ERuDIte: The Educational Resource Discovery Index for Data Science Learning

Jose-Luis Ambite, University of Southern California Information Sciences Institute; *Kristina Lerman, University of Southern California Information Sciences Institute; Lily Fierro, University of Southern California Information Sciences Institute; Jonathan Gordon, University of Southern California Information Sciences Institute; Florian Geigl, University of Southern California Information Sciences Institute; Knowledge Technologies Institute, Graz University of Technology; Gully A.P.C. Burns, University of Southern California Information Sciences Institute; John Darrell Van Horn, University of Southern California Mark and Mary Stevens Neuroimaging and Informatics Institute; Jeana Kamdar, University of Southern California Mark and Mary Stevens Neuroimaging and Informatics Institute; Sumiko Abe, University of Southern California Mark and Mary Stevens Neuroimaging and Informatics Institute; Avnish Bhattrai, University of Southern California Mark and Mary Stevens Neuroimaging and Informatics Institute; Xiaoxiao Lei, University of Southern California Mark and Mary Stevens Neuroimaging and Informatics Institute; Crystal Stewart, University of Southern California Mark and Mary Stevens Neuroimaging and Informatics Institute*

Data Science is a rapidly evolving field that draws techniques from many disciplines. There is a large number of resources available online for learning Data Science. However, these materials are highly heterogeneous, ranging from Massive Open Online Courses (MOOCs), to videos of tutorials and research talks presented at conferences, to books, blog posts, and standalone webpages. Consequently, any general search for “data science” will yield results ranging in difficulty, format, and topic, making the field intimidating to enter and difficult to navigate. In order to facilitate learning in Data Science, we are developing ERuDIte, the educational resource discovery index that powers the BD2K Training Coordinating Center Web Portal, which seeks to address the multiple issues involved in gathering and organizing heterogeneous learning resources. The development of ERuDIte has focused on three key areas: collection, integration, and organization of training resources. In the collection stage, we have manually identified sources of high-quality content, and we have built an automated web-scraping system to extract rich data from these sources. In the integration stage, we have designed a unified schema to integrate heterogeneous resource data into a single model that serves as a source for faceted search and may be expressed as linked data to facilitate information sharing. In the organization stage, we use methods from machine learning, information retrieval, and natural language processing to tag resources with concepts from a hierarchical, multi-dimensional taxonomy designed for the Data Science field. In summary, ERuDIte not only serves as a resource collector and aggregator, but also as a system powered by Data Science to intelligently organize resources and eventually provide a dynamic and personalized curriculum for biomedical researchers interested in learning about Data Science.