Geophysical Research Abstracts
Vol. 17, EGU2015-2436, 2015
EGU General Assembly 2015

Results and Activities in RDA and their Potential for Efficient Data Processing

Peter Wittenburg (1), Herman Stehouwer (2), and Rob Pennington (3)

(1) Computer Center, Max Planck Society, Germany, (2) Computer Center, Max Planck Society, Germany, (3) NCSA - National Center for Supercomputing Applications, USA

A large cross-disciplinary survey on data practices in Europe with about 120 interviews and intensive meetings revealed that working with data is extremely inefficient and costly. In addition our methods in the departments are such that only a small percentage of papers resulting from data intensive research is reproducible. A workshop organized by Research Data Alliance and MPS with leading scientists from various disciplines came to similar conclusions.
This is the reason why the global and cross-disciplinary Research Data Alliance started working on aspects of data management, description/ annotation,access, re-use, interoperability etc. Already after 18 months the first working groups came up with their results that have the potential to help overcoming the current situation. Agreeing on a common basic data model would help in many data operations, re-using best practices for practical policies would improve reproducibility, making use of a common API for registering and resolving persistent identifiers would make usage of persistent identifier services much simpler and thus increase trust in data and making use of data type registries would help us to deal with unknown data types which is a usual phenomenon in data science. In addition, after 2 years of intensive discussions new groups have been formed such as Data Fabric which is analyzing data life cycle phases and the scientific data processing machinery to identify essential common components and services and place the activities of current working and interest groups into this language. Many data scientists are involved in these discussions which gives us hope in quick convergence. Based on a few projects that started to uptake these results and the broad interest in the new activities we can see the huge potential of them. In a presentation we will describe these first results and their potential impact for data intensive science and we will describe the objectives behind discussions of Data Fabric and the impact it already has on cross-disciplinary brainstorming.