Open Science, Business Analytics and FAIR Digital Objects

George O. Strawn

Introduction

- Claim: data can be made considerably more useful with the right research and use case development

- Application pull exists: the Internet of Things is just around the corner, deep learning is dependent on evermore data, and calls for a more Open Science are growing

- More technology push is needed: accelerate the emergence of the new world of interoperable data

Data Wrangling

- We must automate activities that currently require much manual intervention

Fifty Years of Data Science by David Donoho

- 1) *Data Gathering, Preparation, and Exploration*; 2) Data Representation and Transformation; 3) Computing with Data; 4) Data Modeling; 5) Data Visualization and Presentation; and 6) Science about Data Science
- Data preparation currently consumes about 80% of the data user’s time

Open Science

- The printing press enabled the Royal Society’s call in the 17th century science for scientists to publish their results
- Networked computers could today enable the publishing of *all* science products: articles, data, software, workflows, etc
- The US National Academy of Sciences consensus report of July, 2018, titled *Open Science by Design,* included the recommendation that *all research products be made available according to the FAIR principles*
- The High Level Expert Group advising on the European Open Science Cloud has made a similar recommendation

Business Analytics

- Big Data at Google: map-reduce, big table, and other tools
- Supervised machine learning processes data that is currently available to ”train” a deep learning algorithm so that new/additional data can be categorized and/or used to direct actions
- Unsupervised learning simply dives into existing data in an exploratory mode called data mining
- In both cases, “understanding” the data, such as can be achieved by utilizing *FAIR Digital Objects*, is a requirement for utilizing machine learning

Enabling Hardware

- In 1970, one thousand transistors could be constructed on a chip, in 1990 it was one million, and in 2010 it was one billion

- Fiber optic/laser communication bandwidth has increased even more rapidly from mega-bits per second in the 1980s, to giga-bps in the 1990s, to tera-bps in the 2000s, to (experimental) peta-bps in the 2010s

- Disk prices have dropped from $500,000 per gigabyte in 1981 to $0.03 per gigabyte today (a four terabyte drive for $100)

- These great increases in performance and the equally important decreases in cost have enabled data-intensive science, machine learning, and other use case advances

Fanciful IT Eras

- The first era (from 1950 to 1995) was one of many computers and many datasets

- The second era (from 1995 to 2025?) has been one of a single computer and many datasets. Recall SUN Computer’s marketing slogan, “the network is the computer”

- The anticipated third era (2025?-) will be one of a single computer and a single dataset. That is, the desired state of the *interoperability of* all *heterogeneous data*

Big, Open, and FAIR Data

- In 2011, the US Federal interagency committee coordinating IT research established a senior steering group for *big data* research. This acknowledged that we could now store more data than we could effectively process

- In 2013, the US President's Science Advisor signed an executive order requiring all research products (articles, data, software, workflows, etc.) produced under Federal support to be *open* for public access

- In January, 2014, a workshop was held at the Leiden University Lorentz Center under the leadership of Professor Barend Mons, to consider what characteristics open data should have to be useful. The result: *FAIR* data

The GO FAIR Initiative

- The FAIR Data Paper resulting from the workshop has been widely read and cited more than 1,000 times in the last four years

- Ministries within the Dutch, German and French governments created the *Global Open (GO) FAIR* office to provide a focal point for groups interested in pursuing implementations of FAIR data

- In 2018, a call went out from the GO FAIR office for proposals for *Implementation Networks* (INs)

- In January, 2019, more than 20 INs were represented in Leiden at the first annual IN workshop

C2CAMP, Digital Objects

- With the intent of producing a system that implements a FAIR data infrastructure, the organizers of the C2CAMP project selected the Digital Object Architecture developed by Bob Kahn and his associates at CNRI

- By the 1990s, CNRI had implemented a global permanent identifier system called the Handle System, which has been adopted by the publishing industry (and others) to create the Digital Object Identifier (DOI) System, which provides a globally unique resolvable identifier for every published article

- By the 2000s, CNRI had also implemented an architecture for digital objects that could be referenced by handles.

- IMHO, the Digital Object System has the same elegance in design for data that TCP/IP has for networks

- In the European scientific world, the concept of DOs has been widely accepted and the European Commission Expert Group report contains the term "FAIR DO"

Interoperable Computing

- Creating interoperable computing elements by introducing *new levels of abstraction* has been a powerful computer science technique

- High level languages and their interpreters solve the interoperability problem for heterogeneous computers

- The Internet solves the interoperability problem for heterogeneous networks

- The Digital Object Architecture could solve the interoperability problem for heterogeneous data

The virtual Internet

- The Internet is *a virtual network interconnecting existing networks*. Two computers attached to different networks can communicate if those networks include an Internet *router* that is attached to both the networks

- Existing networks did not have to be replaced in order to connect to the Internet. “Only” a router had to be added to each network (and Internet software to each host).

- This approach greatly simplified (perhaps enabled) the original job of “connecting to the Internet.”

Abstract-er Data Types

- In the 1970s *abstract data types* were defined. The implementation of the data structure was “hidden” from the programmer (with named operations defined on it)

- Digital objects carry information hiding one step further: the operations themselves are hidden until *the digital object is queried by the programmer to find out what operations it has*

- *Additional metadata* beyond the operations will be required to capture fully the meaning of the data

Digital Object Architecture

- Two elements, *Digital Objects* (including *repositories* and *registries)* and *Handles*; and two protocols, the *Handle resolution protocol* and the *Digital Object Interface Protocol* (DOIP)—used for accessing digital objects

- The Digital Object System does not require that existing data systems be discontinued and “converted” to digital objects. That is, Digital Objects provide an virtual layer above an existing data system

FAIR Digital Objects

Digital objects have the built-in capability to implement FAIR data, but *that capability has to be put to use via the “right” metadata*

The following example illustrates at a high level how it might work:

- Find: The ***program*** searches metadata registries for data of the desired characteristics
- Access: Via DOIP
- Interoperate: utilize the relevant data elements from each selected data instance *via metadata*
- Reuse: Process and combine the retrieved data elements according the the requirements of the job

Web Objects

Compare this process with a typical web search:

- Find: The ***human*** enters keywords that relate to the desired subject.
- Access: the found web pages
- Interoperate: Read the selected pages to gain knowledge of the subject. *Note bene* that when the human is performing this action, the “interoperability semantics/metadata” are in the mind of the searcher.
- Reuse: Cut and paste (and process) segments of the selected web pages

The big difference between these two activities is clearly the automation involved in *enabling a program to do what a human had been doing*. This automation is at the core of interoperable data and its greatly increased efficiency and effectiveness

Conclusions

- IT is poised to do for data what the Internet did for networks
- And it could be the GO FAIR/C2CAMP project that does it
- Interoperable data will become a core infrastructure, like the Internet has already become
- We will wonder how we ever got along without it!