D2.1 Evaluation Criteria and Methods

LinkedUp: Linking Web Data for Education Project – Open Challenge in Web-scale Data Integration

http://linkedup-project.eu/

Coordination and Support Action (CSA)

Grant Agreement No: 317620

D2.1 Evaluation Criteria and Methods

Deliverable Coordinator: Drachsler, Hendrik

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

Other Authors: Wolfgang Greller, (OUNL)
Slavi Stoyanov, (OUNL)

Document Identifier:	LinkedUp/2013/D2.2.1/v1.	Date due:	30.04.2013
Class Deliverable:	LinkedUp 317620	Submission date:	26.04.2013
Project start date:	November 1, 2012	Version:	v1.1
Project duration:	2 years	State:	
		Distribution:	Public

LinkedUp Consortium

This document is a 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 Panton Street 37, CB2 1HL Cambridge United Kingdom Contact person: Sander van der Waal E-mail address: sander.vanderwaal@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	**EXACT Learning Solutions SPA** Viale Gramsci 19 50121 Firenze Italy Contact person: Elisabetta Parodi E-mail address: e.parodi@exactls.com

Work package participants

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

- LUH, Besnik Fetahu, Constantin Rex, Eelco Herder
- OU, Enrico Daga, Alessandro Adamou
- ELS, Elisabetta Parodi, Marco Mosca
- OUNL, Slavi Stoyanov, Wolfgang Greller

Change Log

Version	Date	Amended by	Changes
0.1	01.03.2013	Hendrik Drachsler	Initial structure
0.2	10.03.2013	Slavi Stoyanov Hendrik Drachsler	Enrichment
0.3	12.03.2013	Wolfgang Greller	Enrichment
0.4	15.03.2013	Alessandro Adamou	Enrichment
0.5	17.03.2013	Elisabetta Parodi	Enrichment
0.6	19.03.2013	Besnik Fetahu	Enrichment
0.7	23.03.2013	Constantin Rex	Enrichment
0.8	25.03.2013	Hendrik Drachsler	Enrichment
0.9	29.03.2013	Hendrik Drachsler	Minor corrections
1.0	23.04.2013	Wolfgang Greller	Reviewers feedback incorporated
1.1	25.04.2013	Hendrik Drachsler	Minor corrections

Executive Summary

The main purposes of the current deliverable D2.1 is to provide the foundations of an Evaluation Framework that can be applied to compare Open Web Data applications and rank them according to their achievements. D2.1 contains the information gained from *Task 2.1 - Evaluation criteria and method review* and *Task 2.2 - Validation of the evaluation criteria and methods* of WP2 (DoW. p. 8). According to those tasks, we conducted an expert survey with the Group Concept Mapping method to identify relevant indicators and criteria for the Evaluation Framework. In a second step, we conducted a focused literature review to extend the outcomes of the expert survey with latest indicators reported in the literature. We finally, present the initial concept of the Evaluation Framework and its criteria and indicators.
This deliverable provides the theoretical foundations for the Evaluation Framework that is further developed into a scoring sheet for the judges of LinkedUp challenge in deliverable D2.2.1. The Evaluation Framework will be further developed and amended according to the experiences collected in the three LinkedUp data competitions during the LinkedUp challenge.

Table of Contents

1. Introduction

The overall objective of WP2 is to develop an Evaluation Framework (EF) that can be applied to compare Open Web Data applications and rank them according to their achievements. We will consider all relevant criteria and methods for the evaluation of Open Web Data applications. It is intended as a comprehensive overview of all possible and relevant evaluation approaches for Open Web Data applications. This should support various domains and can be customised to specific needs. It is intended as an instrument to standardise the evaluation of Open Web Data applications.

For the LinkedUp challenge specifically, the EF will be tailored to evaluate software applications that are submitted in the educational domain. The challenge consists of three data competitions that call for innovative tools in the educational domain (see FP7 LinkedUp deliverable D.1.1).

As the EF is one of the main outcomes of the FP7 LinkedUp project, this deliverable provides the foundations for deliverable D2.2.1 that presents the first version of the EF for the LinkedUp data competition. The deliverable contains the information gained from *Task 2.1 - Evaluation criteria and method review* and *Task 2.2 - Validation of the evaluation criteria and methods* of WP2 (DoW. p. 8). The EF will be further developed and amended according to the experiences collected in the three LinkedUp data competitions.

1.1 Problem zone and evaluation issues

In order to scope the task and objectives at hand, we need to take a closer look at today's challenges in the field of open data and open data applications.

Data is everywhere! It is manifested through an abundance of computer systems fulfilling various tasks in an organisational or individual context. It also emerges from collective behaviours of users and automated agents that use these systems. In a rather recent trend to openness, many datasets are being exposed to third-party use as linked open data (LOD). This, however, is not without challenges. Especially in education, harmonisation and comparability are key to successful application of LOD in a quality-assured way as demanded by general educational principles of equal opportunities and standard qualifications, but also guided by elementary research principles, including transparency and replicability.

In describing the challenges faced by LOD for education, we can distinguish three main hierarchical areas for evaluation:

(1) raw datasets
(2) applications built upon such datasets
(3) added value the applications bring to education

Regarding point (1), issues related to the raw datasets of open linked data concern the data quality and general usefulness, as well as the exchangeability of data. Quality criteria concern, among other issues, the "cleanliness" of datasets, i.e. the state of meaningful authentic data items free of erroneous and testing data. Quality criteria on the datasets also concern the legal clarity on the possibilities to use them in an anonymised uncompromising way, free of legal risks. While international laws protect the intellectual property rights (IPR) of database structures, the ownership

of content is less clear and may pose certain risks with respect to data protection and privacy laws, and even competition laws.

With respect to the exchangeability of datasets, one of the biggest challenges for the reuse of educational datasets is the inconsistency in descriptions and meta-information available. In order to make LOD sets searchable, retrievable, re-usable, and open to peer-evaluation, proper description schemas are required as an essential precondition. Such curation needs to include, among other things, information about the size of the dataset, quality criteria (as above, including licence information), contextual information, and general metadata information (including time and validity info). The development of a standard edu-data schema, therefore, is paramount to making LOD more widely used, both in practical contexts as well as in research.

The elaboration of such curation criteria may lead to establishing a gold standard of datasets and support proper comparison and transparency that use the same operationalised metrics.

Regarding point (2), similar questions can be raised at the level of data applications. Here too, the contextual information needs to be more explicit for re-use. What is required to make data tools comparable and open to impartial evaluation is, among other things, an assessment framework for how the underlying data is handled by the application. Notably, it is important for replication and evaluation to know exactly which input data is used to produce the output, as well as the algorithmic approach. Note that data absent (i.e. being ignored by the system) can in many cases be as important for evaluation of an application as the data that is being used. Evaluators need to know what is in and what is out in order to being able to assess the metrics in use (e.g. weightings) and the accuracy of the output. In this, LinkedUp can learn from approaches in adaptive hypermedia systems that categorise the application architecture into content (learning resources), the domain (domain ontology), and the user (user model) (Manouselis, et al. 2012).

Regarding point (3), on the highest level, the value these LOD data application bring to education faces a number of challenges. Just because an application can be built and used, does not guarantee its usefulness for real-life operations. Especially, the added value they bring to end users in terms of efficiency or enrichment of their learning are areas that need to be focused on by designers using open datasets. This is an area where indicators are largely lacking, but, return-of-investment metrics may play an important part in establishing the ratio of added benefit to input efforts.

For competitions like the LinkedUp challenge, but also for ordinary use, it is important to have explicit statements and feedback, both contrastive and comparative, in order to, e.g. distinguish between two applications built on the same dataset. The method for developing an evaluation framework that spans across the three levels is described in the next parts of this document.

1.2 Potential methods for developing the EF

There are different ways for deriving criteria for evaluating open educational data: literature review and expert consultation. However, both methods also have their drawbacks. The issues that need to be addressed in a literature review are the adaptation of available criteria for the purpose of educational data and the operationalisation of these criteria with a set of indicators, to keep reviewers from assigning different meanings to each criterion. The most used expert consultation methods are individual and group interviews, affinity diagram and the Delphi method. A major issue with online expert consultations is getting an agreement on the list of evaluation criteria and how much emphasis should be put on each learning criterion. The experts represent different professional domains and as individuals they could have rather different thinking styles. Additionally, during live meetings there

is always the phenomenon of 'groupthink' (the negative effect of the group on the opinions of the individual members). The analysis of individual and focus groups interview in most of the cases is based on pre-determined classification schemas, which can be either non-exhaustive or impose bias. In affinity diagram sessions, participants typically would suggest different solutions, both in terms of groups of criteria and the content of these groups, which makes it difficult for researchers to come up with a unified vision on how best to structure the information. The Delphi method requires several iterative rounds before claiming consensus in the group. The consensus is more or less forced and the subjective approach is always there.

We, therefore, chose the method of Group Concept Mapping (Trochim, 1989; Trochim & Kane & Trochim, 2007). This research methodology, while building on the strengths of interviews, affinity diagrams and the Delphi method, mitigates some of their weaknesses.

2. Group Concept Mapping

Group Concept Mapping (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 a few simple activities that most of the people are familiar with: idea generation, sorting of ideas into groups and rating the ideas on some values (e.g. priority and applicability). 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. Group Concept Mapping produces three main types of visualisations: conceptual maps, pattern matching and go-zones.

In contrast to the Delphi method, in GCM, there is only one round of structuring the data as the participants work independently and anonymously for each other. Unlike interviews, GCM does not rely on pre-determined classification schemas. The method does not need intercoder discussions to come up with an agreement. When sorting the statements into groups, the participants, in fact, 'code' the concepts themselves. Then multivariate statistical analysis aggregates the individual coding schemas across all participants. Consensus is not forced but 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.

Group Concept Mapping in the LinkedUp project supports a bottom-up approach of building an evaluation framework. The top-down approach typically defines a set of criteria. The problem, however, is that definitions are brief using quite general terms, which may lead to people getting different understanding of the criteria. In addition, the chance of providing a non-comprehensive set of criteria is high.

The GCM approach generates first a number of indicators, which are then structured and weighted in more general categories (evaluation criteria). Each evaluation criterion is operationalised through the set of concrete indicators in a particular cluster. A numeric scale (e.g 1 to 5) can be attached to each

indicator and judges are asked to evaluate the extent to which the educational data application covers each of the indicators. A compound score for each criterion can then be calculated.

2.1 Participants

In total, 122 external experts have been identified for the GCM study. The candidates were selected according to two criteria: 1. holding a PhD degree, 2. a publication list that demonstrates experiences in developing and evaluating data driven applications (for education). 74 experts responded positively to the invitation to participate in the study. They registered to the Concept System Global system (Concept System Global, 2012) for online data collection by creating a username and password. All participants gave their research informed consent. Of all participants assigned to the study, 58 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 122 experts that were invited to the GCM study.

	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

2.2 Procedure

The procedure consisted of four phases, namely: 1. idea generation, 2. sorting of ideas into groups, 3. rating on two values (priority and applicability), and 4. analysis of the data and interpretation of the results. These results from the GCM were then used for determining evaluation methods and metrics.

All LinkedUp project members were invited through an email to participate in the evaluation framework GCM study. In addition, 122 external experts were also invited to participate in the study. 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 (Concept System Global, 2012). The participants were able to visit the web site for as many times as they needed, using the credentials they had created. 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 ..."

The ideas should be short phrases or statements expressing one thought. The whole brainstorming instruction is given in Appendix A.

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. The detailed instructions for sorting

and rating are also given in Appendix A. The participants got two weeks for completing both sorting and rating. A reminder was sent every week. As in the brainstorming, the participant could save their work and return later to continue. The analysis included multidimensional scaling and hierarchical cluster analysis for the structured data and mean, standard deviation, and correlation for the rating data.

3. Results

3.1 Point map

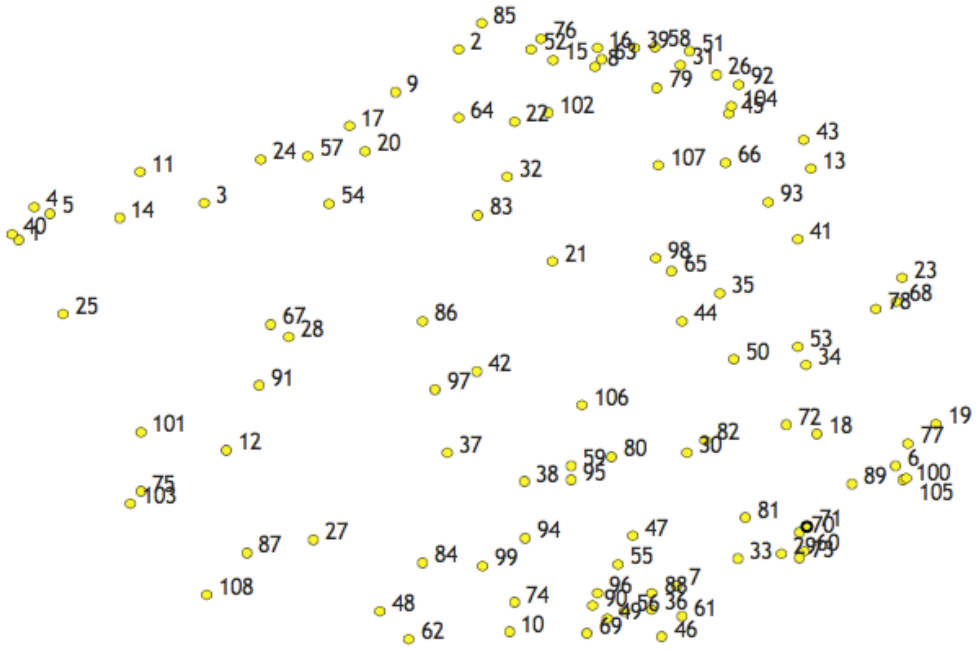

Fig. 2: Point map

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 1. 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 is applied.

3.2 From the point map to most suitable 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 '20-to-5' to '15-to-4' (Kane & Trochim,

2007) because 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.

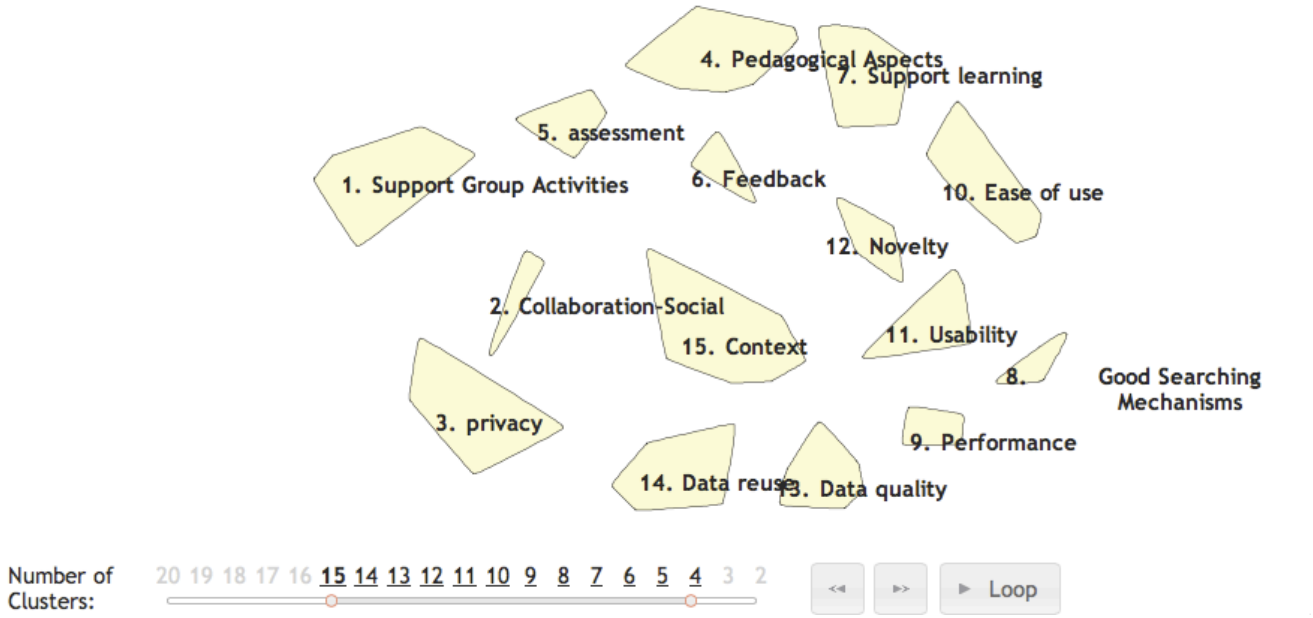

Fig. 3: A replay scaling 15-to-4 cluster solutions, currently shown 15 clusters

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).

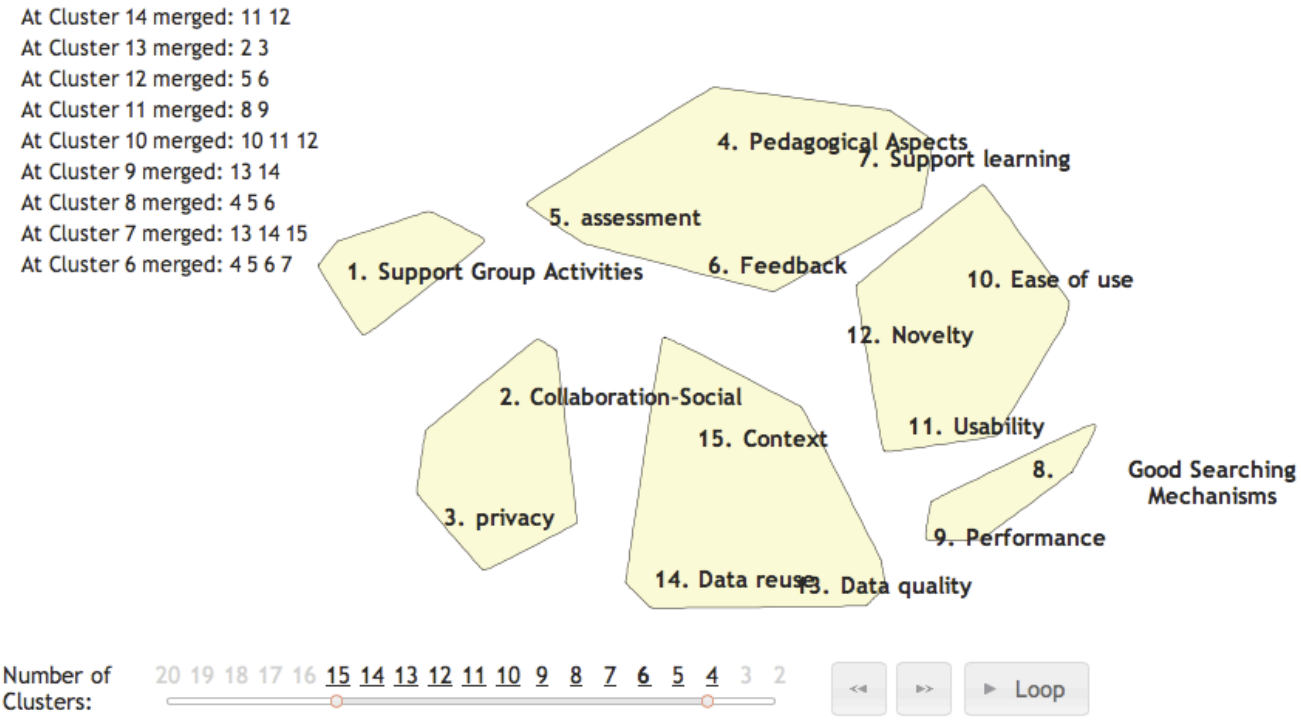

Fig. 4: The same replay map showing 6 clusters.

From Figure 4 it can be seen that there is a very stable *Data* (south) 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 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 system puts on the top of suggestions the label of a cluster named by a participant whose centroid is the closest to the centroid of the cluster formed by the aggregation of the data from all the participants. 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, what the cluster wants to tell us. To define the clusters (categories/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, 6. Data (see Figure 5).

The most coherent clusters (with the lowest bridging value), meaning that they had the highest agreement rate from experts for the statements contained within, are ‘Usability’ (0.17), followed by ‘Data’ (0.21), ‘Educational Innovation’ (0.22), and ‘Performance’ (0.39). The clusters with the highest bridging value are ‘Support Group Activities’ (0.50), and ‘Privacy’ (0.81). Appendix B presents all statistics regarding the sorting data. In the following paragraphs we will shortly characterise each of the cluster and their triggering statements.

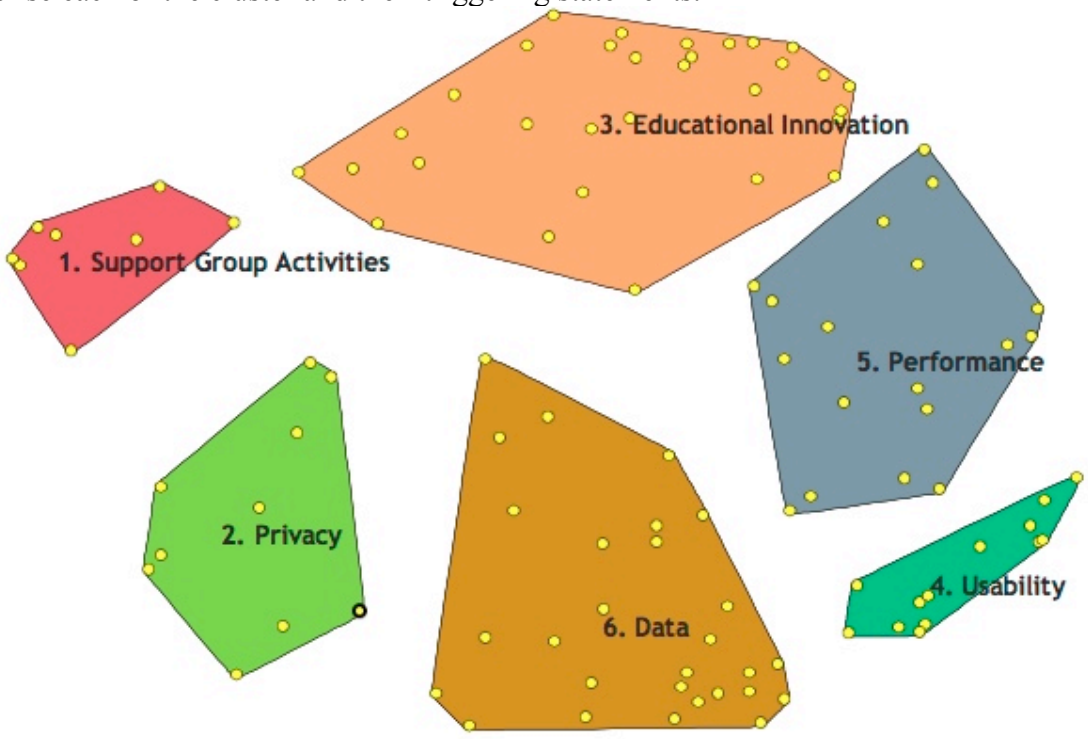

Fig. 5: Cluster labels

'Support Group Activities' is about the vision that future Open Web data tools should provide facilities to connect educational stakeholders (learners or teachers) while they are learning on the web. Representative statements for this cluster are: "enables (students and teachers) to connect with people based on shared interests and information"; "to better support group activities (both long term course level and short term project level)"; and "to build on online collaborative system to support awareness of the users (e.g., social- , task- , and or group awareness)".

'Privacy' contains legal and privacy aspects that are affected by Open Web data applications. Representative statements for this cluster are partly aiming on privacy protection of the individuum and partly on legal and copyright aspects of data providers. For individual privacy protection statements like this are added to the cluster: 'to enable users to define their own privacy regulations', or 'the application enables users to define their policies of data sharing to assess the trade-off between privacy and system performance' whereas the legal aspects are covered in statements like: "to automatically detect some kind of risks of personal data presence and warn the data provider before the release".

'Educational Interventions' describes a large list of features Open Web data applications should support to improve the learning and teaching. Representative statements for this cluster are "increase ability of learners to learn better and faster", "the application enables students to gain new insights into a study topic", or "the application detects different points of view or contrasting facts for a specific topic".

'Usability' is clearly about usability aspects of the applications. The external experts had a very clear view on usability as it is a well-known concept in computer science. Here are statements clustered like: "the application supports an easy navigation", or "general principles of usability should be considered", or "it can easily be used by users who are not technology savvy".

'Performance' is a very stable and consistent cluster emergent from the indicators contributed by the external experts. It refers to technical aspects that describe the scalability of an application such as: "the application runs stable and does not crash", or "a fast response time also with huge amount of data".

'Data' is also a very stable cluster through almost all cluster maps starting from 15 to 6. It encapsulates all statements that are related to data aspects, from quality of data until repurposing of data. Representative statements are for instance: "that the data is provided in an interoperable format", or "that is can handle unstructured data", or "that the used datasets are sufficiently semantically described.

3.3 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 evaluation framework: *Priority* and *Applicability*. The former refers to the importance of a particular cluster; whereas the latter identifies the ease for reviewers to assess LinkedUp applications (cf. Appendix A).

As Figure 6 shows, the clusters 'Usability' received the highest rating on priority followed by and 'Educational Innovation' and 'Data' with three layers each. 'Support Group Activities' and 'Privacy' received the lowest score (one or two layers). Appendix C presents the ratings values of all statements and clusters on priority.

