Session Summary: The RDAP12 Panel on Training Data Management Practitioners
by Xiao Hu

The Training Data Management Practitioners panel was truly instructional. Kirk Borne from the School of Physics, Astronomy & Computational Sciences at George Mason University started the session by telling a compelling story on how data inspired students’ interests and how the data science (informatics) curriculum was developed in GMU during the last decades. Borne emphasized that science is data-driven and that “data smartness” should become a standard in all education programs. This theme is in accordance with the information literacy and digital literacy themes that have been advocated for years: data literacy has become a necessity for every citizen in the 21st century.

Peter Fox from Rensselaer Polytechnic Institute (RPI) presented the curriculum development on data science at RPI’s Tetherless World Constellation, which explores the research and engineering principles that underlie the web. Based on its strong research themes and centers that are data-intensive, RPI established a center for data science research and education. At RPI, the curriculum related to data science is fairly complete, which seemed to excite the audience. Topic wise, the curriculum ranges from web science and information technology (IT) to discipline-oriented informatics (short named “xInformatics”) as well as data science. The curriculum offers bachelor’s, master’s and multidisciplinary Ph.D programs, all emphasizing methodology and ways of thinking over specific technology or skills. RPI educators advocate that data science should be a natural skill set for future scientists, similar to using instruments or writing code. Teamwork is also integrated into the curriculum as there are rarely individual projects in the workforce. One issue Fox raised caused a stir in the audience: theoretical foundations of data, which needs to be taught to prepare students further down their career path, not just during the first days. But, do we really have a theory on digital data?
From a different perspective than that of hard sciences, Jian Qin from the School of Information Studies at Syracuse University reported experience on developing curriculum to train librarians specializing in data management. The e-Science Librarianship Curriculum project (eSLib) is a collaboration between the iSchool in Syracuse and Cornell University library. The curriculum emphasizes the combination of eScience project planning, IT competency, librarianship and soft skills on collaboration and communication. Qin also brought good news to many RDAP attendees: the iSchool in Syracuse is offering a certificate of advanced study (CAS) program in data science in a completely online format. Interested readers can find more information at http://eslib.ischool.syr.edu.

After the panel’s presentations, the audience was very actively engaged in discussions. The first question was on sharing educational materials, and the answer was highly positive. Nearly all materials are online (except for lecture recordings). Educators highly appreciate feedback from readers who find the materials useful or otherwise. Opinions were also raised regarding training a larger workforce: training should not be limited to schools; data-driven projects should allocate part of the grants on training. As mentioned above, a lot of discussions were around whether there were theories in data science. Interestingly, two students from Syracuse gave seemingly contradictory opinions: one shared that her experience in the eSLib program was very hands-on, whereas the other told the audience that he had learned a lot of theories and theoretical thinking. On the same topic, Karen Wickett from the University of Illinois shared extremely positive feedback from students taking her theoretical course on information modeling. Bill Anderson, the panel moderator, also commented that conceptual modeling isn’t easy; it needs teamwork and close connections to applications in the real world.

This panel broadened my horizon on training in data science and data management. In the emerging field of RDAP the curricula presented in this panel are definitely pioneering work. We look forward to more efforts of this kind in helping train a strong workforce to meet the increasing demand for data management practitioners.