Geophysical Research Abstracts
Vol. 16, EGU2014-10027, 2014
EGU General Assembly 2014

Data Science: History repeated? – The heritage of the Free and Open Source GIS community

Peter Löwe (1) and Markus Neteler (2)

(1) Technische Informationsbibliothek TIB, Development, Hannover, Germany (peter.loewe@tib.uni-hannover.de), (2) GIS and Remote Sensing Unit, CRI-DBEM, Fondazione Edmund Mach S. Michele all'Adige, Italy

Data Science is described as the process of knowledge extraction from large data sets by means of scientific methods. The discipline draws heavily from techniques and theories from many fields, which are jointly used to furthermore develop information retrieval on structured or unstructured very large datasets. While the term Data Science was already coined in 1960, the current perception of this field places is still in the first section of the hype cycle according to Gartner, being well en route from the technology trigger stage to the peak of inflated expectations.

In our view the future development of Data Science could benefit from the analysis of experiences from related evolutionary processes. One predecessor is the area of Geographic Information Systems (GIS). The intrinsic scope of GIS is the integration and storage of spatial information from often heterogeneous sources, data analysis, sharing of reconstructed or aggregated results in visual form or via data transfer. GIS is successfully applied to process and analyse spatially referenced content in a wide and still expanding range of science areas, spanning from human and social sciences like archeology, politics and architecture to environmental and geoscientific applications, even including planetology.

This paper presents proven patterns for innovation and organisation derived from the evolution of GIS, which can be ported to Data Science. Within the GIS landscape, three strategic interacting tiers can be denoted: i) Standardisation, ii) applications based on closed-source software, without the option of access to and analysis of the implemented algorithms, and iii) Free and Open Source Software (FOSS) based on freely accessible program code enabling analysis, education and ,improvement by everyone. This paper focuses on patterns gained from the synthesis of three decades of FOSS development. We identified best-practices which evolved from long term FOSS projects, describe the role of community-driven global umbrella organisations such as OSGeo, as well as the standardization of innovative services. The main driver is the acknowledgement of a meritocratic attitude.

These patterns follow evolutionary processes of establishing and maintaining a web-based democratic culture spawning new kinds of communication and projects. This culture transcends the established compartmentation and stratification of science by creating mutual benefits for the participants, irrespective of their respective research interest and standing. Adopting these best practices will enable the emerging Data Science communities to avoid pitfalls and to accelerate the progress to stages of productivity.