Building an Open, Large-Scale Research Data Repository of Initial Programming Student Behaviour

Michael Kölling
University of Kent
Canterbury
UK
mik@kent.ac.uk

Ian Utting
University of Kent
Canterbury
UK
I.A.Utting@kent.ac.uk

SUMMARY

Many initiatives in improving initial learning of programming are based on gut instinct, guesswork, or localised experiences. Gathering real data as a basis for interventions and development work is rare, and doing so on a large scale is hard.

The BlueJ environment is currently being instrumented to allow users to opt in to a large scale research data acquisition project, which is intended to collect data useful to a wide variety of educational programming researchers.

The BlueJ project and its development team are ideally placed to collect such data: The development team has working contacts to various educational research groups, and the annual users of BlueJ number in the millions, with users in almost all parts of the world. This scale of deployment of BlueJ, and hence the volume and diversity of the data which can be gathered, are unique in the history of such investigations, and present a significant opportunity for researchers.

It is intended to provide open access to this data to any interested research project. We expect that the availability of a large scale, real world data set documenting the behaviour of learners of programming will enable a wide variety of investigations that were previously impractical.

This session aims at presenting the ideas so far, and getting input from interested researchers about the design of the data collection details.

Categories and Subject Descriptors

K.3.2 [**Computing Milieux**]: Computer and Information Science Education – *Computer science education.*

General Terms

Measurement, Experimentation.

Keywords

CS1, student behaviour, initial programming, BlueJ, data collection.

SIGCSE'12, February 29–March 3, 2012, Raleigh, North Carolina, USA.
ACM 978-1-4503-1098-7/12/02.

1. OVERALL OBJECTIVES

The main objectives of the session are two-fold:

- To inform educational programming researchers of this project and the resulting opportunities for large scale, data-driven programing education research; and
- To gather input for the design of the data gathering exercise, including the kind of data to be collected, possible research questions, and practical considerations.

The session will introduce the facilities which have been developed to gather data about beginners' behaviour in learning to program using Java and BlueJ. The theoretical opportunities of the availability of a large scale data set of student programming behaviour are significant, and have the potential to inform many different kinds of further research and development. However, details of the type of data being collected place constraints on the research questions which can be addressed.

While a framework for the data gathering is in the prototyping stage, details in many respects have not been fixed. Input from researchers potentially interested in such a data set is sought here, to ensure that as many users as possible can participate in the collection process, and that data is gathered that is useful to address a wide variety of diverse research questions.

This session gives researchers the chance to influence the type of data being collected and the mechanisms which facilitate the collection. Apart from the data design, important practical, political and ethical questions have to be addressed to make this project feasible. These include:

- How to deal with dissemination and installation in home and lab settings, including dealing with firewalls and locked-down systems.
- How to ensure anonymisation of data.
- How to enable longitudinal studies with anonymous data.
- How to gather informed consent, including how to disseminate all relevant information.
- How to enable local studies to add additional contextual (possibly personal) data to the centralised anonymous data.
- How to manage access to the data repository.

Input is sought in this session on all of these areas. We will also look at how contextual data can be gathered locally about subsets of the students using the system to complement the central anonymous data, and how these data sets can be combined to enrich the possible research.

2. OUTLINE OF THE SESSION

The session will start with a presentation of work undertaken to date, a description of the collection framework prototype, and a list of issues identified to far. The majority of the time will be dedicated to a structured discussion of the areas where input from the community is sought.

- Presentation of the data gathering framework; technical details and practical considerations
- Discussion of technical aspects: type of data gathered, technical issues
- Discussion of potential research questions that can be addressed using the data set
- Discussion of political and ethical considerations, and solutions to address them
- Description and discussion of a framework for additional contextual data gathering
- Discussion of managing access to the research data
- Discussion of the process from here: How to get and stay involved

3. EXPECTATIONS

The intended audience is Computer Science Education researchers with an interest in investigating beginning students' interactions with the environment in which they are learning to program.

We expect that the availability of a large set of research data can inform many different research questions, improve the quality of existing research and enable the investigation of new research questions. Making this data openly available to any interested research group creates the potential of wide ranging impact. As such, we expect that this project, if successful, can be of significant benefit to the CS Education community as a whole.

The success, however, depends on the extent to which practical and contextual issues are addressed in the project. It is expected that this session contributes fundamentally to ensuring that the research design is able to gather the right data and address the right questions for many independent research groups.

We expect that attendees will leave with an understanding of how the data-gathering framework could be useful in their research, at both world-wide and classroom-scale. We also expect that attendees will form the nucleus of a developing community around this data.

4. SUITABILITY FOR A SPECIAL SESSION

This session describes a technical work-in-progress which is intended to serve the CS Education research community. To ensure maximum usefulness, the project now depends in direct input and involvement of education researchers in our community.

The intention of this session is to form a community around this work based on the attendees' research questions. As such, it needs time for initial presentation of the framework, but mainly scope for discussion of what the emerging community needs. This time for discussion is not available in a traditional paper presentation, and the possibility of forming a community is not possible with a poster presentation.

The special session format is the only format that allows the type and scope of presentation and discussion required for these goals.