Computer Games and Serious Games

With the growing interest and growing popularity of serious and learning games as educational tools, there is also an increasing need for the evidence of the impact of those games on learning. Since educational games are fundamentally different from traditional learning environments or other software products, evaluation approaches valid for those applications may fall short when used in serious games evaluation. Universal evaluation frameworks for e-learning or training programmes (e.g. Attenwell, 2006; Cho, Park, Jo, Jeung, & Lim, 2009; Kirkpatrick, 1976; Pawson & Tilley, 1997) may only serve as a starting point for assessing
serious games. Given the complexity of digital game environments and the embedding of non-leisure and learning purposes in the game, there is a need to select and adapt suitable evaluation methodologies. A range of evaluation models or frameworks have been suggested in the literature to specifically frame the research and evaluation of serious games (e.g. Connolly, Stansfield, & Hainey, 2009; GALA\textsuperscript{18}; Hainey, Connolly, & Boyle, 2010; De Freitas & Oliver, 2006; Kriz & Hense, 2006; Mayer et al., 2013; Law, 2012).

The evaluation of serious games must consider both core aspects of this kind of game, the fun and enjoyment part, as well as the educational aspect. This corresponds to the assessment of a games’ effectiveness in terms of reaching their goals of learning and engagement (in a wider sense). Besides, an evaluation goal of serious games may consist of a measurement of the software quality of the game. As a result, game enjoyment, learning effectiveness, and usability are the evaluation variables commonly addressed. The fundamental difference of serious games to other productivity tools or learning technologies thereby imposes specific challenges to usability evaluation and complicates the use of traditional usability instruments (Moreno-Ger et al., 2012). Usability in the context of (serious) games is referred to as the degree to which a player is able to learn, control and understand a game (Pinelle, Wong, & Stach, 2008). Techniques applied for usability evaluation cover heuristics, think-aloud user testing (e.g. Desurvire, Caplan, & Toth, 2004) and observational methods (e.g. Moreno-Ger et al., 2012). A further instrument, which may be applied for evaluating usability aspects as well as other evaluation variables through end-users or experts are evaluation grids providing criteria for characterising and evaluating serious games (e.g. Boughzala, Bououd, & Michel, 2013).

Learning, i.e. the educational effectiveness of games, is typically evaluated by applying a pre- and post-test design, i.e. the performance assessment of learning outcomes of a certain unit of study (e.g. Ebner & Holzinger, 2007). Alternative approaches consist in the use of self-reports (self-assessments), where people are asked to indicate what they feel they have learned from undertaking an activity (e.g. Whitton, 2007). Besides, built-in assessment procedures of the educational game simulation (e.g. Wesiak et al., 2014). Michael and Chen (2005, cited after Bellotti, Kapralos, Lee, Moreno-Ger, & Berta, 2013) differentiate between completion assessment (summative – e.g. game completion, student assessment) and in-process assessment (formative – usually incorporated in the game), and teacher assessment. The latter focuses on an instructor’s observation and judgement of learners in action and aims at capturing factors that in-game assessment is not able to evaluate (Bellotti et al., 2013).

User engagement, flow, satisfaction and motivation are aspects subsuming a range of attributes related to the subjective experience and enjoyment of games (e.g. Boyle et al. 2012; Law, Hvannberg, & Hassenzahl, 2006). These aspects are relevant for evaluating the effectiveness of serious games. The engagement and enjoyment that learners experienced in gaming may also be an indicator for the learning effectiveness of a serious game. Common approaches to evaluate engagement, motivation and other aspects of user experience are through questionnaires or interviews eliciting subjective feelings of enjoyment and game experience (e.g. Song, & Keller, 2001), attendance rates, measurement of (voluntary) time-on-task (e.g. Chapman, 2003). More sophisticated techniques include observations and screen recording (Law & Sun, 2012) or non-intrusive assessment based on users’ interaction with the system (e.g. detection of disengagement, like Cocea & Weibelzahl, 2009). Besides, physiological responses as correlates of emotional responses while playing games have been used for investigating the engagement in games (for an overview see Boyle et al., 2012). In addition, emerging technologies for unobtrusive data tracking and sensing enable real-time, in-game data collection and may also be exploited for evaluation purposes (e.g. Serrano-Laguna, Torrente, Moreno-Ger, & Fernández-Manjón, 2014). This kind of technologies are, in fact, represented in RAGE by the WP2 assets. Evaluation work will take advantage of the game-based user data collected and made available by those assets and relevant in terms of evaluation variables (compare Table 4).

Evaluations of educational games usually apply posterior approaches, taking an applied game as is. In practice, applied games should be well designed, i.e. incorporating learning

\textsuperscript{18}http://www.seriousgamessociety.org/download/FrameworkToolkit.pdf
theories/principles and instructional design approaches in their design and development. Theory-based research designs aim at validating these pedagogical foundations in applied games and further investigate gaming and learning activities against this background (e.g. Wesiak et al., 2014). Ideas of design-based research are taken up in the applied game field (e.g. Earp, Catalano, & Mortara, 2015). A design-based approach follows a research methodology where an intervention is conceptualised and iteratively implemented in natural settings in order to investigate, validate, and develop theories and frameworks in learning sciences. In particular, this approach highlights the importance of the context, and the interdependence an educational intervention and its context. The didactical or learning context in which a game is deployed has, in fact, been identified as one main influencing factor when using applied games (Arnab et al., 2012; de Smale et al., 2015).

### A5.2.2 Cost-Benefit Analysis

The development and introduction of learning technologies oftentimes requires high levels of investment. With this there comes the need to justify investments and demonstrate cost-efficiency (Wentling & Park, 2002), and to compare costs for developing and applying a learning technology against its benefits for the organisation doing this kind of investment. The goal of cost-benefit analysis is to identify the benefit to the investor resulting from the investment in some resource.

Cost-benefits analysis (e.g. Boardman, Greenberg, Vining, & Weimer, 2001) may be done from the perspective of educational institutions (i.e. justifying investments of introducing a learning technology), but also from the perspective of e-learning developers/providers (i.e. justifying the investment in developing a learning technology) — it is therefore relevant for evaluation questions G5 and G6. Most analysis of the costs and benefits of learning technologies, though, take into consideration the implementation, use and outcomes of a system rather than focusing on the development process (Berge & Donaldson, 2008). In a typical approach of a basic cost-benefit analysis of a learning system costs are considered in terms of fixed costs and variable operating costs – for the total costs these are multiplied by the number of students the system serves. The costs involved in using e-learning technologies consist of the costs for developing the new training material, the costs of e-learning delivery, and the overhead costs of introducing e-learning (Rumble, 2001). Fixed costs are capital costs of equipment and content development, while variable costs are associated with the students and instructors as courses offered (Bramble & Panda, 2008). A SWOT analysis may also be used as an approach for cost-benefit analysis; thereby the aim is to bring down the average cost per student, whereas the threats can be the scale and scope as resources can be invested into developing one good system. Another method is Return on Investment (ROI), which may also be used to determine the benefit of training to an organisation. It takes the form of \% ROI = (benefits/costs) x100 (Bramble & Panda, 2008). A high ROI means a favourable comparison of investment gains to investment costs. A ROI analysis requires the comparisons of the costs with benefits on a monetary level. Since it is often difficult to translate training outcomes into monetary training benefits, there are few ROI reports in the field of e-learning (Attwell, 2006). Instead, approaches of cost-effectiveness analysis are more commonly applied, which is a less complex approach for obtaining insightful results and identifying economically beneficial investments in e-learning (Giertz, 2010). For this purpose, the analysis of costs (calculations on variable costs, fixed costs, and the number of participants) is combined with effectiveness analysis (for example using measures of learners’ satisfaction with a training program or changes in knowledge/skills) and compared on a cost-effectiveness pane (a graphical representation of the changes in cost an effect) to previous traditional training.

In all approaches, costs are dependent on the type of system and context under investigation. Costs and benefits of a learning technology may vary significantly with the characteristics of the institution and learners in question. The costs, the operating costs, development costs, and opportunity for economies of scale will differ based on the instructional design employed, the process for developing the instruction, the degree of interactivity and support and the role of the faculty in the training (Bramble & Panda, 2008). Analysis from the point of view of educational organisations is useful as a basis for planning of and decisions on instructional offers and approaches. Analysis from the perspective of developers is important as it provides a better
understanding of the problems facing a particular development area and by examining the trade-offs an analysis will then allow the designer to examine the cost/benefit relationship (Kacprzynsky & Hess, 2002).

While a lot of research has been carried out on the cognitive and affective impact of using serious games, the integration of serious games technologies in a corporate context has rarely been investigated in a systematic manner (e.g. Azadegan et al., 2012). This relates to the integration of games as educational tools in training organisations, as well as to the use of software components for the development of serious games.

### A5.3 Elaboration of Evaluation Questions

The evaluation questions on applied games in the application scenarios are summarised in Table 4 in section 7.1. Further elaboration and background for each of these questions is presented below.

Most of the evaluation questions address the educational/pedagogical effectiveness of the applied games developed for the RAGE use cases, i.e. questions G1-G5. Effectiveness is thereby understood in terms of the effect of games on learning, users’ (primarily learners, but also training providers) perception of games (G1-G4), and the benefits and costs for educational institutions (G5). Studies investigating these questions on pedagogical effectiveness will therefore involve educational stakeholders, i.e. end-users and training providers. Evaluation question G6 addresses the (cost) effectiveness of the applied games from the perspective of game development. The process of game creation and the use and integration of assets is already covered to some extent by the evaluation of game development and assets in the context of T8.3 (see section 5). Investigating evaluation question G6 will target the analysis of the estimated costs and benefits incurred for developers for the applied games in the individual use cases.

#### Usability

**G1: Are users able to interact easily with the applied games?**

This question refers to the usability of the applied games, which can be considered a key aspect in applied game design and development. Since applied games address a potentially broad target audience of learners, including end users who may not play games very regularly, usability issues may affect the learning experience (Moreno-Ger, Torrente, Hsieh, & Lester, 2012). In evaluation it is therefore important to consider the usability of an applied game in addition to its learning effectiveness and pedagogical value and to correlate them. Interaction with a game should not be hindered by poor interface design; instead it needs to be made sure that users can successfully interact with and navigate the game. In RAGE the usability of games shall not only be assessed by learners interacting with a game, but also by training providers. Usability assessment by the latter stakeholder group is especially relevant if there are different ways to interact with a game, i.e. games providing specific interfaces and tools for training providers (e.g. dashboard of game and learning performance for a class; interfaces for adding content).

#### Game experience

**G2: How do end users experience the use of the applied games?**

Successful games, as well as other software products and services, need to ensure that they impose a sufficiently high/pleasant user experience (game experience). The evaluation of user experience relates to the users’ experience of what a product is designed for. Games and applied games are designed for providing enjoyable and engaging activities. User experience thereby can be understood as end users’ perceptions and responses that result from the use of the applied games (e.g. Schrepp et al., 2014), the enjoyment of or engagement in games (e.g. Boyle, Connolly, Hainey, & Boyle, 2012). It is a multi-dimensional construct that may be related to a range of different attributes or quality aspects, like satisfaction, motivation, pleasure, or challenge (Law, 2012). The evaluation of game experience therefore needs to be based on a foregoing systematic definition and operationalization of the user experience attributes addressed (e.g. digital gameplay experience model – Callejy, 2007) and the related user experience evaluation methods (Law, 2012).
Considering learners’ interaction with a game, the usability of a game (see above) and game experience are related and outcomes on these aspects will be correlated to identify any interdependencies. While in the evaluation of usability a focus on aspects of interface design and navigation is put, game experience evaluation will address in more detail how learners actually perceive the game on a more affective and enjoyment level.

Relevant conceptualisations or variables of user experience that may be addressed in RAGE evaluations are listed below. While consistency in the operationalization of game experience is aimed at over use cases for the sake of comparability, some variables might be especially relevant and therefore considered only in specific use cases.

- motivation (e.g. Wouters, van Nimwegen, van Oostendorp, & van der Spek (2013)
- flow (Csikszentmihalyi, 1990) – which in turn subsumes attributes like immersion, challenge
- pragmatic and, in particular, hedonic quality aspects of user experience (for example ‘stimulation’ and ‘novelty’, according to Rauschenberger et al., 2013)
- game experience according to IJsselsteijn et al. (2008) – comprising immersion, tension, competence, flow, positive/negative affect, challenge
- social believability (e.g. Johansson et al., 2013) of the game as a whole (as opposed to the evaluation of asset-specific effects of social believability, which is covered in the evaluation of assets, see Annexes A3.2.4 and A3.3.2)
- presence – the subjective experience of ‘being there’ when interacting with a virtual environment (e.g. Sanchez-Vives & Slater, 2005; Witmer & Singer, 1998)

### Learning effectiveness

**G3: Do the applied games effectively support learning?**

This refers to the question whether the learning objectives of the applied games are achieved, i.e. whether end-users actually acquire the addressed employability skills. By investigating this evaluation question, learning effects of the applied games shall be proven. For a meaningful investigation of this question a comparative approach is needed, to demonstrate a gain in knowledge and competence from pre- to post-test (before and after playing the game) or/and to demonstrate an advantage over an existing/traditional training approach (without game or using different instruction methods), if available. Learning effectiveness may thereby be operationalized in terms of shorter-term knowledge acquisition or longer-term retention (e.g. Wouters et al., 2013). In addition, the differentiation of levels of cognitive demands may be used as a conceptual framework for evaluating learning (e.g. CRESST model of learning – Baker & Mayer, 1999; Bloom’s taxonomy of educational objective – Anderson & Krathwohl, 2001). Since learning effectiveness may be related to or affected by learners’ subjective perceptions on usability and game experience aspects, the existence of any interdependencies shall be considered and analysed in evaluations. The evaluation of learning may furthermore include the analysis and validation of learning activities/episodes against the background of the learning theories/principles and instructional models implemented in game design.

### Transfer effect

**G4: Do the applied games support transfer of acquired knowledge/skills to the performance context?**

This question refers to the aspect of transferring knowledge and skills acquired in the context of an applied game (i.e. learning context) to other contexts and to the real-world. This means that the knowledge can be translated to situations different from those of the game. The aspect of transferability is particularly important in case of applied games that use a metaphor context for conveying learning objectives, like the game being developed for T5.5, which will use a sports metaphor to teach employability skills. Optimally, this question is assessed in terms of the application or behavioural change in the performance context, in which learners are expected to apply and demonstrate the targeted skills and knowledge (e.g. Hirumi & Stapelton, 2008). A possible example would be the context of an internship (e.g. using assessment by internship tutor or self-assessment). This means, in the end the goal is to investigate the effect of the game-world experience on the end-users’ real-world behaviour (Van der Kooij, Hoogendoorn, Spijkerman, & Visch, 2015). In RAGE the assessment of transfer might also be approached by the use of games from other use cases.
Pedagogical costs and benefits

**G5: What are the benefits and what are the costs/disadvantages of applying the applied games for training?**

This question puts the evaluation of the effectiveness of applied games to the higher, institutional level of the training provider/intermediary organisation. This should include the more in-depth consideration of the educational context, course/training integration, debriefing and support of instructors/trainers, which constitute important aspects and enabling factors of using applied games (cf. de Smale, Overmans, Jeuring, & van de Grint, 2015). The investigation of this evaluation question implies an analysis of the pedagogical benefits that the application of games for training in the use cases brings to intermediary organisations. On the other hand, any disadvantages, barriers, additional efforts or expenses of integrating applied games into organisations’ existing training approaches and cultures (e.g. change management, training of instructors, necessary investments for infrastructure, delay between student assignment and feedback, flexibility in deploying games etc). The analysis of pedagogical costs and benefits incorporates the idea of second-order learning effects (Mayer et al., 2014). This refers to influences of applied games in the large, i.e. changes on the organisational level; they need to be investigated on a mid- to longer-term level.

Besides, in the context of this evaluation question also the degree and perception of customizability of applied games will be considered – the possibility of customising a game will likely demand additional effort/cost for training providers, but will also provide the opportunity/benefit of tailoring a game to the needs of an end user group. The understanding of the pedagogical barriers, gains, and benefits of integrating applied games in training/educational organisations obtained from evaluations may help to improve benefits and overcome barriers towards using applied games for training (Azadegan, Riedel, & Baalsrud Hauge, 2012).

Costs and benefits for game development

**G6: How cost effective is the application and integration of RAGE applied game technologies and methodologies for game development?**

This evaluation question takes the viewpoint of the game industry and aims at analysing for each application scenario the benefits of using and integrating the RAGE technologies and methodologies (the reusable assets and the Ecosystem) in game development and whether these outweigh the additional costs of introducing these technologies. To evaluate the benefits, it is important to consider the also current market of assets and the related costs, in order to analyse the advantage of the RAGE assets compared to alternative assets available on the market. The investigation of this evaluation question serves as evidence of the impact on game industry, the congruency of the RAGE approach with market demands and existing organisational and game development cultures, and eventually the actual market and exploitation potential of RAGE.

A5.4 Instruments

**A5.4.1 Instruments for Evaluation of Educational Effectiveness**

A mix of evaluation instruments will be applied in the studies on educational effectiveness, combining qualitative and quantitative data (e.g. open answers and ratings) as well as subjective and objective measures (e.g. self-reports and actual test performance). Importantly, also a combination of post- or retrospective assessments (e.g. questionnaire feedback after game episode) with continuous data or in-game data (e.g. observations, game-based user data) will be used. The mixed-method approach in evaluation instruments will provide the possibility of data triangulation and a more comprehensive understanding of educational effectiveness.

Evaluation instruments for each evaluation study will be selected in line with the evaluation and research questions addressed and the specific conditions and aims of the use cases. Where possible, a consistency in measurement instruments will be strived for, i.e. applying shared data collection instruments.
The evaluation instruments intended to be applied in the context of the evaluation studies are summarised below. Not all instruments will be used in a given study, but the most suitable ones will be selected and specified.

Heuristic evaluation
For assessing applied games related to different evaluation variables on G1-G3, especially for usability inspection, heuristic evaluations and expert reviews may be applied. A range of heuristics for the evaluation of games and applied games are available (e.g. Jerzak & Robelo, 2014). Candidate heuristics that may be used are the game usability and game playability heuristics suggested by Law, (2012, p. 144) based on previous work. Alternatively, the event categories for the system and the user dimension to annotate game sessions for the identification of usability issues used by Moreno-Ger et al. (2012) may be used.

Questionnaires
Questionnaires are a very popular and easy to administer evaluation instrument and are considered as a key instrument for collecting a sufficient dataset for validating educational effectiveness of applied games.

Standardised instruments and questionnaires from related/prior work are available for a range of usability and game experience attributes and will be used for evaluation data collection. The definition and application of a shared set of construct will be pursued for all use cases; some constructs may be relevant and targeted only in specific application scenarios. Candidate questionnaires (depending on the operationalization of relevant game experience variables) are:
- EgameFlow to measure enjoyment of learning games (Fu, Su, & Yu, 2009)
- Game Engagement Questionnaire (Brockmyer, Fox, Curtiss, McBroom, Burkhart, & Pidruzny, 2009)
- AttrakDiff – for assessment of hedonic and pragmatic quality (Hassenzahl, 2003)
- Game Experience Questionnaire (IJsselsteijn et al., 2008)
- Presence questionnaire (Witmer & Singer, 1998) or Social Presence in Gaming Questionnaire (de Kort, IJsselsteijn & Poels, 2007) – for measuring presence in the virtual world of the applied games
- Questionnaire for measuring immersion in games (Jennett et al., 2008)

In addition to standard and existing questionnaires, home-grown questionnaires may be used for gathering RAGE specific feedback, for example for self- or peer-assessment of transfer effects or for a preliminary analysis of pedagogical costs and benefit. As far as possible and appropriate, such newly constructed instruments will be based on existing literature.

Interviews and focus groups
Conversations with learners and, respectively, training providers in the context of interviews or focus groups will be suitable for gathering in-depth qualitative feedback on all evaluation questions, in particular game experience (from learners) and pedagogical costs and benefits (from training providers). A semi-structured approach will be applied, using an interview guide with predefined questions as a basis for discussion and feedback collection, while still providing flexibility to probe for further detail or explanation (for example to complement observational or questionnaire data or to clarify open issues).

Observational approaches
For a deeper analysis of game experience and learning, observations and screen recordings may be applied. In case of pilot studies, which will usually involve group field tests, informal observations may be realised, preferably by the presence of at least two researchers. Evaluators’ observational notes can meaningfully supplement data from post-gameplay questionnaires.

Evaluation studies implementing dyadic or individual settings of end users interacting with the applied games would enable for systematic observations, which may be combined with think alouds for a deeper understanding of the users reasoning and thinking processes. The running comments from users are assumed to be more valid than retrospective reporting (Law, 2012).
Dyadic settings with pairs of users collaboratively dealing with an applied game usually facilitates verbalisation of thoughts.

**Learning performance metrics**
For assessing learning performance (G3), learning tests adopted from existing assessment tests or specifically constructed for the purpose of the validation studies will be used. These tests will cover the learning objectives addressed by the applied games and will be used to assess learning achievements. For pre-post test designs targeting knowledge growth optimally parallel versions of learning tests should be used. Moreover, any built-in learning performance assessments (ongoing or at the end) of the applied games will be used as supplement learning measures for evaluation purposes.

In addition, the availability of assessments from the training and use case context (e.g. test data, grades) in which the game is embedded will be sought and used, as available. These will likely address a broader subject area and go beyond the discrete learning objectives covered by the applied game, nevertheless this kind of data will be useful for analysing the effectiveness of the games.

In addition to explicit learning assessment, also subjective judgements of learning performance in terms of self- or peer-assessment, as well as teacher/trainer judgement may be applied. This may be a backup approach for gathering data on knowledge/skill transfer in cases where obtaining other measures is difficult or impossible.

**In-game data tracking and analytics assessment**
RAGE assets developed in the context of WP2 provide different kinds of learning and user analytics. These assets make use of game-based user data. Some of these assets will be built into each of the applied games. Apart from their use for reporting learning success and dynamic adaptation within the game they will also serve assessing the effectiveness of the educational game itself. For instance, behavioural patterns indicating successful strategies may be identified, or assets for emotion detection may be used to incorporate in-situ data like facial expressions during gaming for the purpose of evaluation.

A dedicated evaluation asset is under development in WP8 and represents an instrument for evaluating the quality of learning games (see Annex A5.4.2 for a more detailed description of this asset). It will use continuous data from the interaction with a game and will also draw from the WP2 assets to provide insights to users' perception of games and their progress towards game goals. This is done by translating log and sensor data into meaningful information about usability, game experience, and learning based on pre-defined, configurable evaluation metrics. The asset will facilitate the use of analytics for game evaluation purposes and will advance evaluation methods for serious games by complementing traditional instruments (such as questionnaires).

The use of unobtrusive data tracking and analytics of assets for the purpose of game evaluation will enable a meaningful triangulation and cross-validation of different data sources and types, to derive more conclusive evidence on the quality and effect of the applied games. In addition, it will leverage evaluation load for end users.

The analytics procedures/algorithms provided by the assets, however, need to be validated in order to meaningfully apply them as evaluation. Besides, assets can only be used as instruments for evaluation data collection to the extent to which they are actually integrated in the applied games.

**Natural language processing tools**
Analysis of open answers collected via questionnaires, interviews, or from texts composed by users is envisaged to be supported by automated analysis through natural language processing tools. Hence, these tools are not an instrument for data collection but rather for data analysis. Applying opinion mining / sentiment analysis tools on open answer questions might be a great addition, as it can for example be used to cross-correlate with quantitative questionnaire feedback or to determine the general evolution of users' opinions and emotions. This means, in
order to obtain a more in-depth perspective of users’ perceptions, automated methods specific to text mining, opinion and emotion extraction will be applied on open-answer questions in order to determine the general tendency of users’ feedback on certain game assets. In other words, summarised opinions grasping the most representative associations of <functionality, sentiment valence> will be automatically extracted and cross-correlated to the answer to closed questions, following a Likert scale of evaluation.

A5.4.2 Evaluation Asset
The evaluation asset currently under development in T8.2 of WP8 supplements the collection of gaming assets developed in WP2 and WP3 of RAGE. This asset will apply analytics to evaluate the game artefact itself (instead of assessing the player) and shall be used as an evaluation instrument in the context of the T8.5 evaluation and validation studies in the application scenarios. This can naturally only be the case for those games that actually integrate the evaluation asset. This section provides an overview description of the asset. The more complete conceptual approach and implementation of the evaluation asset will be delivered in D8.2 in M22 of the project.

The evaluation asset allows continuous, in-game evaluation of applied games while users are playing. In this way, traditional evaluation of the game quality and users’ perception through explicit post-assessment can be meaningfully complemented. The evaluation asset will collect data in a non-invasive and continuous manner; it will monitor the player's usage behaviour and interactions in a game and translate it into meaningful information about different aspects of game quality and user experience.

The evaluation asset uses relevant tracking and logging data from the game that is sent to the evaluation asset. The types of actions that are tracked are dependent on the type of game. On the one hand, the asset provides general usage data in terms of frequency of actions, which may provide initial useful information for game evaluation. In addition, the evaluation asset will translate and interpret the interaction logs into more sophisticated measures on evaluation qualities, like variables on usability or user acceptance. One simple example is to use the frequency of using the help function of a game and transform it into a measure/score on the ease of use. Data on different types of interactions may also be combined to form a new measure. Linking up to WP2 assets on competence and on motivation assessment the evaluation asset will allow to use those data for the assessment of learning effectiveness and motivational potential of an applied game.

Since the definition of meaningful measures of evaluation variables from interaction logs is a complex task and may require too much effort from game developers, a set of generic semantics and interpretations applicable with most games will be made available with the evaluation asset. Specifying the calculation of these measures game-based user data is subject of research in the context of T8.2 of WP8. The definition of evaluation variables covered by the evaluation asset will be done in line with the evaluation framework set out in the present deliverable and in relation to the evaluation questions defined for the validation studies on the applied games (see section 7.1).

A5.4.3 Instruments for Evaluation of Cost-Effectiveness
The main instruments used for this evaluation task will consist in surveys and semi-structured interviews with game developers. These are planned to be integrated with the work on business modelling and carried out as part of the of stakeholder consultation in WP7.

Apart from the consultation of game developers via interviews and surveys, ways of gathering additional data on development costs and savings will be explored. For example, project management tools used by game developers in their game projects (e.g. for definition of milestones, expectation models of man hours, recording of hours spent etc.) are considered as a potential source for gathering meaningful data in terms of costs and benefits. In addition to data on the game projects for the RAGE use cases that can be derived from these tools, metrics from previous game projects would be relevant for identifying savings or additional costs with
vs. without the use of RAGE technologies. It will be explored whether and which kind of project management tools and data are used and available by WP4 partners and whether they can be used for the purpose of cost effectiveness analysis of developing and integrating applied game assets in RAGE. While this kind of data would be highly valuable for a deeper understanding of the cost efficiencies in game development, it has to be recognised that there may be concerns from the perspective of game developers and commercial sensitivities, which may hinder the accessibility and use of this kind of data. Besides, a variance in project management tools can be assumed for different partners. This will most probably mean the availability of different kinds of data and at different granularity levels, which may compromise also the comparability of the outcomes of cost effectiveness analysis.

For a graphical representation of the changes in cost and effect based on the use of and investment in applied gaming technologies, the approach of using a cost-effectiveness pane, as suggested by Giertz (2010) is suggested. This graphical representation consists of a cost and effectiveness/benefit dimension representing the x-axis and y-axis and will allow the analysis of changes of using RAGE technologies compared to the previous traditional approach of game development.