Common Patterns in Revolutionary Infrastructures and Data

Peter Wittenburg, Max Planck Computing and Data Facility
George Strawn, US National Academy of Sciences
February 2018
http://doi.org/10.23728/b2share.4e8ac36c0dd343da81fd9e83e72805a0

1. Summary

Societies have seen large infrastructures emerge when new technologies become available. From history we see that such infrastructures can have a huge influence on all aspects of societal life. Moreover, some patterns appear to reoccur in the evolution of such infrastructures. *Early visions* about the possibilities of a new technology lead to a phase of *creolization*[1] of approaches resulting in a deeper knowledge of the technology's pros, cons and limitations. A huge "solutions space" emerges and fragmentation results. Some solutions are more *attractive* than others, but a final phase transition occurs where the experts converge towards broadly accepted principles and specifications that lead to *exploitation and standardization*.

It appears that the "data infrastructure" is evolving into such a large infrastructure, with a potentially large influence on societies, industry and science. In order to gain new insights about complex relationships in nature, societies and minds, by integrating data from different silos we have seen an explosion of (non-interoperable!) solutions for data management, access and processing, i.e. we have entered a phase of *creolization*. Also, we have an increasingly clear view of the current inefficiencies in working with data. These inefficiencies retard innovation and broad participation, which will become even more important as billions of smart devices produce the data deluge of the Internet of Things. Stakeholders have begun looking for steps toward *convergence* that would increase efficiency without hampering innovation.

Comparing the evolution of the data infrastructure with the evolution of the infrastructures of electrification, computer networking and of information networking (WWW), we can observe that, despite all initiatives already taken, we have not reached convergence on a set of universals that would boost developments and create a momentum towards an efficient and interoperable data infrastructure. We propose that such a set of universals could be based on the concepts of "Digital Objects" (DOs), persistent identifiers (PIDs), and metadata (including data typing). These concepts could greatly reduce current inefficiencies in data processing and open the way towards automatic processing. In particular, the Core Data Model of the Research Data Alliance (RDA) provides a design for a universal Digital Object Access Protocol (DOAP, comparable to IP for the Internet or HTTP for the Web) which can interconnect the many organizations of data in use today, such as cloud systems, files systems, SQL databases, no-SQL databases and so forth. The agreement on fairly simple but potentially universal commonalities such as PIDs, DOs, and a DAOP could create the confidence for many developers to invest in data infrastructure building. We believe that it is time to take this step towards convergence.

Acknowledgements
We would like to acknowledge the many contributions to this discussion from close collaborators during the last months. In particular, we should mention here Robert Kahn, Larry Lannom, Tobias Weigel, Barend Mons and various colleagues from the Research Data Alliance (RDA)[2] and the C2CAMP[3] initiatives.

[1] *Creolization* is a term used to describe the development of culture and languages. It describes a process in which continuously new cultures/languages emerge and are mixing resulting in a broad spectrum of them.
[2] http://rd-alliance.org

2. Introduction

Large infrastructures can fundamentally change societies, cultures and economies. Whether we look at the Roman Water Supply System which opened the way to building the largest capital in ancient times, the evolution of the railroad systems which allowed to exchange people and goods at unknown speeds and facilitated the first industrial revolution, the global electrification which fundamentally changed the availability of power and facilitated the second industrial revolution or the introduction of the Internet with its Web application which changed the availability of information and facilitated new kinds of businesses, we can observe the huge impact of such massive infrastructures. There is no doubt that they have an effect on the ordering, integrating, coordinating and systematizing nature of modern human societies.

The questions we raise are the following:

- Are there re-occurring patterns to be found in the evolution of large infrastructures and in ordering complexity?
- What qualifies new developments to be categorized as new infrastructures changing the world?
- Can we extract patterns that could guide us in establishing new infrastructures, knowing that the pace of innovation is increasing?
- Can we use these patterns to decide whether "data" will be qualified as a new infrastructure and if so, which high priority steps need to be taken?

Answering these questions would require writing a thick book. Here we limit ourselves

- to describing a number of assertions about the evolution of infrastructures
- to commenting on the dynamics of infrastructure evolution
- to comparing this with the developments in the data domain

3. Assertions about Infrastructure Evolution

Infrastructures develop in characteristic phases which we will sketch briefly first. Then we will describe a number of contextual aspects of relevance.

3.1 Phasing Aspects

Thomas Hughes delineates four phases of infrastructure development in his book *Networks of Power* [1] where he describes the evolution of the electricity infrastructure.

- Phase 1 is characterized by inventions and development of start-up systems, driven by "inventor-entrepreneurs"
- Phase 2 is characterized by technology transfer between regions and also society driven now also by organizers and financiers, which can be called creolization.
- Phase 3 is characterized by the planning for system growth where "reverse salients"[4] need to be tackled
- Phase 4 is characterized by substantial momentum (mass, velocity, direction) where governments, business, professional societies, educational institutions join the activities

At the beginning, some holistic conceptualizers generate a vision based on results from theoretical considerations and especially laboratory work. In some cases these brilliant minds are not aware of

[3] http://www.c2camp.org

[4] T. Hughes introduces the notion of "reverse salients" to describe the fact that in developing complex systems there are always components that are behind, hampering the overall success and thus need to be overcome with highest priority.

the combination of many factors (technology, organization, politics, economics, entrepreneurship) that are necessary to turn such a vision into an infrastructure. The original ideas indicate that something is possible rather than perhaps immediately needed, but there is a conviction that new methods/technologies can help solving big societal and economic problems and lead to new unseen opportunities. Holistic views can act as guiding frameworks, although practice shows that some initial assumptions may need to be revised at later stages.

With respect to scope and size, systems usually start with test installations, followed up first by small size installations as local islands, and then being extended stepwise to interconnected systems. Since such infrastructures are complex systems containing many interacting components, their evolution is characterized by uneven growth and continuous decisions of tackling reverse salients. Thus, new components are continuously being introduced in particular in the start phases. This leads to a large variety of solutions preventing economy of scale effects and a reduction of costs. A universal solution promises exactly that.

In the course of this trajectory, partial and restricted solutions will be transformed into universal solutions that can be applied massively. Universals are often introduced in a later phase based on needs and insights, but they are essential to create a momentum by overcoming fragmentation and achieving economy of scale. These universals are often "simple" principles that only influence directly a specific part of the overall infrastructure. Universals can be seen as the "persistent fundament" of an efficient eco-system, enabling efficiency at the top-layers and they are characterized by the fact that they are broadly supported, and that people then believe it makes sense to invest. By turning to universals a momentum can be achieved since there will be supportive context, changed culture and a broadly understood demand. Various stakeholders join, such as big companies, educational institutions, research institutions, etc.

The last phase is characterized by solving huge organizational and logistic challenges and a proper planning for the available resources. Once fully established, these complex infrastructures inherently develop a certain inflexibility against major changes or new paradigms. Reasons for this can be found in the comprehensive "system culture" which has been established and in the huge investments that have been made.

3.2 Contextual Aspects

When a new technology is being invented, it is often seen as a continuation of something which already exists. It is its massiveness and its impact that makes a new technology a distinguishable one leading to a new infrastructure. When electricity was new, it was first seen as a subdivision of mechanical engineering or of physics, dependent on the view. The methods to be applied, the roadblocks and reverse salients to be overcome, and the impact, however, were of an order of magnitude such that a new discipline was required to enable the dynamic evolution of the global electricity infrastructure.

Such complex infrastructures reflect and influence context, but also develop internal dynamics. They are both causes and effects of social and cultural change. They are open systems in so far that the environment (society, economy, nature) has much influence on their development, making it impossible to fully design them, but requiring a stepwise, iterative and agile approach. Processes are not linear due to environmental influences (regulations, competition, funding streams, etc.) and inherent dynamics. Governmental regulations can have huge impact - hampering or fostering progress - on the development of infrastructures. However, the inherent drive for order in infrastructure building must be tempered by tolerance of messy vitality, which requires a balanced approach in regulations.

Another important aspect when building large infrastructures is that some reverse salients will not be seen as long as people don't focus on concrete steps to achieve the goals. It is finally economics

and efficiency that define the direction of system development. This also raises questions about the success of a top-down process when building infrastructures.

3.3 Dynamics of Infrastructure Evolution

Based on what was discussed in the previous section, we can summarize some special phenomena in infrastructure evolution by looking from two specific points of view: divergence and convergence (see figure 1).

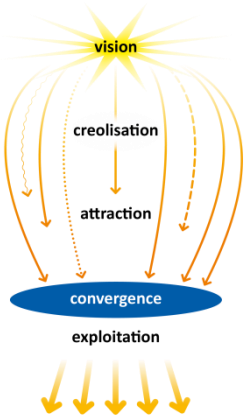

Figure 1 indicates the usual process of creolization, attraction, convergence and exploitation that can be observed when new visions are emerging.

New technological inventions or concepts with high potential for new types of complex infrastructures driven by early **visions** and often disruptive approaches are followed by a phase of **creolization**, i.e. the landscape of possible choices is being explored by many actors. Increasingly more options of the new concepts are investigated allowing the experts to better understand the underlying rules, principles and limitations. These developments are driven by competition, overcoming reverse salients and efficiency considerations.

Some tracks in this complex landscape turn out to be promising candidates and get more **attraction,** still heterogeneity and fragmentation hamper economy of scale effects. Therefore, some experts exploring the landscape start looking for **universals** that can act as a bridge between the different remaining solutions and as a stable basis for new developments and massive exploitation. Attraction and convergence are driven mainly by efficiency and economic concerns. The benefit of convergence is the belief of stakeholders that a stable fundament has been built, on top of which new investments and developments can be made to fully **exploit** the new technologies and infrastructures.

Successful standardization and governance activities often start in parallel with the convergence step and guide the exploitation phase, when the phase of inherent dynamics of scouting the landscape of solutions is over.

The Example of Electrification[5]

Thomas Edison had the crucial **vision** for electrification: to provide light by electricity, replacing older technologies based on gas and kerosene. His belief was that electric light could be deployed massively and at cheaper costs. He based his tests and deployments on direct current (DC), which leads to problems with long-distance transmission. Therefore, early systems were restricted to local areas. Other companies and labs started working on alternating current electricity (AC) in hopes of overcoming this reverse salient. After many test installations it became clear that indeed AC systems driven at high voltages can overcome the long-distance transmission cost problem. After having redesigned all components (generators, lamps, transformers, etc.) AC systems of various characteristics were deployed leading to a **creolization** with respect to multi-phasing, frequencies, voltages, generator and transformer characteristics. Despite some attraction, for some time, DC and AC systems of various flavors existed in parallel which turned out to be one reason for fragmentation.

[5] This paragraph is mainly based on the book Network of Powers from Thomas Hughes.

Ultimately, Westinghouse started to build AC systems based on **universals** which meant building a transmission system based on defined frequencies and voltages. Over the years DC systems were restricted to special local systems (such as power supply for our personal equipment). The deep knowledge about pros and cons of AC and DC electrical technology allowed the engineers to deploy each type of system where most appropriate. The standardization of AC systems with specific parameter choices (frequencies, voltages) created a huge **momentum,** ultimately revolutionizing all aspects of human life. Massive **standardization** by specifying all kinds of components down to small ones such as switches and connectors followed as a consequence of agreeing on universals.

T. Hughes compared the evolution of electrification in three communities (Chicago, London, Berlin) to study the involvement of **political and social factors** in the dynamics. He found that in the US little political influence interfered with the utterly dynamic evolution, while political influence in the UK was used to hamper fast evolution of the disruptive electrical technology[6], causing delays in deployment and a lack of knowledge. In Germany, political leaders listened to influential proponents of this new technology and supported it vigorously.

The Example of the Internet

Rudimentary computer networking began in the 1950s to connect computers to remote terminals. The technology utilized was a continuation of the principles of telephone communication systems. Broad thinking began in the early 1960s when J.C.R. Licklider proposed his **vision** of an "Intergalactic Computer Network" [2]. More modestly, he also proposed that he should only need one terminal on his desk to connect to any computer. That vision was realized in 1969 when the ARPAnet[7] began service. The ARPAnet utilized the novel technology of packet switching as opposed to the telephone technology of circuit switching.

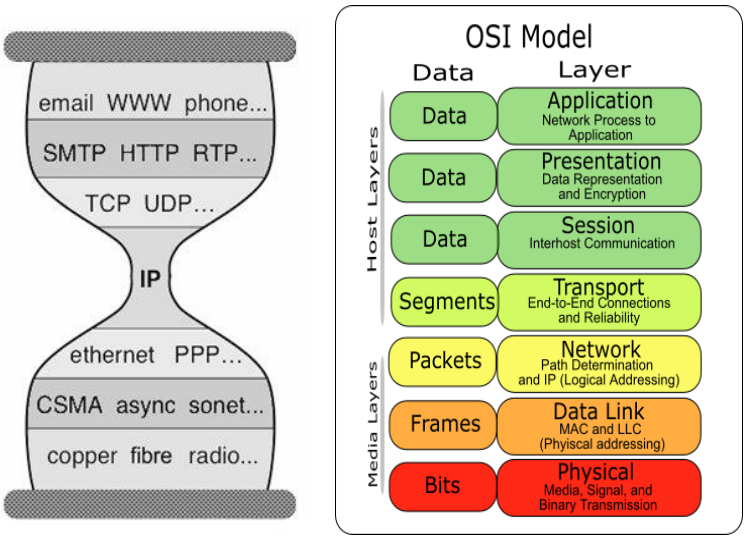

Figure 2a (left) indicates the well-known hourglass model of the Internet where the IP numbering system is in the core together with message exchange protocols such as TCP. Figure 2b shows the ISO/OSI reference model with its carefully specified 7 layers which however led to a complexity difficult to implement.

Phases of **creolization and attraction** could be seen since various systems were developed over the years such as X25 based on circuit switching[8], Ethernet for local computer communication, ARCNET[9] as a token-passing network first used to share storage devices, and others[10]. Based on these early studies and increasing knowledge it became obvious that computer networking was not just a dream, but could become reality with big impact. Several large computer companies such as IBM and DEC built their own proprietary networks. The consequence of this creolization phase was the existence of many different types of

[6] Companies producing gas-based lighting were strongly opposed to electricity.
[7] ARPAnet: https://de.wikipedia.org/wiki/Arpanet
[8] X25: https://de.wikipedia.org/wiki/X.25
[9] ARCnet: https://en.wikipedia.org/wiki/ARCNET
[10] Even in scientific organisations people started with building their own networks to solve advanced scientific challenges requiring computer communication.

computer networks hampering easy interconnection and exchange across them and thus hampering economies of scale.

This situation led experts to think about **convergence** and here we will mention two major approaches. R. Kahn and V. Cerf started working on version two of the ARPAnet in the early 1970s. The result was the "Internet" based on the TCP/IP protocols and Internet “numbers” (globally unique integers referring to Internet hosts—computers attached to the Internet). This demonstrated that using software adapters not only could computers from various companies could talk to each other, but more significantly different computer networks could interoperate.

A second effort driven by ITU and ISO that aimed to formalize the Internet approach was the 7-layer ISO-OSI reference model. It was developed in the late 1970s and specified in the early 1980s. The above diagrams indicate the essentials of the two approaches were moving towards a universal. The famous hourglass characterization of the TCP/IP approach (see figure 2a) indicates that in the core of the idea there was a hierarchical numbering system and a protocol specification in which the IP protocol was at the narrow waist of the hour glass with implementation protocols in the lower half and application protocols in the upper half. The ISO/OSI reference model[11] was more complex, defining seven protocol layers as opposed to the Internet’s five layers (see figure 2b).

Finally, “running code” and simplicity motivated the decision to move towards using TCP/IP as **universal** for implementing global computer networking. The ISO/OSI model continued to be used as a way to describe the different activities that need to be taken care of in networking. Big companies first created adapters to connect to TCP/IP, but stepwise changed their internal systems to native Internet code.

This global agreement on TCP/IP led to a highly dynamic **exploitation** phase with new kinds of machinery (routers, bridges, etc.) and services taken up by new innovative industries. With the Internet Society (formalized in 1992) a new agile forum was created to foster discussions about all aspects of TCP/IP-based computer networking, including the development of advanced applications (see figure 3). The Internet Engineering Task Force was incorporated into the Internet Society and became the most important group driving specifications of further protocols, procedures etc.

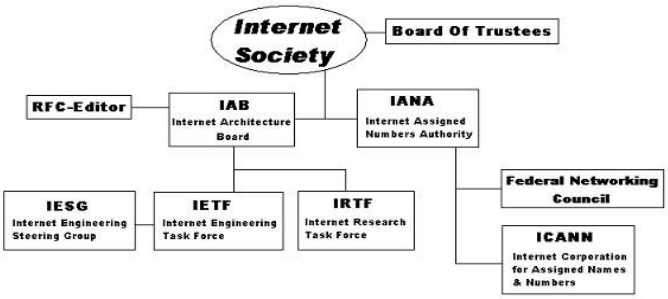

Figure 3 shows the kind of organization that was realized to make the Internet happen and to guide its future development. The approach taken for IETF can be compared with the one chosen by RDA.

At the beginning of the Internet age, traditional experts from telecommunications saw the Internet simply as an extension of the existing telephone networks. But it turned out to be an entirely new way of connecting people, informations and computers and required disruptive thinking to realize its potential. Thus **standardization** was first done completely outside of the traditional standardization bodies.

The Example of the World Wide Web

What started as one application on top of the Internet protocols became another major infrastructure in itself boosting global information exchange. T. Berners-Lee formulated the HTTP

[11] ISO-OSI-Model: https://en.wikipedia.org/wiki/OSI_model

(hypertext transport protocol) protocol "on top of" TCP/IP and the (initially) simple HTML (hypertext markup language) allowing anybody to format information and expose it on the Internet. Information (web pages) were uniquely identified by URLs (uniform resource locators), which were based on Internet names/numbers. Next, the development of the Mosaic browser demonstrated users and experts alike how this technology could be exploited. This knowhow spread with enormous speed and was used globally by all stakeholders that wanted to openly disseminate information. The technology was extended dynamically to give access also to any data located on the Internet. Boards were quickly established to govern the development of the World Wide Web and to promote interoperability[12].

Seeing the huge amount of information that was being created, three major developments occurred:

- Various companies and other stakeholders developed search engines, portals and complex content management tools to offer access to the information universe of the web (eg, Google).
- Various companies developed web-based retail stores (eg, Amazon)
- Various companies and other stakeholders developed new types of tools to facilitate different forms of communication based on the Internet (eg, Facebook).

Next, the *Semantic Web* was developed in the early 2000s to extend the web into the domain of knowledge representation. In particular, the Semantic Web triggered much conceptualization and development of frameworks and tools such as RDF, OWL, SKOS, and LOD[13] which began to deal with the important topic of semantics in the web.

The phasing of the World-Wide-Web was slightly different compared to the previous examples. At its basis was the definition of HTTP and HTML - a simple and clearly defined specification, i.e. **convergence** was stated at its beginning, if we ignore experience with some earlier markup languages. However, it should be noted the the Gopher system developed at the University of Minnesota was in use before the Web was popularized on the Internet by the University of Illinois' Mosaic browser. Minnesota made the bad decision to charge for Gopher and before they could rescind that decision, the Web had taken over. The governance structure safeguarded the Web's specifications and a phase of enormous **exploitation** was started.

4. The Data Domain

In this section we will analyze the state in the domain of data and compare it with what we can extract from other examples. First note that what is called the World Wide Web could have been called the Information Domain, in that Information is "data with semantics." In the case of the web, implicit semantics is provided by the human reader of web pages. The Semantic Web provides explicit semantics enabling computers as well as humans to "read" the semantic web pages. In this sense, both the Web and the Semantic Web are special purpose cases of the Data Domain.

4.1 Data as a New Infrastructure

Documents such as "Riding the Wave"[14] created by a high level expert group of the EC in 2011 and articles like "The world's most valuable resource is no longer oil, but data"[15] in The Economist from

[12] Advisory Committee, Advisory Board, Technical Architecture Group and the W3C Director

[13] https://www.w3.org/standards/semanticweb/

[14] 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

[15] https://www.economist.com/news/leaders/21721656-data-economy-demands-new-approach-antitrust-rules-worlds-most-valuable-resource

2017 shed light on the fact that we are faced with new types of challenges and opportunities. A **vision** has been formulated in so far as everyone should have seamless access to all kinds of relevant data and be able to reuse that data in different contexts to exploit their inherent richness for the benefit of societies, science and economy. Looking forward to the development of the Internet of Things, billions of smart devices are being installed everywhere, each of which will be creating continuous data streams. This example alone demonstrates that "data in general" is indeed a new challenge requiring a new type of infrastructure and new methods. The task will be to determine patterns of relevance for humans and societies in this endless data space using state-of-the-art methods, such as machine learning, to increasingly automate workflows, to extract knowledge from these findings in the form of exploitable annotations and assertions. These annotations and assertions can be subject to semantic processing and statistical analysis.

Some argue that the World-Wide-Web is ready to meet these "data challenges" and thus a new type of infrastructure is not needed. We argue that the Web was made for more limited data challenges and lacks a number of essential features. For one thing, the Web is “ephemeral” with continuously changing information. And it was not made for the management of huge amounts of data associated with metadata (the Semantic Web addresses this issue in a limited way). We need to turn to a domain of *registered Digital Objects,* each of which has a clear identity, and which is associated with metadata and a *persistent identifier* that is independent of any protocol. In the future, highly automated and reproducible domains of data will be key for building trust in all kinds of results and thus stable references to clearly identified digital objects (data, software, configurations, schemas, semantic categories, etc.) will be crucial (in contrast to typical web information data that is not self-explaining).

Results from surveys and interviews indicate that current data management and processing mechanisms are highly inefficient. An RDA survey from 2013[16] stated that typically a data scientist is spending 75% of his time on “data wrangling”[17]. M. Brodie reported about an MIT study [3] indicating that data scientists spend 80% on data wrangling and a recent study from CrowdFlower[18] also came up with 79% of the time being spent on data wrangling in industry. Yet we cannot express this inefficiency in actual costs, but it should be noted that the biggest factors for inefficiencies are to be found in the lack of proper and explicit data organization and bad data quality. This indicates that most of the time of highly qualified data scientists is spent on all those steps that are needed before starting the real data analytics work. IoT data volumes and complexity will increase this inefficiency, if we do not take urgent action. Industry is discovering a fundamental change in approach, speaking about a move from the "data warehouse" concept to a "data lake" concept, accepting data heterogeneity and dynamics.

4.2 The Phases of Data Infrastructure Development

Utilizing Thomas Hughes’ vocabulary, a new substantial infrastructure for data appears to be emerging. The first question to address is where the evolution of the data domain is at this moment. For years a number of researchers have been aggregating data to add another dimension to their scientific work, a process now called data driven or data intensive science. Research projects increasingly explored the new methods and built accessible databases and repositories such as the Human Genome Database[19], the Protein Database[20] and the Database on Endangered Languages[21]. Anticipating the coming "exaflood" of data and its impact on science, J. Gray, a computer scientist at

[16] RDA EU survey: http://hdl.handle.net/11304/6e1424cc-8927-11e4-ac7e-860aa0063d1f
[17] "Data Wrangling includes all preparatory steps necessary to finally start the analytics.
[18] Crowdflower: https://visit.crowdflower.com/WC-2017-Data-Science-Report_LP.html
[19] http://gdbwww.gdb.org/
[20] http://www.rcsb.org/ established as a web resource in 2003
[21] http://dobes.mpi.nl/ established as a web resource in 2000

Microsoft, called this shift the "Fourth Paradigm" of science [4]. In Europe the ESFRI[22] process led to the building of about 48 large research infrastructures in many different disciplines with a clear task to systematize and harmonize data-oriented work in the disciplines. Researchers working at the cutting edge of their disciplines started to look beyond discipline boundaries and to aggregate data from other disciplines, for example, to study the influence of climate and geography on the evolution of languages.

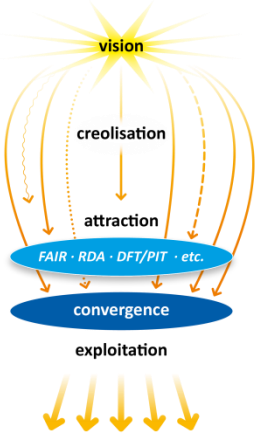

Figure 4 is copied from figure 1 but includes in addition an indication of where we believe the data domain currently is: man initiatives have tried many ways to solve the challenges, now it is time for a convergence step.

These projects and initiatives built their own tools and defined their own standards for almost all aspects of the data work and due to a lack of common agreements; new projects are still looking for the best way to solve their specific challenges. The result of this **creolization** phase which we are still in is, on the one hand a deep knowledge about possible solutions for the different problem areas and their pros and cons, and on the other hand a huge fragmentation of the solutions space. It should be noted that in industry a similar trend can be observed, which can best be explained as the trend to move towards a "data lake" approach compared to the traditional "data warehouse," and the many non-interoperable architectures that have been built on the top of “cloud systems.”

This fragmentation led to today’s inefficiencies and led some experts to start thinking about ways to reduce the solutions space. At policy level, initiatives such as CODATA[23] were founded and OECD started a data group[24] to work out recommendations. At data science level, initiatives such as Research Data Alliance (RDA)[25], OAI[26] and FORCE11[27], were started to work on recommendations for component specifications, principles and procedures. One of the great challenges in the ongoing discussions was the question how to tackle the overwhelming complexity of the data space, ranging from issues about the organization of data to issues about deep learning and semantics. L. Lannom gave this discussion a direction when he introduced four layers (Searchability, Accessibility, Interoperability and Re-usability) at a workshop in March 2012[28] which was at the start of RDA and which led to RDA's Core Data Model finally endorsed in 2014[29]. Later OECD and the G8 data group came up with similar statements. Then a Workshop at Leiden Medical Center in 2013 defined the FAIR principles of Findability, Accessibility, Interoperability and Reusability[30] which were published in 2014, and which now define a common language in worldwide use. The FAIR principles can be seen as a milestone since they clearly form **attractors** in this endless solutions space. However, the experts widely agree that the FAIR principles express policy goals, not blueprints for data infrastructure building. However, if data infrastructures adhere to the FAIR principles, they may provide a major step towards **convergence**.

---

[22] ESFRI: http://www.esfri.eu/
[23] CODATA: http://www.codata.org/
[24] OECD: http://www.oecd.org/
[25] RDA: https://www.rd-alliance.org/
[26] OAI: https://www.openarchives.org/
[27] FORCE11: https://www.force11.org/
[28] International DAITF Workshop at the ICRI 2012 Conference: http://www.icri2012.dk/www.ereg.me/ehome/index06e1.html
[29] Core Data Model: http://hdl.handle.net/11304/5d760a3e-991d-11e5-9bb4-2b0aad496318
[30] FAIR: https://www.force11.org/group/fairgroup/fairprinciples

The phrase “cloud computing” is often used, incorrectly, as a metaphor for the integrative solutions we are aiming at. Technically, cloud computing currently comes in three flavors: IaaS, PaaS and AaaS (infrastructure, platform and application as a service)[31]. The availability of the increasingly high performance Internet and increasingly inexpensive disk storage has led to the availability of a number of cloud software solutions from large companies. And this has led to the development of an increasing number of “cloud-based solutions,” each requiring their own administrative layer. However, today’s clouds can be seen only as a means to aggregate digital objects, not as a solution to the important interoperability challenge.

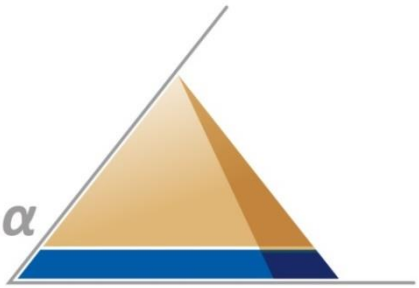

Figure 5 indicates the re-occurring pattern for infrastructure building: simple principles such as geometrical ones and a solid, well-designed fundament are sufficient to define a new commodity layer stimulating a boost of innovation. This figure does not indicate the huge logistics effort that was required and the many "reverse salients" that had to be overcome in ancient Egypt.

We should also mention that for more than 15 years some scientific communities and industries have utilized persistent identifier systems, such as Handles and DOIs[32], and this has had an enormous effect on **convergence**. Recently, delegates from 47 large European research infrastructures agreed on a paper for the use of persistent identifiers[33]. This gives weight to the assertion that the usage of protocol independent persistent identifiers such as Handles[34] and DOIs[35] is now widely accepted in many research and some commercial communities. But protocol dependent PIDs have also been important. For example, Internet numbers are PIDs that are closely associated with the TCP/IP protocols. Further experience will determine whether protocol independent or dependent PIDs will be more important in the Data Domain. More speculatively, the W3C is investigating decentralized PIDs (DIDs), which utilized peer to peer technology and might remove the need for central registration for PIDs, but would not remove the need for PIDs themselves.

4.3 Need of Convergence

The phase, data infrastructure building, is where **creolization** is counter-balanced by some **attraction** measures such as the FAIR principles, the implicit compliance with RDA's Core Data Model and the usage of persistent identifiers. Yet we have not reached the phase of **convergence,** which will be required to create a broadly supported momentum towards unification but that still enables an **exploitation** phase full of innovation. The need for **convergence** becomes more and more apparent as new data streams emerge from smart devices, crowd-sourcing initiatives, the simulations being run on big machines, and so forth. Experts know that only automatic procedures will help us to cope with this data deluge and to extract meaningful messages for human interpreters. Machines will require a systematic and systemic approach based on proper data specifications.

What kinds of measures are we looking for to make a change? Here we can learn from the examples previously analyzed. A common pattern for reaching **convergence** and kicking off an **exploitation** phase is a set of simple specifications such as "ac current at 110/220 V and 50/60 Hz", "IP numbering system with the TCP/IP protocols and some registries" and "HTTP protocol and HTML mark-up

[31] Cloud Computing: https://en.wikipedia.org/wiki/Cloud_computing
[32] DOIs are Handles with specific prefix and a community of practice within the Handle domain now governed by the Swiss DONA foundation. There are currently more than 3000 service providers to serve data science
[33] https://zenodo.org/record/1116189
[34] Handles: https://en.wikipedia.org/wiki/Handle_System
[35] DOIs: https://en.wikipedia.org/wiki/Digital_object_identifier

language on top of TCP/IP". At a recent workshop[36] J. Hendler called this “Find commonality and grow on top of that". The task ahead seems to be comparable to designing the essential geometrical principles, building a proper fundament and solving gigantic logistic problems which were necessary for the construction of the ancient Egypt pyramids (see figure 5).

Three other common patterns found are to be mentioned as well: (1) Such moments of convergence require political, organizational, administrational, economic and technological efforts to turn technology into an infrastructure accessible for all, i.e. it is about taking decisions to invest, aggregating the required resources and convincing minds to accept. (2) Only massive efforts based on a step of convergence will push the whole data community to focus on tackling reverse salients to make the final goal happen. (3) Too much political influence and too many regulations hampering the developments are counterproductive and lead to disadvantages (economic, technical).

4.4 A Possible Solution

As has been shown recently, there is wide agreement across disciplines and sectors to use persistent identifiers, since experts now are relying on the existence of global and reliable data. The scholarly communication community relies on DOIs and the data management and processing community widely relies on Handles which are now being served by more than 3000 service providers all over the world. Since the DOIs are Handles, both are now under the international guidance of the DONA Foundation based in Geneva[37] which takes care of a global and reliable system of connected root resolvers. But stable PIDs that are resolvable to useful information about the state of the Digital Object are not sufficient, they are the precondition.

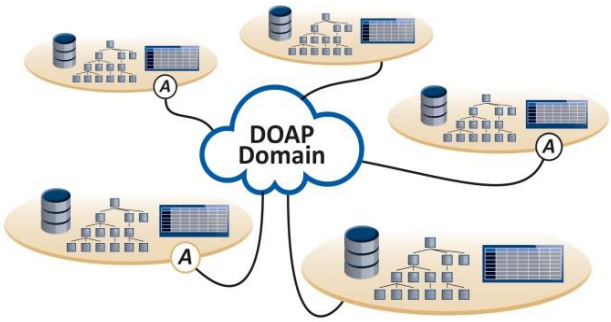

Figure 6 characterizes the unifying role of DOAP as the glue that binds together the many different repositories all using different data organizations and storage systems which make data integration currently so inefficient. The DOAP domain can be compared with the Internet where some connect by adapters (A).

What we are suggesting is to build a global virtualized domain of Digital Objects that are Identified by PIDs and characterized by different types of metadata (descriptive, administrational, data types, rights, transactions, etc.). Moreover, RDA's Core Data Model and the results of other related RDA groups indicate how these can be combined to achieve a globally unique way of organizing data, which would allow both humans and machines to find and process the appropriate entities. The following simple statements are fundamental and establish a universal guideline for an implementation of the data infrastructure:

- a Digital Object has a structured bit sequence that is stored in trustworthy repositories
- a Digital Object has assigned a PID and metadata
- the PID of a Digital Object is associated with all relevant kernel information that allows humans and machines to enable findability, accessibility, interoperability and re-usability[38]
- kernel information and digital objects have types allowing humans and machines to associate operations with them

[36] BRDI workshop: http://sites.nationalacademies.org/PGA/brdi/PGA_182878
[37] DONA: https://www.dona.net/ , one of the authors is member of the DONA board
[38] Essential kernel information are amongst others references to the locations of the bit sequences, to the metadata, to provenance records, to access permission records, to transaction records, earlier versions, etc.

Following these simple guidelines, we could specify a generic protocol to access digital objects (DOAP) independent of the storage system (files, clouds, relational databases, no-sql databases, etc.) and independent of the concrete organization of the data being applied in repositories (see figure 6). This will be the decisive point creating a momentum, since all repositories could invest in writing adaptors or adapt their internal organization. Interoperability is being created by an "Interdata" system, just as TCP/IP created an Internet between networks or as comparatively simple assumptions allowed ancient Egypt to build the various pyramids (see figure 5). It should be noted that these ideas go back to early papers on Digital Objects from R. Kahn and R. Wilensky published in 1995 and updated in 2006[39].

Some may argue that such a unique protocol system may hamper innovation. In contrast, we believe that just as in the case of the discussed examples we believe that such unification would lead to a boost of innovation, since it would create trust of scientists and industry that many will invest the huge amounts of funds needed to create an interoperable global eco-system of data infrastructures. It would pave the way to the same kind of hourglass approach that was achieved by TCP/IP, which also was basic enough to not hamper but to boost innovation. In the core of the hourglass model[40] are the PIDs and the DOAP which are to data like Internet numbers and TCP/IP are to networks and URLs and HTTP are to web pages (see figure 7). Access to all Digital Objects having a PID will be governed by the open DO Access Protocol being specified in RDA where DOs can range from individual files to complex collections of different DOs of different types.

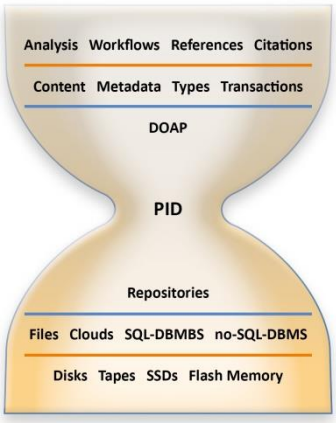

Figure 7 shows the hourglass model used to describe the key message for the Internet transformed to the data domain with PID numbers and the Digital Object Access Protocol in its center.

The broad engagement in various bottom-up driven interaction platforms such as the RDA, the Open Archives Initiative, the World Wide Web and recently the GO FAIR initiative and their increased understanding of the need to collaborate demonstrate that there is an excellent basis to further work on the concept of DO-based data infrastructures.

4.5 Problems being Addressed

The DO approach does not solve all challenges, of course; it is just the trigger that could synchronize minds to create a new momentum that can be catalyzed by disruptive technologies and processes. Just as TCP/IP didn't solve all network problems but did provide a solid base for many further developments, including the Web, Google, Amazon and Facebook. Which problems does the DO approach solve and which does it facilitate, but not solve? The DO approach will help overcome the differences in data organizations since the protocol comes with a unified approach allowing repositories and registries to either create adaptors or to adapt their internal organizations. As has been indicated these differences are responsible for a large fraction of the costs in current data management and processing. It will also help to establish trust in data since it will enable proper tracing of DOs and help creating proper provenance records. In addition, together with an improved "typing system" where types are also DOs as has been suggested by RDAs Data Type Registry[41] it will be the fundament for automatic processing of data. Machines will be guided by clear and stable

[39] https://www.doi.org/topics/2006_05_02_Kahn_Framework.pdf
[40] This diagram is based on the first version presented by L. Lannom at the ICRI workshop in 2012.
[41] DTR: https://www.rd-alliance.org/group/data-type-registries-wg/outcomes/data-type-registries

references to data, metadata, schemas, semantic categories, vocabularies, software components, repositories[42], registries and many more.

This DO approach does not directly address and solve the huge challenge of interoperability. But DOs facilitate this work by giving all entities a stable reference and the associated type of the entity will enable knowledge engineers to produce semantic metadata that will facilitate interoperability. Through the systematic implementation of the DOs, **scientific users** will be enabled to work at a level of abstraction which is their domain of conceptualization. Thus through this virtualization they are relieved of the need to deal with the (increasing) complexity of the data domain and can focus on the scientific aspects of their work.

Also, **data industry** can take advantage of the DO approach to increase "data wrangling" efficiency and thus increase the value of data trading. The DO approach also supports transaction records in a stable way, which implies that blockchain technology[43] (another application of peer to peer technology) could be utilized to keep track of re-usage and other dimensions of the data applications. As it was the case of Internet, new opportunities based on disruptive innovations will be opened to IT industry to provide tools and services based on this unified approach.

The broader public will also profit from the rich data domain when the costs are reduced and when tools and services become available that support easy and interoperable access based on a broadly supported fundament. Many **policy makers** seem to recognize the dimension of this new data infrastructure and are looking for ways to create the momentum. In Europe the European Open Science Cloud (EOSC)[44] and the GO FAIR initiates have been launched with high expectations. However, they lack a concept for convergence without which they will not succeed. The DO concept with its implicit push towards systematic virtualization could be this missing stone in the puzzle.

As represented in the pyramid diagram, the DO approach would provide a stable base on which to make large investments in an efficient ecosystem of interoperable and reusable data. The DO approach could be a "gate opener" for a prosperous data economy which urgently needs a convergence step to create the required momentum.

References

[1] **Thomas Hughes**: Networks of Power, Johns Hopkins University Press, 1983
[2] **J.C.R. Licklider**: "Intergalactic Computer Network",
https://en.wikipedia.org/wiki/Intergalactic_Computer_Network
[3] **M. L. Brodie**, Understanding Data Science: An Emerging Discipline for Data-Intensive Discovery, keynote, Proc.of the XVII Int'l Conf Data Analytics and Management in Data Intensive Domains (DAMDID'2015), Obninsk, Russia, October 13-16, 2015.
[4] **T. Hey,S. Tansley, K. Tolle**: The Fourth Paradigm: Data-Intensive Scientific Discovery, October 2009,
https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/

[42] It should be noted here that from a pure data science point of view a repository is also a DO, since it can be seen as hosting a complex collection all entities of which can be accessed by DOAP.
[43] https://en.wikipedia.org/wiki/Blockchain
[44] EOSC: https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud