D2.2.1 Evaluation Framework

Citation for published version (APA):

Drachsler, H., Greller, W., & Stoyanov, S. (2013). *D2.2.1 Evaluation Framework.*

Document status and date:
Published: 01/11/2013

Document Version:
Peer reviewed version

Document license:
CC BY-NC-SA

Open Universiteit
www.ou.nl

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

D2.2.1   Evaluation Framework

Deliverable Coordinator: Drachsler, Hendrik

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

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

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 the elaboration of this document, even if they might not have directly contributed to the writing of this document or its parts:
- LUH
- OU
- EXT
- ELSV

Change Log

Version	Date	Amended by	Changes
0.1	15.03.2013	Hendrik Drachsler	Initial structure
0.2	25.03.2013	Hendrik Drachsler	Enrichment
0.3	26.03.2013	Wolfgang Greller	Minor corrections
0.3	01.04.2013	Hendrik Drachsler	Minor corrections
0.4	23.04.2013	Slavi Stoyanov	Reviewers feedback incorporated
0.5	25.04.2013	Hendrik Drachsler	Minor corrections

Executive Summary

The main purpose of the current deliverable D2.2.1 is to hold the current version of the Evaluation Framework and to operationalise it for the LinkedUp challenge judges into a concrete evaluation instrument. This deliverable is not intended as a very elaborated report rather than a summary of the current version of the Evaluation Framework based on the extensive studies in deliverable D2.1 – *Evaluation Methods and Metrics*. D2.2.1will be reconsidered in the final report of WP2 to demonstrate the development of the Evaluation Framework during the life cycle of the LinkedUp project. For this purpose it is supportive to have the first version of the Evaluation Framework as a tangible outcome and an own entity as conducted in this deliverable.

Table of Contents

1. Introduction

The deliverable D2.1 – Evaluation Criteria and Metrics of the Task 2.1 of WP2 describes the foundations for the first version of the LinkedUp Evaluation Framework (EF). This first version of the EF is based on Group Concept Mapping approach that identified consensus about criteria and methods for the evaluation of Open Web Data applications in Education and a state-of-the-art analysis of available evaluation metrics.

The main purpose of the current deliverable D2.2.1 is to freeze the current version of the EF and to operationalise it for the LinkedUp challenge judges into a concrete evaluation instrument. The EF is one of the main outcomes of the FP7 LinkedUp project and will be further developed and improved throughout the duration of the project, especially after each round of a data competition in the LinkedUp Challenge (see D1.2). Therefore, this deliverable is not intended to be an elaborated report but rather a summary of the current version of the EF that will be reconsidered in the final report of WP2 to demonstrate the development of the EF during the life cycle of the LinkedUp project. For this purpose it is important to have the first version of the EF as a tangible outcome and an own entity as conducted in this deliverable.

In *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. These upcoming content validation steps of the EF after each data competition cycle is the main responsibility for WP2 in the LinkedUp project. Each of the content validation reviews will be reported in an amended version of D2.2.1 (D2.3.1, D2.3.2) and reported respectively in the final version of the EF in deliverable D2.2.2.

2. Overview of the first version of the Evaluation Framework

The information shown in this section is based on the extensive analysis reported in D2.1. Before reporting on main findings we briefly describe the procedure for deriving the set of evaluation criteria and indicators to enable readers who are unaware of D2.1 to get an idea about the background of the EF. The evaluation framework is based on an empirical study applying the Group Concept Mapping approach. 57 experts generated 212 evaluation indicators. 26 experts then sorted the ideas generated into groups of similarity in meaning and rated the indicators on two values: priority and applicability. The statistics of multidimensional scaling and hierarchical cluster analysis identified 6 criteria. The Linkedup Consortium discussed the results of the study. The final, shared vision of the Consortium is presented in Figure 1. The six criteria are: 1. Educational Innovation, 2. Usability, 3. Performance, 4. Data, 5. Legal aspects, and 6. Audience. In the following we will shortly introduce each evaluation criterion and it aligned evaluation method.
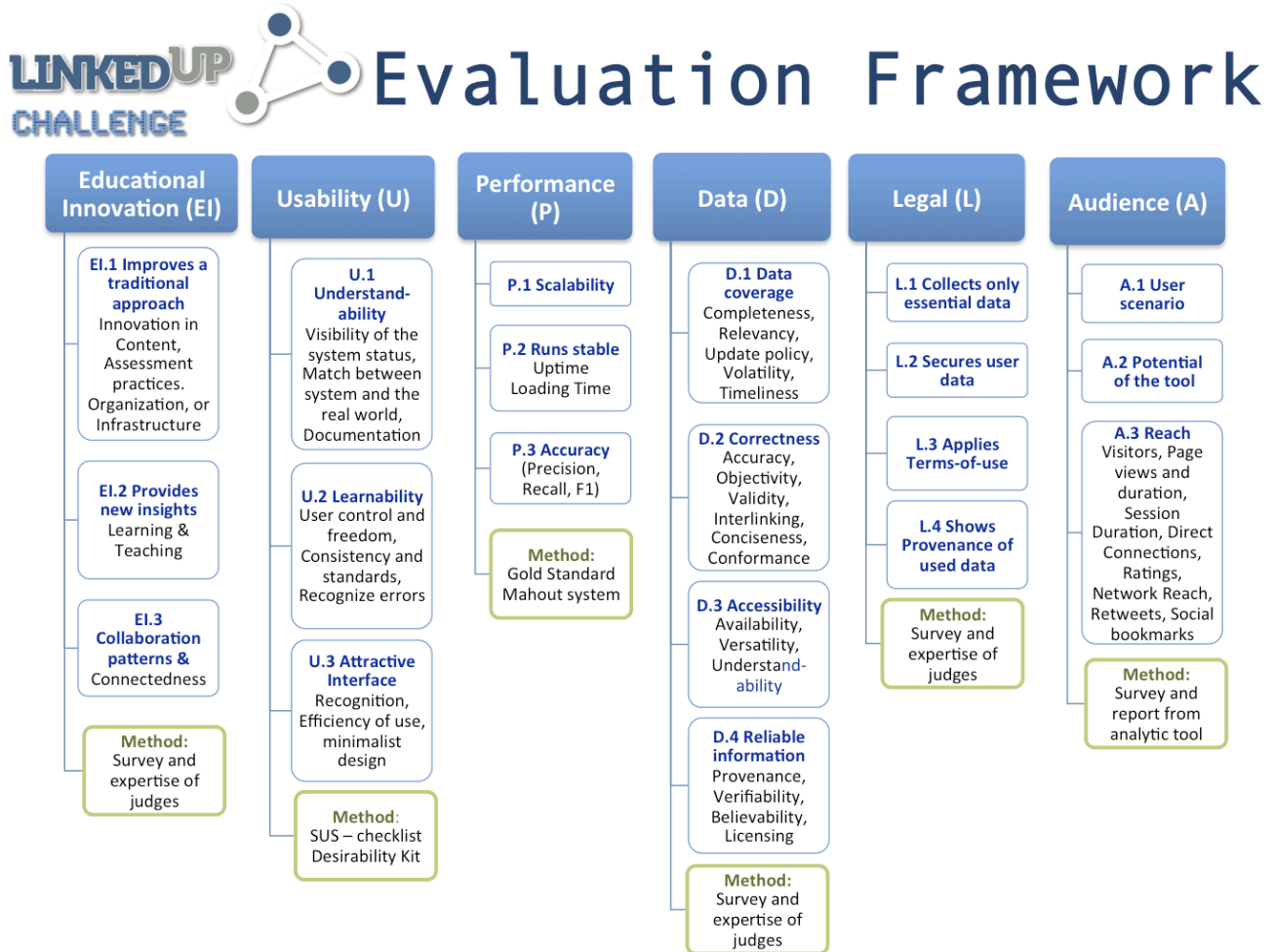
Figure 1: Comprehensive version of the LinkedUp Evaluation Framework based on the deliverable D2.1 of the LinkedUp project.

Educational Innovation

'Educational Innovation' is based on a list of indicators that innovative educational tools should support based on an expert survey and a recent report of Institute for Prospective Technological Studies (IPTS), an EC research institute. In the first version of the EF, judges of the data challenge will be able to check whether data applications address the set of indicators composing this criterion In addition, we will ask the judges to provide a short statement for how innovative is the application and a rating on a scale from 1-5 stars.

Usability

'Usability' is a very well known and elaborate concept with clear evaluation indicators. There is also a wide range of standardised tools that can be applied to measure this criterion. The two most applicable methods for the evaluation of the LinkedUp challenge are the Open Source Desirability Kit (Storm, 2012), and the SUS method (Tullis and Stetson, 2004). SUS is often used in carrying out comparisons of usability between software, it is quickly done, and yields a single benchmarking score on a scale of 0–100 that provides an objective indication of the usability of a tool. This makes it highly relevant for the LinkedUp challenge especially in the later stages of the data competition where more advanced systems are expected to be entered into the competitions.

The Desirability Kit is relatively easy to apply by the judges. However, it provides more a general description of the user satisfaction with the tool rather than a comparison score. Nevertheless, this

approach might be very helpful to evaluate participants especially in the open track, where no clear task is provided.

Performance

The 'Performance' criterion provides very clear measuring indicators derived from both the GCM study and the literature review. For the first version of the EF we will ask participants to report suitable indicators for their systems and asks the judges to review those descriptions.
For a future version of the EF we are considering to develop a gold standard benchmark based on the data pool of the LinkedUp project. Such a benchmark could be based on standard algorithms as they are part of the Mahout system[1] and provided clear metrics to the participants where improvements by their tools are expected.

Data

The indicators of the 'Data' criterion can be partly evaluated by providing statistics about the used data sources, a description of some of the indicators by the participants, and an evaluation of the same indicators by the judges of the LinkedUp challenge. For the first version of the EF, we are considering to provide tick boxes if certain information is provided, open review fields, in addition, and a rating scale from 1-5 stars for the judges.

Legal and Privacy

Privacy was a very consistent cluster in the GCM study and was also rated as important by the LinkedUp consortium. We can inform the scoring sheet with some specific questionnaire items reported in related literature. The judges will then need to rate these question items on ordinal and nominal scales.

Audience

Audience is a very relevant aspect of the LinkedUp competition, as we are aiming to promote Linked Data applications that have potential to change current educational practices. An application can score very high on technical aspects of data and user interface but if it does not address educational problems learners, teachers and educational managers have, then it is useless. Users characteristics simply can not be ignored when developing a linked educational data application. In addition, when looking at the impact of applications, we tend to appreciate more those that address issues of larger user groups. The analysis can easily be done by reports gained from common analytics tools (e.g. Google analytics) or indicators from social media applications. Thus, for the evaluation of this criterion we expect the participant to provide indicators from analytic tools and describe their future development and marketing plans. Finally, we will rely on the expertise of the judges to estimate the potential of the tool and the user scenario for the near future of 1-3 years.

3. Evaluation of the LinkedUp scoring sheet

Based on the first version of the LinkedUp EF, we created a scoring sheet in Google forms[2] that allows an effective and efficient comparison of the judges' ranked reviews of the participants performance in the LinkedUp challenge. The scoring sheet will support the members of the review board in evaluating the participating projects and award the cash prices. Another advantage is that we can integrate survey-based system such as SUS for Usability and directly compute the SUS score for

---

[1] http://mahout.apache.org/
[2] https://docs.google.com/forms/d/1-LhIS_wmoQNKFHZvod1JFMCqm-o9EevaL7ABD6-aSl4/edit