About Building Data Infrastructures

Peter Wittenburg (Max Planck Computing & Data Facility, Garching/Munich)
George Strawn (US National Academy of Sciences, Washington)

request for comments

1. Executive Summary

The great relevance of data for coming to new scientific insights, making progress in tackling the grand challenges and in making commercial profit has widely been commented [A1]. Some of the big questions that are currently being discussed are a) who will own the relevant data and b) what kind of facilities need to be developed to make data available. With respect to the first question there is no doubt that data from publicly funded research should in principle be open for broad usage [A2]. This implies an answer to the second question in so far as data infrastructures should avoid dependencies on commercial services and interests. This is the reason that huge investments are currently being made, in particular in Europe, towards the development of an eco-system of data infrastructures that build on public investments that have previously been made.

Building such data infrastructures[1] (DI) should respect a balance between three components: scientific interest (S), technology advancement (T), and organizational form (O); the O dimension should follow the high dynamics in the T and S dimensions. In contrast to the US, where we see a reluctance to invest in DI building, the EU and its member states invest large amounts of funds in DI building. However, we can observe that the three dimensions (S, T, O) are not well balanced, a situation which bears high risks. After having overcome the historically motivated split between "research infrastructures" and "e-Infrastructures"[2] the EC launched the European Open Science Cloud, focusing on the O dimension first. This strategy opens the door to all kinds of lobbying which has frequently been shown to lead to failure. In Germany a national program with high ambitions has been launched which focuses on the S dimension instead; this focus ignores the fact that DI building needs to be based on an interplay between IT-driven core infrastructure components and science-driven discipline-specific components.

Summarizing, we can see that a) the current DI programs have a broad mobilization effect amongst researchers, which is important to change cultures, and that b) there seems to be a deep distrust in IT-driven approaches, although commercial providers show that generic services can be successful. Decades of funding in e-Infrastructures in Europe have not led to solutions that are convincing to all stakeholders.

Based on these impressions we can make a few recommendations:

- The investments being made are so huge and the challenges so fundamental that we need to agree that the basics of the DI we are building should be designed to remain operative for centuries and not just a few decades.
- Given this time frame we need to focus our debates on generic core components that can act as a fundament, and not merely on applications and tools although they are very important to convince the users.
- We are convinced that a balanced and open analysis, free of conflicts of interest, of the investments already made in research infrastructures and eInfrastructures can provide guidance for determining such core components. Taking the establishment of the Internet as

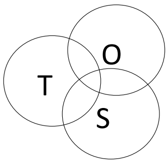
[1] Data Infrastructures point to one important aspect research infrastructures we need to address.
[2] In the US this distinction was not made and the term "cyberinfrastructure" was used covering all flavours of RDI.

an example, we need to dare to invest in IT-driven components and interfaces in parallel to inspiring an increasing group of researchers, so as to anticipate new opportunities and to driving the science-driven components to develop the eco-system we have in mind.

- We are convinced that the FAIR principles and the FAIR Digital Object concept can guide us towards building the fundament of such DI with the requested long-term perspective. These guiding factors can easily be aligned with industry approaches that focus on reference architectures built on top of open standards.

Thus, we are lobbying for two pillars of future DIs that are open goods and compliant with the "hourglass" promises (that is, a small number of core requirements logically between implementations and applications, as was TCP/IP in the Internet case, see also chapter 4.4). There are no particular or commercial interests involved, but the conviction that it is time for convergence as we argued in another paper [?].

2. Factors driving Data Infrastructure Building



Research data infrastructure building is driven by the three closely related and interdependent dimensions already mentioned above: scientific needs (S), technological innovation (T) and organizational aspects (O). When also catering to industry, the scientific dimension would be complemented by development and production dimensions that promote economic interests. In general these three dimensions need to be balanced to come to optimal solutions. The scientific and industrial needs for developing new types of data infrastructures can be described in terms of four major goals:

- being able to do competitive data intensive science by combining large amounts of data of different types from different sources;
- tackling the complex grand challenges with new analytical and simulation methods on large volumes of data that is mostly heterogeneous;
- making data intensive work much more efficient to reduce the enormous costs, to tackle more problems and to engage more scientists;
- supporting mechanisms that enable cutting-edge science, a goal that drives nations to fund the required infrastructures[3].

Technological innovation is a product of research insights and socio-economic interests [1]. Innovation therefore follows non-linear paths making it difficult to predict the next successful innovations that will create a real momentum. Obviously technological innovations have effects on science and partly require new organizational forms. In general, organizational forms are driven by the changes required by new capabilities and insights.

It is the uncertainty resulting from this non-linear path that prevents us from taking more radical actions. Everyone involved knows that we will need a radical change in data practices, but the responsibility for decision taking is shifted: scientists know that policy actions are required to overcome the fragmentation, policy experts expect, perhaps optimistically, that the scientific community will present converging solutions. Most probably one reason for shifting responsibility for decision taking is that we still lack a clear and convincing message of how to move ahead.

[3] The long-time relation between the wellness of modern societies and their scientific strength which will direct policies has been shown [2].

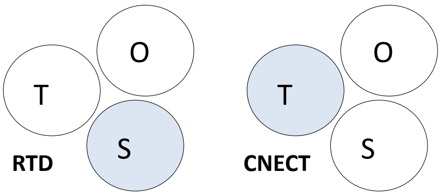
3. Analysis of current approaches

Currently, many technologically advanced countries are investing large amounts of funds to implement data infrastructures. Different approaches have been chosen which we want to compare with what was stated above.

3.1 European Approaches

Past EC Approach

The past approach by the EC was characterized by a split of two major funding streams via DG RTD supporting mainly research-driven projects (ESFRI, etc.) [3] and DG CNECT focusing on IT-based approaches (e-Infrastructures)[4]. The three related dimensions S, T, O were separated as each of the two DGs set different foci. The simplified diagram illustrates the two approaches from DG CNECT and DG RTD. Initial investments in modern infrastructures, took innovative approaches, since they basically helped in exploring the landscape and offering services from two different points of view. In the early phase of IT-based infrastructures[5], however, the focus was on compute grids where relevant knowledge could be gathered. At that time no one could foresee that finally the hard challenge would be to develop a persistent and interoperable "data infrastructure". Consequently, early discussions about organizational aspects were widely premature or not addressed seriously. The contribution of the IT-driven infrastructure work towards establishing a FAIR data infrastructure was rather limited due to different agendas and little mutual understanding between the involved stakeholders.



After the appearance of the "Riding the Wave" report [4], when awareness about the need for rethinking became apparent, the two-track approach started to become counter-productive. On the one hand, where cross-fertilization appeared necessary, attempts by e-Infrastructures to do "consumer binding" could be observed based on services to ensure continuous funding streams. On the other hand, the ESFRI projects[6] started with "silo" approaches to satisfy the needs of their specific research community without getting substantial help from the e-Infrastructures in setting up scalable solutions based on urgent needs that soon became apparent. Later, cluster projects [5] were started to bring together various communities to jointly develop cross-disciplinary methods and services, but again without the explicit request to look for "hourglass" type of solutions [6] (see also 4.4).

Although the Research Data Alliance (RDA) [7] was funded by CNECT, the engaged Europeans mainly came from the research infrastructures and research libraries. The e-Infrastructures did not yet take a major and driving role, since for many of the questions they were interested in they believed to have designed the solutions already. Again, chances for cross-fertilization were missed.

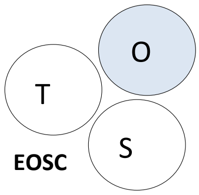
EOSC Approach

To overcome this dilemma that resulted from the two different funding streams the EC decided in ?? to launch the EU Open Science Cloud initiative [8] which is now a gigantic endeavor covering many different dimensions and formulating the clear goal to channel funding streams that are related with the development of the data infrastructure eco-system through the emerging EOSC mechanisms. Some are legacy mechanisms, e.g., services designed under the former CNECT regime such as from OpenAIRE [9] and the EOSC Portal [10], for example, are seen as defining core aspects of EOSC. This

[4] For the purposes of this note we ignore other important activities from RTD and CNECT.
[5] In Europe these were also called eInfrastructures.
[6] In Europe this led to a first wave of mobilization of research communities.

is exactly the opposite of what is required to achieve breakthroughs. A large infrastructure as aimed at by EOSC needs to be inventive in the sense of the hourglass model, since the investments in the coming years will be so huge that the components and their interfaces must be specified so as to be universally adopted and survive for a long time. Service development, in contrast, should be an area of competition for small innovative teams continuously looking for better solutions overruling the previous ones. Offering services as basic pillars of infrastructure solutions will hamper progress, in particular if competition is prevented[7].

The experts of the EC have managed to convince all member states to invest in a powerful federative research infrastructure with the intention to overcome at the end disciplinary and national boundaries. The first result was the establishment of an organizational solution, after which the next steps in the planning process were taken. This meant the separation of the organizational part from the scientific and technological discussions and developments. The intention for the coming phase is to involve the large infrastructures organized for example in the ESFRI projects and technological experts in the further planning. This would enable the desirable close interaction between all three pillars. Currently, working groups are being set up that will assist the boards in managing the required interaction in collaboration with a secretariat. The process has just started and we need to see whether existing and functioning interaction platforms will be used or whether other mechanisms will need to be invented[8].



A huge challenge to address is how to come to a vision about hourglass-type of core elements without which a large and complex federated infrastructure cannot survive. All prior investments, for example, by the ESFRI program, cannot be ignored in this conceptual process. Focusing on applications and tools in such a debate would mean wasting money. *Applications and tools are essential to demonstrate progress and finally a usefulness of the infrastructure for the users, but they cannot be the fundaments of an infrastructure.* Therefore, some of the current discussions around EOSC seem to lead into wrong directions. Some benefits can be expected from forums such as RDA and GOFAIR [11] to continue and even intensify a global interaction on specifications, but there is still a lack of agreement to use such open and bottom up driven interaction platforms.

As usual, the two boards established to guide EOSC, the governance board with representatives of the member states and the executive board which must execute this top-down process, will be the focus point for all lobbyism. The governance board will need to mediate between the differing interests of the member states, some of which are ready to invest large amounts of funds while others will be fully dependent on EC funds. This may prevent deep discussions about the needs for advanced data intensive science which finally must be the motivation for such an expensive data infrastructure. Given all these factors, it is hard to assume that discussions about open hourglass type of mechanisms will be in the primary focus.

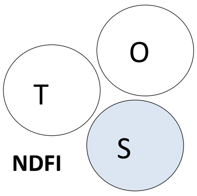
German National Research Infrastructure

Germany has launched its plans for a national research data infrastructure (NFDI) [12] implying large investments over the coming years. Other countries such as France, Switzerland, the Nordic countries, etc., will follow as they also see the need to invest in this. Some other countries have ambitions, but will depend on EC funds to get national data infrastructures started.

[7] Services after short seed funding periods need to show that they are wanted and self-financed.

[8] GEDE is such a platform where about 47 research infrastructures are collaborating without almost any larger funding support to come to agreements.

The German authorities followed a different approach as compared to EOSC. They put the research communities in charge of organizing themselves and making suggestions for further developments. A second large wave of mobilization and awareness raising within the research communities about possible new opportunities can be observed, which is crucial for success. It was not surprising that most of them suggested to work along the lines they were already doing within ESFRI and other projects to satisfy the immediate needs of their respective communities. It was also noticeable that almost all share the need to make use of PIDs, to associate metadata, to improve the FAIRness [13] of their data, to talk about standardization of their data/metadata formats, to optimize legal & ethical interoperability, etc. That is, a large part of the intended work in each of the proposed initiatives will be devoted to improved data management and stewardship which is re-occurring in different flavors across most of them. Postponing technology discussions about infrastructure commons, however, means that technological decisions and developments will be taken within the projects after having been granted. This raises the question whether there are better and more efficient ways to go ahead. It would be a feasible task to identify commonalities[9] already now and to see where common infrastructure components could be realized based on advanced IT concepts, allowing the community-specific work to be carried out with more emphasis.



We can hardly expect from this approach that hourglass type of issues common to all will be put in focus, since the immediate scientific needs will have highest priority. Cross-disciplinary work will be given second priority in the start-up phase. Following this approach, the risk will be high that even more funds will be devoted to silo approaches, making it even harder and more costly to do the transformation toward common components which without any doubt will have to emerge. In addition, much money will be spent on parallel work developing flavors of the same basic infrastructure components.

3.2 US Approaches

In 2013, the US President's science advisor issued a directive requiring all Federal agencies that provide at least $100 million in yearly support to research and development to require their awardees (and agency researchers) to develop data management plans that describe how they will make their research data results available for public access. The agencies were then required to submit to OSTP their plans for implementing this requirement. All required plans were completed by 2017 and agencies have been actively implementing their policies.

The US agencies have also been focusing on data infrastructure *research*, as well as research data *infrastructure*. NSF has supported RDA-US and several data infrastructure pilot projects, but not building a general research data infrastructure (as it did in the past when building a research networking infrastructure—NSFnet). NIH considered building a "research data commons," but that project was put on hold when the responsible chief data scientist left government service. And as stated above, no interagency research data infrastructure project has emerged. On the other hand, a Federal law was recently enacted that requires Federal agencies to make their own data (but not Federally-supported *research* data) open as the default. It would be interesting if US government administrative data infrastructure emerges before US science data infrastructure.

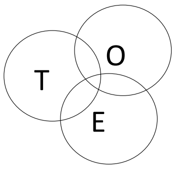
Compared to Europe, the US lacks a broadly supported conviction that concrete steps towards building a *common* national research data infrastructure should be taken.

[9] Analyses of commonalities between various research infrastructure projects have been made earlier, but are widely ignored.

3.3 Industrial Approaches

We devote a paragraph on industrial approaches in order to understand whether we can profit from cross-fertilization. Recent meetings at the RDA 11th plenary in Berlin [14], at the IoTWeek [15,16] and various personal interactions with industry experts indicated that there are basically four approaches at the moment:

- Big players such as Google, Amazon and Facebook (GAF) try to push everyone onto their platforms and systems; they dictate the rules of the game and the standards to be followed. They have an incredible amount of capital to further extend their strong almost monopolistic market position.
- Other IT players such as Oracle, HP, Telekom, IBM, SAP, etc. try to define their market position beyond the influence of the GAFs. They are making huge investments to build proprietary platforms with adapters to all kinds of tools and standards which are very expensive to build and maintain. They would be willing to rely on open standards since it will be the only way to compete with GAF, but their priorities are driven by their customer wishes and their prior investments.
- Big consortiums are being formed, in particular in production industry, that work on reference architectures (RAMI [17], IDS [18], IIC, etc.) and protected data spaces. The ideas are a) to come to well-designed specifications, including data flows and processes, based on a top-down approach and b) to build secure data spaces where data can be exchanged and reused in trusted environments.
- Smaller companies, in particular SMEs and startups, are dependent on access to data platforms offered by larger companies and thus develop their business plans by adopting specific platform solutions.



Whatever the chosen approach is, companies will need to establish a good balance between the three dimensions to guarantee business success. Only the invention of open standards will prevent a monopolized data market hampering innovation. For the research domain the most promising approach in industry is the one working on reference architectures. Their intention is to rely to open standards and long-term solutions. Trends towards encapsulated data spaces will not be successful. Methods to protect data and trace their usage need to be developed on top of open protocols.

4. Aspects of Research Infrastructures

This chapter will highlight a few aspects that are underrepresented in the current discussions about building research infrastructures. These are aspects which we believe require an integrative approach, i.e., the consequences for the three pillars (S, T, O) need to be looked at. It is widely agreed that the big challenges ahead of us can be sketched as 1) extracting knowledge from the huge amounts of FAIRified data requiring stepwise-increasing automation, 2) integrating knowledge to achieve scientific insights and conclusions and 3) building the key pillars which should continue to exist for a long period of time given the huge investments that are needed.

4.1 Complexity of RI

There is no doubt that the complexity of the task ahead of us is huge, much more complex than, for example, establishing the Internet. The latter was established by small cooperating research groups. Nowadays, data researchers are part of large, globally acting research communities defining specific cultures and what is seen as cutting edge research. They are also part of research organizations that need to show competitiveness as a whole to get continued funding streams. Research organizations are layered constructions from research groups all the way up to countries and continents that want to ensure competitiveness at global level. All these different layers have particular interests which are affecting infrastructure building and making it a complex sociological and technological enterprise. Top-down and bottom-up driven initiatives such as RDA, GOFAIR and CODATA have been set up and having their particular strengths and weaknesses. All three agreed to closely collaborate and use the strengths of each other, yet this has to work out in practice and show that they are indeed able to help overcoming the social and technological challenges and to send convergent messages. *One concrete result is the shared concept of FAIR Digital Objects (FAIR-DOs) which indicates that the FAIR principles and the DO concept indeed are complementary and could lead to the required momentum.*

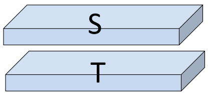
Recently, a survey was carried out by RDA GEDE[10] [19] on the expectations with respect to the usefulness of the FAIR-DO concept [20, 21] from a scientific point of view which resulted in 31 different use cases from different research infrastructures (mostly ESFRI projects). 10 of these use cases mentioned the need to manage the evolving complex landscape of digital entities as the main reason to make use of FAIR-DOs. Scientific disciplines and domains are faced with an increasingly large domain of digital entities of various types on top of which a variety of systems of relationships are being established representing the state of domain knowledge. The major concern is that this knowledge needs to be stable and preserved over centuries. In the view of the proposers only the FAIR-DO approach will guarantee that we will not end up in a dark age.

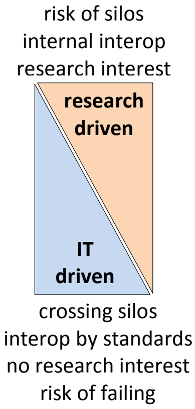
Thus, the primary challenge in building the eco-system of infrastructures is to be able to manage complexity in an area where many bright people have worked out many brilliant solutions, which we now see as causing fragmentation. Managing complexity requires close interactions between the three pillars S, T and O and decisions on specifications of core components and their interfaces.

4.2 Science versus IT Driving

Any research infrastructure building can roughly be divided in 2 principle layers of which the border line in general is fuzzy and changes over time. Defining the border line is very much a social challenge[11]:

- components that are closely related with scientific challenges will be carried out and thus driven by scientists to be successful;
- components that have a generic character should be driven by IT insights based on abstractions.





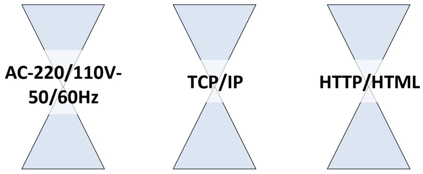
It is obvious that research data infrastructures that are built around a specific "project" are still the most successful ones in terms of scientific impact (DOBES [22], NOMAD [23], PDB [24], etc.). They are based on a shared and clearly specified goal within the project and have the possibility to define standards within the projects right from the beginning to establish interoperability. Domain-specific research infrastructures (ELIXIR [25], CLARIN [26], EPOS [27], and many others) are already steps away from the direct needs of the researchers and are confronted with integration and interoperability tasks between different silos — tasks in which researchers in general have little interest. Even further away from the direct researcher interests are generic infrastructures driven by IT insights where interoperability needs to be achieved by defining common standards, which is a time-consuming process. Researchers are, however, interested in using read-made infrastructures, as the use of professionally designed commercial products shows.

[10] Currently, a paper is being written to give a broad view on the usefulness of the FAIR-DO concept for deep scientific matters. Here we refer to a first analysis made.

[11] In early discussions on the Internet, natural scientists argued that they would need a special email system compared to the other sciences to meet their needs. Today no one would understand these debates. After 20 years of experience some sciences still believe that they need a special system for persistent identifiers.

4.4 Interfaces versus Applications

Major breakthroughs in most recent large infrastructures were promoted by agreements on very simple basic standards [6]. General electrification caught on when technological experts agreed on AC ~220V at 50 Hz or 110V at 60Hz[12]. The Internet was founded when people agreed on TCP/IP and the Web was born when people saw the potential of HTTP/HTML. Many more examples also for smaller infrastructures could be mentioned. The simple fact about the shown hourglasses is that they reduce complexity from n*n to 1*n problems, define a stable basis to stimulate huge investments which in general opens new possibilities not considered beforehand, and are pre-competitive agreements which do not hamper innovation for substantial periods of time.



New and scalable infrastructures are not primarily defined by applications and tools since these change at higher frequencies, but by identifying the right layers to define these minimal interfacing standards, thereby reducing overall complexity. We hope, and in fact, expect that innovative infrastructure enterprises will develop the required data infrastructure. Investing in user oriented tools and applications is a means to better understand the landscape in such early phases and to identify the right layers, but *they are not the solutions for the core of the infrastructures*. Quite the opposite, focusing on applications and tools will only hamper breakthroughs and will instead lead to methods of binding users to their silos.

4.5 Evolution versus Disruption

In this context it is important to note that researchers, with the exception of a few visionary ones, are fairly conservative with respect to disruptive technologies, since disruptions are accompanied by phases of decreased productivity. Therefore, most researchers were, for example, skeptical about the Internet and the Web when they were introduced. Only a few immediately saw their huge potential to optimize scientific methods. Most researchers like to take evolutionary steps allowing them to follow the methods they are used to. It took a while until new possibilities such as email, ftp, http, etc. were widely accepted in science.

The Internet and the Web were clearly guided by IT concepts and did not evolve naturally from researchers' fantasies. In both cases selected researchers in advanced scientific fields saw the need for networking and information exchange, but IT concepts were generalized from early examples and turned wishes into disruptive steps, which later had an impact on all researchers. Much effort was invested in Compute Grids and this experience led to Cloud systems, which opened the doors to efficiently store large amounts of data and to arrange compute units in a way which allows researchers to tackle even larger computational problems. However, clouds themselves cannot be seen as a mechanism to solve interoperability.

4.6 Key Message and FAIR Digital Objects

With the help of Internet, people were able to connect to remote computers, to remote people (SMTP) and to files with information (FTP, HTTP); in this way, the original vision spelled out by Licklider [28] was made true. When we address the question of the subsequent revolutionary vision we suggest to widely replace “people” by machines as actors:

Digital entities (data, services) will find each other to perform widely automatic extraction of knowledge while humans will operate in a stable domain of assertions capturing machine conclusions to come to decisions.

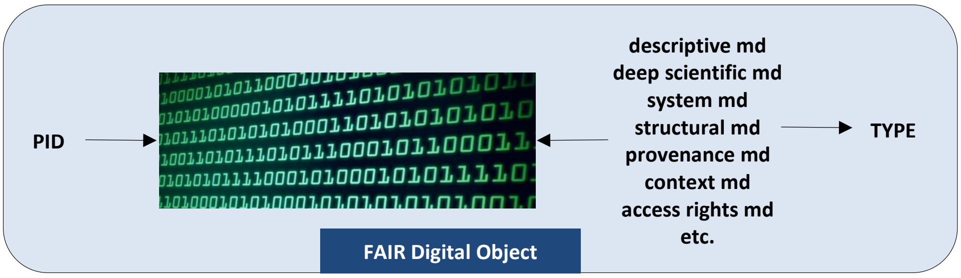
[12] Note that once appliances for these standards were constructed, further development towards a single standard was blocked

Therefore, any infrastructure supporting such a domain needs to focus on machine actionability and thus perfect identification, metadata including provenance and persistence to establish trust and enable tracing whenever this is needed.

On digital facilities (computers, networks, storage devices) we need to create *meaningful* "**sequences of bits,**" which have a structure and content that may not be obvious from the bit sequence itself. In contrast to the world of physical objects, form and content are separated in the digital world, i.e. *we need assertions about the bit sequences to make them findable, accessible, interoperable and reusable* (FAIR). These assertions may associate structural descriptions, documentation, methodological motivations, provenance, access rights and many other types of information with the bit sequences. This is what we in general call "**metadata**". There is an increasing tendency to summarize a set of structural assertions into what we call "**types**" allowing us to associate operations with classes of bit sequences. In the world of physical objects, inherent properties are sufficient to clearly identify them, but in the digital world where objects do not have visible characteristics, identifiers (visualized as strings) are a must. Identifiers are at the base of a stable and reliable domain of digital entities that we can build on in the coming decades. This is what we call **unique and persistent identifiers (PID)**.

We suggest calling this complex entity, i.e. a bit sequence associated with a PID and metadata, a **FAIR Digital Object (FAIR-DO)**. Clearly, at this level it does not make a difference whether the bit sequence contains data, metadata, software, semantic assertions or anything else. The metadata assertions tell us what it is and how to interpret it.

Why do we assert that FAIR-DOs are the next step of disruptive development? The current world of digital entities is not organized in form of FAIR-DOs. The necessary information is often existing, but it is hopelessly fragmented and hardly FAIR-compliant. Identities and references are not explicit; the structural and semantic descriptions are not persistently linked with the bit sequence; etc. It is the "**binding together of bit sequences encoding the content with a PID and metadata**" that allows accessing all necessary information to process the bit sequence, i.e., FAIR-DOs are thus semantically enabled.

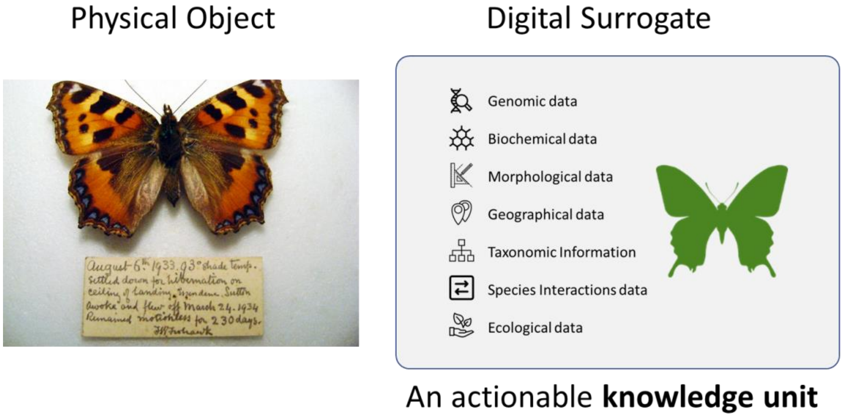


Changing our world of digital entities to become FAIR-DOs is a disruptive step: on the one hand, much effort will have to be invested to make the current digital entities FAIR-DOs, and many software stacks developed in the last two decades will have to be adapted or even re-written; on the other hand, new opportunities of networking and automatic processing such DOs will evolve. The current loss of 80% of efficiency [29] when dealing with data might be reduced to 40%, for example, which would save a gigantic amount of costs across all sectors[13]. To come to a flourishing shared data economy, it will be necessary to make this transformative step just as it was necessary to come to a flourishing electricity economy by agreeing upon AC voltage standards centuries ago.

[13] BDVA estimated the volume of the data economy in Europe with 360 billion € [31] of which about 290 billion € of the costs are due to data wrangling. In increase in efficiency of for example 40% would save about 145 Billion € which could then be used for advanced analytics.

4.5 Cumulative Domain of Knowledge

Recent discussions in various meetings gave an impression of how we can expect the knowledge domain to emerge due to digital science. Here we like to refer to the concrete example of biodiversity which has been nicely worked out by the DISSCO colleagues [30]. In their digital world the anchor points of the crucial knowledge of the field are the "observed specimens" which are represented by a "virtual box" with links to all digital information about a physical object. This virtual box has a relationship to the physical object, but more importantly, it will have complex and cumulatively increasing relationships to all kinds of knowledge evolving over time. As indicated in the following diagram, this evolving knowledge will come from various types of analyses, from classifications according to some characteristics, from places of occurrences and many more activities which are unknown yet. Creating this knowledge domain involves thousands of researchers and millions of citizen scientists, and thus it represents an enormous value.


Physical Object
Digital Surrogate
August 6th 1933. 93° shade Temp.
settled down for hibernation on
ceiling of landing, Tyndene, Sutton
awoke and flew off March 24. 1934
Remained motionless for 230 days.
F.W. Frohawk
Genomic data
Biochemical data
Morphological data
Geographical data
Taxonomic Information
Species Interactions data
Ecological data
An actionable **knowledge unit**

Such an effort can only be undertaken if the research community has trust in the stability of the mechanisms replacing books as main carriers. In many research areas we see this move towards developing such complex domains of knowledge where persistence of all digital entities and relations will be crucial for success and where ephemeral technologies would lead to disaster, i.e., the loss of years of work by the community which can hardly be reconstructed.

It is therefore not surprising that the biodiversity community decided to use FAIR DOs with technology independent persistent identifiers as the basis for their work.

4.6 Impulse for Semantics

Advanced semantic processing in such complex knowledge domains will be a big challenge in future, since the massive amounts of data will have to be processed (increasingly automatically), first of all to extract knowledge, which is then to be described using formal mechanisms such as augmented RDF assertions, for example. Furthermore, these assertions, defined in some ontological space, need to be combined and processed to extract scientific conclusions.

Recent information processing research in the realm of W3C has produced excellent tools used by many people worldwide. These include, among others, XML to express structures in a uniform way, RDF to express semantic assertions, SKOS as a simplified version to define the semantics of concepts, SPARQL to formulate queries on heaps of semantic assertions, etc. However, ontologies are still heavily underused in daily scientific work and hopes for seamless semantic processing have not materialized.

We see three major challenges to make semantic operations more of a commodity:

- The semantic domain needs to be built on top of a stable knowledge management layer as described above, defined by stable identifiers to ensure a long-term return on the huge investments that will be made.
- A flexible infrastructure is needed to allow researchers to quickly create, share and integrate relations between concepts without the need to look for semantic experts who deal with complex ontologies. Relationships are very much dependent on the cultural and scientific context, are also dependent on the current scientific task[14] and are a means to do flexible bridge-building, which will be required to do cross-walks.
- Much better tools are required to efficiently work in the semantic domain such as to flexibly identify and share knowlets, which are clusters of highly related concepts, and to use them for knowledge integration and exploitation.

Comment [KDS1]: This metaphore could be explained a bit.

5. Conclusions

For understandable reasons, some large programs on infrastructure building have postponed the establishment of a close interaction between the three pillars (O, T and S) which is a precondition for success. This strategy bears high risks of failures, includes a potential to continue fragmentation and ignores the two decades of discussions and practical work on distributed research data infrastructures of different types.

Basic Message and Time Scale

There is a growing insight that the huge investments which are currently being made need to lay the foundation of a new phase of data/information/knowledge management to prevent a dark digital age [32]. We need an infrastructure where digital entities (data, services) will find each other to perform widely automatic extraction of knowledge and humans will operate in a stable domain of assertions capturing machine conclusions to come to decisions. Such an infrastructure will be based on clear and stable identities of all entities, on a high level of abstraction, on ways of binding allowing machines to find all required information and on encapsulation to easily exchange technologies. Current strategies are too ephemeral and researchers who are dealing with increased masses of automatically generated digital entities and their many relationships feel the need to fundamentally change the approaches which will include disruptive steps. Therefore, it is time to understand that the core of the data infrastructures to be built must hold for centuries and not just a few decades. This is about preserving deep and fundamental scientific knowledge for future generations.

Approach

It is obvious that the type of infrastructure we need to design requires extensive conceptualizing. After having identified the FAIR principles as guidelines, we need to define high priority components and their interfaces as indicated by the hourglass paradigm that may have the potential to stay for long periods. Focusing on applications and services that may lead to successes for short time periods will blur the discussions about essentials. Eventually they will populate such an infrastructure in ways we cannot yet anticipate and a new phase of exploitation and competition will lead to ongoing innovation. It is time to make use of global bottom up forums such as RDA and GO FAIR to work out specifications and to implement them in form of extensible testbeds.

Lessons Learned

Many investments have already been done in the past two decades in building data infrastructures - some directly emerging from research questions, others from more generic approaches and some also based on commercial interest. It is time to analyze the experiences made, based on the goals outlined above, to determine generic components and their specifications and to also determine ways how existing infrastructure solutions could be integrated. We are still suffering from a "we do it

[14] For specific scientific analyses researchers may want to ignore fine semantic differences.

our own way" thinking which will eventually increase the fragmentation which we need to overcome urgently.

FAIR-DOs

The concept of FAIR Digital Objects (FAIR-DOs) emerged from intensive discussions in several RDA groups and in the interaction with the GOFAIR initiative abstracting from many scientific use cases on the one hand and all discussions about insights that resulted in the FAIR principles on the other hand. It has been shown that the FAIR principles and the concept of DOs are complementary and that FAIR-DOs actually are a way to implement the FAIR principles. The recently suggested DO Interface Protocol (DOIP) V2.0 [33] in combination with a global system to register and resolve unique and persistent identifiers seems to be the basic building block to bring FAIR-DOs to work. Currently, FAIR-DOs are the only suggested architecture that combines the required potential of abstraction, binding and encapsulation that we can see that could be used as fundamental building block for the data infrastructure as described above.

FAIR-DO and Research Perspectives

The biggest challenges for establishing a data infrastructure eco-system are of sociological nature: How to convince a broad group of researchers to adopt new methodologies and technologies? FAIR-DOs are about building a solid fundament for the evolving domains of knowledge, thus, not of primary interest to the researchers. FAIR-DOs will not solve, for example, the challenges of semantic mapping, but they can give researchers trust that their efforts will not be wasted after years. Since "trust" is a big concern of all infrastructure builders, much effort is still spent across disciplines on fundamental aspects which are covered by the FAIR-DO concept. Wide agreement on FAIR-DOs as common basis for all will save a lot of effort which could then be devoted to aspects that are closer to science. Agreeing on FAIR-DOs as a basis would also show a direction to overcome the gap in reproducibility which is currently one of most serious crisis on confidence in research results.

Acknowledgements:

Larry?, Per?, Barend?, Hilary? Koenraad

References

[1] Thomas Hughes: Networks of Power, Johns Hopkins University Press, 1983
[2] https://www.esfri.eu/
[3] Y. N. Harari, Sapiens: A Brief History of Humankind; https://en.wikipedia.org/wiki/Sapiens:_A_Brief_History_of_Humankind
[4] https://www.google.de/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0ahUKEwibwrPj4InZAhXQ3KQKHX-qCfAQFggoMAA&url=http%3A%2F%2Fec.europa.eu%2Finformation_society%2Fnewsroom%2Fcf%2Fdocument.cfm%3Faction%3Ddisplay%26doc_id%3D707&usg=AOvVaw0hiiGhm2KDmaoDsuUvosqa
[5] ESFRI Roadmap 2018; http://roadmap2018.esfri.eu/
[6] P. Wittenburg, G. Strawn: Common Patterns in Revolutionary Infrastructures and Data; http://doi.org/10.23728/b2share.4e8ac36c0dd343da81fd9e83e72805a0
[7] Research Data Alliance https://www.rd-alliance.org
[8] European Open Science Cloud; http://www.esfri.eu/ri-world-news/implementation-roadmap-european-open-science-cloud
[9] OpenAIRE; https://www.openaire.eu/
[10] EOSC Portal; https://www.eosc-portal.eu/
[11] GO FAIR; https://www.go-fair.org/
[12] NFDI; https://www.dfg.de/en/research_funding/programmes/nfdi/index.html
[13] Wilkinson, M.D., et al. A design framework and exemplar metrics for FAIRness. Sci. Data 5:180118 doi: 10.1038/sdata.2018.118 (2018)

[14] Peter Wittenburg et. Al.: Report Industry Side Meeting at the 11th RDA plenary in Berlin; https://www.rda-deutschland.de/intern/dateien/20180319-20_rda_p11_industry-side-event-report-final.pdf
[15] Report Big Data Workshop IOTWeek Geneva 2017; http://doi.org/10.23728/b2share.370397196404436691e2dc4979977cba
[16] Report Big Data Workshop IOTWeek Bilbao 2018, https://www.rd-alliance.org/making-data-revolution-happen-joint-rda-iot-forum-workshop-june-4-bilbao-spain
[17] RAMI4.0; https://ec.europa.eu/futurium/en/system/files/ged/a2-schweichhart-reference_architectural_model_industrie_4.0_rami_4.0.pdf
[18] Industrial Data Space; https://www.fraunhofer.de/en/research/lighthouse-projects-fraunhofer-initiatives/industrial-data-space.html
[19] GEDE; https://www.rd-alliance.org/groups/gede-group-european-data-experts-rda
[20] E. Schultes, P. Wittenburg, GOing FAIR and DOing FAIR: The complementary roles of the Digital Object Framework and the FAIR Principles in emerging data infrastructures; to appear
[21] G. Strawn, Open Science, Business Analytics, and FAIR Digital Objects http://doi.org/10.23728/b2share.6ceeed13eb6340fcb132bcb5b5e3d69a
[22] DOBES, http://dobes.mpi.nl/
[23] NOMAD, https://nomad-coe.eu/
[24] PDB, https://www.rcsb.org/
[25] ELIXIR, https://elixir-europe.org/
[26] CLARIN, https://www.clarin.eu/
[27] EPOS, https://www.epos-ip.org/glossary/eric
[28] J.C.R. Licklider: "Intergalactic Computer Network", https://en.wikipedia.org/wiki/Intergalactic_Computer_Network
[29] P. Wittenburg, G. Strawn, B. Mons et al.; Digital Objects as Drivers towards Convergence in Data Infrastructures: http://doi.org/10.23728/b2share.b605d85809ca45679b110719b6c6cb11
[30] DISSCO; https://dissco.eu/
[30] BDVA; http://www.bdva.eu/
[31] GOFAIR, https://www.go-fair.org/
[32] Vint Cerf: Google's Vint Cerf warns of 'digital Dark Age'; https://www.bbc.com/news/science-environment-31450389
[33] DOIP V2.0 : https://www.dona.net/sites/default/files/2018-11/DOIPv2Spec_1.pdf

[a1] Ritchie wrote about the relevance of the information society in 1991: http://dx.doi.org/10.1177/009365091018003006

[a2] Cite: OECD (2007) "OECD Principles and Guidelines for Access to Research Data from Public Funding", OECD Publications, Paris. http://www.oecd.org/sti/inno/38500813.pdf.
Cite: OECD (2015), "Making Open Science a Reality", OECD Science, Technology and Industry Policy Papers, No. 25, OECD Publishing, Paris. http://dx.doi.org/10.1787/5jrs2f963zs1-en