A First Metadata Schema for Learning Analytics Research Data Management

Ian Wolff, Otto-von-Guericke University Magdeburg
David Broneske, German Centre for Higher Education Research and Science Studies
Veit Köppen, Zentral- und Landesbibliothek Berlin

Abstract
In most cases, research data builds the ground for scientific work and to gain new knowledge. Learning analytics is the science to improve learning in different fields of the educational sector. Even though it is a data-driven science, there is no research data management culture or concepts yet. As every research discipline, learning analytics has its own characteristics, which are important for the creation of research data management concepts, in particular for generalization of data and modeling of a metadata model. The following work presents our results of a requirements analysis for learning analytics, in order to identify relevant elements for a metadata schema. To reach this goal, we conducted a literature survey followed by an analysis of our own research about frameworks for evaluation of collaborative programming scenarios from two universities. With these results, we present a discipline-specific scientific workflow, as well as a subject-specific object model, which lists all required characteristics for the development of a learning analytics specific metadata model for data repository usage.

Zusammenfassung
Forschungsdaten bilden die Grundlage für wissenschaftliches Arbeiten und um neue Erkenntnisse zu gewinnen. Learning Analytics ist die Wissenschaft zur Verbesserung des Lernens in verschiedenen Bereichen des Bildungssektors, doch obwohl die Datenerhebung zum größten Teil mittels computergestützter Verfahren durchgeführt wird, besitzt die Disziplin zum jetzigen Zeitpunkt noch keine Forschungsdatenmanagementkultur oder -konzepte. Wie jede Forschungsdisziplin hat Learning Analytics ihre Eigenheiten, die für die Erstellung von Forschungsdatenmanagementkonzepten, insbesondere für die Generalisierung von Daten und die Modellierung eines Metadatenmodells, wichtig sind. Die folgende Arbeit präsentiert Ergebnisse einer Anforderungsanalyse für Learning Analytics, um relevante Elemente für ein Metadatenschema zu identifizieren. Zur Erreichung dieses Ziels führten wir zunächst eine Literaturrecherche durch, gefolgt von einer Untersuchung unserer eigenen Forschung an Softwareumgebungen zur Evaluierung von kollaborativen Programmierszenarien an zwei Hochschulstandorten. Aus den Ergebnissen lassen sich ein disziplinspezifischer wissenschaftlicher Workflow sowie ein fachspezifisches Objektmodell ableiten, das alle erforderlichen Merkmale für die Entwicklung eines für Learning Analytics spezifischen Metadatenmodells für die Nutzung von Datenbeständen aufzeigt.

Zitierfähiger Link (DOI): https://doi.org/10.5282/o-bib/5735

Autorenidentifikation:
Wolff, Ian: GND: 1238114954; ORCID: https://orcid.org/0000-0002-0413-0035
Broneske, David: GND: 1190135663; ORCID: https://orcid.org/0000-0002-9580-740X
Köppen, Veit: GND: 1054320845; ORCID: https://orcid.org/0000-0002-6068-3275

Schlagwörter: Metadaten, Forschungsdatenmanagement, Open Science, Learning Analytics

1. Introduction

Learning analytics and data mining methodologies for extraction of useful and actionable information became a trend during the past decades.[1] Learning analytics is the science involved with the evaluation of produced data by learners, with the aim to reveal information and social connections by using analysis models for learning behavior prediction. These methods bear considerable potential for the higher educational sector.[2] When research data is generated, a sustainable research data management is necessary that gives information on data provenance and makes data traceable and replicable. Data management of learning analytics data comes with several challenges that hinder the development of a learning analytics open science culture and prevent researchers from sharing their data so far.[3] For a neighboring discipline educational data mining,[4] a data repository exists for the community in the American area,[5] but without matching all FAIR criteria.[6] Currently, there is neither a data repository nor a metadata schema for learning analytics that fits the discipline-specific characteristics.[7] Nevertheless, it is shown that with the existence of a research data infrastructure, researchers tend to store their data properly.[8]

1 Lee, Lap-Kei; Cheung, Simon K. S.; Kwok, Lam-For: Learning analytics: current trends and innovative practices, in: Journal of Computers in Education 7 (1), 2020, pp. 1–6. Online: <https://doi.org/10.1007/s40692-020-00155-8>.

2 Leitner, Philipp; Khalil, Mohammad; Ebner, Martin: Learning Analytics in Higher Education – A Literature Review, in: Pena-Ayala, Alejandro (ed.): Learning Analytics. Fundaments, Applications, and Trends, 2017, pp. 1–24. Online: <https://doi.org/10.1007/978-3-319-52977-6>.

3 Biernacka, Katarzyna; Pinkwart, Niels: Opportunities for Adopting Open Research Data in Learning Analytics, in: Tomei, Lawrence; Azevedo, Ana; Azevedo, José Manuel et al. (ed.): Advancing the Power of Learning Analytics and Big Data in Education, 2021, pp. 29–60. Online: <https://publications.informatik.hu-berlin.de/archive/cses/publications/Opportunities-for-Adopting-Open-Research-Data-in-Learning-Analytics.pdf>.

4 Calvet Liñán, Laura; Juan Pérez, Ángel Alejandro: Educational Data Mining and Learning Analytics. Differences, similarities, and time evolution, in: RUSC. Universities and Knowledge Society Journal 12 (3), 2015, p. 98. Online: <http://dx.doi.org/10.7238/rusc.v12i3.2515>.

5 DataShop Homepage, <https://pslcdatashop.web.cmu.edu/>, last accessed 21.06.2021, Koedinger, K.R.: A Data Repository for the EDM Community. The PSLC DataShop, in: Romero, C., Ventura, S., Pechenizkiy, M., Baker, R. (ed.): Handbook of Educational Data Mining, Boca Raton 2011, pp. 43–55. Online: <https://doi.org/10.1201/b10274>.

6 FAIR stands for Findability, Accessibility, Interoperability, and Reusability. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, I. Jsbrand Jan et al.: The FAIR Guiding Principles for scientific data management and stewardship, in: Scientific data 3, 2016. Online: <doi: 10.1038/sdata.2016.18>.

7 For a repository, see the Registry of Research Data Repositories, <https://www.re3data.org/>, last accessed 21.06.2021, The metadata directory of the research data alliance has no specific metadata schema for learning analytics listed, https://rd-alliance.github.io/metadata-directory/, last accessed 21.06.2021.

8 Kowalczyk, Stacy T.: Where Does All the Data Go. Quantifying the Final Disposition of Research Data, in: Proceedings of the American Society for Information Science and Technology 51 (1), 2014, pp. 1–10. Online: <https://doi.org/10.1002/meet.2014.14505101044>.

In our project DiP-iT,[9] we use learning analytics approaches by evaluating collaborative programming scenarios for the development of an educational concept for higher education. To make our data reusable, we also focus on the development of research data management solutions for the generated data during the research processes. In a first step, we develop a discipline-specific metadata schema. For this, we follow Robert Allan's lifecycle for e-research data and information.[10] Initially, we conducted a literature survey to identify general learning analytics approaches and methods for data collection in the field of the evaluation of collaborative programming learning scenarios. In a next step, we analyzed the process around our first generated data and the data itself. From these findings, we deduced a scientific workflow specific to learning analytics. Finally, all findings lead to a model of objects, which maps all relevant criteria for the evolution of a metadata schema specific to learning analytics. In the future, library repositories can make use of this, and data librarians can give researchers assistance in terms of data management to make learning analytics data findable and reusable.

2. Information Lifecycle in Research Data Management

Our approach to develop a metadata schema and concepts for learning analytics research data management follow the e-research data and information lifecycle of Robert Allan (see Fig. 1).[11] The lifecycle fits the DiP-iT data management purpose, because it focuses on metadata creation and data sharing. Driven by the idea that activities involved in conducting research leads to knowledge creation, the generated data comes from observations, experiments, and computation. Knowledge is gained from collected information and their relationships. When it comes to data sharing, especially in the context of collaborative research, the lifecycle meets the problems by creating metadata. The creation of metadata results first from finding resources to the proposed work area, in our case for learning analytics, and second by reviewing literature, whereupon follows the generation of new data. After this, data analysis and sharing takes place. Here, the lifecycle can be supported by the creation of a data repository, which accompanies the aforementioned steps, and makes data sharing possible. Additionally, research discussion about data and the created metadata become vital with a repository. Therefore, researchers can discuss metadata from a practical point of view.

Here, the lifecycle matches the idea of the knowledge pyramid. The pyramid is defined by the tiers of data, information, and knowledge.[12] Data is generated by observation, but unusable until they get meaning by becoming information, in our case through metadata. Knowledge is the next tier and makes it possible to transform information into instructions; this is represented in Allan's lifecycle by publishing reports or papers. The lifecycle claims furthermore for publishing the metadata, which allows future researchers to use, reuse, and to gain new insights upon them as well as extending the schema. Because of the tasks inside the information lifecycle, suitable metadata is a necessity for most tasks inside the lifecycle. Hence, an early creation of a metadata schema makes results reusable for other researchers, also by meeting the FAIR principles.

9 Digitales Programmieren im Team, see project homepage, <http://dip-it.ovgu.de/>, last accessed 21.06.2021.
10 Allan, Robert: Virtual Research Environments. From portals to science gateways, Oxford 2009.
11 Allan, Robert: Virtual Research Environments. From portals to science gateways, Oxford 2009.
12 Rowley, Jennifer: The wisdom hierarchy. Representations of the DIKW hierarchy, in: Journal of Information Science 33 (2), 2007, pp. 163–180. Online: <https://doi.org/10.1177%2F0165551506070706>.

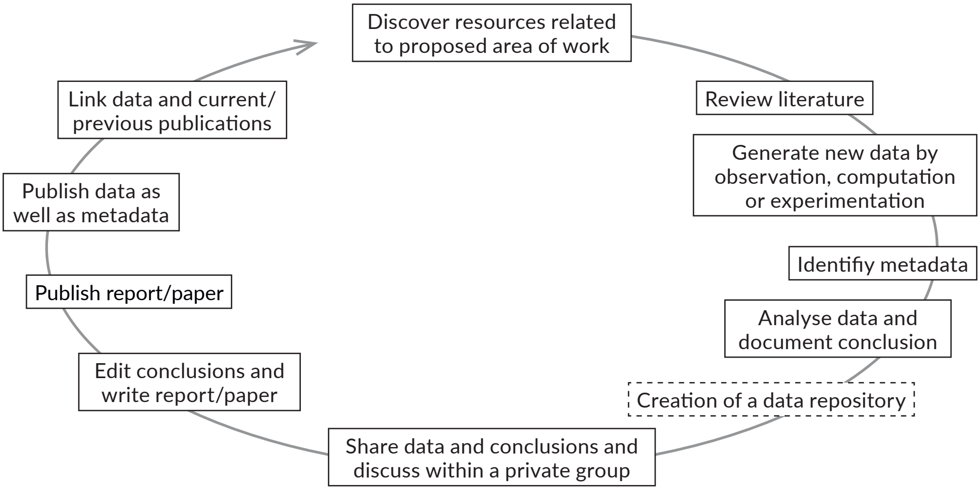

Fig. 1: e-research data and information lifecycle adapted to Allan 2009 and extended by creation of a data repository

3. Requirements Analysis for a Learning Analytics Metadata Schema

Metadata describe objects, they provide administrative information on rights and property rights and describe the structure of an object, which is particularly important when there are several files. Metadata are still essential when search engines seek for data sets or people want to understand the data. Hence, they are still an elementary part of the FAIR principles to make data findable, accessible, interoperable, and reusable.[13] Library infrastructure takes up these principles and plays an important role to adopt the metadata and give access to the information.[14] In the following, we show the results of a literature review, with the aim to determine peculiarities that are to be developed for the learning analytics metadata schema and especially for the collaborative learning of programming languages. This will be incorporated into a research data repository later on in the project.

3.1. Requirement Analysis from Literature Survey

We executed a literature review along the same methodology used in Hawlitschek et al.[15] They have a quite similar research interest (the analysis of pair programming or collaborative programming scenarios in higher education), but from a didactical point of view, while we were looking from the

13 Go FAIR Initiative, <https://www.go-fair.org/fair-principles/>, last accessed 29.06.2021.
14 For example, the PHAIDRA repository of the University of Vienna is following the FAIR principles. Blumesberger, Susanne: Repositorien als Tools für ein umfassendes Forschungsdatenmanagement, in: Bibliothek Forschung und Praxis 44 (3), 2020, pp. 503–511.
15 Hawlitschek, A., Berndt, S., Schulz, S.: Current Landscape of Empirical Research on Pair Programming in Higher Education. A Literature Review, in: Computer Science Education, 2021. [submitted]

information technical side and with a specific focus on metadata. The search was carried out in databases for sociology and computer science – Web of Science, ScienceDirect, ACM Digital Library, IEEE Xplore, Springer-Link, WorldCat, JSTOR, Fachportal Pädagogik and SocioHub. Criteria for database search were the keywords “pair programming” or “pair-programming” and “collaboration” or “cooperation” in abstracts, titles and keywords. The finding period was limited to the years between 2010 and 2020 and only full text paper have been considered from studies within the higher education context. To identify requirements for a metadata schema, we took the corpus of 61 articles (see Appendix) and examined it regarding the question how the learning analytics data are collected, and which methods are used for data collection. This means on the one hand, which measuring instruments are used, and on the other hand, how researchers collect the data in detail. The last one includes the number of attributes measured and for the measurement instrument questionnaire, what scales are used. Furthermore, we examined the computations within the articles to get an idea how data are generated and predictions are made by the researchers. The analysis of the corpus was executed by hand for all relevant criteria described above.

Most studies examine more than one learning characteristic, accordingly the aggregated percentage of the result is greater than hundred. Hawlitschek et al. classify these characteristics into eight categories. The most examined characteristic is the effects of pair programming on students’ perceptions (N = 32; 52.5 %), followed by effects on programming performance (N = 25; 41.0 %), and learning outcome (N = 24; 39.3 %), plus the perceived learning outcome (N = 7; 11.5 %). A medium amount of research studies considered the effects of pair programming on learning behavior (N = 14; 23.0 %) or programming behavior (N = 11; 18.0 %). A small number of studies focusses on persistency (N = 7; 11.5 %) and efficiency (N = 5; 8.2 %).

We found out that for measuring these learning characteristics, different measurement instruments are used (see Fig. 2). Here as well, more than one measurement instrument is used and accordingly, for the aggregated percentages, the result is greater than hundred. Questionnaires (N = 49; 80.3 %) are used by the majority of articles as the method for data collection. Performance tests are made in 35 papers (57.4 %) and used as a variable and calculation together with the results from the questionnaires. Furthermore, although technical tools (N = 20; 32.9 %) like video recording apply as measurement instruments a few times, the ones mostly used are activity log files. In a few of the reviewed articles, interviews (N = 5; 8.2 %) and student reflection papers (N = 4; 6.6 %) are executed. As a result of the plurality of characteristics and measurement instruments given in learning analytics research, a discipline specific metadata schema should provide information about the used measurement instrument. The aim is that another researcher can see whether a secondary analysis with the data set is possible regarding a new research question and different methods.

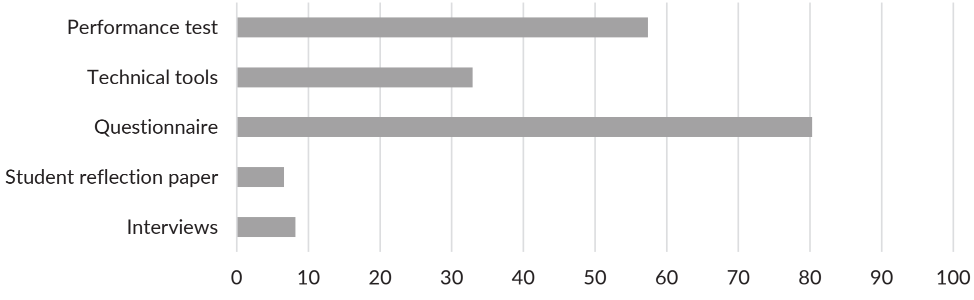

Fig. 2: Distribution of measurement instruments for learning analytics data collection in percentage

Tests and questionnaires are the most common measurement instruments used for measuring items. An item is a task or a question within these tests, e.g., a psychological test. Items relate to one specific characteristic, such as programming experience. In the case of our literature analysis, we used the term attribute for all measured characteristics by all different measurement instruments. The chart in Fig. 3 presents on the x-axis the number of attributes measured within the number of papers on the y-axis. Depending on the research design, several attributes can build a dimension that groups the attributes into a common concept. For example, time needed for programming and lines programmed per hour build the dimension programming performance. Furthermore, every attribute can be a variable itself or counted together as a variable by using codes. These are mostly presented in charts as research results within a publication. Therefore, it is important to map the number of attributes as well as a definition about what was measured within a metadata model, in order that a third-party scientist can assess their suitability for a probable reuse case.

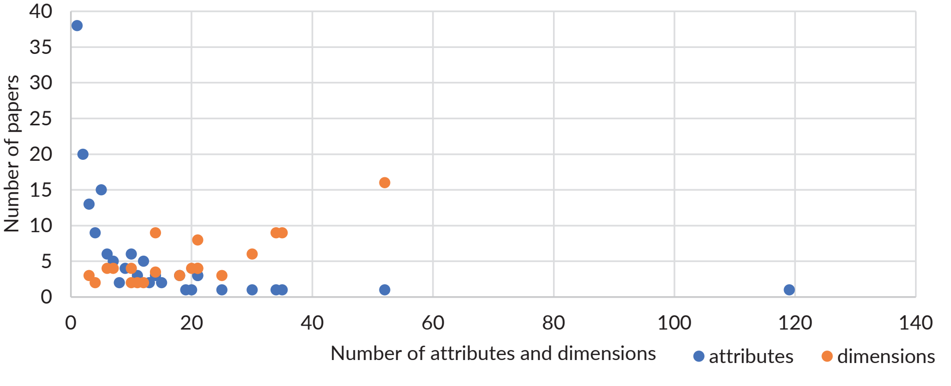

Fig. 3: Number of attributes used in research articles and dimensions used on number of attributes

Every item in a test or questionnaire needs a measurement scale. Scale implies that the questionnaire or test has been scored. Several different scale types are found in the literature with different ranges. Some of the articles do not reveal all details about their research design. Therefore it was not possible to derive the scales used for their test or questionnaires. Mostly used is the 5 point scale and other point scales with different graduations. Among these scales, the Likert scale (N = 29; 47.5 %) appeared most often: a rating scale with different levels (points) that give information about frequencies and intensities that are stepwise graded and that possess an ordinal scale niveau. Besides the point scales, single choice, multiple choice, and open-end questions are used in the questionnaires. To know the exact form of scale in a metadata schema is quite important for the reuse. Third-party scientists can see whether the dataset fits their own, newly devised research questions and can be useful for pretests or whether they can even build their own research on the provided dataset. Furthermore, the variables which resulted from the research and the measured attributes become obvious and this can be described by metadata.

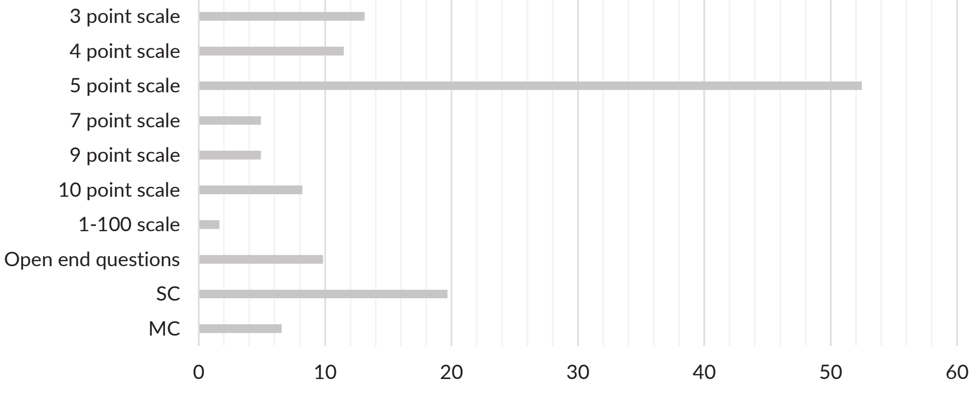

Fig. 4: Scales and kinds of questions used for measurement in questionnaires in percentage

Once the measurement of attributes and characteristics is completed, learning analytics rely on a wide range of statistical methods to test the examined predictions. In our literature review, we found 27 methods. First, reliability is mostly tested with the Cronbach's alpha (N = 7; 11.5 %) and two more statistical methods, which are attributed to an instrument like scales and measure their consistency. To measure the difference between measured variables mostly the t-test (N = 17; 27.9 %), Mann-Whitney-U test (N = 16; 16.4 %), ANOVA (N = 14; 18 %), mean value (N = 7; 11.5 %), Chi-squared test (N = 4; 5.6 %), and ten more statistical calculation models are used. Relationships between variables are measured by two statistical computations models. One time a factor analysis is executed to capture the correlation between variables. Tests for normal distribution appears two times, frequency distribution one time and three more computations related to the aforementioned are used just a few times.

3.2. Frameworks for Evaluation of Collaborative Programming

In the DiP-iT project, we use learning analytics to assess the acquisition of programming skills by collaboration inside teams. To this end, we base our assessment on the *personal characteristics* of students (i.e., their demographic background, motivation, attitude) from questionnaires and combine these with their *skill development* in collaborative programming tasks.

While there is a considerable number of systems for devising questionnaires that assess the status and motivation of students, explicit frameworks are missing that allow for monitoring collaborative behavior of students when programming. To this end, a big challenge in collaborative programming supported by learning analytics is the identification and implementation of suitable frameworks to capture the interaction of students in collaborative programming tasks. Our current investigations in two universities are based on the usage of the SQLValidator[16] and the usage of the version control system GIT.[17]

The SQLValidator is an in-house development from the University of Magdeburg. It features a database backend that allows to run and check user-defined database queries (i.e., in the structured query language SQL).[18] SQL is a technical query language to request, change, delete, and add data from database systems. Furthermore, we extended the system with a team functionality in such a way that collaborative tasks can be posed to a group of students. To assess the students' learning behavior, the SQLValidator logs student interactions with the system (i.e., one action is the submission of a user query and stores the timestamp of the submission and the encountered error), which can be subsequently analyzed. As a result, this log data is needed to be stored and shared with other researchers to draw conclusions.

Similar to the SQLValidator, student interactions and their collaborative learning behavior can be assessed by their interactions with the version control system GIT, which is used at TU Freiberg to submit solutions of collaborative programming tasks in C. Using GIT, we log different actions, which are (1) *issues* that relate to student discussions on a given exercise task, (2) *commits* that represent an intermediate task solution shared among the group, and (3) *pull requests* that represent a task submission. In addition to that we designed a dashboard, which is integrated into GIT that visualizes student interactions and aims at motivating student groups to collaborate more.

Because of the two environments for collaborative learning of programming languages, the metadata model needs to capture different types and formats of data. Therefore, the metadata schema has to list metadata for the *measurement instrument* and the context of environment where it is performed including the *software* used. Additionally, the results in form of the *measurements* are defined as well as the *variables* used for later *calculations* and predictions.

3.3. Learning Analytics Scientific Workflow

The general research process for analyzing collaborative programming scenarios always follows a similar pattern: a researcher or a team of researchers are working on a project that examines the learning

16 Obionwu, Victor; Broneske, David; Hawlitschek, Anja et al.: SQLValidator – An Online Student Playground to Learn SQL, in: Datenbank-Spektrum, 2021. Online: <https://doi.org/10.1007/s13222-021-00372-0>.

17 Zug, Sebastian; Dietrich, André; Rudolf, Galina; Treumer, Jonas: Teamarbeit lernen – im Team lernen. Gruppenorientiertes Arbeiten in der Informatik, in: ACAMONTA 27, 2020, pp. 84–87. Online: <https://tu-freiberg.de/sites/default/files/media/freunde-und-foerderer-der-technischen-universitaet-bergakademie-freiberg-ev-6089/pdf/acamonta2020/acamonta_webversion_verlinkt.pdf>.

18 Saake, Gunter; Kai-Uwe Sattler; Andreas Heuer: Datenbanken. Konzepte und Sprachen, Frechen 2018[6].

process within a teacher-learner scenario. In our scenario, learners and teachers come from the university sector, i.e. they are students and lecturers. As the practical survey about our research and the literature analysis has shown, data are mostly collected by using various measurement instruments such as interviews, questionnaires, performance tests, technical tools and student reflection papers.

In general, learning analytics adapt the idea of measurement, e.g., psychological, to collect results needed for improving the learning process. Psychological measurement incorporates a definition of a construct, then figuring out how to analyze and account for different sources of error by using a specified measurement model and a reliable instrument, to frame a valid argument for particular uses of the outcome at the end.[19]

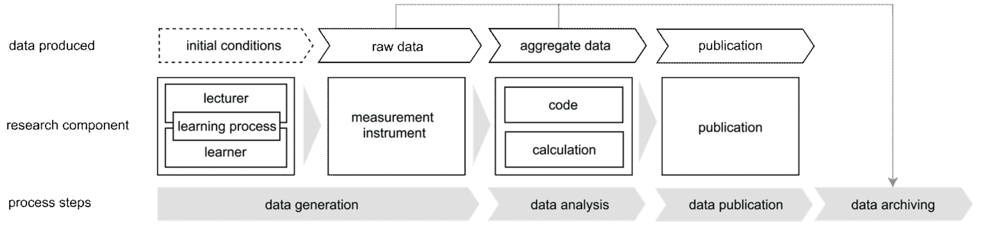

Fig. 5: Learning analytics scientific workflow

Data Generation

The learning analytics scientific workflow shown in Fig. 5 starts with initial conditions, where lecturers observe certain actions and behaviors of the learners to make the learning process visible. In the first step, learners are classified by questionnaires whereupon characteristics of the test groups become apparent. The next step is the measurement of learners' actions with a measurement instrument, which, in our case, means to evaluate (1) the collaborative programming scenarios and (2) the use of technical tools like e-learning platforms combined with performance tests. When measurement instruments observe the learners' actions, the first raw data sets are generated, which provide information about the actions of the learners.

Data Analysis & Publication

As third step, researchers conduct data selection and further processing according to the research question. In this step, data are aggregated by using different computations, transformations, mappings, or codes. These are not only part of the data aggregation, but also part of the final publication. For example, researchers include individual aspects from the questionnaires as variables or compare the measurement data with the results of the performance test. Afterwards the results are summarized in graphics and published in publications. Note that a considerable part of raw data from the measurement instruments are not used for research, as it does not depict the proper behavior or due to

19 Bergner, Yoav: Measurement and its Uses in Learning Analytics, in: Lang, Charles; Siemens, George; Wise, Alyssa et al.: Handbook of Learning Analytics, 2017, pp. 34–48. Online: <https://doi.org/10.18608/hla17>.

incorrect assumptions, sensor errors, misconduct of learners, missing data, inconsistencies, or outliers. This data cleansing is subject specific and thus also included in the data analysis part of the process.

Privacy issues are a further matter of data cleansing due to the learning analytics data generation from students and lecturers. To ensure data publication and archiving, an informed consent from the students is needed, in which they get information about the kind of data usage after data generation.[20] The informed consent takes place bevor the data generation phase, and can be withdrawn from the students at every point in time. Accordingly, a trustful relationship between students and researchers is needed, while the researcher decides about the access to the data. Hence, data must be anonymized when used for a public use case and the used transformations on the raw data need to be documented. Storage of the raw data is possible, but only with regulations for access, because this data could allow a disclosure of identities even after proper anonymization.

Data Archiving

In a project, several observations are usually carried out. Once the project has been completed, some of the data are stored on the institute's data storage servers and often a considerable amount of the extensive raw data is deleted or stored in an archive system in the university data center. However, this does not follow a standardized protocol. An internal exchange of data within an institute would be possible in this way, but there is still no subject-specific learning analytics repository in which research data can be published for data sharing with the broader community. The evaluation of the literature used for the literature review has shown that none of the articles had a reference to any research data. Nonetheless, the specialist community of the learning analytics is constantly developing,[21] which is also evident by the LAK conference, which started in 2010.[22] The resulting data have a certain value because, based on a small group of examined learners, the results of different measurements could be combined to increase the plausibility of the examined objects. Older data does not lose its topicality very quickly. Accordingly, if a subject-specific data management plan is used and a sustainable data management is executed during the research process, the collected data can remain usable for a longer period.[23]

Libraries and their service portfolio can play a crucial role in this process. When the metadata schema is derived from the following model of objects, already existing library infrastructure and staff can adapt the schema. Especially data librarians, embedded in the research process, can assist by managing data collection and files for primary and secondary analysis. Furthermore, technical assistance and reference services can be provided to reach an entire community of users. Therefore,

20 The DELICATE checklist for learning analytics can be applied for the creation of an informed consent. Drachsler, Hendrik; Greller, Wolfgang: Privacy and analytics, in: Gašević, Dragan; Lynch, Grace; Dawson, Shane et al. (eds.): Proceedings of the Sixth International Conference on Learning Analytics & Knowledge – LAK '16, New York 2016, pp. 89–98.

21 Lee: Learning analytics, 2020, pp. 1–6.

22 Learning Analytics and Knowledge Conference, <https://www.solaresearch.org/events/lak/lak21/>, last accessed 29.06.2021.

23 Kowalczyk, Stacy T.: Where Does All the Data Go. Quantifying the Final Disposition of Research Data, in: Proceedings of the American Society for Information Science and Technology 51 (1), 2014, pp. 1–10. Online: <https://doi.org/10.1002/meet.2014.14505101044>.

open access data sets provided by the already existing library research data repository infrastructure can be used. Data librarians can also bridge the gap between the stored data sets and the learning analytics research community by providing education on access and use of the data sets to develop further researcher awareness.[24]

3.4. Model of Objects

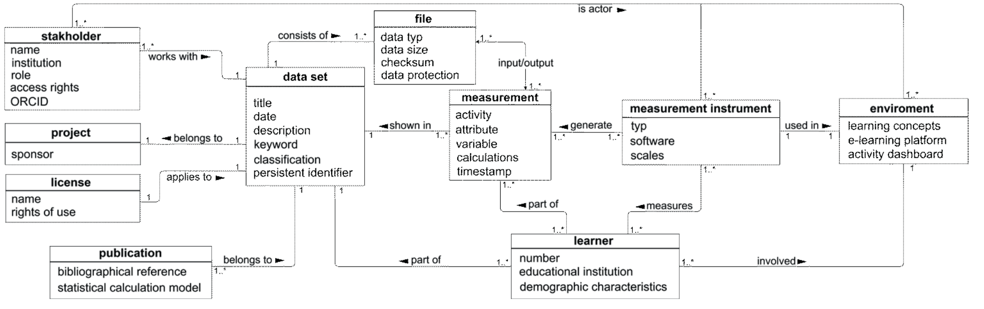

Fig. 6: Model of objects with domain specific characteristics for learning analytics

Based on the findings of the requirements analysis, we deduced an object model in UML (Unified Modeling Language) class diagram notation, depicted in Fig. 6.[25] It is a first step towards developing a metadata schema. The model connects all relevant domain-specific characteristics of the learning analytics scientific workflow. Like the *stakeholder*, who represents the lecturer from the workflow; the *learner* and all domain specific entities like *environment*, *measurement instrument*, and *measurement* that could be derived from the data production phase of the scientific workflow. Descriptions of the measurements like *variables* and *calculations* result directly from the data analysis and data publication phase of the learning analytics workflow, while the entity *file* and their specifics are important in the phase of archiving.

The data set marks the result data after observation of the learner's learning behavior. It is the central entity in the object model and consist of multiple *files* with file attributes. These are *name* and *data type* and furthermore a *checksum* per file and a *persistent identifier* for the data set, which is important for preservation. These entities are the technical metadata. Descriptive metadata like *title*, a *date*, a *description*, *keywords*, and a *classification* describe the data set in more detail. The last one classifies

24 See Semeler, Alexandre Ribas; Pinto, Adilson Luiz; Rozados, Helen Beatriz Frota: Data science in data librarianship. Core competencies of a data librarian, in: Journal of Librarianship and Information Science 51 (3), 2019, pp. 771–780, Ohaji, Isaac K.; Chawner, Brenda; Yoong, Pak: The role of a data librarian in academic and research libraries, in: Information Research 24 (4), 2019. Online: <http://informationr.net/ir/24-4/paper844.html>, Shumaker, David: The embedded librarian. Innovative strategies for taking knowledge where it's needed, Medford 2012, Cremer, Fabian; Engelhardt, Claudia; Neuroth, Heike: Embedded Data Manager – Integriertes Forschungsdatenmanagement. Praxis, Perspektiven und Potentiale, in: Bibliothek Forschung und Praxis 39 (1), 2015.

25 Kecher, Christoph; Salvanos, Alexander; Hoffmann-Elbern, Ralf: UML 2.5. Das umfassende Handbuch, Bonn 2018[6].

the data set into existing library classification systems. For example, considering the Dewey Decimal Classification, our project's data would fit the class 300 for social science and the division 370 for education.[26] Furthermore, in this context embedded is the *stakeholder*, who is the actor for the measurement instrument and observes the learner. Described with a *name* and a *persistent identifier* like ORCID,[27] a related *institution*, he or she possesses *access rights* to the data set for repository usage. A *role* is also defined, which marks the position within a *project* which has a *funding context*, and *publications* is also related to this type of metadata. Administrative metadata as the *license* give information about the *rights of use* for the data.

Moreover, our object model focuses on the discipline-specific metadata. Considering the *learner* that possesses characteristics like the *number* and the attended *educational institution* and who has also as observed object *demographic characteristics*. This information already results from measurements made by the measurement instrument questionnaire. It marks mostly the first step of learning analytics data generation and builds variables for later transformation and research summaries. In the special learning analytics field that evaluates a collaborative programming scenario, the environment is the place of actions. The e-learning platform tracks learners' actions and the stakeholder is observing them. A software entity describes the used version of the e-Learning platform. Results of the used measurement instruments within the environment are measurements. They represent the general *activity* exercised by the learners and in detail the measured *attributes*, which can also be a *variable* by using *calculations*, or variables are used for *statistical calculation models* in a publication. The raw data handling with data selection and processing takes also place here, which is described by the calculation entity and by the measurements in form of attributes and variables. Resulting aggregated data is then stored as files in the data set and defined by the file entity. Especially important in the learning analytics context are information about data protection. Here the kind and degree of anonymization is defined, which has been applied to the files.

The model of objects maps all relevant criteria for our research in the evaluation of collaborative programming scenarios. Moreover, the model of objects applies to other learning analytics research approaches and gives researchers a structure for their data. It matches the FAIR principles,[28] depicted in Fig. 7, by taking community standards into consideration, to make data interoperable and reusable, but also findable with the help of persistent identifiers and library classifications standards. Hence, libraries can use the resulting schema within their repositories and make the data findable and reusable, especially by taking long-term preservation standards into consideration.

26 Dewey-Dezimalklassifikation und Register: DDC 22 (german edition), founded by Melvil Dewey, Joan S. Mitchell et al. (eds.), 2 volumes, München 2005.

27 ORCID, <https://orcid.org/>, last accessed 01.07.2021.

28 Wolff, Ian; Broneske, David; Köppen, Veit: FAIR Research Data Management for Learning Analytics, in: Lingnau, Andreas (ed.): Proceedings of DELFI Workshop 2021, Bottrop 2021, pp. 158–163. Online: <https://repositorium.hs-ruhrwest.de/frontdoor/index/index/docId/733>.

Findability:	use of a persistent identifier, e.g. DOI[29]
Accessibility:	definition of *access rights* and *roles* for *stakeholders*
Interoperability:	the model of objects refers in terms of standard vocabulary to the LAEP (learning analytics educational policy) glossary[30]
Reusability:	use of domain specific metadata like *learner*, *environment*, *measurement instrument* and *measurements* to provide detailed information about data provenance attached *license* to show what usage rights belong to data use of data types that meet community standards

Fig. 7: FAIR principles adapted by the model of objects

4. Conclusion and Future Perspective

In this paper, we focus on the development of a metadata schema for learning analytics, which is currently lacking for research data management solutions. For this, we executed a requirement analysis. First, we made a literature survey following the questions how data are generated and what methods are used in learning analytics. In a second step, we observed the research process of our own research in the evaluation of collaborative programming scenarios. These led us to a detailed learning analytics specific scientific workflow, which pictures all relevant steps of data generation.

The result of the requirement analysis is an object model that shows all relevant criteria of the learning analytics workflow and relates all entities by considering the FAIR Principles. Metadata can be seen as the key for third-party scientist to find, get access, and reuse scientific data. In a next step and derived from these findings, the metadata schema will be developed by using parts of already existing metadata schemas to increase the findability and extend it by adding discipline specific metadata. After the subsequent implementation of the schema in a test repository, a usability test of the schema will be executed within our project context. When the metadata schema is completed, it builds the ground for research data management of learning analytics data and can be recommended by Open Science specialists in libraries and implemented in already existing repository infrastructures. A final version of the metadata schema will be shared with the learning analytics and library community on the project website[31], and permanently published in a data repository.

Acknowledgements

Funding: This work was supported by the German Federal Ministry of Education and Research [grant number 16DHB 3008].

29 DOI stands for Digital Object Identifier, <https://www.doi.org/>, last accessed 01.09.2021.
30 LAEP Glossary, <https://iet-ou.github.io/cloudworks-ac-uk/cloud/view/9781/links.html>, last accessed 01.09.2021.
31 Project DiP-iT,<http://dip-it.ovgu.de/>, last accessed 01.09.2021.

References

- Allan, Robert: Virtual Research Environments. From portals to science gateways, Oxford 2009.
- Bergner, Yoav: Measurement and its Uses in Learning Analytics, in: Lang, Charles; Siemens, George; Wise, Alyssa et al.: Handbook of Learning Analytics, 2017, pp. 34-48. Online: <https://doi.org/10.18608/hla17>.
- Biernacka, Katarzyna; Pinkwart, Niels: Opportunities for Adopting Open Research Data in Learning Analytics, in: Tomei, Lawrence; Azevedo, Ana; Azevedo, José Manuel et al. (ed.): Advancing the Power of Learning Analytics and Big Data in Education, 2021, pp. 29–60. Online: <https://publications.informatik.hu-berlin.de/archive/cses/publications/Opportunities-for-Adopting-Open-Research-Data-in-Learning-Analytics.pdf>.
- Blumesberger, Susanne: Repositorien als Tools für ein umfassendes Forschungsdatenmanagement, in: Bibliothek Forschung und Praxis 44 (3), 2020, pp. 503–511.
- Calvet Liñán, Laura; Juan Pérez, Ángel Alejandro: Educational Data Mining and Learning Analytics: differences, similarities, and time evolution, in: RUSC. Universities and Knowledge Society Journal 12 (3), 2015. Online: <http://dx.doi.org/10.7238/rusc.v12i3.2515>.
- Cremer, Fabian; Engelhardt, Claudia; Neuroth, Heike: Embedded Data Manager – Integriertes Forschungsdatenmanagement. Praxis, Perspektiven und Potentiale, in: Bibliothek Forschung und Praxis 39 (1), 2015, pp. 13–31.
- Dewey-Dezimalklassifikation und Register: DDC 22 (german edition), founded by Melvil Dewey, Joan S. Mitchell et al. (eds.), 2 volumes, München 2005.
- Drachsler, Hendrik; Greller, Wolfgang: Privacy and analytics, in: Gašević, Dragan; Lynch, Grace; Dawson, Shane et al. (eds.): Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16, New York 2016, pp. 89–98.
- Hawlitschek, A., Berndt, S., Schulz, S.: Current Landscape of Empirical Research on Pair Programming in Higher Education. A Literature Review, in: Computer Science Education, 2021. *[submitted]*
- Kecher, Christoph; Salvanos, Alexander; Hoffmann-Elbern, Ralf: UML 2.5. Das umfassende Handbuch, Bonn 2018[6].
- Koedinger, K.R.: A Data Repository for the EDM Community. The PSLC DataShop, in: Romero, C., Ventura, S., Pechenizkiy, M., Baker, R. (ed.): Handbook of Educational Data Mining, Boca Raton 2011, pp. 43–55. Online: <https://doi.org/10.1201/b10274>.
- Kowalczyk, Stacy T.: Where Does All the Data Go: Quantifying the Final Disposition of Research Data, in: Proceedings of the American Society for Information Science and Technology 51 (1), 2014, pp. 1–10. Online: <https://doi.org/10.1002/meet.2014.14505101044>.
- Lee, Lap-Kei; Cheung, Simon K. S.; Kwok, Lam-For: Learning analytics. Current trends and innovative practices, in: Journal of Computers in Education 7 (1), 2020, pp. 1–6. Online: <https://doi.org/10.1007/s40692-020-00155-8>.
- Leitner, Philipp; Khalil, Mohammad; Ebner, Martin: Learning Analytics in Higher Education – A Literature Review, in: Pena-Ayala, Alejandro (ed.): Learning Analytics. Fundaments, Applications, and Trends, 2017, pp. 1–24. Online: <https://doi.org/10.1007/978-3-319-52977-6>.

- Obionwu, Victor; Broneske, David; Hawlitschek, Anja et al.: SQLValidator – An Online Student Playground to Learn SQL, in: Datenbank-Spektrum, 2021. Online: <https://doi.org/10.1007/s13222-021-00372-0>.
- Ohaji, Isaac K.; Chawner, Brenda; Yoong, Pak: The role of a data librarian in academic and research libraries, in: Information Research 24 (4), 2019. Online: <http://informationr.net/ir/24-4/paper844.html>.
- Rowley, Jennifer: The wisdom hierarchy. Representations of the DIKW hierarchy, in: Journal of Information Science 33 (2), 2007, pp. 163–180. Online: <https://doi.org/10.1177%2F0165551506070706>.
- Saake, Gunter; Sattler, Kai-Uwe; Heuer, Andreas: Datenbanken: Konzepte und Sprachen, Frechen 2018[6].
- Semeler, Alexandre Ribas; Pinto, Adilson Luiz; Rozados, Helen Beatriz Frota: Data science in data librarianship. Core competencies of a data librarian, in: Journal of Librarianship and Information Science 51 (3), 2019, pp. 771–780.
- Shumaker, David: The embedded librarian. Innovative strategies for taking knowledge where it's needed, Medford 2012.
- Wolff, Ian; Broneske, David; Köppen, Veit: FAIR Research Data Management for Learning Analytics, in: Lingnau, Andreas (ed.): Proceedings of DELFI Workshop 2021, Bottrop 2021, pp. 158–163. Online: <https://repositorium.hs-ruhrwest.de/frontdoor/index/index/docId/733>.
- Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, I. Jsbrand Jan et al.: The FAIR Guiding Principles for scientific data management and stewardship, in: Scientific data 3, 2016. Online: <doi: 10.1038/sdata.2016.18>.
- Zug, Sebastian; Dietrich, André; Rudolf, Galina; Treumer, Jonas: Teamarbeit lernen – im Team lernen. Gruppenorientiertes Arbeiten in der Informatik, in: ACAMONTA 27, 2020, pp. 84–87. Online: <https://tu-freiberg.de/sites/default/files/media/freunde-und-foerderer-der-technischen-universitaet-bergakademie-freiberg-ev-6089/pdf/acamonta2020/acamonta_webversion_verlinkt.pdf>.

Appendix: Reference List from Literature Survey

- Ahmad, Mazida; Abd Razak, Ainul Husna; Omar, Mazni et al.: The Impact of Knowledge Management in Pair Programming on Program Quality, in: Abraham, Ajith; Muda, Azah Kamilah; Choo, Yun-Huoy (ed.): Pattern Analysis, Intelligent Security and the Internet of Things, Cham 2015 (Advances in Intelligent Systems and Computing), pp. 159–168.
- Al-Jarrah, Ahmad; Pontelli, Enrico; Zaphiris, Panayiotis et al: On the Effectiveness of a Collaborative Virtual Pair-Programming Environment, in: Zaphiris, P.; Ioannou, A. (ed.): Learning and Collaboration Technologies, 2016 (Lecture notes in computer science), pp. 583–595.
- Alvarado, Christine; Lee, Cynthia Bailey; Gillespie, Gary: New CS1 Pedagogies and Curriculum, the Same Success Factors?, in: Proceedings of the 45th ACM Technical Symposium on Computer Sciences Education, 2014 (SIGCSE'14), pp. 379–384.
- Anderson, Nicole; Gegg-Harrison, Tim: Learning Computer Science in the "Comfort Zone of Proximal Development", in: Proceeding of the 44th ACM Technical Symposium on Computer Science Education, New York 2013 (SIGCSE '13), pp. 495–500.

- Benadé, Trudie; Liebenberg, Janet: Pair Programming as a Learning Method Beyond the Context of Programming, in: Proceedings of the 6th Computer Science Education Research Conference, New York 2017 (CSERC '17), pp. 48–55.
- Bowman, Nicholas A.; Jarratt, Lindsay; Culver, K. C. et al.: Pair Programming in Perspective: Effects on Persistence, Achievement, and Equity in Computer Science, in: Journal of Research on Educational Effectiveness, 2020, pp. 1–28.
- Bowman, Nicholas A.; Jarratt, Lindsay; Culver, K. C. et al.: How Prior Programming Experience Affects Students' Pair Programming Experiences and Outcomes, in: Proceedings of the ACM Conference on Innovation and Technology in Computer Science Education 2019, New York 2019 (ITiCSE '19), pp. 170–175.
- Braught, Grant; MacCormick, John; Wahls, Tim: The Benefits of Pairing by Ability, in: Proceedings of the 41st ACM Technical Symposium on Computer Science Education, New York 2010 (SIGCSE '10), pp. 249–253.
- Braught, Grant; Wahls, Tim; Eby, L. Marlin: The Case for Pair Programming in the Computer Science Classroom, in: ACM Trans. Comput. Educ. 11 (1), 2011, pp. 1–21.
- Celepkolu, Mehmet; Boyer, Kristy Elizabeth: The Importance of Producing Shared Code through Pair Programming, in: Proceedings of the 49th ACM Technical Symposium on Computer Science Education, New York 2018a (SIGCSE '18), pp. 765–770.
- Celepkolu, Mehmet; Boyer, Kristy Elizabeth: Thematic Analysis of Students' Reflections on Pair Programming in CS1, in: Proceedings of the 49th ACM Technical Symposium on Computer Science Education, New York 2018b (SIGCSE '18), pp. 771–776.
- Chen, Kuanchin; Rea, Alan: Do Pair Programming Approaches Transcend Coding? Measuring Agile Attitudes in Diverse Information Systems Courses, in: Journal of Information Systems Education 29 (2), 2018, pp. 53–64. Online: <http://jise.org/Volume29/n2/JISEv29n2p53.html>.
- Choi, Kyungsub S.: Evaluating Gender Significance within a Pair Programming Context, in: Proceedings of the 46th Hawaii International Conference on System Sciences 2013, 2013 (HICSS '13), pp. 4817–4825. Online: <https://doi.org/10.1109/HICSS.2013.209>.
- Coman, Irina D.; Robillard, Pierre N.; Sillitti, Alberto et al.: Cooperation, Collaboration and Pair-Programming. Field Studies on Backup Behavior, in: Journal of Systems and Software 91, 2014, pp. 124–134.
- Demir, Ömer; Seferoglu, Süleyman Sadi: A Comparison of Solo and Pair Programming in Terms of Flow Experience, Coding Quality, and Coding Achievement, in: Journal of Educational Computing Research, 2020, pp. 1–19.
- Demir, Ömer; Seferoglu, Süleyman Sadi: The Effect of Determining Pair Programming Groups According to Various Individual Difference Variables on Group Compatibility, Flow, and Coding Performance, in: Journal of Educational Computing Research, 2020, pp. 1–30.
- Dongo, Tendai; Reed, April H.; O'Hara, Margaret et al.: Exploring Pair Programming Benefits for MIS Majors, in: Journal of Information Technology Education-Innovations in Practice 15, 2016, pp. 223–239.
- Dou, WanFeng; He, Wei: Compatibility and Requirements Analysis of Distributed Pair Programming, in: Second International Workshop on Education Technology and Computer Science, 2010 (ECTS '10), pp. 467–470.

- Edwards, Richard L.; Stewart, Jennifer K.; Ferati, Mexhid: Assessing the Effectiveness of Distributed Pair Programming for an Online Informatics Curriculum, in: ACM Inroads 1 (1), 2010, pp. 48–54.
- Estácio, Bernardo; Oliveira, Roberto; Marczak, Sabrina et al.: Evaluating Collaborative Practices in Acquiring Programming Skills Findings of a Controlled Experiment, in: Proceedings of the 29th Brazilian Symposium on Software Engineering 2015 (SBES '15), pp. 150–159.
- Estácio, Bernardo; Valentim, Natasha; Rivero, Luis et al.: Evaluating the Use of Pair Programming and Coding Dojo in Teaching Mockups Development. An Empirical Study, in: 48th Hawaii International Conference 2015 (HICSS '15), pp. 5084–5093.
- Faja, Silvana: Evaluating Effectiveness of Pair Programming as a Teaching Tool in Programming Courses, in: Information Systems Education Journal 12 (6), 2014, pp. 36–45. Online: <https://eric.ed.gov/?id=EJ1140923>.
- Ghorashi, Soroush; Jensen, Carlos: Integrating Collaborative and Live Coding for Distance Education, in: Computer 50 (5), 2017, pp. 27–35.
- Goel, Sanjay; Kathuria, Vanshi: A Novel Approach for Collaborative Pair Programming, in: Journal of Information Technology Education: Research 9 (1), 2010, pp. 183–196. Online: <https://www.learntechlib.org/f/111363/>.
- Govender, Desmond W.; Govender, T. P.: Using a Collaborative Learning Technique as a Pedagogic Intervention for the Effective Teaching and Learning of a Programming Course, in: Mediterranean Journal of Social Sciences 5 (20), 2014, pp. 1077–1086.
- Harsley, Rachel; Di Eugenio, Barbara; Green, Nick u. a.: Enhancing an Intelligent Tutoring System to Support Student Collaboration: Effects on Learning and Behavior, in: Andre, E.; Baker, R.; Hu, X. et al. (ed.): Artificial Intelligence in Education, 2017, pp. 519–522.
- Harsley, Rachel; Fossati, Davide; Di Eugenio, Barbara et al.: Interactions of Individual and Pair Programmers with an Intelligent Tutoring System for Computer Science, in: Proceedings of the ACM SIGCSE Technical Symposium on Computer Science Education 2017, New York 2017 (SIGCSE '17), pp. 285–290. Online: <https://doi.org/10.1145/3017680.3017786>.
- Jarratt, Lindsay; Bowman, Nicholas A.; Culver, K. C. et al.: A Large-Scale Experimental Study of Gender and Pair Composition in Pair Programming, in: Proceedings of the ACM Conference on Innovation and Technology in Computer Science Education 2019, New York 2019 (ITiCSE '19), pp. 176–181.
- Jurado, Francisco; Molina, Ana I.; Redondo, Miguel A. et al.: Cole-Programming: Shaping Collaborative Learning Support in Eclipse, in: IEEE Revista Iberoamericana de Tecnologias del Aprendizaje 8 (4), 2013, pp. 153–162.
- Karthiekheyan, Kavitha; Ahmed, Irfan; Jayalakshmi, Jalaja: Pair Programming for Software Engineering Education: An Empirical Study, in: The International Arab Journal of Information Technology (IAJIT) 15 (2), 2018, pp. 246–255.
- Kavitha, R. K.; Ahmed, M. S. Irfan: Knowledge Sharing through Pair Programming in Learning Environments: An Empirical Study, in: Education and Information Technologies 20 (2), 2015, pp. 319–333.
- Kongcharoen, Chaknarin; Hwang, Wu-Yuin; Ghinea, Gheorghita: Synchronized Pair Configuration in Virtualization-Based Lab for Learning Computer Networks, in: Educational Technology & Society 20 (3), 2017, pp. 54–68.

- Lewis, Colleen M.; Titterton, Nathaniel; Clancy, Michael: Using Collaboration to Overcome Disparities in Java Experience, in: Proceedings of the Ninth Annual International Conference on International Computing Education Research, New York 2012 (ICER '12), pp. 79–86.
- Maguire, Phil; Maguire, Rebecca; Hyland, Philip et al.: Enhancing Collaborative Learning Using Pair Programming: Who Benefits?, in: All Ireland Journal of Teaching and Learning in Higher Education 6 (2), 2014, pp. 1411–1419. Online: <https://ojs.aishe.org/index.php/aishe-j/article/view/141>.
- McChesney, Ian: Three Years of Student Pair Programming. Action Research Insights and Outcomes, in: Proceedings of the 47th ACM Technical Symposium on Computing Science Education, New York 2016 (SIGCSE '16), pp. 84–89.
- Nawahdah, Mamoun; Taji, Dima: Investigating Students' Behavior and Code Quality when Applying Pair-Programming as a Teaching Technique in a Middle Eastern Society, in: Proceedings of IEEE Global Engineering Education Conference 2016, Piscataway 2016 (EDUCON '16), pp. 32–39.
- O'Donnell, Clem; Buckley, Jim; Mahdi, Abdulhussain et al.: Evaluating Pair-Programming for Non-Computer Science Major Students, in: Proceedings of the 46th ACM Technical Symposium on Computer Science Education, New York 2015 (SIGCSE '15), pp. 569–574.
- Oliveira, Caio Matheo Campos de; Canedo, Edna Dias; Faria, Henrique et al.: Improving Student's Learning and Cooperation Skills Using Coding Dojos (In the Wild!), in: IEEE Frontiers in Education Conference 2018, 2018 (FIE '18), pp. 1–8.
- Radermacher, Alex; Walia, Gursimran; Rummelt, Richard: Improving Student Learning Outcomes with Pair Programming, in: Proceedings of the Ninth Annual International Conference on International Computing Education Research, New York 2012 (ICER '12), pp. 87–92.
- Reckinger, Shanon; Hughes, Bryce: Strategies for Implementing In-Class, Active, Programming Assessments: A Multi-Level Model, in: Proceedings of the 51st ACM Technical Symposium on Computer Science Education, New York 2020 (SIGCSE '20), pp. 454–460.
- Rodrigo, Ma. Mercedes T.: Exploratory Analysis of Discourses between Students Engaged in a Debugging Task, in: 25th International Conference on Computers in Education 2017: Technology and Innovation: Computer-based Educational Systems for the 21th Century, 2017 (ICCE '17), pp. 198–203.
- Rodriguez, Fernando J.; Boyer, Kristy Elizabeth: Discovering Individual and Collaborative Problem-Solving Modes with Hidden Markov Models, in: Conati, C.; Heffernan, N.; Mitrovic, A. et al. (ed.): Artificial Intelligence in Education, 2015, pp. 408–418.
- Rodríguez, Fernando J.; Price, Kimberly Michelle; Boyer, Kristy Elizabeth: Exploring the Pair Programming Process. Characteristics of Effective Collaboration, in: Proceedings of the ACM SIGCSE Technical Symposium on Computer Science Education 2017, New York 2017 (SIGCSE '17), pp. 507–512.
- Rong, Guoping; Zhang, He; Xie, Mingjuan et al.: Improving PSP Education by Pairing. An Empirical Study, in: 34th International Conference on Software Engineering 2012, 2012 (ICSE '12), pp. 1245–1254.
- Salleh, Norsaremah; Mendes, Emilia; Grundy, John: Investigating The Effects of Personality Traits on Pair Programming in a Higher Education Setting through a Family of Experiments, in: Empirical Software Engineering 19 (3), 2014, pp. 714–752.

- Satratzemi, Maya; Xinogalos, Stelios; Tsompanoudi, Despina et al.: Examining Student Performance and Attitudes on Distributed Pair Programming, in: Scientific Programming, 2018, pp. 1–8.
- Sennett, Joshua; Sherriff, Mark: Compatibility of Partnered Students in Computer Science Education, in: Proceedings of the 41st ACM Technical Symposium on Computer Science Education, New York 2010 (SIGCSE '10), pp. 244–248.
- Tsompanoudi, Despina; Satratzemi, Maya; Xinogalos, Stelios: Exploring the Effects of Collaboration Scripts Embedded in a Distributed Pair Programming System, in: Downie, J. Stephen (ed.): Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, New York 2013, pp. 225–230.
- Tsompanoudi, Despina; Satratzemi, Maya; Xinogalos, Stelios: Distributed Pair Programming Using Collaboration Scripts. An Educational System and Initial Results, in: Informatics in Education 14 (2), 2015, pp. 291–314.
- Tsompanoudi, Despina; Satratzemi, Maya; Xinogalos, Stelios: Evaluating the Effects of Scripted Distributed Pair Programming on Student Performance and Participation, in: IEEE Transactions on Education 59 (1), 2016, pp. 24–31.
- Tsompanoudi, Despina; Satratzemi, Maya; Xinogalos, Stelios et al.: An Empirical Study on Factors related to Distributed Pair Programming, in: International Journal of Engineering Pedagogy 9 (2), 2019, pp. 61–81.
- Urai, Tomoyuki; Umezawa, Takeshi; Osawa, Noritaka: Enhancements to Support Functions of Distributed Pair Programming Based on Action Analysis, in: Proceedings of the ACM Conference on Innovation and Technology in Computer Science Education 2015, New York 2015 (ITiCSE '15), pp. 177–182.
- Xinogalos, Stelios; Malliarakis, Christos; Tsompanoudi, Despina et al.: Microworlds, Games and Collaboration. Three Effective Approaches to Support Novices in Learning Programming, in: Proceedings of the 7th Balkan Conference on Informatics Conference, New York 2015 (BCI '15), pp. 1–8.
- Xinogalos, Stelios; Satratzemi, Maya; Chatzigeorgiou, Alexander et al.: Student Perceptions on the Benefits and Shortcomings of Distributed Pair Programming Assignments, in: IEEE Global Engineering Education Conference 2017, 2017 (EDUCON '17), pp. 1512–1520.
- Xinogalos, Stelios; Satratzemi, Maya; Chatzigeorgiou, Alexander et al.: Factors Affecting Students' Performance in Distributed Pair Programming, in: Journal of Educational Computing Research 57 (2), 2019, pp. 513–544.
- Yang, Ya-Fei; Lee, Chien-I; Chang, Chih-Kai: Learning Motivation and Retention Effects of Pair Programming in Data Structures Courses, in: Education for Information 32 (3), 2016, pp. 249–267.
- Ying, Kimberly Michelle; Pezzullo, Lydia G.; Ahmed, Mohona et al.: In Their Own Words: Gender Differences in Student Perceptions of Pair Programming, in: Proceedings of the 50th ACM Technical Symposium on Computer Science Education, New York 2019 (SIGCSE '19), pp. 1053–1059.
- Zacharis, Nick Z.: Measuring the Effects of Virtual Pair Programming in an Introductory Programming Java Course, in: IEEE Transactions on Education 54 (1), 2011, pp. 168–170.

- Zarb, Mark; Hughes, Janet; Richards, John: Evaluating Industry-Inspired Pair Programming Communication Guidelines with Undergraduate Students, in: Proceedings of the 45th ACM Technical Symposium on Computer Sciences Education, 2014 (SIGCSE'14), pp. 361–366.
- Zarb, Mark; Hughes, Janet; Richards, John: Further Evaluations of Industry-Inspired Pair Programming Communication Guidelines with Undergraduate Students, in: Proceedings of the 46th ACM Technical Symposium on Computer Science Education, New York 2015 (SIGCSE '15), pp. 314–319.
- Zhang, Yue; Yang, YeBo; Yang, Yang et al.: Research on Pair Learning Method and Pattern Based on Pair Programming, in: Yang, D. (ed.): Proceedings of the 2017 3rd Conference on Education and Teaching in Colleges and Universities (CETCU 2017), 2017, pp. 58–61.