D2.2.2 Final Version of the LinkedUp Evaluation Framework

LinkedUp: Linking Web Data for Education Project
http://linkedup-project.eu/

Coordination and Support Action (CSA)

Grant Agreement No: 317620

D2.2.2 Final Version of the LinkedUp Evaluation Framework

Deliverable Coordinator: Hendrik Drachsler

Deliverable Coordinating Institution: Open University of the Netherlands (OUNL)

Other Authors:
Slavi Stoyanov (OUNL), Marieke Guy (OKF), Maren Scheffel (OUNL)

Document Identifier:	LinkedUp/2014/D2.2.2/v1.0	Date due:	31.10.2014
Class Deliverable:	LinkedUp 317620	Submission date:	27.10.2014
Project start date:	November 1, 2012	Version:	v1.0
Project duration:	2 years	State:	Final
		Distribution:	Public

LinkedUp Consortium

This document is part of the LinkedUp Support Action funded by the ICT Programme of the Commission of the European Communities by the grant number 317620. The following partners are involved in the project:

Leibniz Universität Hannover (LUH) Forschungszentrum L3S Appelstrasse 9a 30169 Hannover Germany Contact person: Stefan Dietze E-mail address: dietze@L3S.de	**The Open University** Walton Hall, MK7 6AA Milton Keynes United Kingdom Contact person: Mathieu d'Aquin E-mail address: m.daquin@open.ac.uk
Open Knowledge Foundation Limited LBG St John's Innovation Centre Cowley Road CB4 0WS, Cambridge United Kingdom Contact person: Marieke Guy E-mail address: marieke.guy@okfn.org	**ELSEVIER BV** Radarweg 29, 1043NX AMSTERDAM The Netherlands Contact person: Michael Lauruhn E-mail address: M.Lauruhn@elsevier.com
Open Universiteit Nederland Valkenburgerweg 177, 6419 AT Heerlen The Netherlands Contact person: Hendrik Drachsler E-mail address: Hendrik.Drachsler@ou.nl	**Lattanzio Learning SPA** via Cimarosa 4 20144 Milano Italy Contact person: Elisabetta Parodi E-mail address: parodi@lattanziogroup.eu

Work package participants

The following partners have taken an active part in the work leading to the elaboration of this document, even if they might not have directly contributed to the writing of this document or its parts:

- OUNL, Hendrik Drachsler, Slavi Stoyanov, Maren Scheffel
- OKF, Marieke Guy
- LUH, Eelco Herder, Stefan Dietze, Ivana Marenzi
- OUUK, Mathieu d'Aquin

Executive Summary

This document (D2.2.2) describes the LinkedUp consortium's experience in developing and on-going improvement of the LinkedUp Evaluation Framework throughout three web open educational data competitions: Veni, Vidi, Vici. D2.2.2 is the final report regarding the Evaluation Framework (EF). It synthesises the work already done in the previous WP2 deliverables (D2.1, D2.2.1, D2.3.1, D2.3.2, D2.3.3) reporting on best practices, providing suggestions for improvements and possible adjustments to additional application areas.

The initial version of the EF was developed by applying the Group Concept Mapping Methodology (GCM). It objectively identified through some advanced statistical techniques the shared vision of experts in the domain of technology-enhanced learning on the criteria and indicators of the EF. The GCM contributed to the construct and content validity of the EF. The first version of the EF was tested during the Learning Analytics and Knowledge Conference 2013 (LAK 13). After each competition round (Veni, Vidi, Vici) usefulness and ease of use of the EF were tested with a number of experts through a questionnaire and interviews. The analysis of the data suggested some improvements. In this final report of the EF we summarise the lessons-learned and provide six main suggestions for future data competitions developers:

1. Designing a data competition starts with a definition of evaluation criteria
2. Test the understandability of your evaluation criteria before publishing those
3. Do not use an 'not applicable' option for evaluation indicators
4. Less (indicators) are more (preferable)
5. Apply an unification of the scale of evaluation indicators'
6. Weighting of important evaluation criteria can be very informative

We finally present the final version of the LinkedUp EF and refer to the LinkedUp toolbox that provides all lessons-learned and further information for future data competition organisers.

Table of Contents

1. Introduction 5

2. Developing states of the EF 6

2.1 First version of the Evaluation Framework. The Veni case. 6

2.1.1 Pilot version of the Evaluation Framework. The LAK13 Case. 8

2.1.2 The Veni Competition 9

2.1.3 Lessons-learned from the Veni version 11

2.2 Second version of the Evaluation Framework. The Vidi Case. 12

2.2.1 The Vidi Competition 13

2.2.2 Lessons-learned from the Vidi version 14

2.3 Third version of the Evaluation Framework. The Vici Case 16

2.3.1 The Vici Competition 17

2.3.2 Lessons-learned from the Vici version 18

3 Conclusions and suggestions for competition organisers 21

4 Reference 25

Appendix A – Veni Evaluation Form 26

Appendix B – Vidi Evaluation Form 32

Appendix C – Vici Evaluation Form 34

Appendix D – Final version of the Evaluation Framework 36

1. Introduction

The document describes the LinkedUp consortium's experience of developing and improving the LinkedUp EF over the course of three web open educational data competitions: Veni, Vidi, and Vici. The overall methodology implements the idea of progressive, spiral refinement through a cyclical prototype development and the reliance on stakeholders' involvement in the design and evaluation of the EF (Holtzblatt, Wendell, & Wood, 2005; Kuniavsky, 2003).

The three data competitions addressed specific objectives and requirements for innovative open linked data submissions. Veni requested mockups and early prototypes that are good examples for Linked Data applications in Education. In Vidi, in addition to the Open Track we introduced Focused Tracks that had specific objectives. Finally, in Vici matured applications were requested and additional Focused Tasks were promoted.

The EF needed to be in line with those specific objectives and requirements. It became gradually more mature throughout the three evaluation cycles. This deliverable will summarise the experience gained in developing and improving the EF and provide some suggestions in the conclusions.

2. Developing states of the Evaluation Framework

2.1 First version of the Evaluation Framework. The Veni case.

We applied the Group Concept Mapping (GCM) research methodology for the development of the first version of the EF. It is a structured, bottom-up approach for facilitating a group of experts to arrive at an agreement on the criteria and indicators of the EF (see deliverable D2.1). GCM has shown to be more effective and efficient than other approaches for defining the set of assessment criteria and indicators (Kane and Trochim, 2007). Very often researchers define the initial list of criteria and indicators in surveys with questionnaires which results in such a set often not being comprehensive. Also, interviews are usually time and resource consuming and the qualitative analysis is often done by an individual researcher. In contrast to this, the GCM does not rely on pre-determined classification schemas in the analysis of the data. The method does not need intercoder discussions to come up with an agreement. In contrast to the Delphi method, the GCM includes only one round of structuring the data as the participants work independently and anonymously of each other. The methodology implements some advanced statistical techniques such as multidimensional scaling (MDS) and hierarchical cluster analysis (HCA) which quantitatively aggregate individual inputs of the participants to identify emerging patterns in the data. Consensus is not forced but rather emerges from the data. Group Concept Mapping supports the researcher in dealing with diverse information, structured in various ways, which is a problem in Affinity diagram sessions. The visualisation through conceptual maps, pattern matches and bi-variate graphs, called go-zones, makes the interpretation of the outputs more beneficial.

The GCM procedure used for the development of the LinkedUp EF required the experts first to generate a list of indicators by completing a focus prompt. Then they were asked to sort the indicators into groups and rate them on two values: importance and priority. The focus prompt read:

"One specific indicator of the evaluation framework for assessing the Open Web Data application in the educational domain is ...".

The GCM approach generated first a number of indicators, which were then structured in more general categories (evaluation criteria) applying the MDS and HCA. Each evaluation criterion was operationalised through the set of concrete indicators in a particular cluster (Drachsler, Stoyanov, d'Aquin, Herder and Dietze, 2014). The Group Concept Mapping review produced six criteria, namely: Group support activities, Privacy, Educational Innovation, Usability, Performance, and Data (see Figure 1).

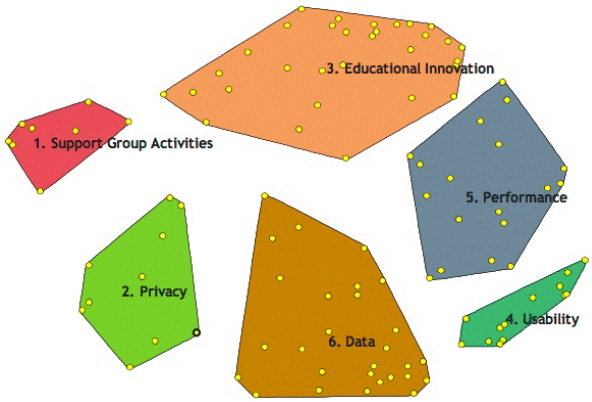


Figure 1. Clusters identified by Group Concept Mapping study

A more detailed analysis of the content of each cluster suggested that *Support group activities* is about target groups, so the consortium renamed it as *Audience*. The cluster *Privacy* referred to a broader scope of legal issues and its name was changed to *Legal aspects*. In addition, indicators and possible methods to test those indicators were identified (see Figure 2).

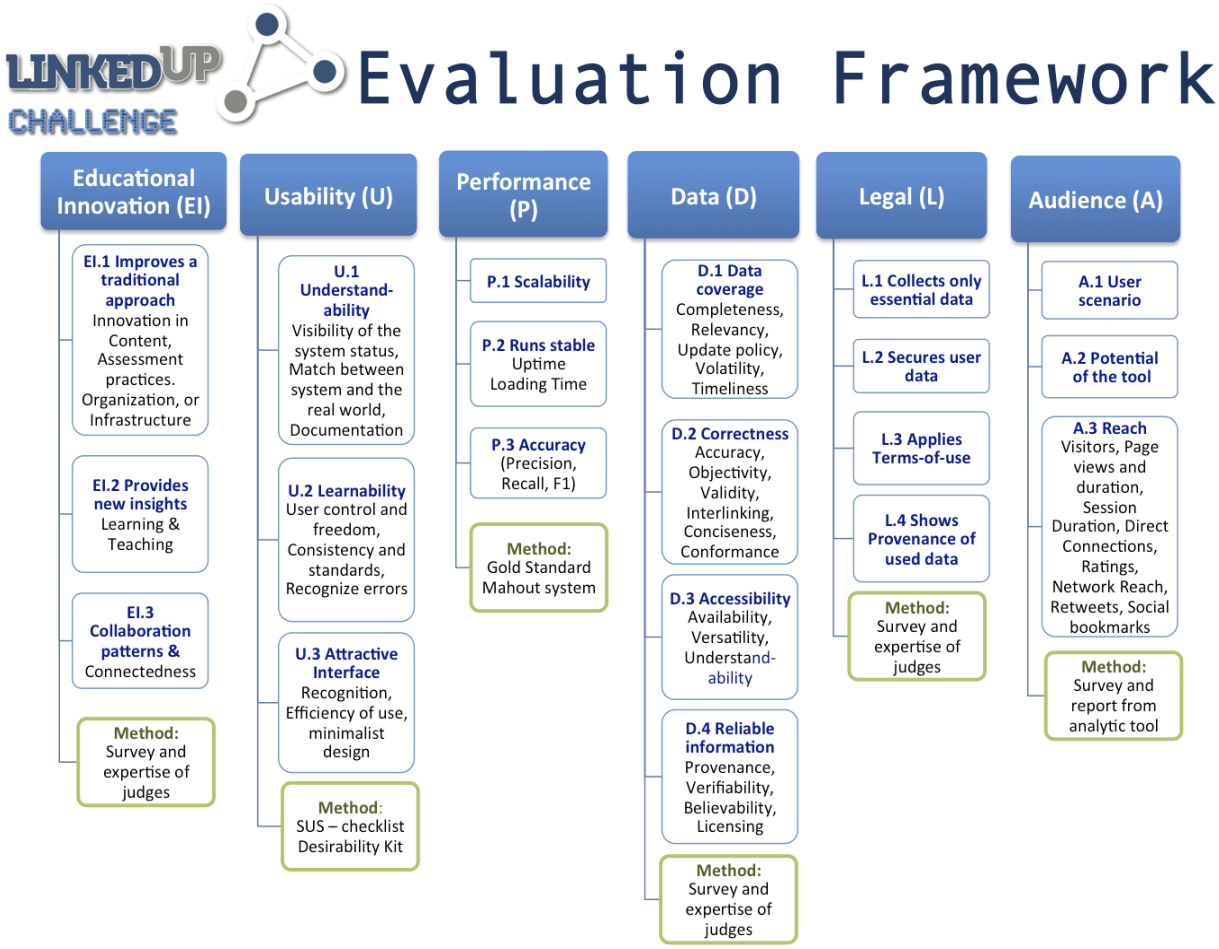


Figure 2: The comprehensive version of the EF derived from the GCM study and a literature review.

A scoring sheet containing the six criteria with a short explanation attached to each was prepared to support reviewers' judgement on the quality of the submissions to the competitions. Each criteria was scored on a 5-stars scale, 1 being the lowest, 5 the highest mark. The zero mark was also

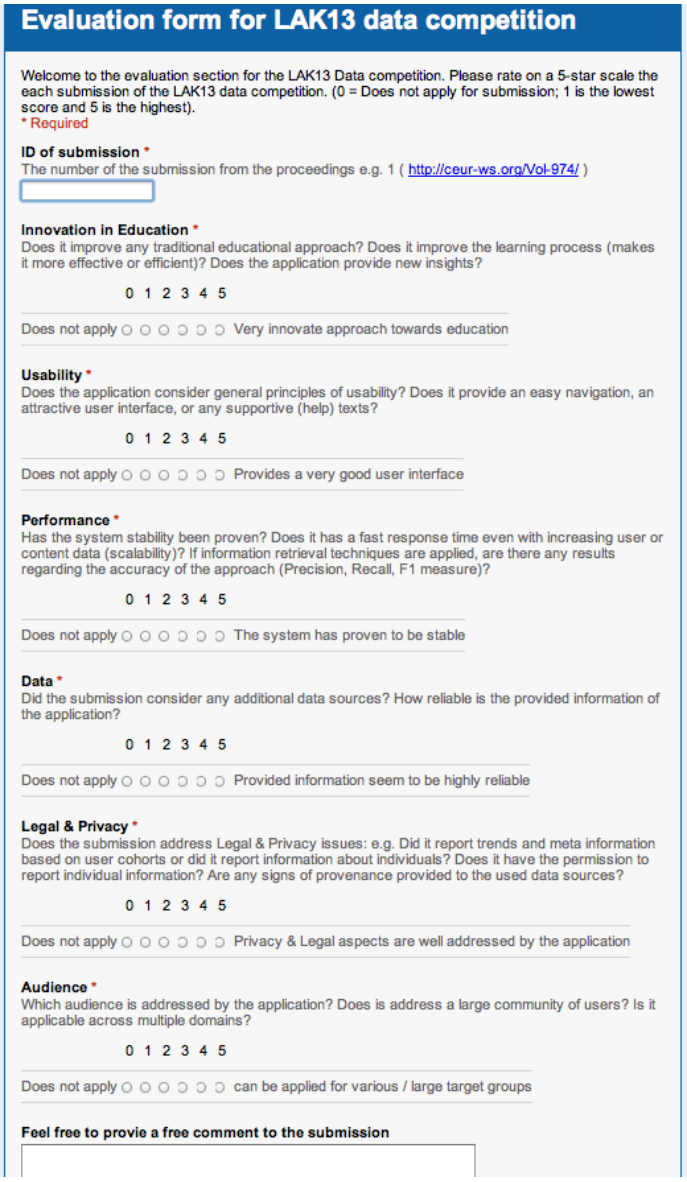
included to indicate that a criteron does not apply to a submission.

2.1.1 Pilot version of the Evaluation Framework. The LAK13 Case.

A pilot version of the EF was tested with the audience of a tutorial / workshop organised by the LinkedUp team at the Learning Analytics and Knowledge conference (LAK13) in Leuven, Belgium in March 2013. At the LAK13 we organised a pre-version of the LinkedUp challenge, i.e the LAK data challenge, as a joint effort between SoLAR (Society of Learning Analytics Research) and several affiliated organisations. The LAK13 dataset provided the scientific papers of the past proceedings of the LAK conference series, the Educational Data Mining conference proceedings series, and a Special Issue of the Educational Technology and Society Journal.
We received 8 submissions in total that can be found in the CEUR proceedings[1]. The audience (n=22) gave 72 ratings for the 8 submissions on the 6 evaluation criteria (see Deliverable D2.2.1). We got very positive feedback from the audience regarding the evaluation criteria. The participants were very satisfied with the criteria and the guidelines for the evaluation of the LAK13 data submissions. They thought the description of the evaluation criteria was clear and the criteria themselves were easy to apply. It was suggested to make the evaluation items more concrete by formulating a few evaluation indicators for each criteria.

The LinkedUp consortium was also pleasantly surprised by the seamless assessment of the submissions. There was a high overlap between the ranking of submissions of the LinkedUp consortium and the rating provided by the audience. The light version of the EF therefore exceeded expectations and smoothly supported the awarding of the submissions.

[1] http://ceur-ws.org/Vol-974/



Evaluation form for LAK13 data competition

Welcome to the evaluation section for the LAK13 Data competition. Please rate on a 5-star scale the each submission of the LAK13 data competition. (0 = Does not apply for submission; 1 is the lowest score and 5 is the highest).
* Required

ID of submission *
The number of the submission from the proceedings e.g. 1 (http://ceur-ws.org/Vol-974/)

Innovation in Education *
Does it improve any traditional educational approach? Does it improve the learning process (makes it more effective or efficient)? Does the application provide new insights?

0 1 2 3 4 5

Does not apply ○ ○ ○ ○ ○ ○ Very innovate approach towards education

Usability *
Does the application consider general principles of usability? Does it provide an easy navigation, an attractive user interface, or any supportive (help) texts?

0 1 2 3 4 5

Does not apply ○ ○ ○ ○ ○ ○ Provides a very good user interface

Performance *
Has the system stability been proven? Does it has a fast response time even with increasing user or content data (scalability)? If information retrieval techniques are applied, are there any results regarding the accuracy of the approach (Precision, Recall, F1 measure)?

0 1 2 3 4 5

Does not apply ○ ○ ○ ○ ○ ○ The system has proven to be stable

Data *
Did the submission consider any additional data sources? How reliable is the provided information of the application?

0 1 2 3 4 5

Does not apply ○ ○ ○ ○ ○ ○ Provided information seem to be highly reliable

Legal & Privacy *
Does the submission address Legal & Privacy issues: e.g. Did it report trends and meta information based on user cohorts or did it report information about individuals? Does it have the permission to report individual information? Are any signs of provenance provided to the used data sources?

0 1 2 3 4 5

Does not apply ○ ○ ○ ○ ○ ○ Privacy & Legal aspects are well addressed by the application

Audience *
Which audience is addressed by the application? Does is address a large community of users? Is it applicable across multiple domains?

0 1 2 3 4 5

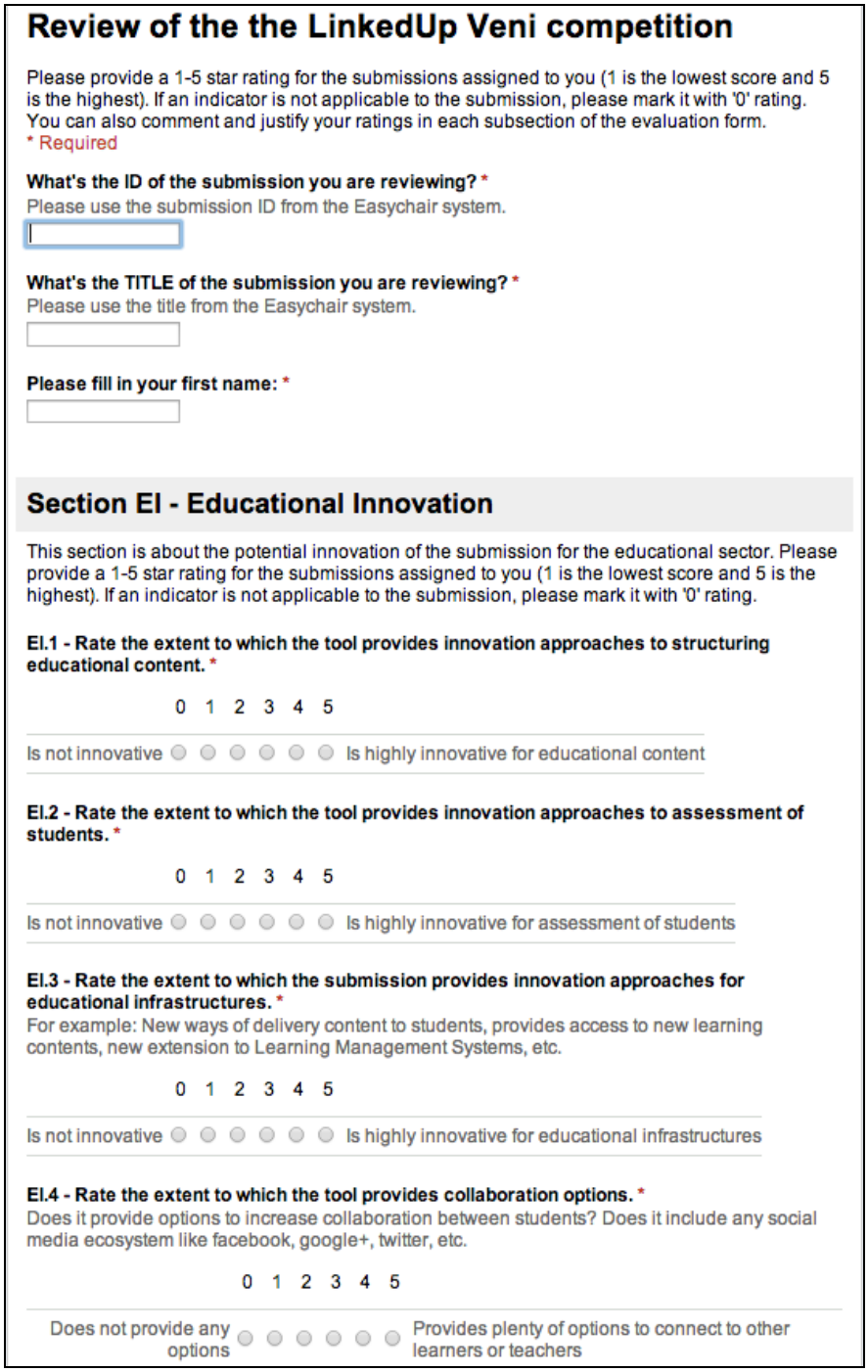
Does not apply ○ ○ ○ ○ ○ ○ can be applied for various / large target groups

Feel free to provie a free comment to the submission

Figure 3: The light version of the LinkedUp EF as used at the LAK13 data competition.

2.1.2 The Veni Competition

Based on the results of the three design processes – (1) an expert validation over the Group Concept Mapping method, (2) a literature review study (both in D2.1), and (3) an initial trial of the EF within the LinkedUp tutorial at LAK13 – a first comprehensive version of the LinkedUp EF was created for the Veni competition see Figure 4 and Appendix A for full details.



Review of the the LinkedUp Veni competition

Please provide a 1-5 star rating for the submissions assigned to you (1 is the lowest score and 5 is the highest). If an indicator is not applicable to the submission, please mark it with '0' rating. You can also comment and justify your ratings in each subsection of the evaluation form.
* Required

What's the ID of the submission you are reviewing? *
Please use the submission ID from the Easychair system.

What's the TITLE of the submission you are reviewing? *
Please use the title from the Easychair system.

Please fill in your first name: *

Section EI - Educational Innovation

This section is about the potential innovation of the submission for the educational sector. Please provide a 1-5 star rating for the submissions assigned to you (1 is the lowest score and 5 is the highest). If an indicator is not applicable to the submission, please mark it with '0' rating.

EI.1 - Rate the extent to which the tool provides innovation approaches to structuring educational content. *

0 1 2 3 4 5

Is not innovative ○ ○ ○ ○ ○ ○ Is highly innovative for educational content

EI.2 - Rate the extent to which the tool provides innovation approaches to assessment of students. *

0 1 2 3 4 5

Is not innovative ○ ○ ○ ○ ○ ○ Is highly innovative for assessment of students

EI.3 - Rate the extent to which the submission provides innovation approaches for educational infrastructures. *
For example: New ways of delivery content to students, provides access to new learning contents, new extension to Learning Management Systems, etc.

0 1 2 3 4 5

Is not innovative ○ ○ ○ ○ ○ ○ Is highly innovative for educational infrastructures

EI.4 - Rate the extent to which the tool provides collaboration options. *
Does it provide options to increase collaboration between students? Does it include any social media ecosystem like facebook, google+, twitter, etc.

0 1 2 3 4 5

Does not provide any options ○ ○ ○ ○ ○ ○ Provides plenty of options to connect to other learners or teachers

Figure 4: The Veni version of the LinkedUp EF (Full version in Appendix A).

.

We provided the Veni judges with a scoring sheet in Google forms that allowed a comparison of the reviews given by the judges in an efficient and effective manner. The outcomes of the scoring sheet supported the members of the review board to evaluate the submissions and award the prizes. Another advantage was that we could integrate survey-based systems such as SUS for Usability and directly compute the SUS score for an application. In addition, further computations could be created in the allocated Google spreadsheet of the Google form and directly shared within WP2. A full version of the Veni EF and the scoring sheet can be found in D2.2.1.

The Veni competition was the first one in the competition series comprising the LinkedUp Challenge[2]. Veni ran from May 22 until June 27, 2013, and requested participants to submit "*an innovative and robust prototype or demo that used linked and/or open data for educational purposes*". The LinkedUp Challenge website defines "educational purposes" by stating that the tools and applications developed must be relevant to education – in the broadest sense of the word. This might mean that they aid learning in some way or that they support educational objectives by expanding knowledge and encouraging critical thinking.

By the closing date, 22 valid submissions had been received from 12 different countries (4 from the UK, 3 from France, 3 from Spain, 4 from the USA, 2 from the Netherlands and 1 from Greece, Bulgaria, Belgium, Italy, Argentina and Nepal). The majority of entries were from teams based at universities or from start-up companies. There were also entries from independent consultants. The submissions varied in terms of number of authors, institutions, countries, and domains.

The participants in the competition had interpreted the specification "educational purposes" in a variety of innovative ways. A number of the submissions, such as *Course Finder*, *LinkedIn MOOCs counselor* and *Moocrank*, had looked at MOOC and course data and offered cross-searching mechanisms. Some, such as *PoliMedia*, *Dr Hoo*, *Neuro-Cloud Free Textbook Project* and *Enrichment of Young Digital Planet's biology lessons*, had focused on discipline-specific data such as political studies, biology, etc. and offered new pedagogical approaches based on data applications for learners to explore and understand discipline-specific content. Others, such as *REthink* and *Learner Journey Navigation System*, also took an exploratory approach using topic maps, but operated on the cross-section of several disciplines. Two of the submissions, *One Million Museum Moments* and *Mismuseos.net*, looked in particular at cultural heritage data and how museum data could be used in an educational context. The remaining submissions covered other educational related areas including use of conference publications, reading lists, mobile learning and annotation.

2.1.3 Lessons-learned from the Veni version

One reason we chose to use the GCM methodology was that its construct and content validity has been proven in projects conducted in various domains (ref. Rosas and Kane, 2012). However, it was tempting to check how a well-known instrument in the technology-enhanced learning domain such as the System Usability Scale (SUS) would work as part of the EF scoring sheet. For the Veni competition the short usability scale in the original EF scoring sheet was replaced by the 10-items SUS instrument. The Veni call attracted 22 submissions, which were evaluated by 25 reviewers. Five reviewers, selected randomly, were interviewed using a consolidated interview script on what worked well and what not so well while they had applied the EF. The interviews were recorded and consequently transcribed in verbatim. The texts were analysed qualitatively (ground theory approach and content analysis) and quantitatively (language technology using the Leximancer software and the free of charge service Text-is-Beautiful (see D2.2.1). The quantitative text mining analysis through concept maps, concept cloud and concept web had independently confirmed the findings from the qualitative text analysis regarding the issues that need to be addressed for the next round of the

[2] http://linkedup-challenge.org/veni.html

competition

In summary, several suggestions for changes can be concluded from the findings. The first deals with the not applicable option. As it has different meaning for different people, the suggestion is to remove it from the survey. The submissions must cover all of the evaluation criteria. This option made it difficult to compare numerically the submissions.

The second suggestion is to return to the short version of the usability criteria. SUS did not affect the final results in any significant way but made it difficult for the reviewers and researchers who consolidated the results of the evaluation process. SUS applies a completely different scoring schema that needed to be adjusted to the scoring on the other criteria. An additional argument is that the original Usability criterion we used includes indicators, which score very high on validity and reliability (no less than 90% explaining the variance in the data) as indicated by other usability measures (e.g. SUS, CSUQ, UTAUT). Jeff Sauro, who popularised the SUS measure recommends using a small number of items, which of course need to be checked for validity and reliability (Sauro, 2010; see also Lewis and Sauro, 2009.). He also advises that in addition to the usual psychometric properties such as validity and reliability, a measurement instrument should be short, easy to respond to, easy to administer, and easy to score.

Reconsidering the indicators of the *Data* and *Performance* criteria was the third suggestion about how the EF could be improved as the reviewers experienced difficulties to interpret them properly.

Another suggestion concerned the indicators of the *Legal* criterion. Those that require a dichotomous answer ('yes/no') should be reformulated so as to conform to the scale style of the other criteria (" Please provide 1 to 5 star rating"..).

Finally, although not related to the EF directly, the reviewers did not feel comfortable with using two tools for managing the evaluation process: Google and Easy Chair.

2.2 Second version of the Evaluation Framework. The Vidi Case.

The second version of the EF tried to incorporate changes that dealt with the suggestions for improvements after the Veni competition, namely: removing the option 'not applicable'; using the short version of the Usability criterion rather than the SUS; a measure to indicate the Overall evaluation of the submission and the level of confidence of the reviewer; and a better formulation of the items operationalising the criteria of Data, Performance and Legal (see Figure 5 and Appendix B for full overview). This time, to gather information about the EF we conducted a survey through a questionnaire. The reasons for using a survey were: (a) organising, conducting and especially transcribing the interviews in the evaluation of the first version of the EF was time consuming; we needed the evaluation of the EF done in a shorter time period; and (b) we were thinking already of the next Vici round when we would have even less time to carry out the assessment of the EF so it would be beneficial to have a measurement instrument ready to use.

Figure 5: The Vidi version of the LinkedUp EF (Full version in Appendix B).

2.2.1 The Vidi Competition

The Vidi Competition was the second one in the LinkedUp Challenge and ran from the November 4, 2013 until February 14, 2014. It requested participants to submit "*an innovative and robust prototype or demo that used linked and/or open data for educational purposes*", with the remark that "*Your tool still may contain some bugs, as long as it has a stable set of features and you have some proof that it can be deployed on a realistic scale*".

Apart from the Open Track, the Vidi competition featured two Focused Tracks, which were selected from eight candidate focused tasks that were developed from the use cases for the Veni competition in WP5, with further guidance from an analysis of the Veni entries. Focused Track 1, *Simplificator*,

called for applications easing access to complex information by summarising them in a simpler form. Focused Track 2, *Pathfinder*, called for applications easing access to recommendation and guidance when choosing an appropriate curriculum of courses and related resources.

By the closing date, 14 valid submissions had been received from 12 different countries. 10 entries related to the Open Track ranging from ways to browse bibliographic records and navigate scientific records, to tools that allow users to build multimedia linked data stories about art or visualise learning materials. Further, we received 4 entries to the *Simplificator* Focused Track, allowing simplification of archeological, historical and health data. Unfortunately, the *Pathfinder* Focused Track did not receive a sufficient number of submissions and it was decided to close this specific track. The entries were heterogeneous, consisting of varying number of authors, institutions, countries etc.

Of the shortlisted Open Track submissions, two submissions provided intelligent search functionality in educational resources: *AGRIS* links bibliographic references from the agricultural domain to external datasets, and *Solvonauts* is an open educational search engine. Three submissions focused on connecting people and things with one another: *Rhizi* is the further development of the Veni submission *KnowNodes* and allows users to interactively create connections; *Konnektid* allows people to connect to others in order to learn or teach something; *LOD Stories* lets users connect artworks, artists and places into a storyboard. Finally, two submissions help users to make sense of data with various visualisations: *DBLPExplorer* is a browsing and exploration interface for publications in the field of computer science. With *TuVaLabs*, students and teachers can explore and visualise datasets and create assignments around them. The shortlisted submissions for *Simplificator* focused on two specific domains: a visualisation of labour conflicts in the Netherlands and an electronic Discharge Letter that makes the lives of patients and doctors easier. Apart from the tool itself, several submissions also provided a SPARQL endpoint for their data.

2.2.2 Lessons-learned from the Vidi version

We evaluated the usefulness of the EF after Vidi with a questionnaire sent to the judges after they finalised their reviews. All reviewers had a strong background in computer science and had been doing research in the technology-enhanced learning domain for several years. The questionnaire contained 15 items on how easy/difficult it was for the reviewers to apply the assessment indicators of the EF. We have two general questions to check how the EF worked in assessing the overall quality of either open or focused tracks of the Vidi competition. After each of the close-ended items, there was also a space available for comments. 25 Vidi reviewers were invited to take part in the survey. 12 of them responded positively.

The analysis of the questionnaire revealed several results (see also Figure 6). 11 out of 16 items received a score higher than M = 3,75 with an overall M = 3.77.
The following items received the highest value:

- Providing a statement on the terms of use (M = 4.33; SD = 1.23) – *Legal*;
- The application has an attractive interface (M = 4.33; SD = 0.65) – *Usability*;

- The application implements an innovative educational concept (M = 4.25; SD = 0,866) – *Educational Innovation*;
- Easy to use (M = 4.17; SD = 0.718) – *Usability*;
- Availability of the tool to its target users (M = 3.92; SD = 0.996) – *Performance*;

The judgment on the overall quality of the focus track submission (M = 3.92; SD = 0.669) and the overall judgment on the open track (M = 3.83; SD = 0.389) got also high scores.
The items that scored relatively low are as follows:

- The application is more effective than existing applications (M = 3.08; SD = 0.996) – *Educational Innovation*;
- Collecting only needed personal information about the user (M = 3.25; SD = 1.138) – *Legal*;
- The application is more efficient than existing applications (M = 3.33; SD = 1.55) – *Educational Innovation*;
- Consuming multiple data sources (M = 3.33; SD = 0.778) – *Data*;
- Exposing new datasets to the Linked Data cloud (M = 3.42; SD = 0.9) - *Data*.

The analysis of the open-ended questions indicated that (a) effectiveness and efficiency of the applications were easy to judge but it became more difficult when submissions needed to be compared to other existing applications and that (b) the comments on *Data* indicators need a better formulation in the next version of the EF.

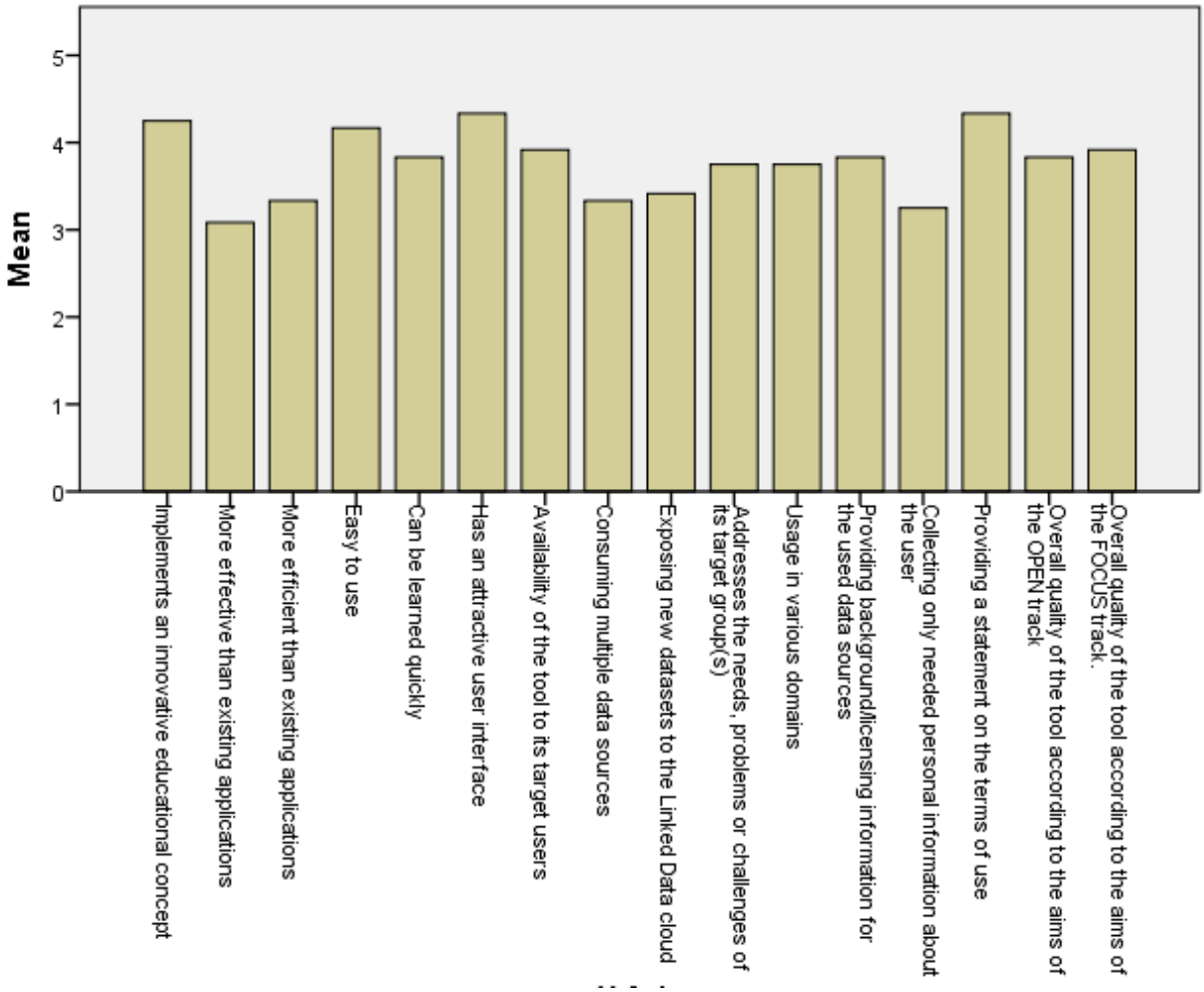


Figure 6. Visualisation of results from the Vidi survey.

Overall, several conclusions can be drawn. Based on the figures produced by the survey and the comments (N =33; see also Appendix C of deliverable WP2- D 2.3.2) made by the reviewers, it seems that the EF worked well for the Vidi competition and does not need substantial changes, except a better formulation of some of the items. Also, the comparison to other existing applications in the formulation of the items referring to effectiveness and efficiency needs to be removed. The two items operationalising the *Data* criterion, namely 'Consuming multiple data' sources' and 'Exposing new datasets to the Linked Data cloud' must be better defined. Finally, the results suggest that the reviewers felt comfortable with the reduced number of the usability indicators, which was the major change in the Vidi EF.

2.3 Third version of the Evaluation Framework. The Vici Case.

The third and final version of the EF incorporates changes that dealt with the suggestions for improvements after the Vidi competition, namely; (a) for Educational Innovation (EI) - removing the text 'Comparison to other existing applications'; (b) Data - 'Consuming multiple data sources' and 'Exposing new datasets to the Linked Data cloud' have been further operationalised with a clear scale and improved text description. Figure 7 shows a screenshot of the evaluation form, a full

overview can be found in Appendix C.

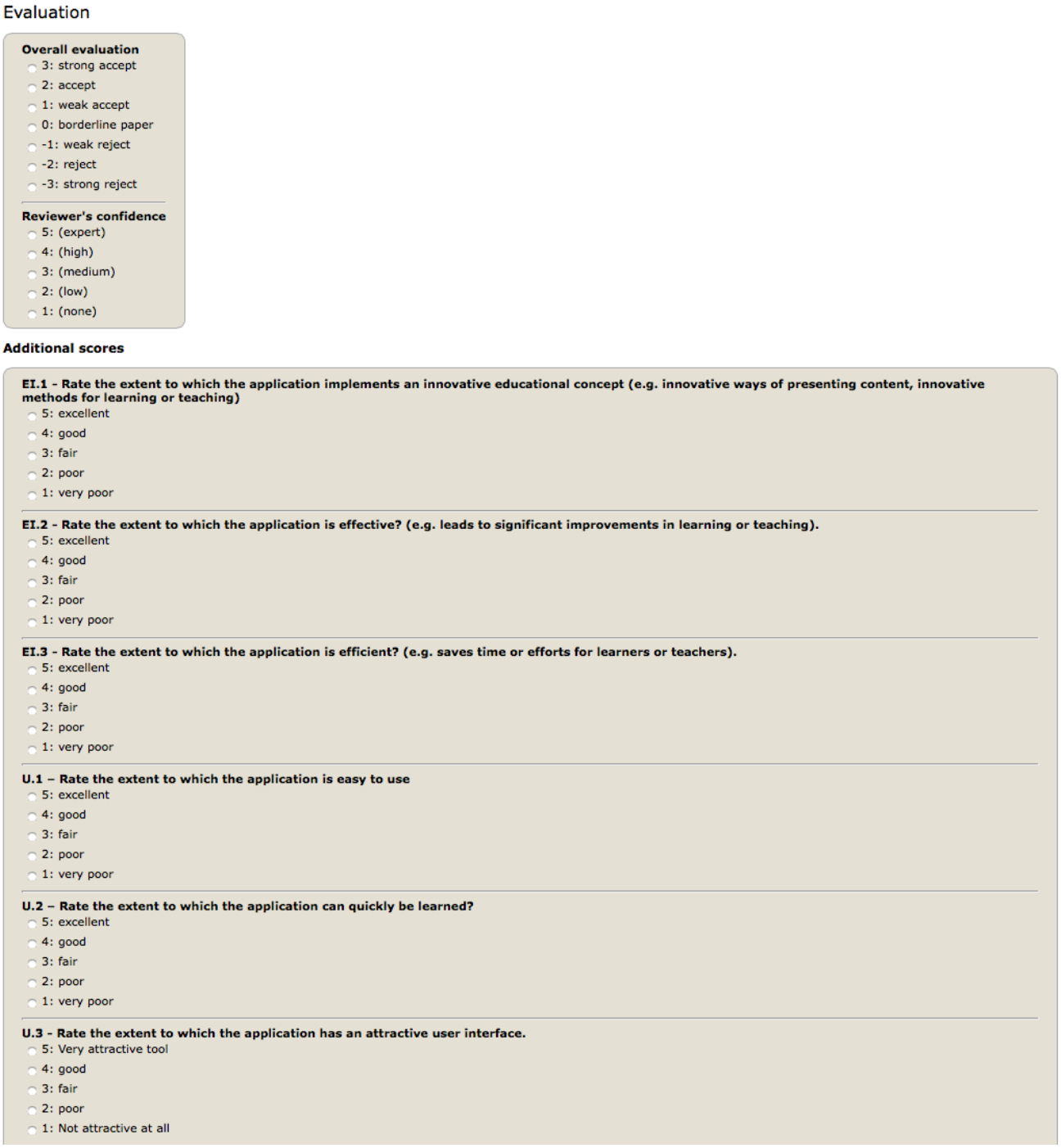


Figure 7. The Vici version of the LinkedUp EF (Full version in Appendix C).

2.3.1 The Vici Competition

The Vici Competition was the third and final competition in the LinkedUp Challenge and ran from the June 4 until September 5, 2014. It requested participants to submit "*advanced prototypes and tools that are driven by linked and/or open data. You can submit your Web application, App, analysis toolkit, documented API or any other tool that connects, exploits or analyses open or linked data and that addresses real educational needs*" and added "*Your tool should be mature and stable; it should be used or have been used by a fair amount of users on a realistic scale.*"

Alongside the Open Track, the Vici competition featured two Focused Tracks. The Focused Track: *Supporting Developing Countries* looked for educational applications that target developing countries, addressing context-specific problems, issues and needs, being technical, societal or environmental. The Focused Track *Water Resources & Ecology* involved enhancing journal article content along with related research statistics and datasets to assist in discovery, learning and interpretation of disparate content and data. A number of Elsevier datasets were made specially available for this track.

2.3.2 Lessons-learned from the Vici version

Due to the positive experiences in Vidi and the short timeframe to evaluate the Vici competition we applied again a questionnaire-based survey. The questionnaire contained closed-ended items on how easy/difficult it was for the reviewers to apply the assessment indicators of the EF. We included two general questions to check how the EF worked in assessing the overall quality of either open or focused tracks of the competition. After each of the close-ended items, there was also a space available for comments. 29 Vici reviewers were invited to take part in the survey. 18 of them responded positively. As already in the previous competition, all of them were experts in the technology-enhanced learning domain with a special interest in semantic web and open linked data.

The analysis of the questionnaire revealed several results (see also Figure 8).
The following items received the highest value:

- 'The application can quickly be learned' (M = 4.28; SD = 0.83)
- 'The application is easy to use' (M = 4.17; SD = 0.92)
- 'The tool provides a statement on the terms of use' (M = 4.17; SD = 0.99)
- 'The application has an attractive user interface' (M = 4.17; SD = 1.1)
- 'The tool provides background / licensing information for the used data sources' (M = 4.06; SD = 1.06).

The items that scored relative low are:

- 'The application is effective. (e.g. leads to significant improvements in learning or teaching)' (M = 2.83; SD = 0.99)
- 'The tool collects only needed personal information about the user' (M = 2.94; SD = 1.26)
- 'The application is efficient (e.g. saves time or efforts for learners or teachers' (M = 3.22; SD = 0.8)
- 'The application addresses the needs, problems or challenges of its target group(s)' (M = 3.28; SD = 0.75)

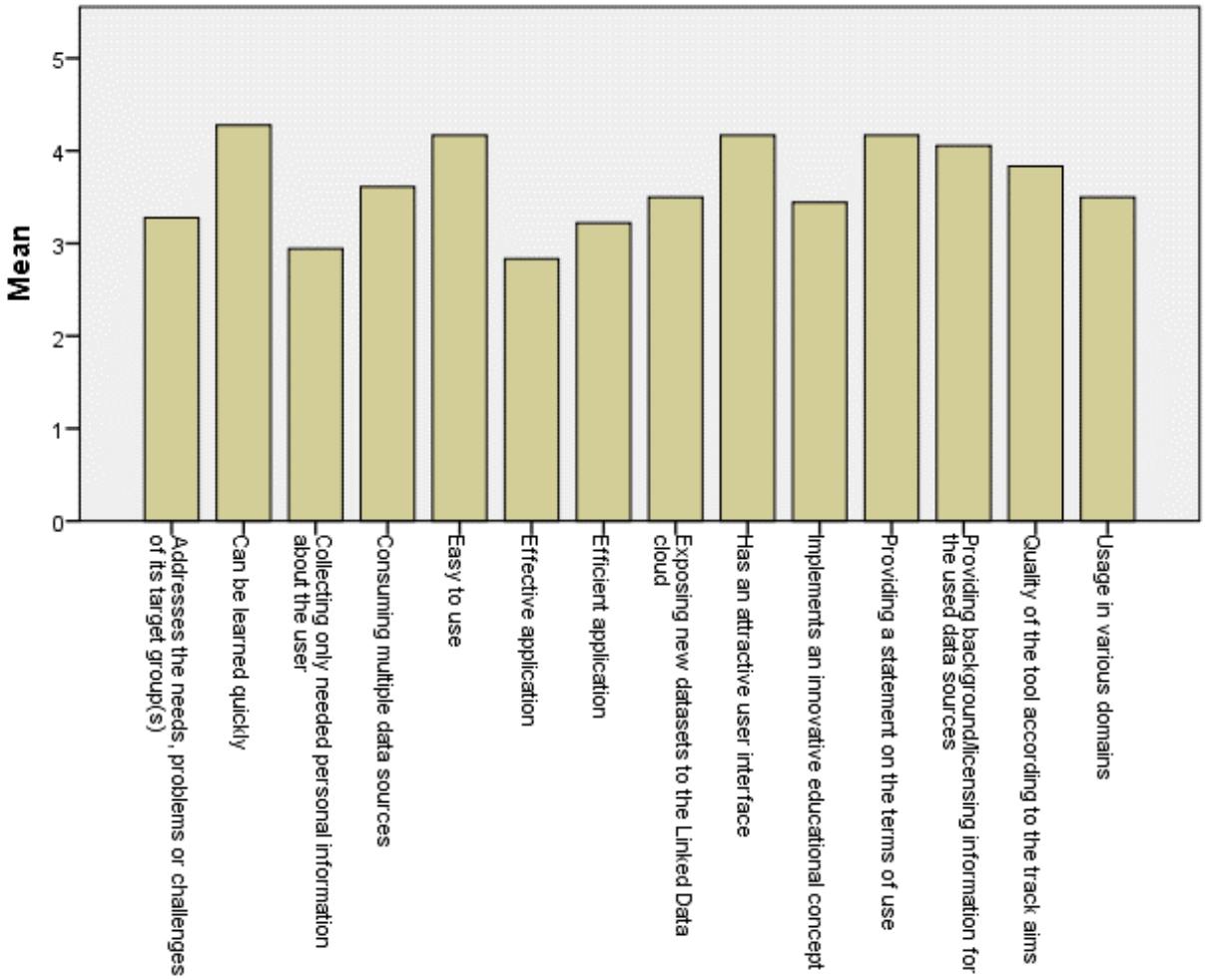


Figure 8. Visualisation of results from the Vici survey.

Some of the reasons why the items concerning effectiveness and efficiency of the tool scored relatively low were given in the comments (open-ended items) of the reviewers to these indicators (see Appendix A of deliverable WP2 - 2.3.3). The reviewers could not make an informed judgement on these indicators as the submissions did not provide sufficient information in this respect. Some of the reviewers claimed that the applications they reviewed did not address an educational problem. The indicators operationalising the criterion *Educational Innovation*, including those referring to effectiveness and efficiency, were meant to be about *perceived* effectiveness and efficiency and the potential of the tool to improve learning and save time and efforts.

According to the reviewers' comments, it is not easy to find out whether the tool implicitly collects personal information (apart from data entered by the users themselves). There was also ambiguity about was meant by 'domain' ('The application can be used in various domains'). In addition, the authors of the submitted tools did not always provide explicit information about the target group of their application.

Some conclusions that follow from the survey close- and open-ended questions are:

- The three items operationalising the *Usability* criterion, again as in the Vidi competition, scored high.

- This time around the reviewers experienced some more difficulties with the indicators included in the *Educational Innovation* criteria, more specifically the items on effectiveness and efficiency. The reviewers could not check that as no information was provided by the authors of the submissions. The formulation of the items should therefore include the word 'perceived'. The reviewers were expected to evaluate the perceived effectiveness and efficiency, that is the potential of the tools to improve learning or teaching and to save time.
- The term 'domain' should be specified ('The application can be used in various domains').
- It was not easy for the reviewers to judge whether the tool collects personal information implicitly (i.e. apart from data entered by the users).
- The criteria *Educational Innovation* and *Data* were weighted as they represent the essence of the project.

3 Conclusions and suggestions for competition organisers

The EF is based upon rigorous empirical data. The Group Concept Mapping (GCM) approach collected information about possible assessment indicators from experts in the domain, who additionally structured it individually through grouping the indicators on similarity in meaning. Then some advanced statistical techniques such as multidimensional scaling and hierarchical cluster analysis identified clusters (criteria) of indicators as a shared vision of the experts participating in the study. In this way, the GCM contributed to the construct validity of the EF.

The initial set of criteria was first a subject of inspection by the experts within the LinkedUp project. Second, we surveyed the opinions of the experts invited to serve as reviewers of the LinkedUp submissions for Veni, Vidi and Vici competitions on how easy or difficult it was for them to apply different criteria and indicators of the EF. In addition we interviewed some of the reviewers to get more details about how useful the EF was for them. We used this information to improve the EF after each competition round. When looking back at the development of the EF over the past 1.5 years, we outline the following suggestions to future data competition organisers in order to have a well grounded and transparent evaluation procedure.

1. **Designing a data competition starts with a definition of evaluation criteria**
   Design your evaluation criteria at an early stage of your data competition. It is crucial to have clear evaluation criteria and indicators prepared before the call for submissions is announced. In that way you can add the 'success criteria' already in your call for submissions. This will make the evaluation process much easier, transparent and convenient for all participating stakeholders (organisers, competition participants, and judges). The evaluation criteria are the backbone of the whole competition and need to be well prepared beforehand.

2. **Test the understandability of your evaluation criteria**
   After having a good selection of evaluation criteria make sure that those are also well formulated and easy to understand by the competition stakeholders. Although we identified important criteria at an early stage within LinkedUp, it was not a trivial task to find a proper formulation that provides a common understanding of each indicator. We therefore highly recommend to provide – if possible – a descriptive text to each indicator with a suitable example to make the meaning of the indicator as clear as possible. Furthermore, each indicator should be read and tested for its understandability by 'external' partners that can represent the competition judges and participants.

3. **Do not use an 'not applicable' option**
   In the first Veni competition we applied quite a lot of indicators for each criteria. We wanted to cover all potential submissions for the Open Track. As a consequence we implemented a 'not applicable' option to enable the judges to flag that some of the evaluation indicators do not fit to a submission. But the '*not applicable'* option was a major source of confusion for the judges. It made the comparison of the scores given to the submissions hardly comparable

as they could have come from different indicators. As a consequence, we clearly described the evaluation criteria to the participants and judges in the Vidi call and agreed that the evaluation criteria are fixed parameters which all submissions would be judged on.

4. **Less (indicators) are more (preferable)**
   Along with the 'not applicable' option from point 3, we also learned that the evaluation items need to be phrased in more general terms. Thus, instead of asking if a tool provides 'new assessment methods' for educational innovation, we better asked for the overall effectiveness, efficiency and innovation of the tool for education. Asking for specific features in an indicator only makes sense for the Focused Track where an additional task is specified to the overall competition goals.

5. **Unification of the scale of evaluation indicators**
   In the Veni competition we still applied the original 10-items SUS scale, as we thought that applying a well-established usability instrument is a good practice. But after the experiences gained from Veni, we needed to apply a shorter usability scale in the Vidi, competition. There are two reasons for that: (a) the SUS applies a completely different scoring scale than the other evaluation indicators, and (b) SUS is not tailored for judges of data competitions. The results of the SUS scale (range from 25 - 100 points) affected the overall evaluation results (the sum of the average score of all evaluation criteria). After reviewing some papers about the SUS scale we found that the original *Usability* criterion includes indicators, which score very high on validity and reliability (no less than 90% explaining the variance in the data) as indicated by other usability measures (e.g. SUS, CSUQ, UTAUT). In addition, high profile experts in the domain such as J.Sauro, and J. Lewis recommend using a few items if possible (Lewis and Sauro, 2009; Sauro, 2010). They also advise that in addition to the usual psychometric properties such as validity and reliability, a measurement instrument should be short, easy to respond to, easy to administer, and easy to score. Furthermore, our assessment of the EF identified that the LinkedUp judges had difficulties to apply the SUS scale that is designed for 'end users' rather than 'judges' of data competitions. Items from SUS such as: "*Would you use this tool more often?*" did not seem to be appropriate to our LinkedUp judges.

6. **Weighting specific evaluation criteria**
   Weighting of criteria can be very informative to provide a different perspective on strong and weak aspects of your submissions. It can for instance be applied to weighting specific objectives of a data competition as we did in LinkedUp with Linked Data and Education. But especially under the view of point 1, weighting of factors should be transparent to the data competition participants and the judges and not be applied later on. An important question is the strength of a weight, as it should amplify a criterion but not diminish the effect of other criteria. Within the Vici competition we decided to apply a weighting factor of 1.5. A factor of 1.5 seems to be strong enough to make a distinction in the ranking (a factor between 1.1 to 1.4 might be too weak), whereas a higher factor (1.6 to 1.9) might be too strong and even diminish the effect of other criteria.

We have incorporated those lessons-learned from the three data competitions Veni, Vidi, and Vici into the LinkedUp toolbox to support others who are planning to organise open data competitions[3]. The toolbox contains several tools that are helpful when setting up academic or industry competitions. The aspects covered by the toolbox are: (1) competition framework, (2) evaluation framework, (3) guidance schedule, (4) data, (5) promotion methodology, (6) and legal and IPR. For the EF the toolbox provides a question and answer page about the evaluation process as well as a visualisation of the comprehensive EF with all criteria and indicators (see Figure 1) and the customised EF for the LinkedUp competitions in its final stage (see Figure 9). This final version incorporates all lesson-learned from the Veni, Vidi, and Vici competitions (see Appendix D).

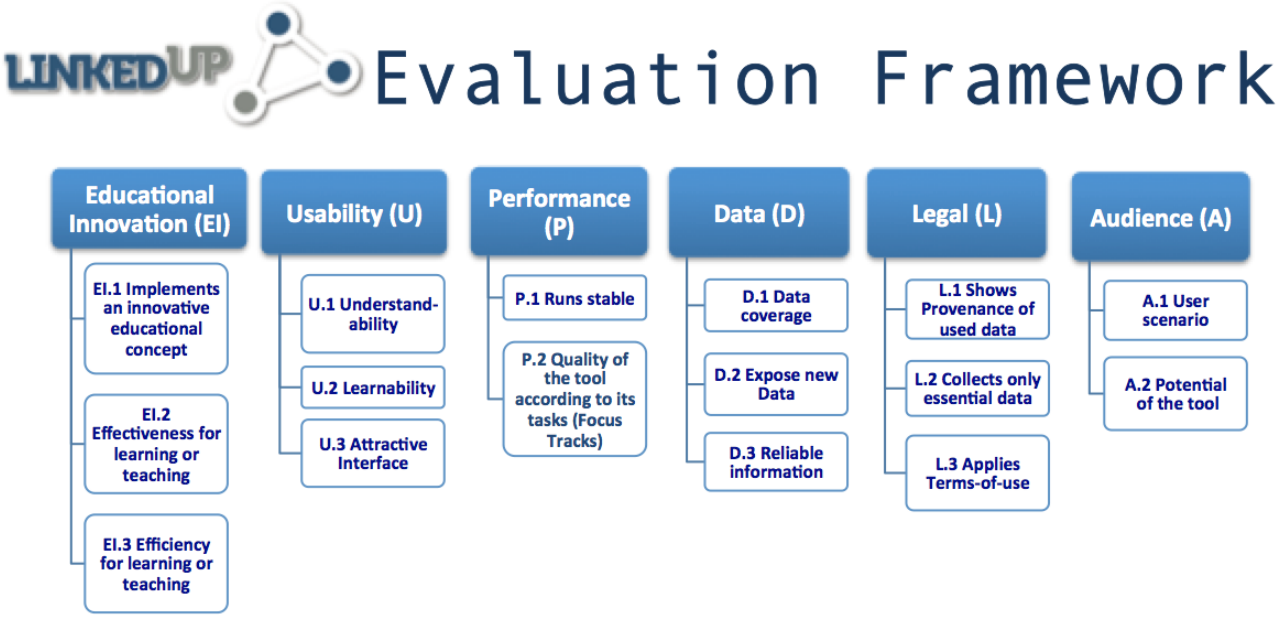


Figure 9. Final version of the LinkedUp EF (Full version in Appendix D).

Among the textual description of the evaluation process and EF, we created a Google spreadsheet that can provide the graphs about the evaluation results in the Veni, Vidi, and Vici competition (see deliverables D2.3.1, D2.3.2 and D2.3.3). The spreadsheet is publicly accessible[4] and filled with simulated data to make it easier to be understood. With this template at hand, future data competition organisers can easily collect the ratings for their submissions, enter those into the spreadsheet, and easily create the overviews of their data competition per criteria. The spreadsheet is the final version of our evaluation tool as it has been used within the LinkedUp competitions.

The final version of the EF and the evaluation spreadsheet could be applied as it is – or with some small adaptations – to LinkedUp-like competitions in the technology-enhanced learning domain. Several organisations such as the Society of Learning Analytics (SoLAR), European Organisation of Technology-Enhanced Learning (EATEL), and the Learning Analytics Community Exchange project (LACE) have already indicated, that they want to use it for own data competitions in 2015.

[3] http://linkedup-project.eu/toolbox/

[4] https://docs.google.com/spreadsheet/ccc?key=0AgBoOSpxxjzYdHpWSTdvbXBGSFlxcE5qRnkxcXFxZXc&usp=sharing

If a more tailored EF is required by future competition organisers, we recommend to follow the procedure for developing and improving the LinkedUp EF as described in the previous WP2 deliverables and the EC-TEL paper (Drachsler, Stoyanov, d'Aquin, Herder, Guy, M. and Dietze, 2014). The Group Concept Mapping provided the empirical basis for defining the initial set of criteria and their operationalisation through a list of more concrete indicators. A mix of both quantitative and qualitative methods for data collection and analysis were applied for subsequently improving the LinkedUp Evaluation Framework.

4 Reference

1. Drachsler, H., Stoyanov, S., d'Aquin, M., Herder, E., Guy, M. & Dietze, S. (2014). An Evaluation Framework for Data Competitions in TEL. In Ch. Rensing, S. de Freitas, T. Ley and P. J. Muñoz-Merino (Eds). *Proceedings of the 9th European Conference on Technology Enhanced Learning*. Springer Lecture Notes in Computer Science (LNCS) series, V. 8719**.** EC-TEL 2014, Graz, Austria, September 16-19, 2014.
2. Holtzblatt, K., Wendell, J., & Wood, S. (2005). *Rapid contextual design*. San Francisco: Morgan Kaufmann
3. Kane, M., & Trochim, W. M. K. (2007). *Concept mapping for planning and evaluation*. Thousand Oaks, CA: Sage Publications.
4. Kuniavsky, M. (2003). *Observing the User Experience: A Practitioner's Guide to User Research*, pp 120-126. San Francisco: Morgan Kaufmann Publishers.
5. Lewis, J., & Sauro, J. (2009). The Factor Structure of the System Usability Scale. Proceedings of the 1st International Conference on Human Centered Design: Held as Part of HCI International 2009, pp. 94 - 103. Springer-Verlag Berlin.
6. Rosas, S.R. & Kane, M. (2012). Quality and rigor of the concept mapping methodology: a pooled study analysis. *Evaluation Program Planning*, 35, 236–245.
7. Sauro, J. (2010). If you could only ask one question use this one. Retrieved from http://www.measuringusability.com/blog/single-question.php

Appendix A – Veni Evaluation Form

Review of the the LinkedUp Veni competition

Please provide a 1-5 star rating for the submissions assigned to you (1 is the lowest score and 5 is the highest). If an indicator is not applicable to the submission, please mark it with '0' rating. You can also comment and justify your ratings in each subsection of the evaluation form.
* Required

What's the ID of the submission you are reviewing? *
Please use the submission ID from the Easychair system.

What's the TITLE of the submission you are reviewing? *
Please use the title from the Easychair system.

Please fill in your first name: *

Section EI - Educational Innovation

This section is about the potential innovation of the submission for the educational sector. Please provide a 1-5 star rating for the submissions assigned to you (1 is the lowest score and 5 is the highest). If an indicator is not applicable to the submission, please mark it with '0' rating.

EI.1 - Rate the extent to which the tool provides innovation approaches to structuring educational content. *

0 1 2 3 4 5

Is not innovative ○ ○ ○ ○ ○ ○ Is highly innovative for educational content

EI.2 - Rate the extent to which the tool provides innovation approaches to assessment of students. *

0 1 2 3 4 5

Is not innovative ○ ○ ○ ○ ○ ○ Is highly innovative for assessment of students

EI.3 - Rate the extent to which the submission provides innovation approaches for educational infrastructures. *
For example: New ways of delivery content to students, provides access to new learning contents, new extension to Learning Management Systems, etc.

0 1 2 3 4 5

Is not innovative ○ ○ ○ ○ ○ ○ Is highly innovative for educational infrastructures

EI.4 - Rate the extent to which the tool provides collaboration options. *
Does it provide options to increase collaboration between students? Does it include any social media ecosystem like facebook, google+, twitter, etc.

0 1 2 3 4 5

Does not provide any options ○ ○ ○ ○ ○ ○ Provides plenty of options to connect to other learners or teachers

EI.5 - Rate the extent to which the tool improves learning. *
Does it save time and effort for the learning process, or makes learning more attractive, or leads to higher learning achievements

0 1 2 3 4 5

Does not improve learning ○ ○ ○ ○ ○ ○ Significantly improves learning

EI.6 - Any comments that were not addressed by the evaluation questions of this section

Section U - Usability

This section is about usability of the tool. It consists of 10 rating items of the SUS method and an open text field for comments.

U.1 - I think I would like to use this tool frequently *

1 2 3 4 5

Disagree ○ ○ ○ ○ ○ Agree

U.2 - I find the tool unnecessarily complex *

1 2 3 4 5

Disagree ○ ○ ○ ○ ○ Agree

U.3 - I thought the tool was easy to use. *

1 2 3 4 5

Disagree ○ ○ ○ ○ ○ Agree

U.4 - I think I would need Tech Support to be able to use the tool *

1 2 3 4 5

Disagree ○ ○ ○ ○ ○ Agree

U.5 - I found the various functions in this tool were well integrated. *

1 2 3 4 5

Disagree ○ ○ ○ ○ ○ Agree

U.6 - I thought there was too much inconsistency in this tool. *

	1	2	3	4	5	
Disagree	○	○	○	○	○	Agree

U.7 - I would imagine that most people would learn to use this tool very quickly. *

	1	2	3	4	5	
Disagree	○	○	○	○	○	Agree

U.8 - I found tool very cumbersome to use. *

	1	2	3	4	5	
Disagree	○	○	○	○	○	Agree

U.9 - I felt very confident using the tool. *

	1	2	3	4	5	
Disagree	○	○	○	○	○	Agree

U.10 - I need to learn a lot about this tool before I could effectively use it. *

	1	2	3	4	5	
Disagree	○	○	○	○	○	Agree

U.11 - Any comments that were not addressed by the evaluation questions of this section

Section P - Performance

This section is about performance indicators of the submission. Please provide a 1-5 star rating for the submissions assigned to you (1 is the lowest score and 5 is the highest). If an indicator is not applicable to the submission, please mark it with '0' rating.

P.1 - Is the response time of the tool acceptable? *

	0	1	2	3	4	5	
Does not respond at all	○	○	○	○	○	○	Does respond very fastly

P.2 - Is the tool scalable? *
E.g. in terms of increasing numbers of users or datasets?

	0	1	2	3	4	5	
Not scalable	○	○	○	○	○	○	Highly scalable

P.3 - Any comments that were not addressed by the evaluation questions of this section

Section D - Data

This section is about indicators for data indicators. Please provide a 1-5 star rating for the submissions assigned to you (1 is the lowest score and 5 is the highest). If an indicator is not applicable to the submission, please mark it with '0' rating.

D.1 - Does the tool use multiple data sources? *

	0	1	2	3	4	5	
The tool uses only one data source	○	○	○	○	○	○	The tool uses multiple data sources

D.2 - Are the used data sources also accessable by third parties? *
Did the tool provide new datasets to the Linked Open Data (LOD) cloud?

	0	1	2	3	4	5	
No new sources have been contributed	○	○	○	○	○	○	Many new data sources have been contributed

D.3 - How reliable is the provided information of the tool? *
Does the tool provide any licenses, signs of believiablity, or provenance of the used data source?

	0	1	2	3	4	5	
Does not indicate how reliable the information is	○	○	○	○	○	○	Does indicate how reliable the information is

D.4 - Any comments that were not addressed by the evaluation questions of this section

Section L - Legal and Privacy

This section is about legal and privacy aspects of the submission. Please provide a 1-5 star rating for the submissions assigned to you (1 is the lowest score and 5 is the highest). If an indicator is not applicable to the submission, please mark it with '0' rating.

L.1 - Does the tool collect only needed personal information about the user? *

	0	1	2	3	4	5	
Does take any data it can get from the user	○	○	○	○	○	○	Does take only the needed data of the user

L.2 - Does the application provide "Terms of use" for the users, where the handling of the personal data is explained? *

	0	1	2	3	4	5	
Does not provide any terms of use	○	○	○	○	○	○	Does provide very detailed terms of use

L.3 - Does the tool offer the possibility to the user to look at all the data (personal and non-personal) which has been collected and stored? *

	0	1	2	3	4	5	
Does not have such an option	○	○	○	○	○	○	Does provide a very detailed overview of all the data stored about the user

L.4 - Can the personal data of a user be deleted? *

	0	1	2	3	4	5	
Does not have such an option	○	○	○	○	○	○	Does have such an option to delete all data stored about the user

L.5 - Any comments that were not addressed by the evaluation questions of this section

Section A - Audience

This section is about the audience of the submission. Please provide a 1-5 star rating for the submissions assigned to you (1 is the lowest score and 5 is the highest). If an indicator is not applicable to the submission, please mark it with '0' rating.

A.1 - Does the tool have a clear target group? *

	0	1	2	3	4	5

A.2 - Can the tool be applied in multiple domains? *

	0	1	2	3	4	5	
Can only be applied in one domain	○	○	○	○	○	○	Can only be applied in multiple domains

A.3 - Does the tool provide any evidence of its usage? *

Does the author provide any web statisitcs from e.g. Google analytics, or social media indicators (amount of retweets, followers, friends)?

	0	1	2	3	4	5	
Does not provide any evidence	○	○	○	○	○	○	Does provide very strong evidence

A.4 - Any comments that were not addressed by the evaluation questions of this section

Submit

Never submit passwords through Google Forms.

Appendix B – Vidi Evaluation Form

Evaluation

Overall evaluation
- 2: accept
- 0: borderline paper
- -2: reject

Reviewer's confidence
- 5: (expert)
- 4: (high)
- 3: (medium)
- 2: (low)
- 1: (none)

Additional scores

EI.1 - Rate the extent to which the application implements an innovative educational concept (e.g. innovative ways of presenting content, innovative methods for learning or teaching)
- 5: excellent
- 4: good
- 3: fair
- 2: poor
- 1: very poor

EI.2 - Rate the extent to which the application is more effective than existing applications? (e.g. leads to significant improvements in learning or teaching).
- 5: excellent
- 4: good
- 3: fair
- 2: poor
- 1: very poor

EI.3 - Rate the extent to which the application is more efficient than existing applications? (e.g. saves time or efforts for learners or teachers).
- 5: excellent
- 4: good
- 3: fair
- 2: poor
- 1: very poor

U.1 – Rate the extent to which the application is easy to use
- 5: excellent
- 4: good
- 3: fair
- 2: poor
- 1: very poor

U.2 – Rate the extent to which the application can quickly be learned?
- 5: excellent
- 4: good
- 3: fair
- 2: poor
- 1: very poor

U.3 - Rate the extent to which the application has an attractive user interface.
- 5: Very attractive
- 4: good
- 3: fair
- 2: poor
- 1: Not attractive at all

P.1 - How is the tool available to its target users?
- 5: The tool is publicly available
- 4: The tool is used for emperical studies with the target users (advanced prototype)
- 3: The tool is available to its target users (early prototype)
- 2: The tool is only available to the developers (mockup)
- 1: The tool is still in conceptual state (paper prototype)

P.2 How would you rate the overall quality of the tool according to the aims of the tracks? Open Track: To what extent integrates the tool open data for education? Focused Track Simplificator: What is the level of simplification reached by the tool?
- 5: excellent
- 4: good
- 3: fair
- 2: poor
- 1: very poor

D.1 - Does the tool consume multiple data sources?
- 5: The tool uses more than eight data sources
- 4: The tool uses more than six data sources
- 3: The tool uses more than four data sources
- 2: The tool uses more than two data sources
- 1: The tool uses only one data source

D.2 - Does the tool expose new datasets to the Linked Data cloud?
- 2: Yes
- 1: No

A.1 – Rate the extent to which the application addresses the needs, problems or challenges of its target group(s)
- 5: excellent
- 4: good
- 3: fair
- 2: poor
- 1: very poor

A.2 - Can the application be used in various domains?
- 2: Can be applied in multiple domains
- 1: Can only be applied in one domain

L.1 - Does the tool provide background / licensing information for the used data sources?
- 5: Does provide background & licence information to the used data sources
- 4: Does provide background information to used data sources
- 3: Does provide links to used data sources
- 2: Does provide some background to used data sources
- 1: Does not provide any background information of the used data sources

L.2 - Does the tool collect only needed personal information about the user?
- 3: Takes only the needed data of the user to provide its services
- 2: Takes quite a lot personal data from the users to provide its services
- 1: Takes as much data as it can get from the target users to provide its service

L.3 - Does the tool provide a statement on the terms of use?
- 2: Yes
- 1: No

Review

Review (*)
Please provide a detailed review, including justification for your scores. This review will be sent to the authors unless the PC chairs decide not to do so. This field is required unless you have an attachment.

Appendix C – Vici Evaluation Form

Evaluation

Overall evaluation
- 3: strong accept
- 2: accept
- 1: weak accept
- 0: borderline paper
- -1: weak reject
- -2: reject
- -3: strong reject

Reviewer's confidence
- 5: (expert)
- 4: (high)
- 3: (medium)
- 2: (low)
- 1: (none)

Additional scores

EI.1 - Rate the extent to which the application implements an innovative educational concept (e.g. innovative ways of presenting content, innovative methods for learning or teaching)
- 5: excellent
- 4: good
- 3: fair
- 2: poor
- 1: very poor

EI.2 - Rate the extent to which the application is effective? (e.g. leads to significant improvements in learning or teaching).
- 5: excellent
- 4: good
- 3: fair
- 2: poor
- 1: very poor

EI.3 - Rate the extent to which the application is efficient? (e.g. saves time or efforts for learners or teachers).
- 5: excellent
- 4: good
- 3: fair
- 2: poor
- 1: very poor

U.1 – Rate the extent to which the application is easy to use
- 5: excellent
- 4: good
- 3: fair
- 2: poor
- 1: very poor

U.2 – Rate the extent to which the application can quickly be learned?
- 5: excellent
- 4: good
- 3: fair
- 2: poor
- 1: very poor

U.3 - Rate the extent to which the application has an attractive user interface.
- 5: Very attractive tool
- 4: good
- 3: fair
- 2: poor
- 1: Not attractive at all

P.1 - How is the tool available to its target users?
- 5: The tool is publicly available
- 4: The tool is used for emperical studies with the target users (advanced prototype)
- 3: The tool is available to its target users (early prototype)
- 2: The tool is only available to the developers (mockup)
- 1: The tool is still in conceptual state (paper prototype)

P.2 How would you rate the quality of the tool according to the track aims? To what extent does it integrate open data: -for education (Open Track) -to improve education in developing countries (FT1) -to better understand issues of Water Resources (FT2)
- 5: excellent
- 4: good
- 3: fair
- 2: poor
- 1: very poor

D.1 - Does the tool consume multiple data sources?
- 5: The tool uses more than eight data sources
- 4: The tool uses more than six data sources
- 3: The tool uses more than four data sources
- 2: The tool uses more than two data sources
- 1: The tool uses only one data source

D.2 - Does the tool expose new datasets to the Linked Data cloud?
- 2: Yes
- 1: No

A.1 – Rate the extent to which the application addresses the needs, problems or challenges of its target group(s)
- 5: excellent
- 4: good
- 3: fair
- 2: poor
- 1: very poor

A.2 - Can the application be used in various domains?
- 2: Can be applied in multiple domain
- 1: Can only be applied in one domain

L.1 - Does the tool provide background / licensing information for the used data sources?
- 5: Does provide background & license information to used data sources
- 4: Does provide background information to used data sources
- 3: Does not provide links to used data sources
- 2: Does not provide some background information of the used data sources
- 1: Does not provide any background information of the used data sources

L.2 - Does the tool collect only needed personal information about the user?
- 3: Takes only the needed data of the user to provide its services
- 2: Takes quite a lot personal data from the users to provide its services
- 1: Takes as much data as it can get from the target users to provide its service

L.3 - Does the tool provide a statement on the terms of use?
- 2: Yes
- 1: No

Review

Review (*)
Please provide a detailed review, including justification for your scores. This review will be sent to the authors unless the PC chairs decide not to do so. This field is required unless you have an attachment.

Appendix D – Final version of the Evaluation Framework

Educational Innovation (EI)	1. Rate the extent to which the application implements an innovative educational concept (e.g. innovative ways of presenting content, innovative methods for learning or teaching). **Scale:** very poor – 5 excellence 2. Rate the extent to which the application is perceived as effective for learners (e.g. leads to significant improvements in learning or teaching). **Scale:** very poor – 5 excellence 3. Rate the extent to which the application is perceived as efficient for learners (e.g. saves time or efforts for learners or teachers). **Scale:** very poor 1 – 5 excellent
Usability (U)	1. Rate the extent to which the application is easy to use. **Scale:** very poor 1 – 5 excellent 2. Rate the extent to which the application can quickly be learned. **Scale:** very poor 1 – 5 excellent 3. Rate the extent to which the application has an attractive user interface. **Scale:** not attractive 1 – 5 very attractive
Performance (P)	1. How is the tool available to its target users? **Scale:** 1: The tool is still in conceptual state (paper prototype) 2: The tool is only available to the developers (mockup) 3: The tool is available to its target users (early prototype) 4: The tool is used for empirical studies with the target users (advanced prototype) 5: The tool is publicly available 2. How would you rate the overall quality of the tool according to the aims of the tracks? Open Track: To what extent does the tool integrate open data to improve education? Focused Track 1: Supporting Developing Countries: To what extent does the tool integrate open data to improve education in developing countries? Focused Track 2: Water Resources & Ecology: To what extent does the tool assist in increasing knowledge and a better understanding of issues on Water Resources & Ecology OR developing countries? **Scale:** very poor 1 – 5 excellent
Data (D)	1. Does the tool consume multiple data sources? **Scale:** 1: The tool uses only one data source 2: The tool uses more than two data sources 3: The tool uses more than four data sources 4: The tool uses more than six data sources 5: The tool uses more than eight data sources 2. Does the tool expose new datasets to the Linked Data cloud

	Scale: very poor 1 – 5 excellent
Legal (L)	1. Does the tool provide background / licensing information for the used data sources? **Scale:** 1: Does not provide any background information of the used data sources 2: Does not provide some background information of the used data sou rces 3: Does not provide links to used data sources 4: Does provide background information to used data sources 5: Does provide background & license information to used data sources **2.** Does the tool collect only needed personal information about the user? **Scale:** 1: Takes as much data as it can get from the target users to provide its service 2: Takes quite a lot personal data from the users to provide its services 3: Takes only the needed data of the user to provide its services 3. Does the tool provide a statement on the terms of use? **Scale:** 1. Yes, 2. No
Audience (A)	1. Rate the extent to which the application addresses the needs, problems or challenges of its target group(s). **Scale:** very poor – 5 excellence 2. Can the application be used in various domains? (E.g. is the tool generalisable or does it only fulfil a task for a specific target group? and can it be applied and used also by other stakeholders with minor modifications?) **Scale:** 1. Can be applied in multiple domain, 2. Can only be applied in one domain