Making data typing efforts or automatically detecting data types for automatic data processing?

Research data consumer on the cloud where very large collections of distributed scientific data are provided

Two major processes in the scientific data use
Data discovery
Data processing

Data processing paradigm shift: from manual to automatic

Data consumer scenario

Automatic data processing

1. Fetch the list of PIDs
2. Query/response for PID KI profile
3. Query/response for data type profile
4. Fetch the data

Manual data processing

1. Fetch and crawl data
2. Manually process the data
3. Query/response for data type profile
4. Fetch the data
5. Automatically process the data

For the future, we aim at an automatic way

Data provider scenario

Step 1

- Data providers build their data in a community standard
  - The data is packed in a commonly used format, i.e. XML, JSON, netCDF, CSV as well as application dependent such as Microsoft EXCEL format.
  - Some data are shipped with a document describing data meaning, data types, and data format.

Step 2

- Data providers use more complicated data format to assert data types
  - A set of Handle server of DOI objects with Kernel Information profile and Data Type Registry is a recommended candidate for a variety of domain community to assert their data types in addition to their data sources in a community standard format.
  - On the other hand, linked data community uses RDF/XML, JSON-LD and other linked data formats, or a kind of mixture format of data type and value.
  - Common vocabularies are provided in a public server, e.g. schema.org.

Well-organized data cloud structure

Kernel information connects data and data types

If the data is not typed, the data consumer still have to process the data in a manual way or use a data type detection tool although current detection technology is immature.

This poster is presented for

11th Research Data Alliance Plenary Meeting
Berlin, Germany
21-23 March, 2018

Acknowledgements: The authors are thankful for RDA KI WG discussions.