Exploring Visualization and Tagging to Manage Big Datasets for DBR: A Modest Proposal With Significant Implications

Kelly J. Barber-Lester, University of North Carolina at Chapel Hill, kelba@live.unc.edu
Sharon Derry, University of North Carolina at Chapel Hill, derry@unc.edu
Lana Minshew, University of North Carolina at Chapel Hill, minshew@live.unc.edu
Janice Anderson, University of North Carolina at Chapel Hill, anderjl@email.unc.edu

Abstract: As design-based research (DBR) approaches increase in popularity, there is greater and greater need to invent database management systems that document the designs and phases of DBR projects as meaningful contexts for archival storage of large databanks that such projects create. Such database systems must also support transparent and efficient search and selection of data to enable rapid analysis by design teams as well as collaborative data sharing and analysis by multiple researchers. This paper outlines one system devised by a team of design-based researchers for documenting their research process and organizing their data. Inspired by the hierarchical workflow model suggested by Hackbarth et al. (2010), this approach employs a combination of visual modeling and data tagging to organize and support rapid access and sharing of a large DBR databank.

Keywords: design-based research, data management, workflow

Introduction

Design-based research (DBR), a currently popular research approach in The Learning Sciences, involves the systematic, iterative design of learning environments in authentic contexts with aims to develop both educational innovations and advance theory (Barab & Squire, 2004; Collins, Joseph & Bielaczyc, 2004; The Design-Based Research Collective, 2003). DBR researchers tend to approach their work with a priori research questions, but also collect data broadly, leaving the door open to unanticipated discoveries (e.g. Derry et al, 2010). Despite its promise and appeal, DBR has been the subject of warranted criticisms. Famously, DBR has been criticized for poorly leveraging the vast amounts of data typically collected, resulting in relatively minor findings (e.g., Dede, 2004). One aspect of this problem follows from the requirement that research from each phase of design should thoughtfully inform the next design phase; thus significant, in-depth data analysis should take place between design iterations. Yet an open question is whether and how DBR projects have achieved the capacity to perform rapid processing of vast amounts of data in sufficient depth to inform continuing design phases within reasonable project timeframes.

This paper shares ways of working seldom discussed in The Learning Sciences while inviting deeper examination of important data management issues that follow from using mixed methodologies and broadly collecting data in authentic contexts during iterative DBR. Little scholarship exists that directly addresses issues of how to adequately organize and conceptualize both research and innovation design and management of DBR data. One exception is work reported by Hackbarth, Derry, Eagan & Gressick (2010), who described how workflow concepts could be utilized to communicate the structure of learning-environment designs and organize the data from DBR projects in terms of those designs. A symposium at the 2013 International Conference of the Learning Sciences (Derry et al., 2013), called for development of a cyberinfrastructure for DBR. This panel focused on solutions involving database standards and development of sophisticated data management tools that might be adopted throughout the community.

In this contribution we share one project's approach to data organization and management for DBR that does not require the existence or development of sophisticated visually-based data management tools (although it may be viewed as a logical step in that direction) and thus represents an approach accessible to all researchers, including those who might not have sophisticated database resources or knowledge. Developed in the context of our curriculum design work in middle school biology to help us clarify and share designs and manage large datasets across multiple iterations of design and implementation, our approach aims to make accessible our data, make apparent and transparent our research and intervention designs, and open the door to the possibility of collaborative data and design sharing as well as big data analytics approaches in the future. In pursuing this work, we attempted to consider not only our own needs, but also raise questions about what forms of standardization and generality would be needed to make this approach viable for researchers in various domains and contexts and with a variety of research structures, methods and questions. For example, the approach should be as suitable for researchers designing museum exhibits as it is for those working on middle school curriculum.

In this paper we share the basics of a data management system designed to meet the following criteria:

1) Visually capture key features of the learning-environment design and design changes that occur over multiple iterations of a project.
2) Visually represent the actual time-stamped workflow of activities that occur during and across multiple design iterations.
3) Organize all data collected (video, assessment, observations, team notes, etc.) and artifacts created (curriculum materials, student work, etc.) so that they are "located" within and thus can be understood in the context of the project's workflows and designs.
4) Allow flexible viewing of the project at multiple levels of specificity and flexible search and access to data at various time points and levels of detail.
5) Offer a structure that is both general enough for use across various projects and domains but also provides sufficient standardization to enable data sharing and analysis across domains.
6) Allow for storing and association of analyses (theses, dissertations, intermediate project work) within the context of the data employed for those analyses, in the manner of the Video Mosaic project (Video Mosaic Collaborative, 2010).

Design

An ideal data management system for DBR might include an adaptable, standardized (by the DBR community) interactive, graphical interface connected to a standardized (by the DBR community) relational database, optimized for usability and able to seamlessly facilitate the organization of and rapid access to fully contextualized and annotated data banks with associated analyses. This is the need and solution argued or implied in some previous work (Derry et al., 2013; Hackbarth et al., 2010). Yet our current approach illustrates a much easier solution and shows how any DBR researcher could build and manage a complex dataset working with readily available programs (e.g., Inspiration), meta-data capabilities inherent in various file types widely in use, and query functionality of most computers' operating systems. Our approach has two aspects: Visual Project Representation and associated Data Tagging System. We illustrate the approach in the context of our own work.

Levels of visual representation

We designed a visual representation system that captured the overall structure of the project as well as the smaller, sub-structures that, on the most detailed level, captured the design of the innovation itself. Our project is a large, multi-site, multi-year endeavor. Researchers at the University of Wisconsin at Madison and the University of North Carolina at Chapel Hill are working together to iteratively develop three different curricular units for middle school biology. For ease of explanation here, we restrict our example to the case of one site, although the ideas are easily extensible.

We designed four levels of organization, each with its own visual representation, as shown in Figure 1. Level 0 is our highest level, which orients us to the curricular unit within which we are operating. Level 1 is organized hierarchically below Level 0, and encompasses major Phases of research in the development of each curricular unit. Following that, Level 2 represents Stages of work within each Phase. Finally, Level 3 provides a work-flow representation of the Activities within each Stage. Following Hackbarth et al. (2010), diagrams adapt standard workflow elements to represent inputs, outputs, beginning and ending activity sequences, activities, technologies, participants and participant structures. Further levels could be added to locate analyses as they develop. The number and names of components within this model, as well as the number of levels in the hierarchy, could be adapted to the needs of other projects. The model serves two purposes. The first is to provide a representation to orient researchers and contextualize data within the overall and detailed structure of the project. The second purpose is to provide an organization for the data tagging system, to be discussed next.

Data tagging

Challenges of DBR research include the vast volume of data collected over time and the complexity of that data. There are many different forms and file types (video, audio, field notes, assessments, student artifacts, etc.), which have meaning and purpose only in the context of complex interrelationships they have with one another, with research questions, and with design features and changes made during a project. For example, our research program produces many hours of video data collected by several cameras running simultaneously within one classroom, recording three separate groups of students engaged in collaborative activity. In addition we have lesson plans, curriculum materials of various file types, technology logs, student work artifacts, photos, assessments, teacher and researcher reflective notes and audio recordings from analysis sessions.

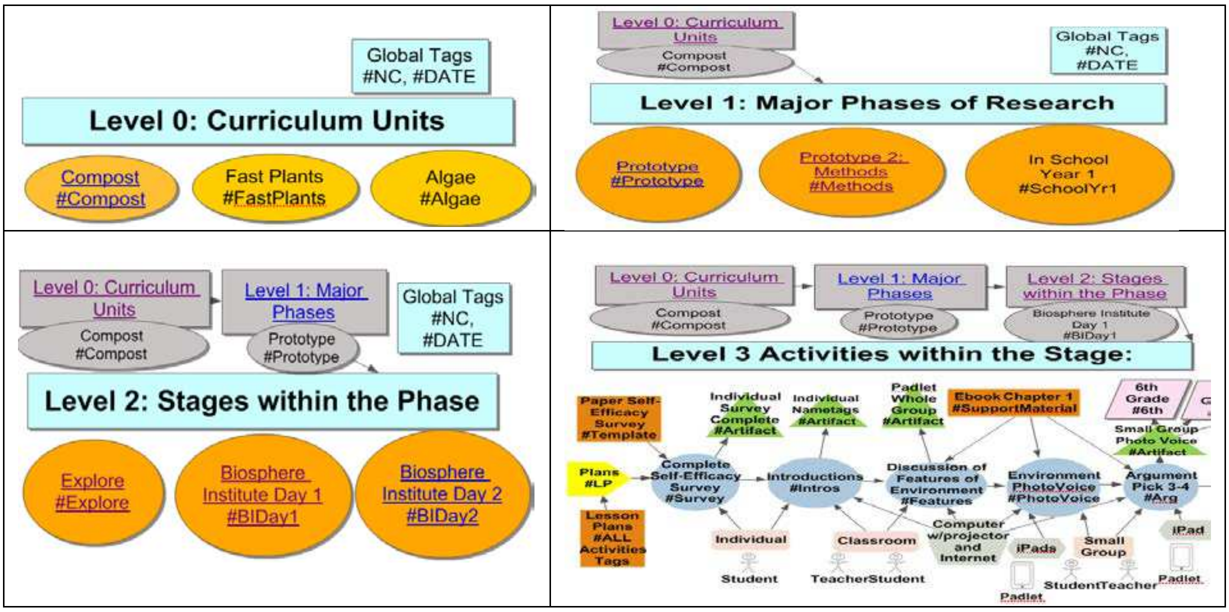

Figure 1. Workflow-Based Visual Representation Sample.

These data forms are typical for DBR projects and our conversations with other researchers indicate that a typical approach to data organization involves the use of various electronic and physical file folder systems, with organizational strategies that are unique to each project and vary widely in structure, quality, and fidelity of implementation. Such systems are difficult to consistently implement even across researchers within a single research team. When using the database to address research questions, projects must crisscross the distributed landscape of their database to locate the specific data required, which must be oriented within both the environmental context (e.g., the classroom and activity) and the context of the project (e.g., the phases of research) in order to be able to appropriately analyze it. Without appropriate organization such search is time-consuming if possible at all. To address these challenges our team devised a system that utilizes the tagging functionality in our files' metadata, based on the visual representation system, as described below:

Tag design

The logic of a tagging system is to provide unique identifiers that can be used alone or in combination in a query to locate data that have certain characteristics. The pound sign (#) in combination with a descriptor (i.e. #Compost), commonly known as a hashtag, has been popularized on social media as a way to index postings on sites like Twitter, Instagram and Facebook so that postings that are thematically related to one another can be located by those interested. Those that post can include one or more hashtags with each post to contribute to any number of online conversations. Inspired by social media hashtagging and in search of a way to label our files with unique identifiers that could be used to sort and filter them, we created all of our file tags in the fashion of hashtags, including first the pound sign (#) then a descriptor that is chosen and used following some basic guidelines that we have developed. Creating tags that begin with the pound sign (i.e. #Prototype), rather than simple key word (i.e. prototype), helps to ensure that tags are unique and will not likely occur naturally within any of our files and erroneously show up in a query. The use of the pound sign to distinguish tags was a convenient extension of familiar social media systems, but any other unique symbol could also be used.

Tag assignment and organization

As discussed above, our visual system is organized around a system of levels and a workflow inspired representation of activities (Figure 1). Tagging is designed to work within this visual system of organization. Each file first receives global tags, which include a location tag, for us #NC, and a date tag (i.e. #5-13-15). From there tags are assigned at each level. Level 0 tag will indicate which curriculum unit the piece of data is associated with (i.e. #Compost). Then a Level 1 tag is applied, which indicates within which phase of research a certain file falls (i.e. #Prototype). Then it is assigned a Level 2 tag, which indicates within which stage that file falls (i.e. #BIDay1). Then the file is tagged based on the appropriate Level 3 tags which typically include activity tags, type-of-data tags and sometimes a group tag. For example, Level 3 tags for a Lesson Plan file for Biosphere Institute Day 1 would include #LP and all of the activity tags for that day (i.e. #Survey, #Intros, #Features, #PhotoVoice, #Arg), because the lesson plan will include information pertinent to all activities of that day. On the other hand a video clip of the 6^{th} grade group's argument task from that same day would include all of the same Global, Level 0, Level 1, and Level 2 tags, but would only include #Arg (activity), #Vid (type of data) and #6^{th} (group indicator) tags for the Level 3 tags. When we query our database with that combination of tags we are able to retrieve a specific video clip. By querying our database with a tag or combination of tags, we can locate specific data, a type of data, or data associated with a specific phase of research or activity. Working together, visual representation and tagging provide a powerful means for data organization and access.

Implications and next steps

Currently our visual representation system in combination with our tagging system allows us quick, flexible access to our data, increasing the speed and efficiency with which we can scan, select and analyze data between design iterations. It provides us with tools for organizing, sharing and communicating about our work amongst ourselves, and with capabilities for sharing both designs and data with other researchers outside our project. By sharing ways of working with data, we hope to inform other researchers about a feasible solution to an important problem that all DBR researchers must grapple with, encourage further conversation about ways of working with the type of "big data" that DBR creates, and contribute to conversations among DBR researchers about data sharing and research collaboration far beyond the scope of what is currently possible. In answering criticism about the large amounts of unused data typically collected in the course of DBR, such systems have the potential to make transparently contextualized data available to multiple researchers, thereby multiplying its usability and the payoff for such resource heavy work.

Moving forward we plan to work to increase the utility and applicability of our design. Working with experts on database creation and management, we aim to explore potential software solutions to make our visual design and data retrieval process more efficient and seamless. We will also consult with other DBR researchers to gain and share insights about other database management solutions and to better understand the possibilities and limitations of our system as currently conceived. Such research may allow us to develop guidelines and a level of standardization for data tagging within the DBR community, further maximizing the collaborative potential and scholarly impact of this innovation.

References

Barab, S.A. & Squire, K. (2004) Design-based research: Putting a stake in the ground. *Journal of the Learning Sciences*, 13, 1-14.

Hackbarth, A. J., Derry, S. J., Eagan, B. R., & Gressick, J.. (June-July, 2010) *Adapting workflow technology to design-based research: Development of a method for organizing the "messiness" of research in technology-rich online learning environments*. Paper presented at the 9th International Conference of the Learning Sciences, Chicago, IL.

Collins, A., Joseph, D., Bielaczyc, K. (2004) Design research: Theoretical and methodological issues. *The Journal of the Learning Sciences*, 13(1), 15-42.

Dede, C. (2004). If design-based research is the answer, what is the question? A commentary on Collins, Joseph, & Bielaczyc; diSessa & Cobb; & Fishman, Marx, Blumenthal, Krajcik, & Soloway. In JLS Special issue on design-based research, *Journal of the Learning Sciences*, 12(1), 105-114.

Derry, S. J., Pea, R. D., Engle, R. A., Erickson, F., Goldman, R., Hall, R.,Koselmann, T., Lemke, J., Sherin, M. & Sherin, B. (2010). Conducting video research in the learning sciences: Guidance on selection, analysis, technology, and ethics. *The Journal of the Learning Sciences, 19*(1), 3–53.

Derry, S.J. (Chair), Hackbarth, A., Gonzalez, C., Sandoval, W., Bielaczyc, K., Leherer, R. (2013), *Cyberinfrastructure for design-based research: Toward a community of practice for learning scientists.* Symposium conducted at meeting of the International Conference of the Learning Sciences, Madison, WI.

The Design-Based Research Collective (2003). Design-based research: An emerging paradigm for educational inquiry. Educational Researcher, 32(1), 5-8.

Video Mosaic Collaborative (2010). Retrieved from http://www.videomosiac.org

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. DRL1418044. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.