An Evaluation Framework for Data Competitions in TEL

Hendrik Drachsler[1], Slavi Stoyanov[1], Mathieu d'Aquin[2], Eelco Herder[3], Marieke Guy[4], Stefan Dietze[3]

[1]Open Universiteit Nederland, Welten Institute, The Netherlands
hendrik.drachsler@ou.nl, slavi.stoyanov@ou.nl,
[2]The Open University, KMI, United Kingdom
Mathieu.Daquin@open.ac.uk
[3]L3S, Research Center, LUH Hannover, Germany
herder@l3s.de, dietze@l3s.de
[4]Open Knowledge Foundation, OKF, United Kingdom
marieke.guy@okfn.org

Abstract. This paper presents a study describing the development of an Evaluation Framework (EF) for data competitions in TEL. The study applies the Group Concept Method (GCM) to empirically depict criteria and their indicators for evaluating software applications in TEL. A statistical analysis including multidimensional scaling and hierarchical clustering on the GCM data identified the following six evaluation criteria: 1.Educational Innovation, 2.Usability, 3.Data, 4.Performance, 5.Privacy, and 6.Audience. Each of them was operationalized through a set of indicators. The resulting Evaluation Framework (EF) incorporating these criteria was applied to the first data competition of the LinkedUp project. The EF was consequently improved using the results from reviewers' interviews, which were analysed qualitatively and quantitatively. The outcome of these efforts is a comprehensive EF that can be used for TEL data competitions and for the evaluation of TEL tools in general.

Keywords. Evaluation Framework, Assessment of TEL tools, Data competition, Group Concept Mapping

1 Introduction

With the raise of data science, there is also a new wave of publications on learning analytics, personalisation and adaptation techniques. But these data-driven research approaches in education are hardly comparable with each other. Most of the reported experiments in the TEL world are not comprehensible and do not provide the underlying research data, neither they describe their learner model, educational reasoning or personalisation techniques in sufficient detail to repeat an experiment. This is a mayor challenge for TEL science, as it is the nature of science to gain general knowledge that can be reproduced under controlled conditions. It is of crucial importance to overcome this issue to make TEL research results more reliable for the community as well as for policy makers and funding bodies. We should strive towards a comprehen-

sive knowledge base about the effects of TEL tools on learning and teaching. For this reason, the dataTEL Theme Team funded by the STELLAR Network of Excellence identified the need for a comprehensive Evaluation Framework as one of the Grand Challenges for technology enhanced learning [1]. In TEL, evaluation has specific characteristics as it needs to take into account technical and educational measures. The technical measures guarantee that the software is working properly and the educational measures indicate the impact of the technology on learning scenarios.

The paper presents an empirical study for the development of an Evaluation Framework for the LinkedUp project that organises three consecutive data competitions in TEL - Veni, Vidi and Vici [2]. The goal of the LinkedUp Veni competition was to gather innovative and robust tools that analyse and/or integrate large scale, open Web data for educational purposes. Veni was open for any end-user application that analyses and makes use of Linked Data or Open Web Data for online learning. The EF has been applied to compare data-driven tools in TEL and rank them according to their achievements in a standardised manner.

The paper will report about this process by first reviewing related evaluation approaches in TEL. Then we introduce the Group Concept Mapping method that was used by the TEL community to identify evaluation criteria and indicators that are suitable for TEL related data competitions. Third, we provide some demographic data about the participants of the study and describe the procedure in further details. Fourth, we present the results of the Group Concept Mapping and summarize its main findings. Fifth, we discuss the first version of the LinkedUp EF as it has been applied to the LinkedUp competition– Veni[1]. Finally, we report how the EF was evaluated and further improved for the Vidi competition.

2 Related work

2.1 Data competitions

Data competitions are of increasing importance as a mean to gain knowledge about data science in various domains. Data competitions enable the data owners to review diverse approaches towards a single dataset and are therefore a strong instrument for innovation purposes.

There are various competitions related to the TEL domain such as the *Elsevier Grand Challenge*[2], where the goal is to improve communication of scientific information. The *Semantic Web Service Challenge*[3] is another example of similar initiative that aims at evaluating Semantic Web Services Mediation, Choreography and Discovery technologies. Collaborative Annotation of a Large Biomedical Corpus (*CALBC*[4]) is a European Support Action addressing the automatic generation of a large, community-wide shared text corpus annotated with biomedical entities. Evalua-

[1] http://www.linkedup-challenge.eu
[2] http://www.elseviergrandchallenge.com/
[3] http://sws-challenge.org
[4] http://www.calbc.eu/

tion will be performed against the harmonized contributions that have been gathered from the participants' contributions to the same challenge.

Although previous examples have demonstrated the suitability and usefulness of organised challenges and competitions to drive innovation, there are still some issues that limit the reusability and impact. Most of these initiatives are *technology-centric rather than outcome-centric*, are based on *artificial, limited test data* and *often lack of real-world scenarios.* One serious drawback in the previous competitions was lack of information about evaluation frameworks used, particularly how the criteria and indicators were identified and with what kind of methods.

In LinkedUp we aimed to overcome those issues by creating TEL related data competitions following a *realistic use-case scenarios,* involving a *large-scale testbed of Web datasets*, and a *transparent evaluation framework* to ensure a high-level of innovation and reusability of project results in education.

2.2 Evaluation approaches in TEL

A systematic literature review that we conducted identified a reasonable number of studies on evaluation of TEL tools and e-learning courses but very little information was returned on evaluation frameworks, criteria and indicators for assessing educational software applications in competitions. Most of the research on TEL evaluation has been focused on usability. Some studies combine usability with specific performance measures for learning of end users (learners). For instance, [3] modified the Nielsen's protocol for the evaluation of an e-learning program. [4] presented a comprehensive usability study that brings together end-user assessments and expert inspections, thus providing a detailed students', teachers' and experts' feedback. [5] developed an integrated model with six dimensions: learners, instructors, courses, technology, design, and environment to evaluate the satisfaction from using an eLearning tool. In another study [6] proposes an usability framework that integrates web usability and instructional design parameters and proposes motivation to learn as a new type of usability dimension in designing and evaluating e-learning applications.

Next to those usability studies there have been different approaches for the evaluation of personalisation and adaption of TEL tools that are also relevant for our EF as they require data to provide their adaptation services. [7] suggested an approach to decompose the adaptation process into two layers that are evaluated separately. This is needed because a 'monolithic' evaluation cannot provide sufficient information at a level of granularity that can be valuable for the system designer to decide which part of the system needs improvement. The layered evaluation approach is still a summative evaluation with two phases rather a formative evaluation process. Simultaneously, two other modular evaluation frameworks have been proposed. The process-based framework presented by [8] consisted of four evaluation layers, the second framework has been presented by [9] and is more detailed in terms of different components involved in the adaptation process. It also addressed the question about methods and tools appropriate for the evaluation of different adaptation modules to yield input for the development process. A merged version of the two frameworks was

finally proposed and has been explored by several studies that evaluated adaptive systems [10].

Another evaluation approach has been suggested by the RecSysTEL community [11]. They propose an evaluation method for Recommender Systems in TEL by using reference datasets to make the findings of the data studies more comparable to each other. They proposed a set of references datasets that could be used to gain comparable evaluation results [13]. Several studies followed this approach since it was mentioned [12] and started to contribute evidences for comparable evaluation results.

Summarising the insights from the related work section we can concluded, that there are various approaches to evaluate TEL tools and that usability is a very common criterion. It also appeared that among usability there is a lack of transparency of used evaluation criteria and how they are operationalized in suitable indicators.

One objective of the LinkedUp project is to address this lack of transparency and develop a framework that can be applied for various domains within the TEL field. The evaluation framework we are aiming for has at least three important differences to the studies discussed. First, it is *not focused on the end user or system designer*. It rather needs to support a *jury of judges to come up with an accurate, comprehensive and transparent assessment* about a submitted tool. Second, it needs to check *if a TEL tool is technically sound* but also *innovative from an educational perspective*. Third, the evaluation cannot run over a longer time period, in fact the jury needs to be able to *make a decision about the submitted tools in a timeframe of 1-2 hours*.

3 Method

We applied the Group Concept Mapping (GCM) method to address the lack of a transparent EF with community-driven quality indicators within TEL [15]. The aim of the GCM is to develop an evaluation framework that is driven by high profile experts from the whole TEL community, rather than a proposal of a single research group. GCM is a structured, mixed approach applying both quantitative and qualitative measures to objectively identify an expert group's common understanding about a particular issue, in our case the evaluation indicators for open educational data. The method involved the participants in three activities: 1. Idea generation, 2. Sorting of ideas into groups, and 3. Rating the ideas on some values (Priority and Applicability of the indicators). The participants work individually, but it is the advanced statistical techniques of multidimensional scaling and hierarchical cluster analysis that quantitatively aggregate individual input of the participants to reveal shared patterns in the data. One of the distinguishing characteristics of GCM is the visualisation, which is a substantial part of the analysis. Visualisation allows for grasping at once the emerging data structures and their interrelationship to support decision-making.

3.1 Participants

In total, 122 external experts have been identified for the GCM study. The candidates were selected according to two criteria: (a) holding a PhD degree and (b) a pub-

lication list that demonstrates experience in developing and evaluating data- driven applications in TEL. 74 experts responded positively to the invitation to participate in the study. They registered to the GCM tool for online data collection by creating a username and password. All participants gave their research informed consent. Of all participants assigned to the study, 57 contributed to the idea generation phase, 26 completed the sorting and 26 finished the rating. Figure 1 shows an overview of the participation of the experts, who agreed to participate. A meta-analytical research including 69 GCM study suggests that 20-30 participants is the optimal number for sorting the ideas [14].

	assigned	started	finished	checked
Project	73	73	*na*	*na*
Questions	2			
Brainstorming	57	57		
Sorting	44	42	26	26
Priority	44	31	26	26
Applicability	44	29	26	26

Fig. 1: Response rate of external experts to the LinkedUp GCM study.

3.2 Procedure

As mentioned earlier the procedure consisted of three phases, namely: 1. Idea generation, 2. Sorting of ideas into groups, and 3. Rating on two values (priority and applicability). Afterwards the researchers analyse the data and interpret the results. The results from the GCM were then used for determining evaluation indicators, criteria and potential methods to measure the indicators for the EF.

All participants were fully informed about the purpose, the procedure, and the time needed for completing the activities. The participants were provided with a link to the brainstorming page of a web-based tool for data collection and analysis. They were asked to generate ideas completing the following trigger statement:

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

During the idea generation phase, the 57 experts contributed a total of 212 original ideas. After cleaning these statements from analogical and vague ideas, and splitting the statements that contained more than one idea we were left with a list of 108 indicators. The final list of 108 indicators was randomised and sent back to the participants. In the next step they were asked to first sort the ideas into groups based on their similarity, giving a representative name to the group, and, second, to rate them on two values – *priority* and *applicability* for the use in the EF.

4 Results

4.1 Point map of the 108 quality indicators

Figure 2 shows the first outcome of the multidimensional scaling analysis – a point map. The closer the statements to each other, the closer in meaning they are, which

also means that more participants cluster them together. Multidimensional scaling assigns each statement a bridging value, which is between 0 and 1. The lower bridging value means that a statement has been grouped together with statements around it; e.g. statements 6, 19, 77, 89, 100, 105 on the right side of figure 2. A higher bridging value means that the statement has been grouped together with some statements further apart from either side (e.g. statement 21 or 86 in the centre of the point map). Some groups of ideas can be detected by eye inspection, but to make the process more efficient a hierarchical cluster analysis was applied.

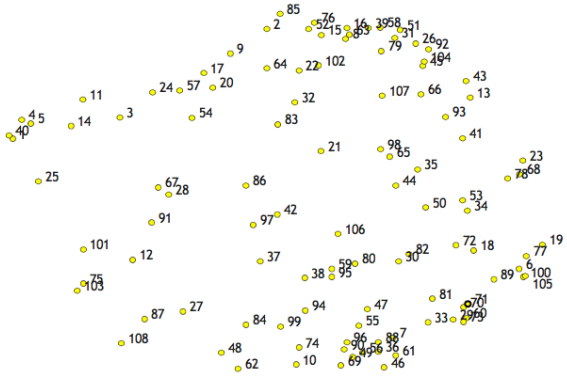

Fig. 2: Point map of the 108 quality indicators contributed by the TEL community.

4.2 From the point map to a cluster map

Several solutions suggested by the hierarchical cluster analysis have been trialed (see Figure 3). For the final decision, we adapted the practical heuristic of '15-to-4' [15] as the average number of clusters per participant was 10. We started from a 15-cluster solution with the idea to arrive at a 4-cluster solution.

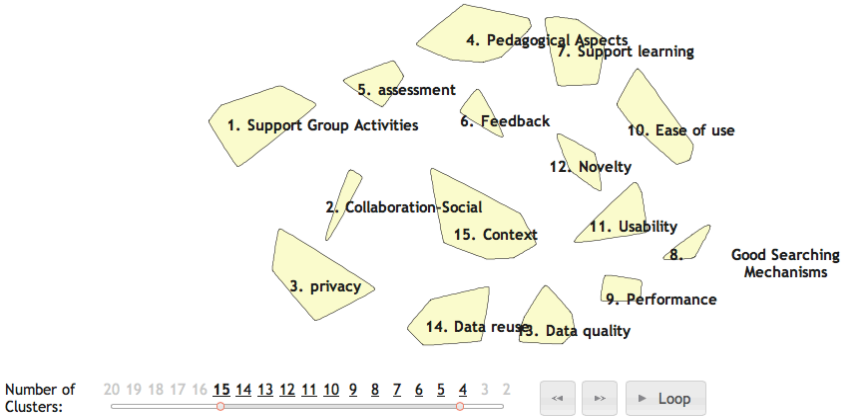

Fig. 3: A replay scaling 15-to-4 cluster solutions, currently shown 15 clusters.

At each step, we checked whether the merging of clusters made sense for the purpose of the LinkedUp project. The six-cluster solution seemed best representing the data and serving the purpose of the study (see Figure 4).

From Figure 4 it can be seen that there is a very stable *Data* (south on the map) and *Education* (north) cluster in the point map that do not share any statements. By contrast, *Performance*, which also includes some Human Computer Interaction statements, is naturally positioned between the *Data* and *Education* clusters. The *Privacy* (west) cluster always remained apart from the other clusters, but it is also a very stable and therefore important entity for the evaluation criteria. Surprisingly, the *Support Group Activities* cluster never merged with the *Educational* clusters, as the external experts see these statements semantically different to the educational aspects of the evaluation criteria. Moreover, it developed as an additional application domain, next to the educational one, which promotes its own indicators for Open Web Data applications.

The next step of processing the clustering results is constructing meaningful labels for the clusters, using the three available methods. The first one is to check what the GCM system suggests. The second way is to look at the bridging values of the statements composing a cluster. The statements with lower bridging value represent better a cluster. The third method is to read through all statements in a cluster and define what is the story behind it. To define the clusters (criteria) we combined the three methods. We finally, chose the following labels for the 6-cluster solution: 1. Support Group Activities, 2. Privacy, 3. Educational Innovation, 4. Usability, 5. Performance, and 6. Data (see Figure 4).

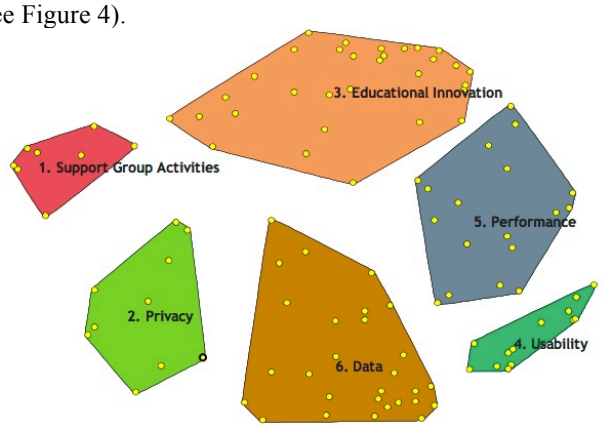

Fig. 4: Cluster labels.

4.3 Six Cluster Rating maps

As described above, the experts applied a rating to the evaluation criteria and their indicators according to two aspects of the LinkedUp EF: *Priority* and *Applicability*. *Priority* refers to the importance of a particular cluster for the evaluation of TEL tools. *Applicability* indicates the perceived ease to apply the indicator and criterion in the review process. Five layers indicate a high rating within the GCM tool, one layer of a cluster visualizes a low rating.

As Figure 5 shows, the clusters 'Usability' received the highest rating on priority followed by 'Educational Innovation' and 'Data' with three layers each. 'Support Group Activities' and 'Privacy' received the lowest score (one and two layers respectively).

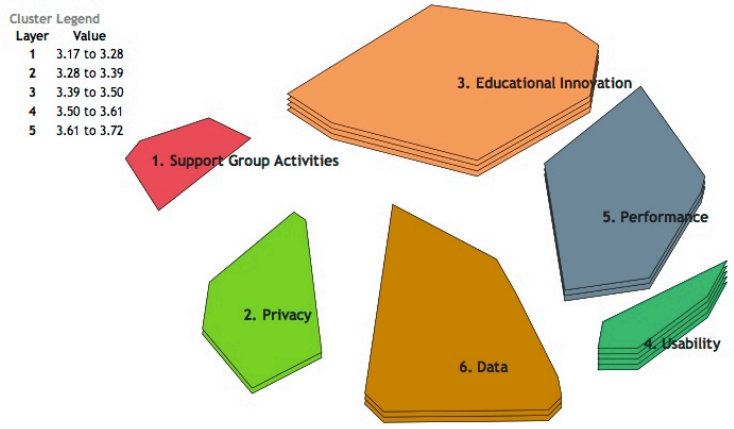

Fig. 5: Rating map on *priority* of the evaluation criteria/indicators for the EF.

A different picture appears for the *Applicability* aspect of the evaluation criteria (see Figure 6). According to the participants, the indicators that are easiest to implement are within the cluster '*Support Group Activities*' and *'Usability'* (four layers). *'Performance'* and *'Privacy'* are both rated with three layers as reasonably applicable indicators of the EF.

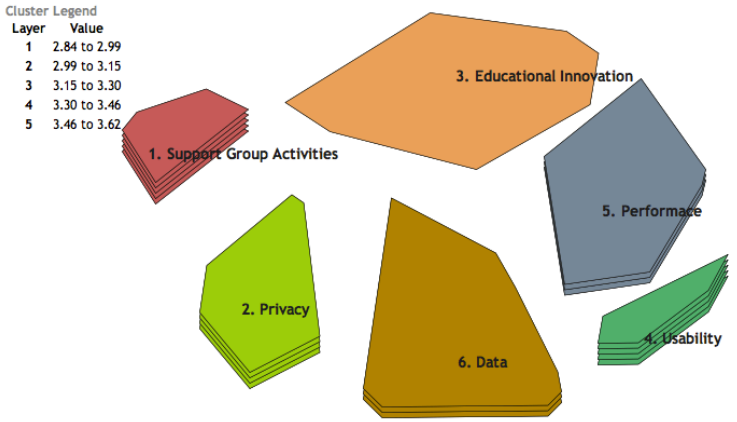

Fig. 6: Rating map on *applicability* of the evaluation criteria/indicators for the EF.

The *'Educational Innovation'* cluster, which has received the highest score on priority, got very low rating here, meaning this is expected to be the most difficult to assess by the judges of a data challenge.

The ladder graph in Figure 7, called pattern match, compares the clusters on the *Priority* and *Applicability* ratings. The lines show how pairs of clusters are related according to their rating values. A Pearson product-moment correlation coefficient (r = -0.16) indicates a weak negative relationship between the two values: priority and applicability. The cluster 'Support group activities' has the biggest margin between the two values. It scores the lowest on priority and the highest on applicability. In contrast, 'Educational innovation' scores relatively high on priority but the lowest on applicability. 'Usability' scores high on both values. There is a relatively small difference between priority and applicability in the clusters 'Data' and 'Performance'.

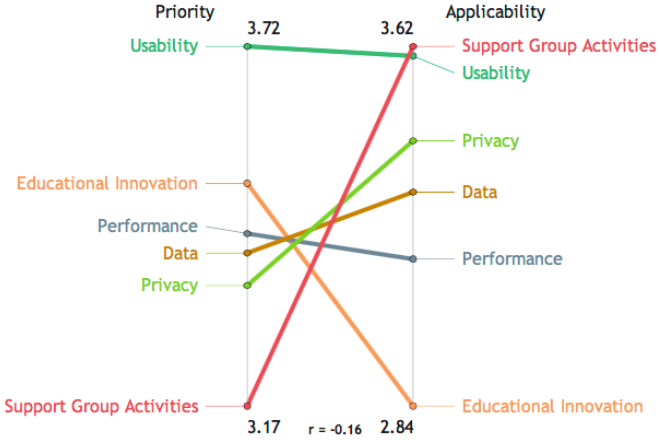

Fig 7: Pattern match *Priority* vs. *Applicability* of the evaluation criteria for the EF.

5 GCM outcomes - a first outline of the LinkedUp EF

A panel of LinkedUp project experts discussed the results of the study in the context of the competitions that are going to be organized and decided on the final set of clusters (criteria), which included: Educational Innovation, Privacy, Usability, Performance, and Data. The sixth cluster, 'Support Group Activities' was disregarded for the first version of the EF, because it mainly contains a list of very specific features that are related to the computer-supported collaborative working (CSCW) field. It could be used later on for the specific Focus Track within a data competition around a specific CSCW use case, but was not relevant to the objectives of the first Open Call of the LinkedUp challenge.

While analysing the bridging values of all statements and their clusters, the consortium also discovered an important omission which is highly relevant for the objectives of public data competitions. The LinkedUp project aims to promote applications with high impact that constitute powerful examples of how to use Linked Data to serve different stakeholders in education. This means that applications that aim for a very narrow target group are less relevant than applications targeting a broader audience. There are some highly rated statements that can be combined in a cluster 'Audience'. Representative statements in support of such a cluster are: "That it addresses a broad

community of users", or "it can be used or tailored to a variety of target groups", and "the calculation of basic metrics on technology usage (like amount of users, browsing sessions, avg. sessions per user)". The project consortium, therefore, decided to add 'Audience' as an additional criterion for the initial version of the LinkedUp EF.

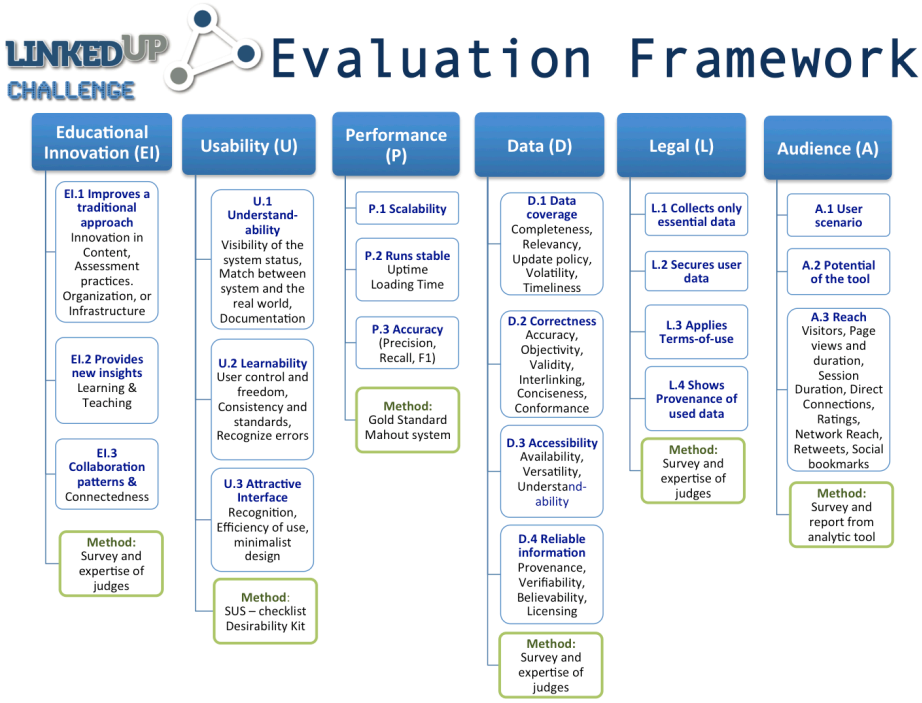

Fig. 8: A first version of the LinkedUp EF after the GCM study.

Figure 8 represents the first version of the EF that was applied to the LinkedUp Veni competition in 2013. It shows the six criteria clusters identified by the GCM study, a set of indicators for each criteria and possible methods to measure those indicators. The indicators have been contributed by the experts participating in the brainstorming of the GCM study. The methods have been identified by another literature review for measuring the evaluation indicators that have been suggested.

6 Implementation of the Evaluation Framework to the Veni competition

The line chart presented in Figure 9 shows an overview of 15 from the 22 participants submitted to the Veni competition and how they got valued according to the evaluation criteria. It indicates also how submissions scored on an individual criterion. The final scores coming form the evaluation framework were used as a basis for a deliberation process conducted by LinkedUp partners. The LinkedUp team was very satisfied with the effortless ranking of the submitted tools according to the scores. It ena-

bled the team to shortlist the submitted tools and identify the three winners of the competition without much more efforts. In the same way it also made the evaluation results transparent to the reviewers and the participants. We could provide the participants with detailed scores about their performance on each of the evaluation criteria and contrast those with the average scores of the Veni competition.

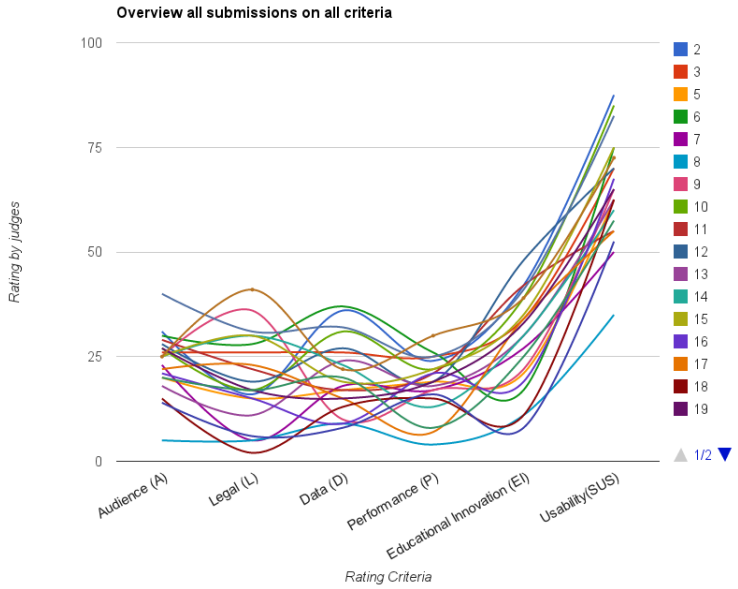

Fig. 9: Overview of the rating results given to the participants of the LinkedUp Veni competition. The high value of usability is affected by the chosen SUS method that has a range for the final score between 20 to 100 points.

6.1 Improvement of the Evaluation Framework

We conducted five individual semi-structured interviews with jury members to identify possible issues that might appear when using the EF. The jury members have been randomly selected. The interviews followed an elaborated script to provide an unified approach among the interviewers. The script contained a suggested sequence of activities, the main questions with possible probes (follow up) and how to ask questions. A letter of inform consent was also part of the document. The interview and the deliberation discussion were transcribed in verbatim before the analysis. The data analysis included both qualitative and quantitative methods.

In general the reviewers were very positive about the evaluation framework. They expressed some concerns regarding the 'not applicable' option in some of the evaluation criteria and suggested either removing this option or providing a clearer guideline how to interpret it and how to proceed with it. In addition, the reviewers felt that some of the indicators, such as 'Assessment' in the cluster 'Educational Innovation' were not relevant for the Veni competition. The reviewers suggested to have a more elaborated instruction to a criterion and possibly providing an example.

To complement the qualitative analysis of the interview data and identify issues not easily detectable in the interviews regarding the EF, we applied a free for use web text mining service called 'Text is Beautiful'[5]. The tool presents the text analytics results in the form of concept web (similar to concept map – see Figure 10). The input was the combined text of the five interviews we conducted.

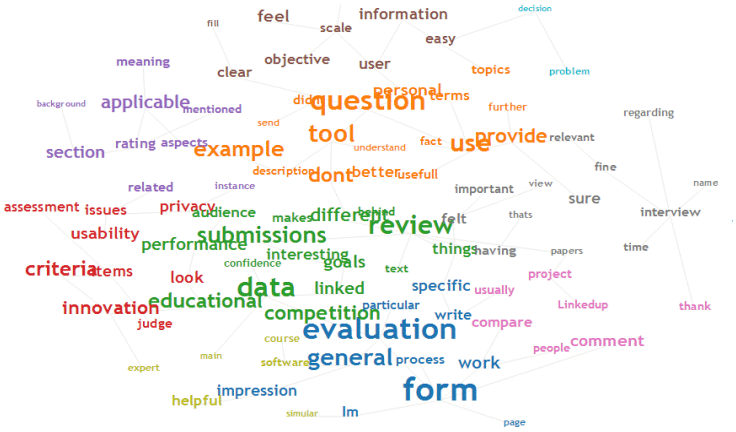

Fig. 10: Concept web.

In the concept web, the concepts are colored-clustered in a two-dimensional space and the concepts that are strongly related are positioned in the map closely to each other. For example, the concepts in orange (e.g. 'question', 'use', 'example', 'provide', 'description', 'better', 'useful', 'description', 'understand') confirm the finding that it would be useful if an example is provided in the description of a criterion to make the instruction better formulated and understandable. The 'violet' concepts ('applicable', 'section', 'rating aspects', 'meaning', 'mentioned') reflect the discussion on the meaning of the rating aspect "not applicable". It seems to be one of the issues (see the concepts in red), specifically related to some of the criteria such as educational innovation and usability. Apparently, assessment as one of the indicators of the criteria educational innovation is an issue ('assessment issue'). The conclusion that can be drawn is that the quantitative text mining analysis had confirmed the findings from the qualitative text analysis regarding the issues that need to be addressed for the next round of the competition (e.g. the issues with the non-applicable option, issues related to the indicator 'assessment', and the need for providing an example to the instruction of each item).

7 Conclusion and further work

In this article we presented the findings of a Group Concept Mapping study for empirically identifying the set of criteria and indicators for evaluating data competi-

[5] http://textisbeautiful.net/

tions in TEL. We tested the first version of the LinkedUp EF during the Veni Competition[6]. The Veni competition required '*an innovative and robust prototype or demo that used linked and/or open data for educational purposes*'. By the closing date, 22 valid submissions had been received from 12 different countries. The LinkedUp judges rated the submitted tools according to concrete criteria and indicators of the EF. The EF enabled the LinkedUp team to shortlist the submitted tools and identify the three winners of the Veni competition applying an unified and empirically validated evaluation framework. We believe that it would have been much more difficult to agree on these results without the application of the EF.

On the basis of this initial version of the EF, we are further investigating suitable evaluation criteria and their specific indicators. We are especially interested in additional metrics, and weighting to evaluate the defined criteria and automated or, semi-automated evaluation tools that can easily be applied by the LinkedUp judges saving their time.

The outcomes of the EF study could be beneficial for organisers of data competitions, not only in TEL. The EF is flexible, specific criteria and their indicators can be selected and combined to address the needs of future data competitions. We will further evaluate and improve the EF during the LinkedUp Vidi and Vici competitions to provide the most comprehensive EF by the end of the LinkedUp runtime. We are aiming to provide a kind of toolbox that can guide data scientists in setting up data competitions and hackathons in their specific domains. The toolbox will provide valuable information, lessons-learned from the LinkedUp competitions, links to suitable public data sources, hints on legal issue and how to solve those, marketing strategies, and also the EF with specific guides and templates' how to organise an accurate and transparent evaluation process for the judges and participants of data competitions. A first exploitation of this work will be done in the context of the EATEL SIG dataTEL, and the Dutch SURF SIG on Learning Analytics by organising specific data competitions as satellite projects of LinkedUp.

Acknowledgements. This work was partly funded by the LinkedUp (GA No:317620) and DURAARK (GA No:600908) projects under the FP7 programme of the European Commission.

References.

[1] Drachsler, H., Verbert, K., Manouselis, N., Vuorikari, R., Wolpers, M., and Lindstaedt, S. (2012). "Preface [Special issue on dataTEL - Data Supported Research in Technology-Enhanced Learning]," *Int. J. Technol. Enhanc. Learn.*, vol. 4, no. 1/2.

[2] d'Aquin, M., Dietze, S., Herder, E., Drachsler, H. (2014). Using linked data in learning analytics. eLearning Papers. Nr. 37. ISSN: 1887-1542. www.openeducationeuropa.eu/en/elearning_papers

[3] Benson, L., Elliot. D., Grant, M., Holschuh, D., Kim., B., Kim., H., Lauber., E., Loh., S., & Reeves, T., (2002). Usability and Instructional Design Heuristics for

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

E-Learning Evaluation, Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications, pp. 1615-1621, Denver, Colorado, USA.

[4] Granic & M. Cukusic: Usability Testing and Expert Inspections Complemented by Educational Evaluation: A Case Study of an e-Learning Platform, Educational Technology & Society. 14 (2011), 2; 107-123.

[5] Sun, P.-C., Tsai, R.J., Finger, G., Chen, Y.-Y., Yeh, D. (2008). What drives a successful e-Learning? An empirical investigation of the critical factors influencing learner satisfaction. Computers and Education. Volume 50, Issue 4, Pages 1183 – 1202.

[6] Zaharias, P. (2009). Usability in the Context of e-Learning: A Framework Augmenting 'Traditional' Usability Constructs with Instructional Design and Motivation to Learn, International Journal of Technology and Human Interaction, 5(4), 38-61.

[7] Brusilovsky, P. and Eklund, J. (1998). A study of user-model based link annotation in educational hypermedia, Journal of Universal Computer Science 4(4), 429–448.

[8] Weibelzahl S. (2001) Evaluation of adaptive systems. In: Bauer M, Gmytrasiewicz, PJ, Vassileva, J (eds), User Modeling: Proceedings of the Eighth International Conference, UM2001, pp. 292294. Berlin: Springer.

[9] Paramythis A, Totter A, Stephanidis C (2001) A modular approach to the evaluation of adaptive user interfaces. In: Weibelzahl S, Chin DN, Weber G (eds) Empirical Evaluation of Adaptive Systems. In: Proceedings of workshop at the Eighth, International Conference on User Modeling, UM2001, pp. 924, Freiburg.

[10] Brusilovsky P, Karagiannidis C, Sampson D (2004) Layered Evaluation of AdaptiveLearning Systems. International Journal of Continuing Engineering Education and Lifelong Learning, Special issue on Adaptivity in Web and Mobile Learning Services, 14(4/5):402-421. Inderscience Pub.B.

[11] Manouselis, N., Drachsler, H., Verbert, K., and Duval, E. (2012). Recommender Systems for Learning. Springer Berlin Heidelberg, pp. 1–61.

[12] Verbert, K., Manouselis, N., Drachsler, H., Duval, E. (2012). Dataset-driven Research to Support Learning and Knowledge Analytics. International Journal of Educational Technology & Society.

[13] Verbert, K., Drachsler, H., Manouselis, N., Wolpers, M., Vuorikari, R., & Duval, E. (2011). Dataset-driven Research for Improving Recommender Systems for Learning. 1st International Conference Learning Analytics & Knowledge. February, 27 - March, 1, 2011, Banff, Alberta, Canada.

[14] Rosas, S.R., Kane, M. (2012). Quality and rigor of the concept mapping methodology: a pooled study analysis. Evaluation and Program Planning, 35, 36–245.

[15] Kane M., Trochim W. Concept mapping for planning and evaluation. Thousand Oaks: Sage Publishing; 2007.