

## European Open Science Cloud

**A recent recommendation that a large number of professional data stewards be trained and employed in all data-rich research projects raises the exciting prospect they will conduct research on data-intensive research itself. It also focuses us on questions about the role of all scientists in data quality and accessibility as well as how best to measure the value of good data stewardship to science and society.**

In April, the European Commission launched the European Open Science Cloud (EOSC; [http://ec.europa.eu/newsroom/dae/document.cfm?doc\\_id=15266](http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=15266)) as its contribution to the Digital Single Market concept, and in June the EOSC High Level Expert Group published a roadmap (<http://ec.europa.eu/research/openscience/>) dealing with implementations to achieve the vision of a European contribution to a global data research commons. This report estimates that half a million data stewards with training and specialized data expertise will be needed to support the work of 1.7 million European Union (EU) scientists and 70 million other workers in innovation. The League of European Research Universities is already in a good position to help train these new data experts via its graduate summer schools (<http://www.leru.org/index.php/public/news/leru-doctoral-summer-school-on-data-stewardship/>) and other curriculum development initiatives.

Professional data stewards working in collaboration with researchers will have an important role in training research groups to standardize and optimize their data sets for research, interoperability, persistent accessibility and reuse. A steward, much like a statistician, can see the properties of data sets that are not limited to one particular research field and thereby reapply best practices from field to field to ensure interoperability of the data sets across disciplines. Whether stewardship becomes a permanent professional specialty or a permanent part of every researcher's job seems to be a very important question, much as academic informatics professionals currently struggle with balancing their service and research roles. Technology again and again moves from specialist centers of intensive instrumentation to distribution of the technology to various users where they work. From the journal's point of view, we see the value of specialized stewardship training to promote better research, but ultimately we would prefer all researchers

to be 5% data stewards rather than have 5% of research salaries go to specialist stewards, as it is only within the research context that data practices can be optimized for better research. If the point of stewardship is to remove barriers to data access by implementing technology and conventions that solve problems, once these problems are solved, it would be good to have the stewards work on research problems to ensure the profession evolves and we gain a new layer of meta-research over data science as a whole.

We hold that data are gathered under specific hypotheses for particular purposes and may have no intrinsic value. However, data reuse, although incidental and secondary to data collection and first analysis, may sometimes be extensive and far more valuable than the original data generators intended. To conserve resources, including the work of data curators and stewards, the differences in utility among data sets therefore need to be monitored. The EOSC proposes that new performance indicators be developed. We propose that data sets should not only be processed for their findability, accessibility, interoperability and reusability (according to the FAIR guidelines; *Sci. Data* **15**, 160018, 2016) but also assessed for a set of metrics related to their usefulness. These can be measured by the range of questions asked of the data set, answers or understanding gained, the data set's influence on research and translation within the field for which it was generated and across disciplines, the number of times the data set is accessed and the economics, not only of storage versus access, but also of relative gains and savings associated with the reuse of FAIR data. The solutions to implementing these metrics are in the technology employed to achieve FAIR data, as linked data conventions and persistent accessible metadata will allow counters to be built in. However, equally important are the social conventions of increased collaboration to agree on standards and greater transparency in research methods that data stewardship engenders. ■