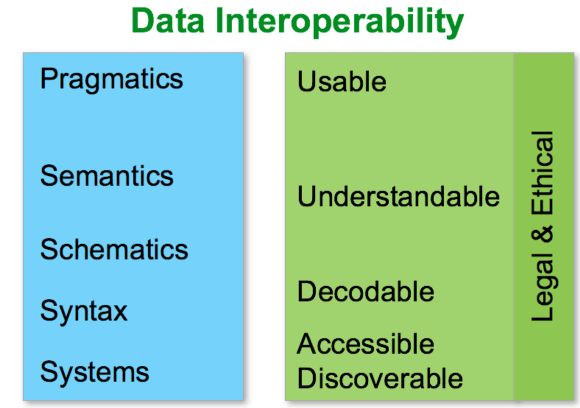
Leveraging Data Science for Geoscience: Experience from the Deep Time Data Infrastructure.

Xiaogang Ma[1], Peter A. Fox[2], Sophie Kolankowski[2], Daniel Hummer[3], Robert M. Hazen[4], Joshua J. Golen[5] and Michael B. Meyer[4], [1] Department of Computer Science, University of Idaho, 875 Perimeter Drive, MS 1010, Moscow, ID 83844-1010, Email: max@uidaho.edu, [2]Tetherless World Constellation, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180, Email: pfox@cs.rpi.edu (P.A.F.), kolans@rpi.edu (S.K.), [3]Department of Geology, Southern Illinois University Carbondale, 1263 Lincoln Drive, Carbondale, IL 62901, Email: daniel.hummer@siu.edu, [4]Geophysical Laboratory, Carnegie Institution for Science, 5251 Broad Branch Road, NW, Washington, DC 20015, Email: rhazen@ciw.edu (R.M.H.), mmeyer@carnegiescience.edu (M.B.M.), [5]Department of Geosciences, University of Arizona, 1040 E. 4th Street, Tucson, AZ 85721, Email: jgolden@email.arizona.edu.

Our ability to model, analyze and understand the Earth's changing environment is hampered by the shortage of interoperability (Figure 1) within and between disciplinary datasets as well as the lack of data synthesis from complementary disciplines. To address this issue, efforts from both data science and domain-specific sciences are needed. A few scientific unions have set up groups on data and information, such as the the International Union for Geological Sciences; the International Union of Geodesy and Geophysical; the International Astronomical Union; and the International Union of Crystallography. However, those disciplinary developments are on an ad hoc basis and there is little coordination between them. International associations such as the Committee on Data for Science and Technology (CODATA) and the Research Data Alliance (RDA) have been taking leads to communicate and coordinate the data standards amongst scientific disciplines.

Such needs of data interoperability and synthesis are also reflected in the ongoing Deep Time Data Infrastructure project (http://dtdi.carnegiescience.edu). The ultimate goal of the project is to study the complex co-evolution between geosphere and biosphere. The areas of interest include mineralogy and petrology, paleobiology and paleontology, paleotectonics and paleomagnetism, geochemistry and geochrononology, genomics and proteomics, and more. Solid progress has been achieved in scientific discovery (e.g., Figure 2) by using datasets across disciplines. Our experience shows that data management and curation deserves the same attentation as data analytics in the implementation of data science for cross-disciplinary sciences.

References: [1] Bishr, Y. (1998) *IJGS, 12*(4), 299-314. [2] Brodaric, B. (2007) *TGIS, 11*(3), 453-477. [3] Ma, X. (2011) *Nat. Geosci., 4*(12), 814. [4] Ma, X. (2014) *Nat. Clim. Chang., 4*(6), 409-413.



Bishr, 1998 [1]; Brodaric, 2007 [2] Ma et al., 2011 [1]; 2014 [2]

Figure 1. Levels of data interoperability from the perspectives of data producers (left) and users (right).

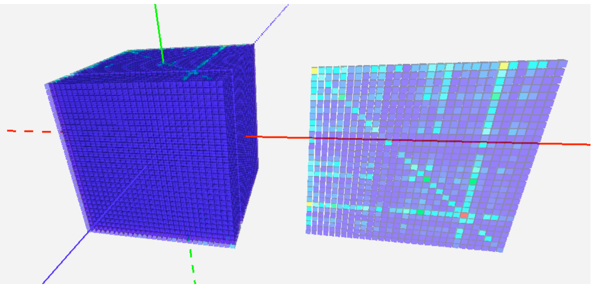


Figure 2. Using a three dimensional cube matrix to study the co-existence of elements in minerals. The same list of 30 key mineral-formaling elements is plotted along each axis. Each matrix element (cube) was first filled with the raw number of minerals in which elements X, Y, and Z coexist, and then rendered with a color according to the value of the number. Here a two dimensional plane for Oxygen is taken out to show details, i.e. O is element on Z axis.