RENGA: an open-source platform fostering cooperation in data science

Rok Roškar*, Eric Bouillet*, Luc Henry† and Olivier Verscheure* (olivier.verscheure@epfl.ch)

* Swiss Data Science Center
EPFL
Station 14
CH-1015 Lausanne
Switzerland

† Presidency
EPFL
Station 14
CH-1015 Lausanne
Switzerland

ABSTRACT

The **Swiss Data Science Center** (SDSC) is a joint venture between EPFL and ETH Zurich. Its mission is to accelerate the adoption of data science and machine learning techniques within academic disciplines of the ETH Domain, the Swiss academic community at large, and the industrial sector. In particular, the SDSC addresses the gap between those who generate data, those who develop data analytics and systems, and those who could potentially extract value from it.

As part of its mandate, the SDSC is developing **RENGA** (http://get-renga.io), a scalable open-source software platform designed to foster cooperation (or coopetition), reproducibility and reuse in data science. The name was borrowed from renga (連句), a traditional form of collaborative poetry. In the same spirit, RENGA's philosophy is to mobilize and connect skills from the data science community, enhance their collective intelligence, and allow for the discovery of knowledge without premeditation.

The platform materializes data lineage automatically therefore seamlessly capturing the workflow both within and across projects, allowing any derived data to be unambiguously traced back to the original raw data sources in a manner that is fully transparent. Because code, data and analysis steps are all recorded in a version control system, all research artifacts are fully reproducible. Interoperability is ensured by using open source tools for workflow descriptions and portable, well-documented metadata schemas.

From the users' perspective, RENGA is a digital canvas where data science projects are conducted and improved, evolving asynchronously into a web of data science threads, where the output of one analytics becomes the input of another. RENGA is governed by a loosely coupled federated model that allows autonomous organizations to share resources without renouncing their authority and ownership. Indeed, they are in full control of their respective platform services and content, and manage access rights according to their own preferences. In this form, RENGA acts as an online ecosystem of data, meta-data, analytics, storage and computing resources. It implements services to securely investigate data science problems and allows data scientists to evaluate, compare, and share each other's methods and results. It can be deployed as a non-circumventable and tamper-proof system, and can thus be used for governance, intellectual property attribution, and audit purposes (e.g. answer the question "who used my data, and for what purpose").

The knowledge representation is fully searchable and represents a one-stop shop to a vast collection of digital content that data scientists can explore to find good quality data they can trust and confidently reuse in order to eliminate redundant efforts and accidental data duplication. We will discuss how the FAIR principles are fully embedded in RENGA in order to meet the requirements for truly searchable and reproducible data science. With all its characteristics, the RENGA platform has the potential to be a stepping-stone to achieve the vision of open science.