Editor's Note: SCIENTIFIC "AGENDA" OF DATA SCIENCE

Shuichi Iwata

Co-Chief Editor
Email: iwata@k.u-tokyo.ac.jp
Former President: CODATA http://www.codata.org
EAJ Member, SCJ Associate Member
Professor: Graduate School of Frontier Sciences, The University of Tokyo
Tel/Fax: +81-3-5841-6985(Hongo Office)
Tel/Fax: +81-4-7136-4604(Kashiwa Office)
Address: Room E222, 5-1-5 Kashiwanoha, Kashiwa City, Chiba, JAPAN 277-8563

For over 40 years, CODATA has been the leading international organization concerned with improving the quality, accessibility, and usability of scientific data. The Information Revolution has provided unprecedented opportunity to ensure that scientific data are fully integrated in the fundamental workings and decision making of our society. Further, these data care critical to improving every aspect of society. In this essay, I describe how data plays these roles and outline an opportunity for the CODATA Data Science Journal to catalyze creation of a Data Commons to provide even greater impact.

Harmonizing Data for a Better Society

Data are used to harmonize different opinions into an action. For example, data on climate changes are linking people to work together for solutions to this global issue. To further the scientific agenda of CODATA, we have added an aspect called "data and society" explicitly through opportunities, such as the World Summit on the Information Society (WSIS), and lessons from tragedies, such natural disasters and epidemics. Not only through global warming problems pointed out clearly by the Intergovernmental Panel on Climate Change (IPCC), but also through other global issues presented partly by the eight UN Millennium Development Goals, we have become aware of the necessity of linking scientific, technical, economic, social, and political agendas through reliable data with important missions for society. Here human-centered reorganization of domain-differentiated sciences - from natural sciences to social sciences - are requested to be done, when and where we need common reliable data to make correct decisions with in a consensus of society.

If we do not share common reliable data on global warming effects, we cannot establish effective remedies against these "Inconvenient Truths." We cannot establish flexible and steady roadmaps for a sustainable society. It is necessary for us experts to propose proper solutions based on proper data by linking associated scientific domains with politics, laws, ethics, economics, environmental sciences, ecology, civil engineering, manufacturing, waste management, and so on. Complementing missing links by reliable data, we may come to create a new scientific domain - "data-driven sustainability science" - to design and manage society properly. So as to find a solution in conflicts, it is essential to start everything from facts and common reliable data. It is our *raison d'être.* It is our identity as "data scientists."

Data are used to integrate parts into a product, link facts into knowledge, and bridge disciplines into new science. Data on nuclear interactions have integrated with data on metals, ceramics, waters and other materials to enable building nuclear power plants. Gathering data about *Arabidopsis thaliana* leads to progress in botany. RNA data link DNA and proteins, which are consequently associated with disease and health. Data on the atomic elements and intrinsic properties of substances are used to link extrinsic properties of substances and structure-sensitive engineering properties, such as for defects.

The fundamental constants have been compiled reflecting advances of precise measurements and basic sciences, as well as higher coherences of scientific models, which in turn have resulted in creating new sciences such as nano-technology, spintronics, and other specialized scientific domains. These have led to several key standards for the information society, namely, radio wave standards, current, voltage and so on. As fundamental constants have been defined in an elaborate network of complicated experiments and models in science and also link different scientific and technical fields coherently, those data have played a central role for many fields.

Data are used to describe the details and diversity of facts and compress information into well-organized knowledge, thereby also creating new knowledge and confidence in our understanding of nature. Spectral and diffraction data have guided us to get micro-structural information on inorganic substances and materials as well as living things. Data obtained through maintenance service in the airports are crafted into safety standards taking into account our understanding on aging and deterioration and creating new service business. Data on tradescantia (spiderwort) and nude mice under controlled irradiation are used to estimate the biological effects of radiation on health, translated available data on casualties in terms of genomics, proteomics, biophysics, and biochemistry.

From Data to Value

It is a challenge to go beyond a collection of data into a full extraction of meaning from raw data - overcoming limitations of epidemiological surveys based on statistics with improvements to health science. Data-oriented statistical approaches are combined with scientific models and practical monitoring, and traditional established safety/risk/reliability standards are changing into proactive and dynamic adaptive standards. Safety/risk issues in medical services, nuclear reactors, aircrafts, company managements, energy resource security and so on can be dealt with in a similar way.

Openness and transparency of many disciplines and scientific domains promoted by e-science projects and so-called global information commons are prerequisites for a new revolution of sciences *by* 7 billion people. Devices for the revolution might take creative forms of emerging wisdom and growing emotions in the internet era, which will be more than such knowledge management approaches as ontology, metadata, object-oriented approach, semantic-web, commonsense reasoning, and so on.

Data are creating value everywhere. Through evaluation of fundamental constants, we are integrating quantum worlds, atomistic worlds and macroscopic worlds quantitatively, where scientific disciplines and domains are networked with a certain consistency. Through

combination of geometric data for parts with property data, we can design an artifact and then assemble the available parts into a product of integrity and cost-effective performance, where domain differentiated engineering disciplines are integrated to establish manufacturing industries. Design and maintenance of landscapes, cities, countries, regional environments and global climates can be carried out in a similar way, namely, sharing data by stakeholders and coordinating different views and opinions by means of the shared common data.

Data quality matters every time and everywhere: gold in, gold out. Reasonable estimations of data uncertainties produce better results and outcomes. The more the problem to be solved is uncertain, the more we should become flexible. Evidence-based deterministic approaches do not work effectively, and adaptive and heuristic approaches work better coupled with *in situ* data capture, evaluation, and quick decision and timely actions. Holistic creativity as a group is a key for success of the group, when practical maintenance of data quality for proper decision is important. The time constants of data life cycle are becoming shorter, and the diversity of stakeholders and complexities of data are increasing. New disciplines are continuously created by taking advantage of available data and devices so as to prepare solutions in a timely fashion. Without proper management of continuously-produced important data and without the productivity of new disciplines based on data, we cannot solve important problems of the world.

Catalyzing the Data Commons

For the scientific agenda of CODATA, we have added or reconfirmed a focus on "data and society." Data and knowledge corresponding missing links of domain-specific sciences have been daily work of CODATA. Data services between "haves" and "have-not" countries are to be carried out timely and appropriately in a self-explanatory manner. Such data sharing is also the key technology to solve global issues such as sustainability, which can be realized by everyone's commitment. I do hope Data Science Journal can be used as "commons" for open publication, a depository of data and knowledge, a saloon for data scientists and experts in other fields, an interface between data producers and users, a window to the real world, and a bed for creating new disciplines including data science. Data science will play a recurring important role here with increasing returns.