Efficient long-term open-access data archiving in mining industries

Saulius Gražulis[a,b], Andrius Merkys[a,b], Antanas Vaitkus[a], Cédric Duée[c], Nicolas Maubec[c], Valérie Laperche[c], Laure Capar[c], Anne Bourguignon[c], Xavier Bourrat [c], Yassine El Mendili [d], Daniel Chateigner [d], Stéphanie Gascoin [d], Gino Mariotto [e], Marco Giarola [e], Arun Kumar [e], Nicola Daldosso [e], Marco Zanatta[e], Adolfo Speghini [f], Andrea Sanson [g], Luca Lutterotti [h], Evgeny Borovin [h], Mauro Bortolotti [h], Maria Secchi [h], Maurizio Montagna [i], Beate Orberger[j,k], Monique Le Guen[j], Anne Salaün[j], Céline Rodriguez[j], Fabien Trotet[j], Mohamed Kadar[j], Karen Devaux[j], Thanh Bui[j], Henry Pillière[l], Thomas Lefèvre[l], Fons Eijkelkamp[m], Harm Nolte[m], Peter Koert[m]

[a] Vilnius University Institute of Biotechnology, Saulėtekio al. 7, LT-10257 Vilnius, Lithuania

[b] Vilnius University Faculty of Mathematics and Informatics, Naugarduko st. 24, LT-03225 Vilnius, Lithuania

[c] BRGM, 3 avenue Claude Guillemin, BP 36009, 45060 Orléans Cedex, France

[d] Normandie Université, CRISMAT-ENSICAEN, CNRS 6508, Université de Caen Normandie, 14050 Caen, France

[e] University of Verona, Department of Computer Science, 37134 Verona, Italy

[f] University of Verona, Department of Biotechnology, 37134 Verona, Italy

[g] University of Padua, Department of Physics, 35131 Padova, Italy

[h] University of Trento, Industrial Engineering Department, 38123 Trento, Italy

[i] University of Trento, Physics Department, 38123 Trento, Italy

[j] ERAMET-RESEARCH-SLN, 1 Avenue Albert Einstein, 78190 Trappes, France

[k] GEOPS, Université Paris Sud, Université Paris Saclay, Bât 504, 91405 Orsay, Cedex, France

[l] Thermo Fisher Scientific, 71 rue d'Orléans, 45410 Artenay, France

[m] Royal Eijkelkamp, Uitmaat 8, 6987 ER Giesbeek, The Netherlands

Key words: Crystallography Open Database, CIF framework, XRD, XRF, Raman spectroscopy, IR spectroscopy, sample characterization, reusable data

Efficient data collection, analysis and preservation are needed to accomplish adequate business decision making. Long-lasting and sustainable business operations, such as mining, add extra requirements to this process: data must be reliably preserved over periods that are longer than that of a typical software life-cycle. These concerns are of special importance for the combined on-line-on-mine-real-time expert system SOLSA (http://www.solsa-mining.eu/) that will produce data not only for immediate industrial utilization, but also for the possible scientific reuse. We thus applied the experience of scientific data publishing to provide efficient, reliable, long term archival data storage.

Crystallography, a field covering one of the methods used in the SOLSA expert system, has long traditions of archiving and disseminating crystallographic data. To that end, the Crystallographic Interchange Framework (CIF, [1]) was developed and is maintained by the International Union of Crystallography (IUCr). This framework provides rich means for describing crystal structures and crystallographic experiments in an unambiguous, human- and machine-readable way, in a standard that is independent of the underlying data storage technology. The Crystallography Open Database (COD, [2]) has been successfully using the CIF framework to maintain its open-access crystallographic data collection for over a decade [3,4]. Since the CIF framework is extensible it is possible to use it for other branches of knowledge. The SOLSA system will generate data using different methods of material identification: XRF, XRD, Raman, IR and DRIFT spectroscopy. For XRD, the CIF is usable out-of-the-box, since we can rely on extensive data definition dictionaries (ontologies) developed by the IUCr and the crystallographic community. For spectroscopic techniques such dictionaries, to our best knowledge, do not exist; thus, the SOLSA team is developing CIF dictionaries for spectroscopic techniques to be used in the SOLSA expert system. All dictionaries will be published under liberal license and communities are encourage to join the development, reuse and extend the dictionaries where necessary. These dictionaries will enable access to open data generated by SOLSA by all interested parties. The use of the

common CIF framework will ensure smooth data exchange among SOLSA partners and seamless data publication from the SOLSA project.

References

1. Hall, S. R. et al. (1991) The crystallographic information file (CIF): a new standard archive file for crystallography, *Acta Crystallogr. A* 47, 655-685.

2. The Crystallography Open Database Web site (2017), http://www.crystallography.net/ [accessed 2017-02-27].

3. Gražulis, S. et al. (2009). *Crystallography Open Database – an open-access collection of crystal structures*, J. Appl. Crystallogr. 42, 726-729.

4. Gražulis, S. et al. (2012). *Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration*, Nucleic Acids Res. 40, D420-D427.