SciDataCon 2016

Conference paper

User Guided Design: Building Confidence in Engineering Data Publication

Authors: Sandra Sweat, Aditi Ranganath, Maria Esteva, Maša Prodanović

Summary

Advances in imaging technology have generated large volumetric datasets in the field of petroleum engineering. To address the need to share this data, a multidisciplinary team developed the Digital Rocks Portal.[1] As part of the portal's iterative design, we conducted a user experience study to understand how and when to provide help to our user groups comprising of engineers, mathematicians, and geologists. The study's talk-aloud protocol focused on users conducting key tasks: uploading and organizing data, entering metadata, and publishing data. The study revealed that users were comfortable conducting data organization and image rendering tasks that relate to the way in which they conduct research. However, preparing their data for publication generated uncertainty. Based on these results we created a standard operating procedure that translates to the portal's graphic user interface as progressive help, aimed to build the confidence of the researcher when curating and preparing data for publication. This paper describes the protocol for conducting the study and analyzing the results, as well as the methods to improve the researcher's experience and enhance the quality of the data publications.

Introduction

As data production increases, repositories have shifted responsibility for data preparation and management onto the data producer (Daniels 2012). A complication of this shift has been the perception that publication increases the burden of effort, resulting in a resistance to data sharing (Volk, Lucero, & Barnas 2014). To assist researchers in the management and publication of data, the Digital Rocks Portal (DRP) provides a platform for organizing, sharing, analyzing, and preserving imaging data in the field of digital rock physics.[2] To map the way in which imaging experiments and simulations are conducted in the field, the portal's data model was designed after interviewing scientists and visiting labs. The model was translated to the user interface to represent a typical imaging research project. For each study conducted on a rock sample, users can upload and describe two sets of data: origin data and analysis data. Origin data are the images produced by computerized tomography, and analysis data consists of image data or other results of the experiments and simulations based on the origin data. Since users may upload origin data as raw files, for the images to render correctly, users have to input parameters that are processed in the background. In addition, users have to describe the physical samples and the experiments, can request DOIs and relate their data to a journal paper or other publication.

Once the first iteration of the DRP interface was finalized, we needed to evaluate if the users understood the data representation, and identify if and what type of help they would need to manage data in the portal. For this, we designed a user experience study focused on learning about the frustrations and successes users encountered when performing tasks in the repository. Based on the results we would modify the portal.

User Experience Study

To understand how researchers use data-sharing infrastructure, developers must engage directly with the user (Betz & Hall 2015). We combined interviews and data management tasks in the form of a user-experience test to identify when users experienced frustration and satisfaction in the course of interacting with the portal (Redström 2008). The main method used to obtain feedback was the 'talk aloud' protocol and observations of nonverbal communication (Wells 2006). We avoided typical usability indicators such as completion times and error rates and collected qualitative data about the experience and emotions of a defined audience (Bardzell 2011).

We recruited seven petroleum engineers practicing in the field of digital rock physics who had their own data to upload. User experience studies render comprehensive results in small test groups, with basic findings detected within the first few subjects (Nielsen 1990). At the beginning of the sessions we obtained background information via a survey and asked the participants to read all the help information available in the 'How To" section of the portal. Next, the participants completed the following tasks: organizing their data into sample, originating and analysis data; entering descriptive and technical metadata; and, publishing a dataset. The latter entailed requesting a DOI. A post-session interview was conducted to understand overall satisfaction or dissatisfaction.

Data collection included recording participant's verbal commentary as they talked aloud while creating a project, and recording observations of participant's behavior during tasks. The comments and behaviors, both negative and positive, were transcribed in a spreadsheet in relation to the task at hand. Data analysis started by compiling and dividing comments and behaviors by issue types. The categorized data was then used to identify when and where users experienced difficulty and could use additional help.

Findings and Applications

Survey responses from the participants captured previous experience storing and publishing data: two of the researchers did not share data at all, and five shared their data informally, via cloud services or by email request and only one had used an open data repository. Their data organization methods varied and included grouping files by date, by topic, or by experiment.

Across participants the main problems recorded converged in the final data publishing steps. While all the researchers understood the data organizational structure, and did not have problems entering experimental metadata for purposes of rendering the raw images, most had difficulty filling descriptive fields such as abstract and keywords. Users reported a generalized frustration related to the scant guidance in the data publication workflow. They perceived a lack of transparency and rigidity in the process of obtaining a DOI and were anxious about the final data presentation. Users indicated that before becoming public they

wanted to preview the data and receive validation from a curator. Additionally, we observed that users did not go to the 'How To' section for guidance, reporting that they would prefer dynamic and integrated help throughout the data management process.

It has been documented that in the fields of science and engineering, data sharing is still not well established, and is often perceived as a career risk (Kim, 2015; Frey, 2014; Antonijeviâc, 2015). In our small group of participants, the majority of their data sharing experiences did not involve publishing data to the world. Interestingly, while users are the experts on this type of data, they have difficulties filling descriptive information and selecting keywords, and need validation from a curator to expose their data to the world. We concluded that a help strategy for DRP had to be designed to build progressive confidence in the researchers that the data that they put out is in condition to be exposed and shared to the world.

To design and implement the help, we created a standard operating procedure (SOP), which highlights the steps involved in publishing the data as a workflow that starts when users first upload data to DRP. The metadata fields, structured to be self-explanatory, include form fields to be either mandatory or optional, text entry or drop down menu, along with the carefully chosen field names to make the process intuitive to the user. We also developed data validation steps in which there are indicators when there is a missing data field or an unsupported image file type. For instance, if a zip file – which is currently not accepted as a part of the image data set – is uploaded, the user will be presented with an opportunity to fix the issues. In addition, help is provided progressively at the points in which participants had difficulties: descriptive metadata creation, project presentation, and DOI request. Help aimed to assure that the data presentation is complete and understandable is provided in the form of questions that guide users and points to specialized vocabularies and exemplars of data publication[3]. To address long tedious form filling procedures for requesting DOIs, we are auto-populating various fields using the stored metadata.

The feedback obtained indicated that users would like to have an option to share only certain data and store other for future publication within a given study, or the desire to add new datasets to a publication. This has been accounted for by assigning ARKs to all the data components within a research project when they are uploaded; hence, providing the flexibility to reuse, share and transition selective ARKs into public DOIs as the research project evolves over time. Though initially the DRP started with a curator handling a major role in the DOI request process, we are marching towards the role of a curator on the loop, called under request by users. In short, the implementation aims at using the progressive help feature to drive the process of curating data, requesting identifiers and publishing work through semi-automated tools and guidance.

Conclusions

The user experience study has played a major role in helping the DRP team understand the users' requirements, observe the problems faced by the target audience, and accordingly design and develop solutions. As data sharing becomes more commonplace and researchers are in charge of publishing their data, curators can focus on building portals that map the researchers' methods and create tools to aid them in tasks that they are not familiar or comfortable. In this case, a major roadblock was data publication, which is reflected in the literature as a data sharing concern amongst scientists and engineers. We designed features and help for data publication that will be integrated progressively in the curation process.

They are directed to build the confidence of the user, thereby promoting them to publish their work more often. Through real-time data validation, direct feedback to the user, and extensible guidance, we aim to maintain high quality standards of publications through the portal, which again enhances the users' experience. Future iterations of the DRP development cycle will continue to employ user experience testing, and an evaluation of the new data publication pipeline is in the plans after development is near completion.

Acknowledgements

This work was funded by grants from the NSF EarthCube 1541008 and NSF EAR Career 1255622.

Competing Interests

The authors declare that they have no competing interests.

Notes

1 Digital Rocks Portal, http://dx.doi.org/10.17612/P7CC7K

2 Digital rock physics uses images to understand rock properties (Andrä 2013).

3 One of the exemplary projects, the Bentheimer Sandstone project, has complete descriptive metadata that meets DRP's guidelines. Bentheimer Sandstone project, doi:10.17612/P77P49

References

Andrä, Heiko, et al. 2013 Digital rock physics benchmarks—Part I: Imaging and segmentation. *Computers & Geosciences,* 50: 25-32.

Antonijeviâc, S. 2015 *Amongst digital humanists: An ethnographic study of digital knowledge production.* US: Palgrave Macmillan Ltd.

Betz, S. & Hall, R. 2015 Self-Archiving with Ease in an Institutional Repository: Microinteractions and the User Experience. *Information Technology & Libraries*, 34(3): 43-58.

Daniels, M., Faniel, I., Fear, K. & Yakel, E. 2012 Managing fixity and fluidity in data repositories. In: iConference, Toronto, Canada, on 7-10 February, pp. 279-286. DOI: http://dx.doi.org/10.1145/2132176.2132212

Frey, J. & Bird, C. 2014 Scientific and technical data sharing: a trading perspective. *Journal of Computer-Aided Molecular Design,* 28(10): 989-996.

Kim, Y. & Stanton, J. 2015 Institutional and Individual Factors Affecting Scientists' Data-Sharing Behaviors: A Multilevel Analysis. *Journal of the Association for Information Science and Technology*, 67(4): 776-779.

Nielsen, J. & Molich, R. 1990 Heuristic evaluation of user interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Seattle, Washington on 1-5 April, pp. 249-256. DOI: http://dx.doi.org/10.1145/97243.97281

Redström, J. 2008 RE: Definitions of use. *Design Studies*, 29(4):410–423.

Volk, C., Lucero, Y. & Barnas, K. 2014 Why is Data Sharing in Collaborative Natural Resource Efforts so Hard and What can We Do to Improve it? *Environmental Management*, 53(5): 883–893. DOI: http://dx.doi.org/10.1007/s00267-014-0258-2