Open Science, Business Analytics, and FAIR Digital Objects

George O. Strawn
National Academies
Washington, DC
gostrawn@gmail.com

Abstract— **The lack of data interoperability is hindering the emergence of Open Science and making data analytics considerably more expensive than it should be. A new technology, FAIR Digital Objects, is seeking to solve this problem along with several others. FAIR is an acronym for findable, accessible, interoperable, and reusable. Digital Object Architecture is a very general virtual layer to ride of top of any data system, which can solve the interoperability problem for heterogeneous data, much as the Internet solved the interoperability problem for heterogeneous networks. A specific project employed these technology will be described.**

Keywords—digital object architecture, FAIR data

I. Introduction

This brief paper is a call to action for computer/data scientists, IT specialists, and domain specialists. I believe that we are at a point where data can be made considerably more useful with the right research and use case development. I will make my case by examining one project which seeks to do just that. With the Internet of Things just around the corner, with deep learning dependent on evermore data, and with the call for "completely" Open Science, there is already more than enough application pull. Readers of this paper are encouraged to help with the technology push that will accelerate the emergence of "the new world of interoperable data."

Data has indeed moved to the center stage as a focus of both IT research and important use case development. This paper will mention several of those use cases (open science and business analytics) and accomplished and active research (the digital object architecture and metadata). Then it will focus on a particular use of that technology (FAIR Digital Objects, defined below) which, if successful, will accelerate use case research by automating activities that currently require much manual intervention (data wrangling)

II. Data Wrangling

In a recent paper titled *Fifty Years of Data Science* (1) [the title refers to John Tukey's call for data science 50 years ago], David Donoho provides a six-step conceptual framework for the effective use of data: 1) Data Gathering, Preparation, and Exploration; 2) Data Representation and Transformation; 3) Computing with Data; Data Modeling; 4) Data Visualization and Presentation; and 6) Science about Data Science. Donoho says, as do many others, that the data preparation currently consumes about 80% of the data user's time, which inflates the cost and lowers the effectiveness of the use of data. Colloquially called data wrangling, this data preparation step needs to be automated and that is one of the goals of creating FAIR digital objects.

III. Open Science

The 17th century science revolution was energized by the Royal Society's call for scientists to publish their results (2) so that others may learn of them and build upon them. This use of the printing press, which had been invented in the 15th century, thus opened science and made it a public endeavor. Today, the printing press has been superseded by networked computers, which could enable the publishing of *all* science products: articles, data, software, workflows, etc. The US National Academy of Sciences released a consensus report in July, 2018, titled *Open Science by Design* (3), which concluded with five recommendations to accelerate the emergence of this new open science: 1) work to create a culture of open science; 2) train students and researchers in the methods and tools of open science; 3) establish procedures to identify research outputs that should be preserved and develop the infrastructure to provide long-term preservation; 4) *ensure that research products are made available according to the FAIR principles* [italics added]; and 5) all segments of the research community should work together to realize open science by design. FAIR Digital Objects provide one potential solution for recommendation four. The High Level Expert Group advising on the European Open Science Cloud has made very similar recommendation regarding open science (4).

Speculating about the future, some observers think that the new open science may be as revolutionary as the original science revolution.

IV. Business Analytics

Business has been making use of "big data" at least as long as Google has been in existence. Their creation of map-reduce, big table, and other tools has enabled them to capture the Web search market and the advertising that supports all their activities. As more businesses adopt machine learning (5) as an analytic tool, the importance of data accelerates. Two types of machine learning both depend directly of the amount and quality of the data available for processing. 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. In so far as this understanding can be automated, the cost of machine learning goes down and its use will go up.

V. Enabling IT Hardware

The progress that information technology (IT) hardware has made in the last decades has been phenomenal. This progress has enabled the rise of data and continues to open up new IT use cases for science and society. This progress has been measured by Moore's Law (6) and similar performance metrics. Regarding Moore's law per se, 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 (7). Last but certainly not least, disk prices have dropped from $500,000 per gigabyte in 1981 to $0.03 per gigabyte today, with corresponding increases in capacity (8). These great increases in performance along with equally important decreases in cost have enabled data-intensive science, machine learning, and other use case advances.

VI. IT Eras

One fanciful way to classify "IT eras" would be: 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; and the anticipated third era will be one of a single computer and a single dataset. Our current era of "a single computer" refers back to SUN Computer's marketing slogan, "the network is the computer," and to a more geeky phrase that my NSFnet friends use to say, "the network is the backplane." The anticipated third era of "one dataset" refers to the desired state of the *interoperability of heterogeneous data,* potentially realized by FAIR digital objects, which is the topic of this paper.

VII. Big, Open, and FAIR Data

As data has become an increasingly important topic in the last decade, we have advanced through the phases big data, open data, and FAIR data. In 2011, the Federal interagency committee coordinating IT research established a senior steering group for big data research (9). This acknowledged that we could now store more data than we could effectively process. The hardware advances listed above could support doing so, but comparable software advances were required. In 2013, the President's Science Advisor signed an executive order requiring that all research products (articles, data, software, workflows, etc.) produced under Federal support should be open for public access (10). Also, 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 (11). An acronym was adopted describing four of those attributes: findable, accessible, interoperable, and reusable: that is, FAIR data. Most recently, the US government created a law called the OPEN Data Act (12), which stipulates that government data be open by default (excepting certain classes of sensitive data).

It has been suggested that most data should be FAIR, including data that cannot be open. That is, FAIR is not equal to open, but associated metadata can be open and can assert how open the data themselves are. As discussed above, much data will be unitized for "analytics," and preparing data for analytics (data wrangling), currently takes about 80% of data scientists' time. Thus, the data analytics activity would be made much more efficient and cost effective if the wrangling time could be reduced to, say, 20%. Also, a new era of open science would be propelled forward by the increased ability to reuse other people's data.

VIII. The GO FAIR Initiative

The FAIR Data Paper referred to above has been widely read and cited almost 1,000 times in the last four years. Based at least in part on this interest, ministries within the Dutch, German and French governments created the Global Open (GO) FAIR office to provide a focal point for countries, disciplines, and other 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). More than 40 INs are now in some stage of development, with more than 20 of the INs represented in Leiden at the first annual IN workshop in January, 2019 (13). The purpose of the workshop(s) is to share experiences, approaches, and perhaps software as the implementations proceed. GO FAIR encourages re-use of existing resources wherever possible, and many synergies were identified including the universal need for FAIR DO's.

The IN project that this author is most familiar with is called C2CAMP (Cross-Continental Collection Access and Management [of data] Pilot), which is led by Peter Wittenburg of the Max Plank Institute, who until recently was also the head of Research Data Alliance-Europe. The next part of this paper will seek to explain the technology and goals of this project.

IX. C2CAMP

With the intent of producing a system that implements a FAIR data infrastructure, the organizers of the C2CAMP project (14) selected the Digital Object Architecture (15) (DO Architecture) developed by Bob Kahn and his associates at the Corporation for National Research Initiatives (CNRI). After Kahn left DARPA in the 1980s (where he co-created TCP/IP with his colleague Vint Cerf), he formed CNRI and began studying digital libraries, knowledge robots (knowbots), and other IT research projects. By the 1990s, they had defined and developed an 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, they had also designed an architecture for digital objects that could be referenced by handles. In this author's opinion, the Digital Object System has the same elegance in design for data that TCP/IP has for networks.

In addition to elegant design, two other reasons were influential in making the choice of the DO architecture. Many European organizations and countries lauded the establishment of the non-profit DONA Foundation (16) as the custodian of the Handle system and the DO Architecture. Also, 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." The Digital Object System will now be described in more detail below.

X. Interoperable Computing Elements and Ease of Implementation

Creating interoperable computing elements by introducing new levels of abstraction is a powerful technique that computer science has employed for a long time. Three examples of using an additional level of abstraction to solve the interoperability of heterogeneous elements are: high level languages and their interpreters to solve the interoperability problem for heterogeneous computers; the Internet, to solve the interoperability problem for heterogeneous networks; and the Digital Object Architecture, to solve the interoperability problem for heterogeneous data.

The Internet example is helpful to understand the digital object solution to interoperability. The Internet is *a virtual network interconnecting existing networks*. Two computers attached to different networks can communicate if those networks include an Internet *router* which is attached to both the networks (and there can be intermediate networks as long as there are also intermediate routers). A great virtue of this solution is that existing networks did not have to be replaced in order to connect to the Internet, "only" that a router had to be added to a network which was also connected to another network which was already part or the Internet system. This approach greatly simplified (perhaps enabled) the job of "connecting to the Internet."

The impact of evolved system admin and what has become known as DevOps is also important. Not only is the network faster and the hardware much more powerful, but it is much easier to deploy the new hardware on the new networks, virtual machines are spun up on demand, software configurations are templated, etc. The software engineering life cycle has likewise been more highly automated and code changes ripple through test cycles and hit production and spread to fleets of servers in minutes. In a way, this increased level of abstraction and automation in sys admin and software engineering presages what we want to do

with objects and data. Breaking down the division between sys admins and developers can be seen as analogous to the distinction between data wranglers and those who do science. We want to get rid of the grunt work and put the people who understand the data in charge of the data.

XI. Digital Object Architecture and Ease Of Implementation

Similarly, 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, which provides the basis for interoperability. Consider the following three-phase development of "information hiding," as it was called in Simula 67 (17). Originally, data structures were constructed out of the rudimentary structures provided by a programming language (typically arrays and records). Then in the 1970s the concept of *abstract data types* was generalized from Simula 67. In an abstract data type, the particular implementation of the data structure was "hidden" from the programmer who could call the operations defined on the data by name without knowing how they were implemented. For example, a linear list of elements (with a key field) with operations: update an element, add/delete an element, list the elements in key order, could be implemented (among other schemes) by a one-dimensional array, a balanced binary tree, or a hash table (depending on which operations were expected to dominate the execution time). If implemented by an abstract data type, the programmer need not know which implementation was chosen by the designer. Just by calling operation subroutines named update, add, delete, and list, he could initiate the particular action. Digital objects carry the information hiding one step further: *each digital object may be queried by the programmer to find out what operations it has*. This is just part of the story: the value of FAIR DATA in general, is that legacy systems can just expose their content as FAIR data and become interoperable. And of course this applies to the broader concept of 'data' as collections of DO's rather than just DO's themselves.

It's high time to say a little more about the elements and structure of the digital object system. At the highest level, there are only two elements, Digital Objects and Handles, and two protocols, the Handle resolution protocol and the Digital Object Interface Protocol (DOIP). Two special types of digital objects are called out for special use: *repositories* and *registries*. A repository is a (logical) place to store (other) Digital Objects and a registry is a place to store metadata for interpreting (other) digital objects. The DOIP is used for accessing digital objects. For example, a DO can be created by calling the DOIP with the handle of a repository where the DO is to be (logically) stored, an existing DO can be accessed by calling the DOIP with the handles of the DO and it's containing repository.

XII. FAIR Digital Objects

Digital objects have the built-in capability to implement FAIR data (18), but that capability has to be put to use via the "right" metadata. How this should be done most effectively is a subject for research and development. The following example illustrates at a high level how it might work:

Find: The *program* searches metadata registries for data of the desired type.
Access: If a data instance of the desired type is open, prepare to access it; if it's behind a paywall, decide whether to make the payment or move on.
Interoperate: Access the relevant data elements from each selected data instance via the DOIP protocol.
Reuse: Process and combine the retrieved data elements according the the requirements of the job.

Compare this process with a typical web search:

Find: The *human* enters keywords that relate to the desired subject.
Access: Prepare to access revenant web pages, paying where necessary (or not).
Interoperate: Read the selected pages to gain knowledge of the selected 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, where not restricted from doing so by copyright.

The big difference between these two activities is clearly the automation involved by 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.

XIII. Conclusions

I believe we are at the time in history where IT is poised to do for data what the Internet did for networks.

Whether the GO FAIR/C2CAMP project, another innovative project, or a hybrid of many of the projects emerges, I believe we will look back at this time the same way we now look back at the time of the emergence of the Internet and wonder how we ever got along without it!

ACKNOWLEDGMENT *(Heading 5)*

Larry Lannom, Barend Mons, Erik Schultes, and Peter Wittenburg Read a draft of this paper and made very helpful comments.

REFERENCES

1. David Donoho (2017) 50 Years of Data Science, Journal of Computational and Graphical Statistics, 26:4, 745-766, DOI: 10.1080/10618600.2017.1384734
2. https://royalsociety.org/about-us/history/
3. https://www.nap.edu/catalog/25116/open-science-by-design-realizing-a-vision-for-21st-century
4. https://ec.europa.eu/research/openscience/pdf/eosc-fair_paper_schouppe-burgelman_2018.pdf#view=fit&pagemode=none
5. https://en.m.wikipedia.org/wiki/Machine_learning
6. https://en.m.wikipedia.org/wiki/Moore's_law
7. https://en.m.wikipedia.org/wiki/Fiber-optic_communication
8. https://www.backblaze.com/blog/hard-drive-cost-per-gigabyte/
9. https://www.nitrd.gov/nitrdgroups/index.php?title=Big_Data
10. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf
11. The FAIR Guiding Principles for scientific data management and stewardship. Mark D. Wilkinson et al. Nature Scientific Data, volume 3, Article number: 160018 (2016) https://www.nature.com/articles/sdata201618
12. https://www.datacoalition.org/open-government-data-act/
13. https://docs.google.com/document/d/1_rCirTaJMEZ5D0HeaF3a-aKte3AI9NDQ634OrPUb9F0/mobilebasic
14. https://www.rd-alliance.org/sites/default/files/CENDI-15.Nov_.17-Lannom-Final-2.pdf
15. https://www.dona.net/digitalobjectarchitecture
16. https://www.dona.net/index
17. https://en.m.wikipedia.org/wiki/Simula
18. https://www.go-fair.org/fair-principles/