A Framework for Informal Learning Analytics - Evidence from the Literacy Domain

Aya Rizk
Luleå University of Technology
aya.rizk@ltu.se

Adrian Rodriguez
Luleå University of Technology
adrian.rodriguez@ltu.se

Abstract

Multidisciplinary approaches to learning analytics (LA) have the potential to provide important insights into student learning beyond interactions within learning management systems (LMS). In this paper we demonstrate the benefits of such an approach by proposing a framework that adds the contextual elements of task design, tools and technologies and datasets to established LA processes. Our framework was developed as a design science research (DSR) artifact, working with teachers of English at two Swedish secondary schools. The results highlight the importance of valid task design for generating relevant, useful insights and provide a basis for simplifying and automating in-situ LA that can be used by teachers in their everyday work. The study also provided important insights for the field of online research and comprehension (ORC) both in relation to methodology and how students engage with a task that requires locating and synthesizing information on the open Internet in a second language.

1. Introduction

This study demonstrates the benefits of a multidisciplinary approach integrating learning analytics (LA) and literacy research. Literacy plays a vital role in development, democracy and equality which is why international organizations such as the OECD invest in regular, large scale international reading assessments. The increasing prevalence of the Internet in everyday life has, however, led to a redefinition of exactly what it means to be literate in the 21st century [1], something that is clearly reflected in public policy response at both international and national levels. In recent years, the OECD's international reading assessments have, for instance, been developed and expanded to include assessments of online informational reading (e.g. PISA 2018, EPIRLS). Countries from Australia to the United States and Norway have all changed school curricula to emphasize the skills needed to successfully locate relevant information in an online setting. In Sweden, in 2017, the National Education Agency (NAE), in keeping with these international tendencies, reformulated the Swedish curriculum to incorporate the concept of *digital competence.* Key skills include searching for information and evaluating sources, as well as being able to efficiently use digital tools and understand digital systems and services. The NAE emphasizes that students should be able to keep their bearings in a complex reality, where there is a vast flow of information and where the rate of change is rapid [2]. The ubiquity of the Internet and laptops or tablets in Swedish secondary schools, means that Swedish students rely on the Internet as an important source of information in a variety of subjects. For independent research tasks and assignments, given that the majority of information on the Internet is in English and that less than one percent of Internet content is in Swedish [3], Swedes who are able to proficiently locate and use online information in English will have access to significantly more information in most areas than their less capable peers. In order for teachers to be able to support students in developing these key skills, research is needed to understand not only how students engage with information in an online setting, but also the role played by language.

Researchers in the field of new literacies highlight the necessity for active, broad scale collaboration that efficiently uses approaches from a range of research fields to address the task of understanding a constantly changing, diverse and widely distributed phenomenon that has undeniably revolutionized the way we communicate and provided new contexts for the traditional literacy practices of reading and writing [1], [4]. Understanding exactly how students engage with information on the Internet to perform a particular task is crucial for teachers to support learning processes [5]. Much literacy research relies, however, on established reading research methods such as think aloud protocols (e.g. [5], [6]) and surveys (e.g. [7], [8]) or examines specific elements of online informational reading using limited or artificial versions of the Internet or a limited selection of Internet texts (e.g. [9], [10]). It is also not uncommon for teachers to suggest that they lack the

URI: https://hdl.handle.net/10125/70795
978-0-9981331-4-0

support necessary to make full use of the affordances of ICT in the classroom (see [11], [12]). With carefully considered design, learning analytics (LA) has the potential to make an important contribution, both to understanding how students actually engage with information online on the open Internet and in providing teachers with ongoing access to the insights generated. In this study, we highlight the benefits of an approach that combines knowledge and methods from different research fields. Knowledge of the specific setting and skills under investigation are integrated into a LA framework to produce results that allow teachers to provide customized feedback and support.

To date, learning analytics is largely dominated by exploratory studies that attempt to discover relevant patterns in data generated through the use of learning management systems (LMS). These types of studies put more emphasis on the technology, systems development and sophistication of the analyses but may fail to take into account other elements specific to the field of investigation (i.e. in the case of our study integrating pedagogy and educational theories). Such techno-centricity limits the potential impact that LA could have on practice, theory and policy [13]. Moreover, by restricting the focus of analyses to LMS most of the "informal learning" that takes place outside the LMS, for instance as students engage with information online, remains invisible to teachers, LA professionals and researchers. A traditional view of informal learning as something that takes place outside of the classroom is blurred by the introduction of classroom tasks that require information gathering carried out on the open Internet. Results of the process may be visible to teachers but not important insights such as how many websites were visited, what language(s) were used, if students have translated information or used tools to support their understanding. Informal learning spaces such as the Internet play a crucial role in students' learning particularly in relation to searching for and synthesizing information for their learning activities. Accordingly, there is a need for an integrated informal learning analytics (ILA) framework that supports teachers in helping them capture and understand (part of) their students' informal learning. The framework needs to utilize accessible tools and technologies to help the teachers drive/conduct the analyses iteratively in a way that would allow them to provide customized, scalable feedback to students and modify tasks to better target intended learning outcomes [14].

1.1 Purpose and research questions

The purpose of this study was to use an approach that incorporated knowledge and methodology from new literacies research and work together with teachers to design a LA framework that could provide teachers with useful insights into their students' behavior when searching for information on the Internet. The questions we were seeking to answer were: 1. How does incorporating relevant pedagogical theory and teachers' experience into LA task design influence the insights generated? 2. How can teachers use subsequent student profiling to provide students with customized feedback to support the development of digital competence and inform future task design?

2. Background

2.1. Learning analytics in an educational setting

Ever since the formalization of learning analytics as a research discipline and domain of practice, the field has witnessed a rapid growth in publications and solutions, respectively [13], [15]. Chatti et al. [16] proposed a reference model for learning analytics systems that explores what data is collected by the LMS, how it is analyzed and to what end. This model is rooted in the analytics domain with emphasis on data mining technology and processes. More recently, Nguyen et al. [17] proposed a methodology for designing and developing learning analytics information systems (LAIS) that integrated knowledge from LA and learning design practices. While this methodology tends to the multidisciplinary nature of LA, the teacher is still regarded as a user that does not influence what and to what end the analysis is conducted. This lack of teacher input to the LA process instance and subsequent analysis may contribute to the minimal impact LA studies currently have on teaching and learning theory and practice [13].

A closer examination of how teachers utilize LA artifacts (e.g. features and systems) reveals a range of use cases; from student modeling to predict knowledge levels [18], monitoring and visual analysis of student progress [19], and improving assessment and feedback - also through statistical analysis and visualization [14]. In the context of institutional education, the majority of LA studies are focused on data collected through LMS or in formal learning settings. Yet, the literature suggests that the majority of learning, including developing digital competence, takes place in informal learning settings [20]. More recent studies address this issue by exploring informal learning environments, whether through the use of specific sensors and tracking tools [21] or through tracking online activity [22]. However, common to these studies and other LA studies in general is the lack of flexibility of the analyses and the use of inaccessible technologies.

In this paper, we acknowledge the role of the teacher as a significant contributor to the analysis of and experimentation with data, as well as the importance of informal learning environments for carrying out formal learning tasks (i.e. the use of the Internet to complete a formal learning task). In the following section, we review the literature on online reading and comprehension on which we build our demonstration of the framework.

2.2 Online research and comprehension

The term online research and comprehension (ORC) was coined by Leu et al. [23] to signify that reading for information in hyperlinked Internet texts involves more than simply *reading*. Leu et al. [24] initially presented a framework of five key skills required to successfully make use of the information available on the Internet: identifying important questions, locating information, critically evaluating the usefulness of information found, synthesizing information to answer questions and communicating information to others (p. 1572). With this framework as a foundation for their research, Coiro & Dobler [5] identified a number of unique elements of the online informational reading process including the activation of prior knowledge of search engines and informational website structure, high incidences of forward inferencing, skimming across multiple texts and reading that was characterized by recursive, self-monitored cycles of plan-predict-monitor-evaluate (p. 235). These findings echo Henry [25] who identified searching and dealing with search engine results as a "gatekeeper skill" (p. 616) in online information gathering.

Coiro [26] demonstrated further that online reading comprehension skills and strategies are distinct from offline reading comprehension ability (as measured by standardized reading test scores) and that higher levels of online reading comprehension can even compensate for lower levels of prior topic knowledge when adolescents read for information on the Internet (p. 374). The increased cognitive load caused by the necessity to manage distraction is also a common theme in research on hypertext reading. Features such as hyperlinks, advertising banners and multimodal features such as animations mean that electronic texts involve higher levels of distraction than print-based media [27], [28]. Cho et al. [6] point to the role of the online informational reader in actively constructing and realizing a coherent, goal-relevant reading path and conclude that proficient online informational readers need to be "constructively responsive" and engage in continuous, strategic self-regulation throughout the entire process.

The findings cited above form a basis for understanding ORC and the field has continued to expand in response to a call for an "open source" approach that combines methodologies from a range of research fields [1]. In this paper we seek to answer that call by demonstrating how LA can incorporate and enfold ORC theory and teachers' knowledge and insights to understand how students engage with information in an online search task and how teachers can use the insights generated.

3. Research method

The study follows a design science research (DSR) methodology in order to design, develop and evaluate the ILA framework [29]. The framework is regarded as an artifact that aims to solve the problem of techno-centric LA artifacts, and propose a more integrated approach to understanding and supporting informal learning – defined as learning that occurs in contexts beyond the confines of the classroom or the LMS. The artifact was developed in 3 phases: a) exploration of the research problem and potential solutions, b) design, implementation & demonstration of the framework through a clustering-based ILA instance, and c) evaluation. The instantiation described below is situated in a research project involving 6 classes with a total of 92 students at 2 Swedish high schools. The project team included a teacher education researcher and a data analytics researcher (the co-authors), working in close collaboration with the 5 high school teachers responsible for the 6 classes. The teachers volunteered to participate in the study after a short presentation of the proposed project by one of the co-authors at a monthly network meeting for teachers in the municipality where the project was carried out. The teachers were involved in adapting the task design to the level of the students and the demands of the curriculum. The research problem was initially identified through desktop research, secondary data from earlier projects and a number of project meetings and workshops.

The design of the artifact draws on two distinct bodies of knowledge: LA processes [16], [30] and ORC theories [5], [23]. While the former discipline provides prescriptive knowledge in the form of a best practice process, the latter delivers the descriptive knowledge necessary to elevate the relevance of the artifact to the intended user and use (i.e. teacher and teaching) [31]. They also inform the two main components of the framework.

The proposed artifact design is provided in section 4 and the detailed implementation of the artifact instance is described in section 5. Evaluation of the framework and ILA insights took place through a 90-minute workshop with the teachers involved in the project. Their input helped us identify the value of the solution, potential limitations and their envisioned use. This input will be further used to formalize some of the design principles for the coming improved instantiations and further abstraction of the framework, both material- and action-oriented principles [32].

4. Proposed ILA Framework

The proposed framework includes two main components: contextual elements and the core LA process (see Figure 1). The contextual elements drive the LA process and include the learning task and its design, the digital tools used during the learning task, and the datasets generated through such use. The LA process, on the other hand, includes four phases that are often conducted iteratively, based on LA process models (e.g. [16]) and more general data science process models [30].

3.1. Driving contextual elements

The importance of integrating learning theory into LA frameworks is stressed repeatedly in the literature [13], [16]. The role of theory varies, and in this framework, theory is expected to be embedded in the artifact through the learning task (which is informed by an understanding of online research and comprehension). The ***learning task***, and its designed instance, should drive the LA process by informing the initiation & planning. It is also informed by the LA results, where the teacher's intervention should be motivated by the insights generated from data. Overall, the integration of the learning task justifies *why* the specific type and form of analysis is necessary.

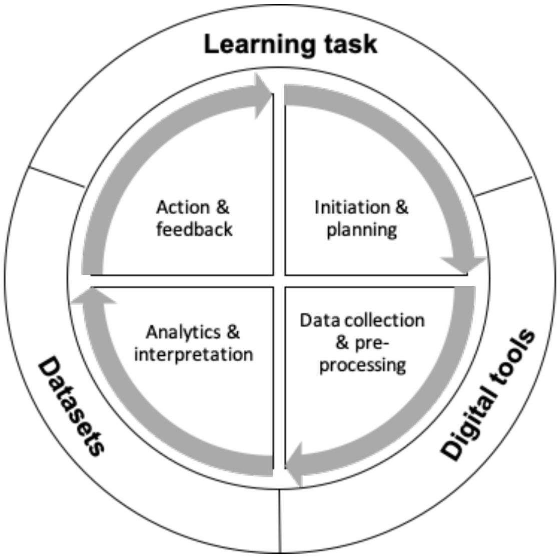

Figure 1. Framework for Informal Learning Analytics (ILA)

The ***digital tools*** are particularly relevant for ILA since the learning task takes place beyond the boundaries of the LMS. By digital tools, we refer to digital products and services that allow teachers to collect and/or access data about their students' learning process during their enactment of the learning task that is not otherwise collected through the LMS. There are various tools that are user-friendly, publicly available and/or already used by the student. We particularly highlight these types of tools that are accessible to teachers and multidisciplinary teams and where no extra development is needed. Deciding on the digital tool to use also depends on the learning task, where minimal intrusion and/or disruption to the task is recommended. The choice should also take into consideration any ethical and/or privacy implications for the student, both during and beyond the learning task. Thus, the integration of the "digital tools" element should address the question of *how* the data will be ethically collected/accessed.

The ***datasets*** element addresses in further detail the question of *what* specific datasets should be collected and used for analysis. These can be generated by the digital tools only, or can be used in combination with other available datasets. While the question of *what* normally precedes the *how*, in practice they are often intertwined. In the same way that it is not uncommon to start by exploring how available datasets can support the learning task, it is similarly not uncommon to begin with digital tools and investigate what datasets they are able to generate. Hence, unlike the process itself, these three driving contextual elements form access points to the ILA instance.

3.2. ILA process

The process starts with ***initiation and planning*** where the main objectives of the analysis are identified, informed by the learning task (e.g. profiling in terms of online search strategies). Accordingly, the relevant datasets are identified, and the data collection procedure is planned. This includes choosing appropriate and available digital tools (e.g. web trackers and scrapers) and methods for data collection and addressing associated issues (e.g. licensing). It is also advisable to conduct a pilot data collection to explore the feasibility of the analysis and identify any data quality issues resulting from the use of the chosen tool(s). Informing the students (and their parents) at this point, if applicable, also provides an opportunity to answer any questions that might arise.

The subsequent phase consists of ***data collection and preprocessing***. The data is collected using a variety of digital tools. In cases where a tool is being used for the first time or in a new context, observing a version of the task carried out by a sample group of students may provide trustworthiness in the data collection procedure. The collected data needs to be pre-processed to be suitable for analysis and the specific LA technique chosen [16]. This includes integration of datasets and necessary semantics (relevant to the task), variable extraction, exploratory statistical analysis, dealing with missing values and outliers, etc. [30].

The ***analytics*** technique is then selected depending on the objective(s) determined in the initiation phase (informed by the learning task) and its suitability for the available dataset(s). Each technique is associated with a set of evaluation criteria to indicate the performance of the technique (e.g. precision or accuracy); however, these criteria indicate the validity of the results given the dataset rather than its relevance to the task. Accordingly, the ***interpretation*** of the generated insights in relation to the specific context is essential for determining their value and consequent action to support student learning. In terms of ***action***, a plan needs to be devised to primarily support the learner through the learning task. For example, this could be in the form of personalized feedback, customized recommendation of resources or even adjustment of the task to better target desired learning outcomes. Two types of ***feedback*** are important here: the team or other teachers' feedback on the proposed action based on the LA results, and the students' feedback on any intervention to assess its effectiveness. These two types of feedback would then inform the learning task design for the following instances (whether with the same student or other students, depending on the learning design).

5. Demonstrating an ILA instance

5.1. Initiation & planning

The classes in which the study was performed are English as a foreign language (EFL) classes for grades 8 and 9. The *learning task* was focused on the students' ability to search, comprehend and synthesize information on an indigenous community from an English speaking country of their choice in order to produce a short text. The students were given up to 6 searching and writing sessions (around 45 minutes each) to complete the task. Since students typically search the internet for this type of information, the ORC literature provided the teachers with a degree of understanding of how their students might search for, locate, assess and comprehend the information they find. Teachers were present at the lessons and provided limited support, where needed, to the students during the online researching process - support was primarily provided prior to the commencement of the task in the form of a clear task description and examples of relevant background information (communicated through the LMS). With teachers, we agreed on the *objective* of this instance, which was to cluster the students' online search behavior to reveal salient ORC strategies used by the students. Subsequently, we identified the following *datasets* as necessary: the web search logs collected during the task and a reliable measure students' second-language reading proficiency. Vocabulary size is a strong predictor of reading ability (see Stæhr (2008); Qian (2001); Nation (2013)) and the Vocabulary Size Test (VST) [31] is a validated measure of vocabulary size (Beglar, 2010; Nguyen & Nation (2011)) consisting of a simple quiz format that is easy to administer. Therefore, the VST (in English) was selected as our measure of reading proficiency.

5.2. Collection & pre-processing

The *digital tool* used to collect the web search logs was a browser extension that exports online search history. The students were given the instructions on how to install it on their school laptop at the beginning of the task, and shared the resulting JSON files with their teacher at the end. The teacher then shared the files with one of the researchers who consolidated all the data files and identified them by school, class, student (masked ID) and session. We chose to collect this type of data for two reasons. First, it captures a high level of detail on the digital activity related to this task, without the need to sample by students or by online activity. Second, it does not interfere with the task and, thus, represents students' learning behavior in a largely unobstructed way. This dataset consists of 330 files and 5883 search instances. The VST scores were collected using Nation & Beglar's [33] VST quiz administered by the teachers with guidance from one of the researchers. The total score was added to the search history dataset to account for reading proficiency in the search strategies.

The first step in the pre-processing was to process the JSON files and convert them to CSV format. This choice, instead of importing to a database, was informed by the solution objective of making this ILA artifact accessible to teachers. Through this conversion the basic features from the visited links were extracted: Visit ID, date, time, URL and number of times visited. From this starting point, other features on the session, task and student level could be extracted, such as the terms or phrases used to search or translate, frequently visited pages and overall sequence of a student's search over the whole task.

To provide a meaningful basis for discussion, given the different research disciplines of the team, semantics about the visited links had to be integrated. Thus, the second step was to automatically classify the web domains (e.g. tyda.se) and related links into categories. During the same step, we identified the domain language (English, Swedish or Other). A total of 409 unique domains were extracted and their categories are presented in Table 1. A sample of 100 links were tested manually for fine-tuning the classification.

Table 1. URL domain classification

Task-related category	**Domain classification**
Search	Search engines
Translation	Translation & synonyms
Content (Text-based)	Indigenous society Government Hosting & blogging – on topic Reference & research Education, newsgroups & forums Press
Content (Multimedia)	Cultural images Travel images
Other	Business & economy Mail & communications Entertainment Social Health & sport Local government Hosting & blogging – off topic Utilities Web advertising Learning management Other

The third step focused on investigating what variables the search log data could provide us with in terms of locating and using information. We began by creating a list of ORC strategies identified in previous studies. By examining what could be extracted from the data, we discussed how the literature-based strategies could be translated to online behaviors manifested during the students' searching sessions. This brainstorming session led us to generate the following list of "computable strategies":

Table 2. Computation of ORC search strategies

Strategy	**Computed as**
Forward inferencing or making predictions	Links opened from a search engine search instance with time variance under 1 minute
Applying prior knowledge of internet locations	A set of "General knowledge" websites such as sorummet.se and ne.se
Applying prior knowledge of the topic on the internet	"Content" pages with lowest ID AND not preceded by "Search" pages
Applying accumulated knowledge along sessions	Starting off from websites already visited in previous sessions
Using keywords effectively	Two or more consecutive "Search" pages with different keywords
Formulating useful searches	
Refining searches and search terms	
Dependence on task-specific websites, Monitoring progress	Ratio between "Content" pages and all pages used in a session + progression of that number over sessions

In addition to these strategies that informed the variable selection in our instance, the role of language was of key interest to the teachers. The initial list contained 30 variables; however, since we were only working with 92 students, this number was too large based on established recommendations [34]. This meant that feature reduction was required and was, subsequently performed using correlation analysis and Principal Component Analysis (PCA), both in the SPSS software package. The final list consisted of 10 features: 9 derived from the search logs and the VST scores (see the full list in the appendix).

5.3. Analytics & interpretation

The 10 factors were used to cluster the students using the K-means algorithm implemented using RapidMiner software, a tool also selected for its user-friendly interface. Numbers of clusters between 2 and 5 were tested with 4 providing the most coherent clusters. Figure 1 below displays the four clusters of student behaviors, interpreted in the following subsections.

Cluster 0: Last-minuters. This group were moderately active online, demonstrating a high focus on search and content sites. Their activity increased gradually, peaking in the final sessions. They used domains in both English and Swedish and translated bidirectionally. They appear to acquire knowledge and forward inferencing capabilities as the task progressed. This group had the lowest VST scores.

Cluster 1: Early, focused achievers. Highest level of online activity concentrated in sessions 1 & 2. This group visited domains in both Swedish and English but used more Swedish domains. They also visited domains in other languages and appeared to use synonyms for comprehension rather than translation. This group had the highest VST scores and demonstrated a relatively high acquisition of online domain knowledge.

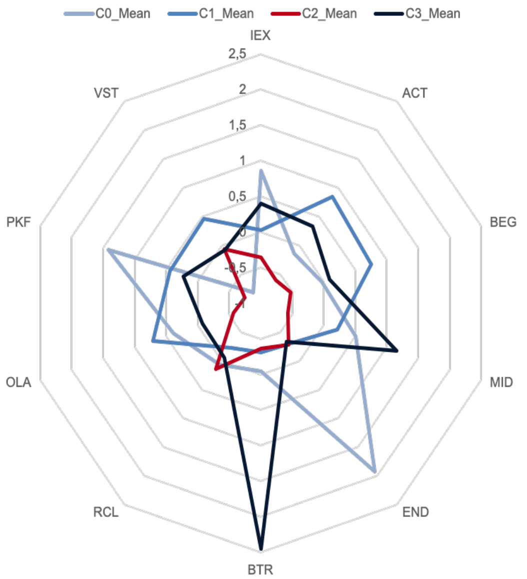

Figure 1. Clusters of search behaviors – comparison

Cluster 2: Dominantly English, stretched task. This group of students demonstrated the lowest level of online activity, taking 4-6 sessions to complete the task. The lowest level of activity was in the middle two sessions (3 & 4). This was the group that scored highest on use of English domains and lowest on translation to English. Cluster 1 also demonstrated minimum use of prior knowledge of online resources.

Cluster 3: Super translators. This group was, by a significant margin, the group that engaged in translation activities both to and from English. They were moderately active online and moderate in levels of using search and content sites. They completed the task in 3-4 sessions. This group scored moderately on the VST.

5.4. Action & feedback

The generated clusters along with their interpretations were then validated with the respective classes' teachers during a 3-hour workshop. The teachers provided the authors with feedback on the process, the variables and factors used, and the resulting clusters. They also discussed the implications of these clusters on their teaching activities and learning support, and the expected benefits of having an automated tool that would allow them to access similar insights for other tasks.

Results from this workshop reveal two important insights. First, the teachers validated the clusters, and as a few noted, they could relate to them based on experience but had not previously seen data-driven evidence that students follow these strategies widely. One of the teachers also noted that different clusters (i.e. behaviors) can be associated with the same student on different tasks and at different times. Being aware of students' information processing behaviors during the task would allow teachers to guide students to ensure that intended learning outcomes, e.g. reading primarily/uniquely in English, focusing on certain types of source text, using translation tools efficiently, refining searches, are met. It would also enable teachers to give more specific, customized feedback post task to help students refine their research approach for similar tasks in the future. Teachers could also use insights to inform future tasks design – to better target intended learning outcomes or encourage students to explore different online searching behaviors. A second iteration of the ILA tool would allow teachers to monitor the impact of feedback that students received or the adapted task design.

The second key insight was that, even though all tools selected have graphical user interfaces and can be regarded as "user-friendly", the workshop highlighted that these criteria alone do not necessarily make the framework and embedded process fully accessible to teachers, especially with regard to some of the preprocessing tasks. The steps of preprocessing and analytics would still need data manipulation skills and parameter selection. Even tools such as RapidMiner that suggests crowd-based parameter values was not perceived by the teachers as "accessible".

6. Discussion and conclusions

The proposed framework extends earlier process-oriented models (cf. [16], [17]) in emphasizing the importance of initiation and planning that is informed by relevant contextual factors. The common practice is that these artifacts start with data collection, or in a more hands-on approach by working with available data. However, the initiation and planning step was important for two reasons: a) contextualizing the LA steps with the teachers' actual needs makes the insights generated from the data more actionable and easier to discuss with colleagues and students, b) since informal learning environments are oversaturated with digital tools, it is necessary to evaluate and select the best tools for collecting data without disturbing students execution of the task. Similarly, our preprocessing phase showed how the literature and extant literacy theories could - and should - influence the selection of features for the analytics phase. Although this is an implicit assumption in LA projects, it is crucial to elucidate how it can be performed to further guide LA researchers towards trustworthy (and actionable) results. The interpretation of results - clusters in this particular demonstration - needs to be carried out in relation to the task and the development of the students. As feedback from teachers in our demonstration revealed, the clusters represented behaviors and not students, and such behaviors may be dynamic. Including teachers in the design process also empowers teachers to experiment and explore with digital tools that are accessible and are easy to use in order to collect and analyze the data they need to support their everyday work.

The proposed framework also adds to existing methods for investigating ORC by proposing a method that is less intrusive than post task, think aloud interviews, provides behavioral data, in contrast to surveys that might capture attitudes and perceptions more accurately than actual task performance, and is able to capture large numbers of students' actual activity on the open Internet. Clustering using a range of carefully selected variables also provides a nuanced view on what shapes students' behavior on a particular task - in our demonstration, for instance, we are clearly able to see the impact of English language reading proficiency and differences in a focus on either language or content across different clusters. An important contribution of this study is to highlight the ORC behaviors and styles of students who are dealing with information in two or more languages, where both information and language play a role in where attention is directed and how a task is performed. The dominance of English on the Internet means that this situation is not unique and remains under-researched.

This study also faced some challenges and limitations. The feedback revealed that there is still a gap between the teachers' competence, the proposed framework and some of the "off the shelf" tools used in the demonstration. This can be tackled in the next iteration of the artifact design by proposing a data-driven design workshop with the teachers for exploring data-generating digital tools that can help them solve learning task problems (e.g. through problem-solution pairing techniques). In the long run, this gap should be investigated to understand if there is a need to integrate LA skills in teacher training curricula and in-service teacher training - in the Swedish context, this would mean equipping teachers with the necessary digital competence required to help students develop their own digital competence. Future work entails the formalization of design principles for the proposed framework based on the improved instantiations currently under work.

References

[1] J. Coiro, M. Knobel, C. Lankshear, and D. J. Leu, *Handbook of research on new literacies*. Routledge, 2014.

[2] Government Offices of Sweden, "Digital kompetens i skolan," 2017.

[3] w3techs, "Usage Statistics and Market Share of Content Languages for Websites," 2020. https://w3techs.com/technologies/overview/content_language (accessed Jul. 15, 2020).

[4] D. J. Leu, C. K. Kinzer, J. Coiro, J. Castek, and L. A. Henry, "New literacies: A dual-level theory of the changing nature of literacy, instruction, and assessment," *Journal of Education*, vol. 197, no. 2, pp. 1–18, 2017.

[5] J. Coiro and E. Dobler, "Exploring the online reading comprehension strategies used by sixth-grade skilled readers to search for and locate information on the Internet," *Reading research quarterly*, vol. 42, no. 2, pp. 214–257, 2007.

[6] B.-Y. Cho and P. Afflerbach, "Reading on the Internet: Realizing and constructing potential texts," *Journal of Adolescent & Adult Literacy*, vol. 58, no. 6, pp. 504–517, 2015.

[7] S. M. Putman, C. Wang, and S. Ki, "Assessing the validity of the cross-cultural survey of online reading attitudes and behaviors with american and south korean fifth-and sixth-grade students," *Journal of Psychoeducational Assessment*, vol. 33, no. 5, pp. 403–418, 2015.

[8] L. Kanniainen, C. Kiili, A. Tolvanen, M. Aro, and P. H. Leppänen, "Literacy skills and online research and comprehension: struggling readers face difficulties online," *Reading and Writing*, vol. 32, no. 9, pp. 2201–2222, 2019.

[9] C. Kiili, D. J. Leu, M. Marttunen, J. Hautala, and P. H. Leppänen, "Exploring early adolescents' evaluation of academic and commercial online resources related to health," *Reading and writing*, vol. 31, no. 3, pp. 533–557, 2018.

[10] D. J. Leu, E. Forzani, C. Rhoads, C. Maykel, C. Kennedy, and N. Timbrell, "The new literacies of online research and comprehension: Rethinking the reading achievement gap," *Reading Research Quarterly*, vol. 50, no. 1, pp. 37–59, 2015.

[11] A. Hutchison and D. Reinking, "Teachers' perceptions of integrating information and communication technologies into literacy instruction: A national survey in the United States," *Reading Research Quarterly*, vol. 46, no. 4, pp. 312–333, 2011.

[12] N. Micic, *Lärares digitala kompetenser: Lärares syn p\aa IKT i skolan*. 2019.

[13] S. Dawson, S. Joksimovic, O. Poquet, and G. Siemens, "Increasing the Impact of Learning Analytics," in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, Tempe AZ USA, Mar. 2019, pp. 446–455, doi: 10.1145/3303772.3303784.

[14] A. Pardo, J. Jovanovic, S. Dawson, D. Gašević, and N. Mirriahi, “Using learning analytics to scale the provision of personalised feedback: Learning analytics to scale personalised feedback,” *Br J Educ Technol*, vol. 50, no. 1, pp. 128–138, Jan. 2019, doi: 10.1111/bjet.12592.

[15] G. Siemens, “Learning analytics: envisioning a research discipline and a domain of practice,” in *Proceedings of the 2nd international conference on learning analytics and knowledge*, 2012, pp. 4–8.

[16] M. A. Chatti, A. L. Dyckhoff, U. Schroeder, and H. Thüs, “A reference model for learning analytics,” *International Journal of Technology Enhanced Learning*, vol. 4, no. 5–6, pp. 318–331, 2012.

[17] A. Nguyen, L. Gardner, and D. Sheridan, “A Design Methodology for Learning Analytics Information Systems: Informing Learning Analytics Development with Learning Design,” 2020.

[18] B. B. Nooraei, Z. A. Pardos, N. T. Heffernan, and R. S. J. de Baker, “Less is More: Improving the Speed and Prediction Power of Knowledge Tracing by Using Less Data.,” in *EDM*, 2011, pp. 101–110.

[19] N. Palavitsinis, V. Protonotarios, and N. Manouselis, “Applying analytics for a learning portal: the Organic. Edunet case study,” in *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, 2011, pp. 140–146.

[20] E. M. Meyers, I. Erickson, and R. V. Small, “Digital literacy and informal learning environments: an introduction,” *Learning, Media and Technology*, vol. 38, no. 4, pp. 355–367, Dec. 2013, doi: 10.1080/17439884.2013.783597.

[21] P. Blikstein and M. Worsley, “Multimodal Learning Analytics and Education Data Mining: Using Computational Technologies to Measure Complex Learning Tasks,” *JLA*, vol. 3, no. 2, pp. 220–238, Sep. 2016, doi: 10.18608/jla.2016.32.11.

[22] R. Jaakonmäki *et al.*, “Understanding Students’ Online Behavior While They Search on the Internet: Searching as Learning,” in *Learning Analytics Cookbook*, Springer, 2020, pp. 75–88.

[23] D. J. Leu *et al.*, “The new literacies of online research and comprehension: Assessing and preparing students for the 21st century with Common Core State Standards,” *Quality reading instruction in the age of common core standards*, pp. 219–236, 2013.

[24] D. J. Leu, C. K. Kinzer, J. L. Coiro, and D. W. Cammack, “Toward a theory of new literacies emerging from the Internet and other information and communication technologies,” *Theoretical models and processes of reading*, vol. 5, no. 1, pp. 1570–1613, 2004.

[25] L. A. Henry, “SEARCHing for an answer: The critical role of new literacies while reading on the Internet,” *The reading teacher*, vol. 59, no. 7, pp. 614–627, 2006.

[26] J. Coiro, “Predicting reading comprehension on the Internet: Contributions of offline reading skills, online reading skills, and prior knowledge,” *Journal of literacy research*, vol. 43, no. 4, pp. 352–392, 2011.

[27] D. B. Daniel and W. D. Woody, “E-textbooks at what cost? Performance and use of electronic v. print texts,” *Computers & Education*, vol. 62, pp. 18–23, 2013.

[28] A. Qayyum and K. Williamson, “The online information experiences of news-seeking young adults,” 2014.

[29] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, “A Design Science Research Methodology for Information Systems Research,” *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, Dec. 2007, doi: 10.2753/MIS0742-1222240302.

[30] F. A. Batarseh, R. Yang, and L. Deng, “A comprehensive model for management and validation of federal big data analytical systems,” *Big Data Analytics*, vol. 2, no. 1, p. 2, 2017.

[31] S. Gregor and A. R. Hevner, “Positioning and presenting design science research for maximum impact,” *MIS Quarterly*, pp. 337–355, 2013.

[32] L. Chandra, S. Seidel, and S. Gregor, “Prescriptive knowledge in IS research: Conceptualizing design principles in terms of materiality, action, and boundary conditions,” in *2015 48th Hawaii International Conference on System Sciences*, 2015, pp. 4039–4048.

[33] D. Beglar and P. Nation, “A vocabulary size test,” *The language teacher*, vol. 31, no. 7, pp. 9–13, 2007.

[34] E. Mooi and M. Sarstedt, *Cluster analysis*. Springer, 2010.

Appendix. Feature list

Group	**Factor**	**Variable(s)**
Task overview	Information extraction (IEX)	Ratio of "Search" pages to total number of pages Ratio of "Content" pages to total number of pages Ratio of "Other" pages to total number of pages
	Level of online activity (ACT)	Average number of pages visited per session Average number of domains visited in the task
Progression	Early sessions (BEG)	Ratio of pages visited in a session to total number of pages (computed for sessions 1 & 2)
	Middle sessions (MID)	Ratio of pages visited in a session to total number of pages (computed for sessions 3 & 4)
	Late sessions (END)	Ratio of pages visited in a session to total number of pages (computed for sessions 5 & 6)
Language	Bidirectional translation (BTR)	Ratio of pages where terms/phrases are translated to English to total number of "Translation" pages Ratio of pages where terms/phrases are translated to Swedish to total number of "Translation" pages Ratio of "Translation" pages to total number of pages
	Reading & comprehension (RCL)	Primary language derived from language of visited pages (Swedish, English or bilingual) Ratio of pages from domains in other languages
	Other language activities (OLA)	Ratio of pages where terms/phrases are translated to a language other than Swedish or English
Prior knowledge	Prior knowledge & forward inferencing (PKF)	General knowledge: Total number of pages not in the Search category, not visited in a preceding session, and not preceded by a Search page Specific knowledge: Same as general knowledge + the page belongs to the subcategories "Indigenous society" or "Relevant blogs" Forward inferencing: Number of non-search pages visited with one minute from a search page
	VST	VST total score