Specifying and Implementing Data Infrastructures Enabling Data Intensive Sciences

© Peter Wittenburg, Herman Stehouwer
Max Planck Data and Compute Center, Garching/Munich
peter.wittenburg@mpi.nl, herman.stehouwer@rzg.mpg.de

Abstract

Examples from Psycholinguistics – a humanities discipline – show that data intensive research is changing all scientific disciplines dramatically. Data intensive sciences pose unprecedented challenges in data management and processing. A survey in Europe showed clearly that most of the research departments are not prepared for this step and that the methods that are used to manage, curate and process data are inefficient and too costly. The Research Data Alliance, as a bottom-up organized global and cross-disciplinary initiative, has been established to accelerate the process of changing data practice. After only two years RDA produced its first concrete results, which have to demonstrate their potential. In particular, the infrastructure builders are requested to act as early adopters of RDA results. The European Commission and its member states have taken serious steps to establish an eco-system of research infrastructures and e-Infrastructures anticipating the challenges imposed by the data deluge which will enable broad uptake of the paradigm of data intensive science. Research organisations have recognised these challenges as well and taken first steps to adapt its structures. However, we need to understand that we are in a phase of gigantic changes which implies that measures currently being taken need to be interpreted as tests on the way to new solid and sustainable structures.

1. Enabling Data Intensive Sciences

Quite a number of scientific institutes have been data oriented for a long time already. For instance, most of the research of the experimental and theoretical institutes of the Max Planck Society was based on data. Even an institute that belongs to the humanities section of the Max Planck Society such as our former affiliation - the Institute for Psycholinguistics [1] was oriented from the start towards the analysis of speech, eye movement and gesture recordings, detecting meaningful patterns, and building models to simulate speech perception. In physics institutes (fusion, astronomy, etc.) of course much larger volumes of data were being processed and they can look back to a much longer history of data oriented work.

It was the book "The Fourth Paradigm – Data Intensive Scientific Discovery" [2] edited by Tony Hey and colleagues that introduced “data intensive science” as the 4th paradigm of scientific discovery by referring to a talk given by J. Gray. It raised much attention for the concept behind this new paradigm. Gray distinguishes 4 paradigms that are co-existing today: (1) Empirical Science describing natural phenomena, (2) Theoretical Science using models to achieve generalizations, (3) Computational Science simulating complex phenomena and (4) Data exploration by unifying theory, experiment and simulation. Indeed, we can observe that science is changing in so far as finding meaningful patterns in data sets becomes an essential approach. Increasingly more powerful and numerous sensors, improved network connections, more powerful and numerous computers and more advanced algorithms are key pillars for this development. The "Riding the Wave" [3] report created by a High Level Expert Group of the European Commission (EC) was one of the documents that summarized the specific data challenges and opportunities, and requested actions by the EC to enable data intensive sciences for a large number of researchers and not only those that have sufficient funding to curate all data and software to be integrated to make use of it.

We see a number of trends which we can summarize as follows:

- An increasing number of research disciplines adopted data intensive methods due to new technological and methodological possibilities. During the last decades these changes were extreme in biological and neurological disciplines.
- The amount of data and its complexity in terms of creation contexts, data types and relations are increasing extremely.
- The Internet allows us to offer data via the web to be re-used by others.
- This enables us to combine data sets in new ways across institutional, national and discipline borders.

Proceedings of the XVII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL’2015), Obninsk, Russia,

- Mathematical methods have advanced to cope with heterogeneous data sets and we see large libraries with statistical, stochastic and machine learning methods becoming available.
- The total amount of available CPU and storage capacity allows researchers to do large amounts of computations on increasingly large data sets.

Despite the increase in compute capacity, however, we can also observe an increasing analysis gap, i.e. the fraction of data we are able to process in a way that we can extract knowledge is getting smaller. The reasons for the analysis gap are many and not subject of discussion in this paper.

Two examples taken from a humanities discipline show the fundamental changes towards data intensive science that could not have been carried out a few years ago. When studying for example the evolution of human languages over thousands of years linguists until recently based their theories on comparing fragmented descriptions of colleagues about several languages. Currently, large feature matrices are extracted describing characteristics of all languages in a particular region such as for example those spoken in Austronesia and these matrices are fed into phylogenetic algorithms to calculate most probable dependency trees that indicate how languages may have influenced each other over thousands of years. For this research a large database is required and also more powerful computers are needed than linguists were using traditionally to let the algorithms generate meaningful optima.

The application of massive crowd sourcing techniques in linguistics for example to understand human communication including multimodal interaction can be used as another example to indicate the dramatic changes in research towards a data centric perspective. These techniques generate many parallel data streams originating from smartphones that need to be annotated immediately by machine processing tools to make them available for scientific studies. This automatic annotation requires smart pre-processing and smart data management. In this setup an increasing number of parallel operating detectors must be trained to detect patterns in speech and video streams in real time with the help of stochastic machines. It is simply the shear amount of data requiring new ways of processing to enable this type of research leading to better assumptions about what guides our interactions.

The basis of such methods as described in the example above is the availability of large amounts of data to estimate the many free parameters of the models and in both examples data cannot come from one project or institute, but from many research labs. Researchers doing this kind of research know how difficult it is to find, access and combine the required data. Such research is very cost intensive and raises the questions whether we can continue without serious changes, and whether the available infrastructures are sufficient.

2. Human Brain Project

An even more extreme example for the shift towards the 4th paradigm is taken from life sciences. The recently started Human Brain Project (HBP) [4] (as an EC flagship project) has as visions (a) to be able to simulate at physiologocial level first rat brains and in a follow up phase human brains (in silico experiments) and (b) to predict brain diseases from patterns found in recorded data sets at an early stage. The main goal of the latter (medical informatics sub project in HBP) is being illustrated in figure 1. Researchers would like to correlate observed phenomena such as specific deficits due to brain diseases with all types of recordings that can be found from corresponding patients such as brain images of different types, gene sequences, protein data and perhaps even reaction time measurements. Without having a model of the human brain at hand this correlation would allow researchers nevertheless to detect co-occuring patterns in the data that seem to cause the observed phenomena. Machine learning methods

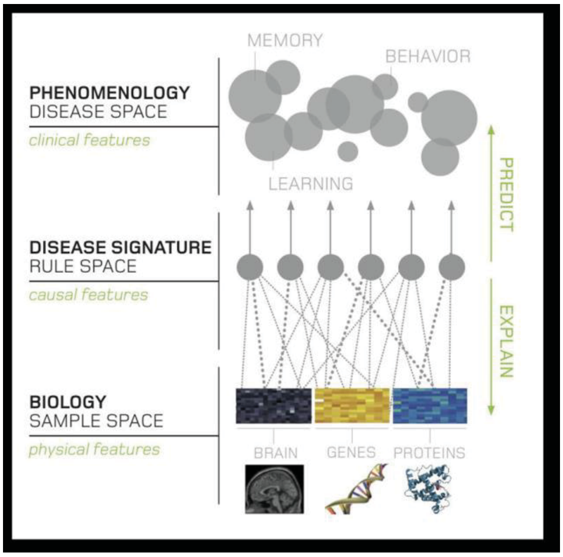

Fig. 1: One example for this new paradigm as it is used in neuro-sciences (HBP) is shown. For example phenomena such as created by specific brain diseases can be observed. Yet there is no chance to model the complexity of the human brain to make statements about their physiological origins. Data from various sources are correlated with the phenomena to find those patterns in the data that are causing the observed deficits.

are used to generate meaningful signatures from physical features in the data that then can be used to predict potential diseases from patients.

No assumptions are made about the structure and functioning of the brain, no assumptions are made how genes may influence brain structure and functioning, etc. since we don't have sufficient knowledge in these areas. Nevertheless, by using a large database of aligned data it is assumed that researchers can relate physical patterns with phenomenological observations first for early prediction, but later also for improved medication. Full brain simulations will typically cover spatial scales from nanometers (proteins) to centimeters (brain) and energy scales from 10 femto Joule at biological (Genome, Transcriptome, Proteome) up to 1 Joule at complex brain level (cognition).

To achieve its goals the HBP defined in total 13 sub-projects each of them having a size of a large project. Here we will briefly describe the new informatics-based platforms that are meant to offer the research community the possibility to work on human brain issues with the help of a set of strong and highly integrated tools:

- Neuroinformatics (searchable atlases and analysis of brain data)
- Brain Simulation (building and simulating multi-level models of brain circuits and functions, incl. for example models of neural microcircuits of up to a million neurons)
- Medical Informatics (see figure 1)
- Neuromorphic Computing (brain-like functions implemented in hardware)
- Neurorobotics (testing brain models and simulations in virtual environments)
- High Performance Computing (providing the necessary computing power by architectures that allow memory intensive applications and new ways of visually interacting with simulations)

HPC facilities at 4 centers can be used for the purposes of the HBP: Jülich (6 petaflops peak, 450 TB memory, 8 PB scratch file system) allowing simulations to up to 100 Mio neurons (scale of mouse brain), Swiss CSCS (836 teraflops peak, 64 T, 4 PB) in particular for software development and optimization, Barcelona SC (1 petaflops peak, 100 TB) for molecular-level simulations, CINECA (2 petaflops, 200 TB, 5 PB) mainly for data analytics. In addition KIT Karlsruhe provides 3 PB of storage. All centers are linked with 10 Gbit/s. In the neuromorphic area SpiNNaker chips are being used that have 18 cores and share 128 MB RAM allowing to simulate 16.000 neurons with 8 Mio plastic synapses with 1 W energy budget.

The goals are ambitious[1] and it is admitted that the gap between physiological modeling and cognition is still huge. However, the HBP indicates how data intensive science is pushed to its extremes in life sciences: (a) huge amounts of data addressing many different levels of brain organization are needed to feed the atlases, to enable analyses needed to feed and test the validity of the models and (b) much computer power will be required to carry out the necessary computations first within the project and afterwards by the interested researchers.

In addition to the problems described in the next section the HBP is confronted with difficult privacy and ethical issues making access to data even more problematic. Distributed data mining solutions are investigated to overcome these problems for example.

3. Data Practices

A large survey about data practices [5], based on some 120 interactions with data practitioners[2] from various disciplines, and two RDA Europe workshops with leading European scientists [22] made very clear that the current data practices are not adequate to support such data intensive science in an efficient and cost-effective way.

The major findings of this survey can be summarized as:

- The ESFRI[3] [6] discussion process and its project initiatives, as well as recent developments in e-Infrastructures, raised much awareness about data issues, the practices and the interaction processes around data management and access crossing discipline boundaries.
- Open Access [7] to publications and now also to data is widely supported but in practice there are so many hurdles that most data is still not available.
- Finding data re-usable for data intensive sciences using the web requires new mechanisms to establish trust. At this moment we are lacking such mechanisms.
- There is much legacy data out there the integration of which in our re-usable data domain will cost an enormous amount of curation and thus funds. In addition, we are

[1] It should be mentioned that there is a broad debate about the question whether the ambitions of the HBP are realistic.

[2] The term "data practitioner" is used here as a term describing skills of data scientists, data managers, data stewards, data librarians, etc. since mostly these terms are not well-defined yet.

[3] European Strategy Forum on Research Infrastructures

still creating legacy-style data despite all advancements since it is not suitably organized and described, which is mainly due to a lack of trained experts and appropriate software.

- There is an increasing pressure for almost all departments to participate in data intensive sciences, but researchers see a lack of expertise in adequate data management and workflow creation/maintenance skills. Currently researchers need to spend a large fraction of their time (partly up to 75%) to find, access and curate data to make it fit for their needs. In addition, the practice of many researchers working with manual steps or with ad hoc scripts does not lead to reproducible science.
- Data management is still widely based on file systems which do not allow capturing the increasing amount of "logical" information about the data (persistent identifiers, metadata, rights, relations, etc.). Ad hoc solutions are being used amplifying the problem of "increasing data entropy" as W. Michener [8] called it (see figure 2).

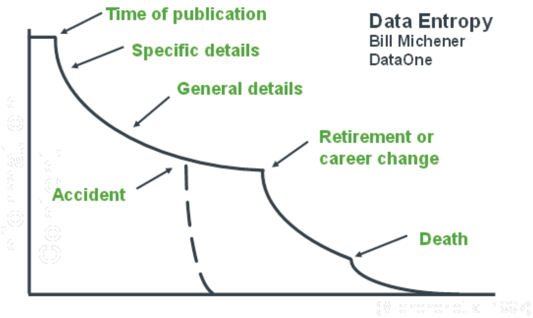

Fig. 2: The typical decrease of available information about data stored over time as described by W. Michener is indicated which results in great problems in making use of data. There are various factors and moments that lead to this decrease of information such as when PhDs leave an institute without having documented their data properly which is a very well-known phenomenon. Assigning persistent identifiers and creating appropriate metadata would help to reduce the speed of losing information.

- The use of persistent identifiers and metadata which would help in identifying, finding and re-using data is still in its infancy. Ad hoc solutions such as handling spreadsheets do only work for the duration of projects and leave chaos afterwards given the increasing amount of data.
- Despite some efforts for specific databases there is in general a lack of explicitness with respect to structure and semantic descriptions of the content of data which creates inefficiencies in particular when users do not have direct relations with the creators.
- There is a clear trend towards using "trustful" centres which offer researchers to host, manage and access their data. However, there are many hurdles for centres to offer cross-border services although economy of scale factors indicate that much can be gained due to the available expertise. Existing certification methods such as defined by Data Seal of Approval [9] need to be applied by the centres to raise the level of trust.
- It is widely agreed that there is a lack of expertise and knowledge about data issues (principles, organization, curation, etc.) and that we need to train a new generation of data practitioners. It is this lack of experts and expertise that hampers progress.

Senior scientists agree that changes in data practices are urgently needed, but they hesitate to take steps for mainly two reasons:

- they lack guidance towards certain agreed solutions which prevents investments,
- they lack the experts that would turn investments into appropriate solutions.

4. Achieving Changes through RDA

This raises the questions who can give guidance in navigating in the huge solution space with respect to data issues and how can we train the new generation towards harmonized solutions that guarantee more efficiency and cost-effectiveness which finally will boost data intensive sciences. Here we would like to refer to the early phases of the Internet where many solutions were suggested with different competing approaches. It took about 15 years until agreements on simple principles such as TCP/IP [10] for global networks were accepted. Basically these agreements led to the boost of connectivity which we can now take profit from.

Quite a number of policy level initiatives have established rules and principles and there seems to be wide agreement [11] about them. An increasing number of funders are also requesting to add so-called data management plans to grant applications which certainly raise the level of awareness about data issues for many researchers. But due to the problems described above there is also great uncertainty how to create such plans that make sense for the many data use cases [12]. An increasing conviction of some data practitioners and some funders emerged that an acceleration of the process to come to agreements that help changing data practices is urgently required. The Internet history seems to offer a possible approach: complement the policy level efforts by an essentially bottom-up driven initiative where data practitioners work on urgent barriers that need to be overcome. To this end a first international workshop was organized at the ICRI conference 2012 [13] under the name "DAITF" which stands for Data Access and Interoperability Task Force. A joint effort from mainly European, US American and Australian experts and funders led then to the birth of the Research Data Alliance (RDA) [14] in autumn 2012. We like to use the similarity of some characteristics with the Internet Engineering Task Force, however, it is obvious that the data domain has many more facets and challenges to deal with.

We would like to cite Naoyuki Tsunematsu (Senior Advisor of Japanese Council for Science and Technology) who pointed to two observations relevant in this context and which motivated Japan to join the Research Data Alliance [15].

- The value proposition for publically funded research is about "stimulating competitiveness" but a new strand needs to be added which is "knowledge discovery on smart data collections" where professional infrastructures and human skills are the key factors for success.
- There seems to be a correlation between a lack of motivation to share data in the Japanese academic world and thus a lack of openness and a decrease in the number of top-level international collaborations and of top-level papers which is a concern for policy makers in Japan[4].

After the workshop at ICRI 2012 the European Commission, NSF and NIST in the US and the Australian Government accepted grant proposals from key experts in their respective regions that allowed the practitioners to start the RDA work, i.e. funding is given to consortiums in the three regions. As one branch the RDA Europe [17] project was funded as a usual EC project, in September 2015 already, the 3rd RDA Europe project will start to allow us to continue the work and EC's new draft work programme 2016/17 indicates future perspectives for RDA. First, a steering board was established between the three funded initiatives to define a governance structure and procedures for RDA, and it started stimulating the practical work.

RDA decided to have a very simple structure where the key roles are given to the ***Working Groups*** and ***Interest Groups***[5] that meet at plenaries and other meetings. Every RDA member can decide to initiate such a group and to be successful a case statement needs to be submitted that must fulfil a number of criteria [18]. A ***Council*** was setup that has an overlooking role to ensure balanced progress and adherence to quality rules and processes. A ***Technical Advisory Board*** that is elected by the RDA members[6] will give advice to all actors on content aspects, i.e. respond on questions such as "do the intentions of the Working and Interest Groups meet the scope of RDA, do they fulfil the established requirements, do they involve existing and relevant initiatives, do they intend to remove practical barriers, etc.". An ***Organisational Advisory Board*** that represents all organizations that are organizational members and thus contribute with some funds to the success of RDA gives advice on organizational and administrative issues. In addition RDA has a ***Secretariat*** that needs to organise the plenaries, keep control on the processes and doing a variety of other administration/ organisational tasks. A ***General Secretary*** has been appointed leading the secretarial work and taking responsibility for managing RDA global.

While RDA global is the platform where agreements are being achieved in form of guidelines, procedures, interface and protocol specifications to overcome barriers, the regional branches such as RDA Europe have the task to raise awareness about RDA in their region, convince experts to participate, interact with many stakeholders to understand the needs and priorities, organize the adoption of RDA results, taking care of training and education and contributing to the costs of RDA Global. RDA Europe for example organises a number of meetings to meet the requirements such as interacting with the EC and member state ministries, European science organisations, European leading scientists, large scale European research infrastructures such as ESFRI projects [19] and e-Infrastructures [20] such as EUDAT [21] and many research communities. The meetings with leading scientists [22] are of great importance and have led to useful recommendations for RDA, most of which will be implemented by RDA Europe from September 2015 on. The interactions with policy stakeholders led for example to the Data Harvest Report [23] setting priorities.

5. Early RDA Results

Thus RDA's mission is about building the many social and technical bridges that are required to make data intensive work much more efficient and thus to allow many researchers to participate in

[4] The recent G8 Open Data Report [16] indicates that in the rating between G8 members Germany and Russia are even behind with respect to openness of data.

[5] It should be noted here that the major difference between the two groups is that the WGs need to come with tangible results after 18 months.

[6] Everyone who agrees with the basic rules of RDA can become a member by registration.

extracting knowledge by processing virtual collections existing of data coming from various providers increasingly often across disciplines and borders. Here we want to briefly indicate the major results of the first working groups that finished after roughly 20 months (or that will finish within the coming few months) and their possible impact on changing practices.

5.1 Data Foundation and Terminology (DFT)

Based on many use cases from various disciplines and countries the DFT Working group [24] came up with a simple core data model and a terminology for registered data. It introduces the notion of the Digital Object which is represented by a bitstream, can be stored in various repositories, is identified by a persistent identifier and described by metadata. The model includes a few further definitions, but important is to note that these definitions are fundamental and independent of disciplines. If scientists worldwide would adhere to such a simple model we could much more easily understand each other when talking about data and would be able to build harmonized software leading to much higher interoperability.

5.2 Data Type Registries (DTR)

The DTR group [25] created a specification for data type registries that allow users to link data types of various sorts with functions (executable code). Data types can be simple types such as semantic categories (temperature, noun, etc.) or complex types such as scientific digital objects (complex annotated images, time series, tables, etc.). DTRs can be used for example to carry out mappings automatically when simple types such as “temperature” occur or start for example visualization software when complex types are found. Such DTRs would overcome the problems we so often have with unknown data types which we receive and where we do not know how to process and interpret them. Thus we see an enormous impact for DTRs in daily practice.

5.3 PID Information Types (PIT)

The PIT working group [26] produced a common API (Application Program Interface) to unify access to Persistent Identifier (PID) service providers. Currently there are different PID systems (Handle/DOI[7], AWK, etc.) and many different service providers all having their own regulations making it very cumbersome to get for example the checksum of a Digital Object to check its identity and integrity. Applying this unified API together with some basic data types such as “checksum” would allow application programmers to simply provide one piece of software allowing them to deal with all PID service providers in the same way. Since PIDs will have such a central role in data management and access the impact of a unified API will be enormous.

5.4 Practical Policies (PP)

In particular data management and curation are guided by specific policies which are then turned into executable procedures such as "replicating a data collection" or "checking digital objects' integrity" that are mostly used in federated environments. The PP group [27] is collecting many such practical policies from various institutions and projects, analysing and evaluating them and suggesting best practices which then can be offered as templates for proven operations. Thus, these templates have the potential to increase the trust level. The work of the group will not end since there are so many areas where best practices can improve the quality and reproducibility of data practices. In collaboration with the EUDAT project the group is working on an open registry standard for such best practice PPs.

5.5 Metadata Standard Registry (MDR)

As has been described the usage of proper metadata is still in its infancy and there are many reasons for this. One reason certainly is that many labs still do not know which metadata they should use, where they can find suitable vocabularies and tools, etc. The MDR group [28] offers a registry which allows researchers to look for most suitable metadata schemas. Therefore this MDR will help data practitioners that are looking for proper metadata solutions. More work in the metadata area is going on within RDA.

5.6 Data Citation (DC)

The Data Citation group [29] worked out suggestions of how to cite so-called dynamic data, i.e. data that changes while people are already working with it and referring to it. All data coming in from seismological sensors for example will immediately be used when it becomes available for processing even if data samples in the sequences are missing due to transmission delays for example. How can researchers refer back to these incomplete versions of data? This is a problem that many disciplines have and this group worked out a suggestion how to solve this citation problem so that it could be implemented in all software and procedures.

5.7 Repository Audit and Certification (RAC)

As indicated above quality assessment of repositories (centres) is increasingly important to raise the level of trust and the RAC group [30]

[7] DOIs are Handles with a special prefix and used to refer to published collections. Handle/DOI services are available worldwide.

wants to come up with a unified standard. A few suggestions have been made such as by Data Seal of Approval [31] and World Data Systems [32]. These two suggestions are already widely used and so similar that the responsible initiatives decided to join forces to make their guidelines compatible with each-other. It is widely agreed that the resulting set of guidelines is a good basis to certify trusted repositories worldwide[8].

5.8 New RDA Phase

At the fifth plenary (P5) we had a first adoption day [33] where experts from different disciplines and institutions presented their way of making use of these early results. The presentations showed that the RDA results were not just an academic enterprise, but indeed fulfil concrete needs of early adopters in particular since in some cases first implementation versions are available and can be used. Currently, RDA Europe is for example looking for further adopters of these results by offering funding for collaboration projects.

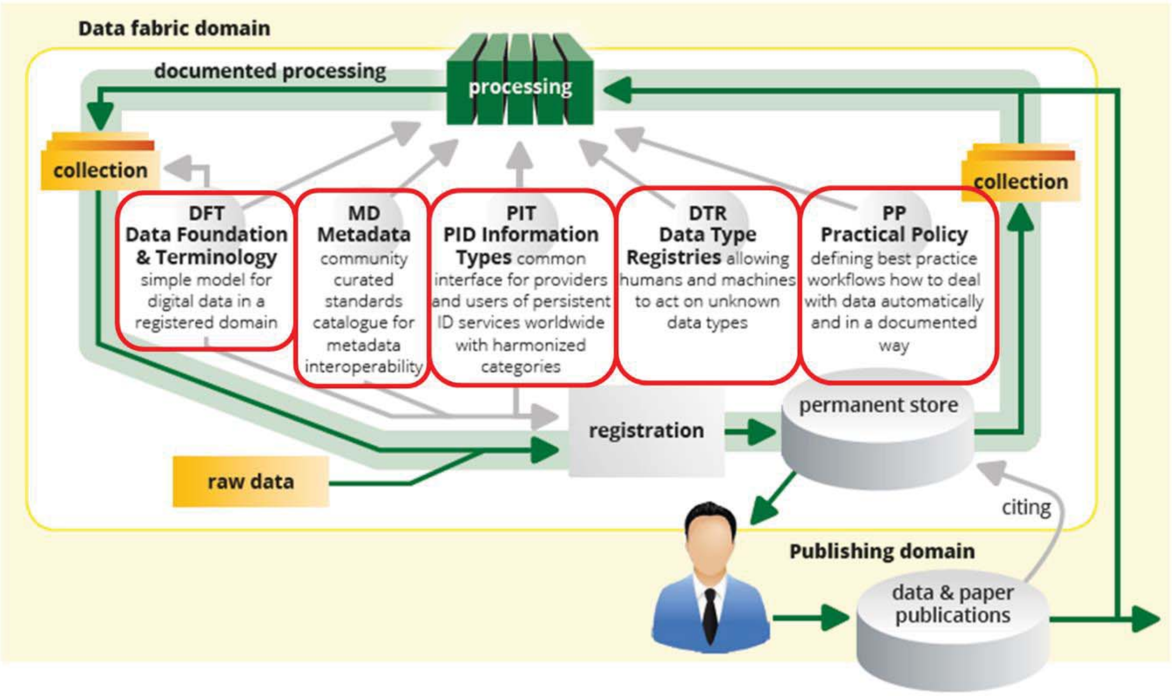
Data fabric domain
documented processing
processing
collection
collection
DFT
Data Foundation & Terminology simple model for digital data in a registered domain
MD
Metadata community curated standards catalogue for metadata interoperability
PIT
PID Information Types common interface for providers and users of persistent ID services worldwide with harmonized categories
DTR
Data Type Registries allowing humans and machines to act on unknown data types
PP
Practical Policy defining best practice workflows how to deal with data automatically and in a documented way
registration
permanent store
raw data
citing
Publishing domain
data & paper publications

Fig. 3: It indicates at an abstract level the typical data creation and consumption cycle as it is being used in the labs doing data intensive sciences. DFIG's questions are now which components are needed to run such a cycle efficient and self-documenting and how these components need to interact. The figure also indicates how the working groups that finished or are finishing fit into this cycle.

We should add here that RDA is obviously entering a new phase. While the first 5 working groups were started at the first plenary in March 2013 each of them focusing on their specific topic under high time pressure, the experts now understand that they need to synchronise more to achieve the needed coherence of all results. One consequence was to set up the Data Fabric Interest Group (DFIG) which is now bundling forces to understand all components that are required to come to efficient and reproducible data intensive sciences. Figure 3 indicates briefly the topic being addressed[9]. Data production and consumption in the daily data driven work can be indicated by a cycle where at a certain moment new raw data is being created and in some form being organised/registered and put into a store. Researchers who want to make use of data define a new (virtual) collection by selecting data from repositories and then carry out some processing steps on it which can be management or analytical operations. The result is a new collection of data which should be registered and stored again. The questions addressed are now which components are needed to run such a "fabric" efficiently and self-documenting and how these components should

[8] We note here that there are several further certification schemes that go more in-depth on specific aspects such as the “Security for Collaborating Infrastructures Assessment and Modification Record” (SCI) for security aspects, or the NESTOR seal (based on DIN 31644) or ISO 16363 certification for general data repository aspects. The DIN and ISO certifications are extremely detailed and thorough, and thus fairly costly to implement.

[9] A White Paper describes DFIG in more detail [34].

interact. Figure 3 also indicates how the finishing working groups fit into this cycle.

Currently the DFIG is collecting many Use Cases to build on what people are already doing and to abstract from these Use Cases to "common components" that are required. Such common components would include for example a global PID system[10] providing PID registration and resolution mechanisms that can be used by everyone. Everyone interested should be motivated to contribute Use Cases that will influence the discussions about common components. A first paper to accelerate discussions has been made available by a number of distinguished experts from various regions [35].

5.9 RDA Summary

RDA is still a very young initiative and its success mainly depends on the willingness of data practitioners to spend time on global and cross-disciplinary[11] problem solving, on the quality of their results, and their uptake by scientific projects worldwide. For TCP/IP in its early days, there was nothing particular that distinguished it from other suggestions. It was its layered approach and robustly running code that finally convinced people worldwide to adopt the standard. RDA needs to do a lot to have similar success and it needs strong infrastructure pillars that provide and maintain services.

6. Infrastructure Pillars

As described, RDA is only working on specifications and it is neither providing services nor maintaining code. It will rely on powerful centres and federations to provide the infrastructures that are finally required to transform specifications into real services that enable efficient data intensive sciences. In the same way we can state that researchers in general are not so much interested in specifications of interfaces for example, but in the services that will facilitate their work. In a simplified way figure 4 indicates the essential relationships between researchers as consumers of facilitating services who would also like to influence specification building to ensure the emergence of useful services, infrastructures that are built compliant to the specifications to ensure interoperability of the services and initiatives such as RDA which establish the specifications as a joint effort of data practitioners, i.e. researchers and infrastructure providers.

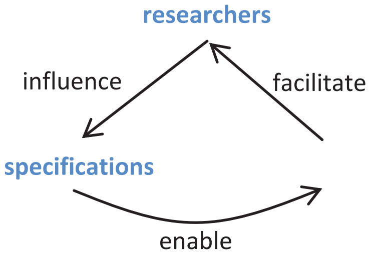

Fig. 4: It indicates schematically the essential relationships between researchers, infrastructures and the specification work such as in RDA.

Information infrastructures in our distributed landscape of data and computational services get very complex and involve several layers, which is sketched in the diagram drawn by the High Level Expert Group on Scientific Data (Figure 5) [3]. This diagram aims to work out the difference between discipline specific and common services that users (top layer) will use probably without noticing who will give the services they are using. Initiatives such as EUDAT were started to offer common services (bottom layer) and thus to complement the typical ESFRI layer (middle layer) with many European research infrastructures in various research disciplines.

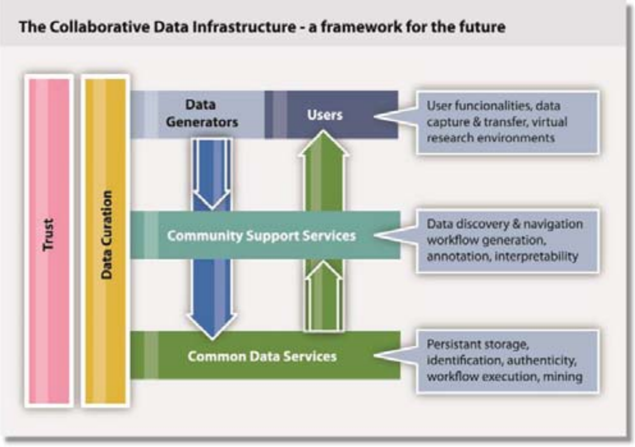
The Collaborative Data Infrastructure - a framework for the future
Trust
Data Curation
Data Generators
Users
User funcionalities, data capture & transfer, virtual research environments
Community Support Services
Data discovery & navigation workflow generation, annotation, interpretability
Common Data Services
Persistant storage, identification, authenticity, workflow execution, mining

Fig. 5; It schematically indicates 3 layers of the so-called Collaborative Data Infrastructure where community based infrastructures offer community specific services and e-Infrastructures offer common discipline crossing services. This was seen by the EC as a blueprint for funding programs.

The first ESFRI roadmap from 2006 [36] led to 44 research infrastructures leading to an intensive and concerted European activity across many disciplines. Most of these infrastructure initiatives are heading towards building persistent distributed information infrastructures.

[10] The Handle System (http://www.handle.net/) is such a global PID system supervised and managed by the international DONA Foundation and it is also basis of the DOI and other service providers such as EPIC in Europe.

[11] RDA also includes some disciplinary groups which are using the global nature of RDA to achieve community agreements.

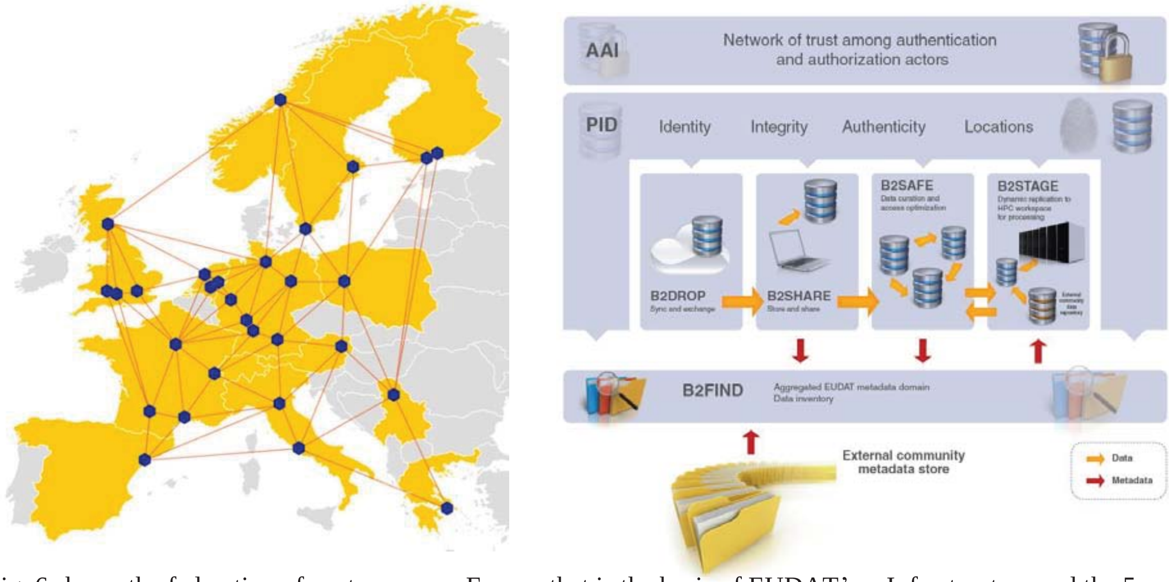
AAI
Network of trust among authentication and authorization actors
PID
Identity
Integrity
Authenticity
Locations
B2DROP
Sync and exchange
B2SHARE
Store and share
B2SAFE
Data curation and access optimization
B2STAGE
Dynamic replication to HPC workspace for processing
External community data repository
B2FIND
Aggregated EUDAT metadata domain
Data inventory
External community metadata store
Data
Metadata

Fig. 6 shows the federation of centres across Europe that is the basis of EUDAT's e-Infrastructure and the 5 basic user services it offers to the research community. In addition to the 5 user services it established system services such as an authentication and authorisation infrastructure and a service to register and resolve persistent identifiers.

One example is the CLARIN initiative [37] in the area of language resources and technology which has recently achieved the status of an ERIC[12]. CLARIN is based on strong and federated centres in a variety of European countries that share the effort in defining standards together with the community, in aggregating digital language resources, and in offering joint services, in managing and curating data with discipline specific knowledge and others. The services offered by CLARIN include deposit possibilities, a joint metadata catalogue called Virtual Language Observatory [38], a distributed workflow tool allowing users to analyse texts in various languages and many smaller services. However, CLARIN centres are not equipped to offer massive compute power to all possible users from all over Europe who may want to execute workflows or use large storage systems to manage large data sets. Therefore research infrastructures such as CLARIN make liaisons with e-Infrastructures such as EUDAT and pay for such common services. All research infrastructures from the different research domains are looking for similar options if they are data and compute oriented.

The ESFRI organisation and the EC are still actively starting new research infrastructures. To come to an optimized eco-system of information infrastructures all ESFRI projects and beyond (such as Human Brain Project) are seeking collaborations with e-Infrastructures such as PRACE [39] and EUDAT to make use of the advanced services that are offered by them.

6.1 EUDAT

EUDAT is a federation of well-resourced and partly national data and compute centres in various countries as figure 6 indicates. Within its first three years EUDAT invested all efforts in developing 5 basic services in collaboration with at the beginning 5 communities[13] (climate modelling, earth plate observation, human physiology, biodiversity and language resources and technology). B2SHARE, B2DROP and B2FIND are services directed to the end users meant for dealing with long tail type data. B2SAFE is a service that allows replicating large data sets between a community centre and the EUDAT centre network. The B2STAGE service is meant to move data sets from the EUDAT store to the workspaces of powerful computers of different types (HPC, etc.) to carry out computations and to return the results. All data in EUDAT are registered, i.e. all digital objects have PIDs and are associated with metadata to make them findable and accessible.

It should be added here that federating data centres and their collections was and is a major challenge and currently not scalable. The reason for this can be found mainly in the data organisations where each centre has chosen a different solution. This lack of interoperability leading to enormous costs is one of the reasons why EUDAT is very

[12] ERIC is a special organisational template invented to allow ESFRI research infrastructures to become European legal entities.

[13] Currently EUDAT is closely interacting with 32 communities.

much interested in harmonised solutions being worked out by RDA, for example in the DFT group. Due to this close interest EUDAT declared that it will try out RDA outputs where possible and thus act as an RDA testbed in Europe.

EUDAT just received its 2nd funding grant for 3 years which needs to be used to stabilize and improve the services being offered, work out a sustainable funding model and look for collaborations with other European e-Infrastructures such as PRACE. This led a to an additional work item which is devoted to improving the exchange of data between EUDAT and PRACE and demonstrating this as an efficient service with the help of concrete data and compute bound projects. Future challenges are anticipated by also strengthening the work on executing automatic workflows. It is understood that data science needs to turn increasingly often to automatic and self-documenting workflows to make its results reproducible. Yet the challenges to let users quickly deploy and execute complex software close to where the data is stored, i.e. operate in a distributed environment, are huge and severe barriers need to be removed. But EUDAT needs to demonstrate that it finally can offer services similar to Amazon and other companies where users can execute their software in a virtual machine environment and basically pay for the cycles used.

In the coming period EUDAT will also be faced by a new initiative and request of the European Commission in the realm of Open Science and Innovation [40] called the European Open Science Cloud. The EC wants to have a “cloud service” for all European researchers without having defined its exact specifications yet. A high level expert group is being formed that will work out the requirements. According to EC experts the term “cloud service” is meant in the broad sense, i.e. it needs to include the necessary structures for persistent identifiers, metadata, relations, etc.

6.2 National Data Service (NDS)

Also in the USA an attempt is being made under the lead of NSCA [41] to setup a National Data Service (NDS) [42] and to offer similar cross-disciplinary data services compared to EUDAT in Europe and ANDS [43] in Australia. The NDS is an emerging vision for how scientists and researchers across all disciplines can find, reuse, and publish data. It wants to build on the data archiving and sharing efforts already underway within specific communities and to link them together with a common set of tools.

Currently the NDS is focusing on collaborations with some communities to find out what kind of services they are expecting. Yet the stakeholders are still discussing which concept will be the best to address the eminent challenges posed by the data deluge and the need to optimize data sharing and re-use in the USA. Recently the leading persons in RDA US agreed to ask NDS to act as national testbed center for RDA results.

7. National Level Pillars

Also at the national level in Europe new organisational structures are being tested and established to meet the challenges of data intensive sciences.

7.1 Max Planck Society

In the Max Planck Society an IT Strategy Committee was founded a few years ago to come up with advice how to reshape the IT service structure in its organisation to maintain competitiveness of its research. With the introduction of parallel computers many years ago the Computer Centre in Garching got the task to provide not only high performance compute capacity but also to provide expertise in parallelising relevant domain specific software codes for simulation and analytics. In collaboration with domain experts such code was optimised allowing optimal use of HPC architectures. The optimal solution for such code parallelization was thus found by bringing together expertise and resources of each institute with central expertise and resources such as storage capacity and compute power. The strategy committee realized that the huge increase of data and the challenges of data intensive sciences require a new approach in so far as it makes sense to also provide central expertise and facilities in data management, curation and analytics.

As a consequence, the centre in Garching got a new name (Max Planck Computing and Data Facility, MPCDF) to indicate the change in focus, and was extended with data experts having expertise in mathematics and algorithms in typical data analytics applications which are widely discipline unspecific. The idea is to carry out collaborations between the centre and the various institutes and their departments that cannot invest in the specific knowledge required and that do not have the local resources to store and manage all data and to carry out the required computations.

We will use the NoMaD (Novel Materials Discovery) Repository project [44] which has been selected as one of the European Centres of Excellence projects as an example for the typical collaboration between a leading research institute in the MPS and its MPCDF centre.

Theoretical material scientists worldwide are doing experiments with a number of well-known chemical software packages (some at petascale performance) to compute possible characteristics for materials. These simulations are typically run on HPC machines after having carried out deep optimization of the software code tuned to certain architectures. Until now the resulting data has been used to write scientific papers, but was not considered valuable as such. This attitude is changing due to the fact that as in other research disciplines the researchers see a value in re-using data in different contexts, in allowing others to do new kinds of computations and to prevent doubling the work. The repository is meant to be a centre for storing results of simulation runs being identified by DOIs and described by proper metadata. Thus proper data organization and stewardship is basis of the work.

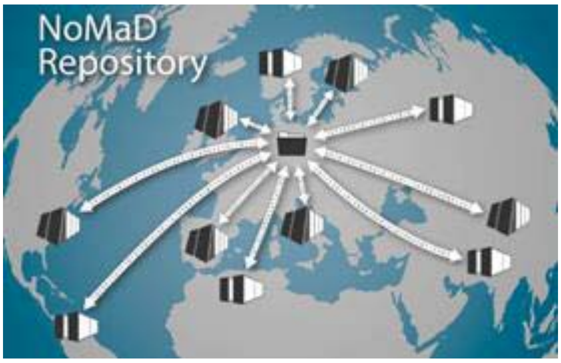

Fig. 7 indicates the intentions of the Novel Materials Discovery project (NoMaD) project to federate and aggregate all data about stemming from material experiments to enable easy access and re-use.

In collaboration with the researchers of the Fritz-Haber Institute the MPCDF experts are developing software to transform the incoming data to a normalized and compressed format, developing the repository software, the user upload, access and search interfaces, and the needed data management tools. In addition, novel analytic tools are being developed in collaboration between the involved centres to allow graphical searches, to carry out machine-learning based comparisons on data sets, to do smart visualizations supporting voyaging methods, etc. Typically all these operations on the aggregated data will be executed by making use of “trivial” parallelization techniques such as enabled by Map-Reduce methods on appropriate Hadoop clusters, i.e. the repository will be hosted at MPCDF and the computations will be carried out on computers offered by MPCDF.

7.2 Approaches in NL

Also in countries such as for example the Netherlands new strategies are being tested. In addition to strengthen domain specific centres of different types new centres have been established to structure the data landscape. DANS [45] and 3TU [46] have received the task to specialise on data management and curation. They should make use of the data services of the national data and compute centre SARA.[47]. In addition the eScience Centre [48] has been established to run collaborative projects where discipline experts and experts with centrally aggregated expertise are shared to meet the challenges of data intensive science. All these national service providers are requested to synchronise their activities to come to an efficiently organised eco-system of infrastructure pillars and services.

8. Conclusions

Data Intensive Science (DIS) is one facet of the digital change which we are currently experiencing and which will change not only science but also societies substantially. DIS which will be open to many to exploit its full innovative power and not exclusive to a few will depend on a change of culture towards open data and accessibility of services. In the European Union and its member states community-driven research infrastructures and e-Infrastructures tackling common cross-disciplinary challenges have been started to address the needs for an efficient eco-system of services enabling data intensive work. The US did not make this distinction, but under the term “cyberinfrastructure” also community-driven and more commons-driven projects were initiated.

After almost a decade of experience in infrastructure building it is obvious that there are still many social and technical barriers prohibiting efficient and cost-effective data usage and reproducible results. In fact one can argue that only active infrastructure building made many of the barriers visible to all stakeholders. The time period between the invention of TP/IP and its broad uptake to enable efficient communication between compute nodes took about 15 years. Several data scientists and infrastructure builders from mainly Europe, US and Australia agreed that it is time to accelerate the process of overcoming the many barriers for efficient data usage since waiting for another decade to overcome the most severe barriers is acceptable. Setting up the RDA based on similar principles as IETF (bottom-up, rough consensus, running code, lean governance) was the preferred choice of the data experts and this choice was supported by the funding organizations.

With this background in mind it is not surprising that almost all strong European infrastructure centres are very active in EUDAT as well as in RDA and that for example also ANDS and NDS engage actively in RDA. The Max Planck Computing and Data Facility for example will coordinate RDA Europe from September 2015, and its members are in the Technical Advisory Board, co-chairing the Data Foundation and Terminology and Data Fabric Interest Groups and are leading a Work Package in EUDAT, SARA and DANS for example are also leading activities in EUDAT and are actively engaged in RDA groups. NDS is co-chairing for example the Data Fabric Interest Group and ANDS is represented in the Council and Technical Advisory Board of RDA.

In addition to accelerating global agreement finding to improve data sharing and re-use and thus to enable inclusive data intensive science two main reasons can be mentioned for the engagement: a) engaging its experts in cutting-edge developments will make them fit for the coming challenges and b) bringing in their expertise will influence decision taking. So far RDA is too young to present final conclusions about the question whether the expectations were met.

We need to accept that the data landscape is changing rapidly and that new structures that have been set up to facilitate data intensive sciences are often still in a test phase. Essential questions in the data domain are still not fully answered yet such as: Which persistent structures need to be funded in addition to libraries that often do not yet have the skills to participate in the emerging data services domain? What is the optimal division between discipline specific and common services? What is the most optimal way to share specialised and expensive data experts that are scarce? Which are the common components that need to be specified to come to global, interoperable and well-maintained services supporting data intensive sciences optimally?

The EU and several of its member states as well as the US decided to take an active role to exploit the possibilities by taking concrete actions and by asking data science experts to develop and test out bottom-up driven models.

9. References

[1] MPI for Psycholinguistics, http://www.mpi.nl

[2] Tony Hey et.al., The Fourth Paradigm - Data Intensive Scientific Discovery, 2009, http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf

[3] John Wood et.al., Riding the Wave Report, 2012, http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

[4] Human Brain Project: https://www.humanbrainproject.eu/

[5] Herman Stehouwer, Peter Wittenburg, RDA Data Practice Report, 2014, http://europe.rd-alliance.org/sites/default/files/RDA-Europe-D2.5-Second-Year-Report-RDA-Europe-Forum-Analysis-Programme.pdf

[6] ESFRI Roadmap 2006, http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri-roadmap

[7] Open Access, http://en.wikipedia.org/wiki/Open_access

[8] http://research.microsoft.com/en-us/um/redmond/events/fs2010/presentations/michener_environ_data_mgmt_rfs_71210.pdf

[9] Data Seal of Approval: http://datasealofapproval.org/en/

[10] TCP/IP Protocol: http://en.wikipedia.org/wiki/Internet_protocol_suite

[11] Herman Stehouwer, Peter Wittenburg, Principles for Data Sharing and Re-use: are they all the same?, 2015 http://hdl.handle.net/11304/1aab3df4-f3ce-11e4-ac7e-860aa0063d1f

[12] Peter Wittenburg, Leif Laaksonen, Hermann Stehouwer, Raphael Ritz, Living with Data Management Plans, 2015 http://hdl.handle.net/11304/ea286e5a-f3d1-11e4-ac7e-860aa0063d1f

[13] ICRI 2012 Conference Copenhagen: http://www.icri2012.dk/www.ereg.me/ehome/index06e1.html

[14] Research Data Alliance: http://rd-alliance.org

[15] Naoyuki Tsunematsu, RDA plenary Keynote, San Diego, 2015: https://rd-alliance.org/keynote-naoyuki-tsunematsu.html

[16] Daniel Castro, Travis Korte, Open Data in the G8, 2015, http://www2.datainnovation.org/2015-open-data-g8.pdf

[17] Research Data Alliance - Europe, http://europe.rd-alliance.org

[18] RDA Case Statements, https://rd-alliance.org/working-and-interest-groups/case-statements.html

[19] ESFRI Projects, http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri

[20] EU e-Infrastructures, http://cordis.europa.eu/fp7/ict/e-infrastructure/

[21] EUDAT e-Infrastructure, http://www.eudat.eu

[22] Bernard Schutz et.al., RDA Europe Science Workshop Report, 2014, http://europe.rd-alliance.org/documents/publications-reports/rda-europe-science-workshop-report

[23] John Wood et.al., The Data Harvest, 2014, https://europe.rd-alliance.org/documents/publications-reports/data-

harvest-how-sharing-research-data-can-yield-knowledge-jobs-and
[24] RDA Data Foundation and Terminology WG, https://rd-alliance.org/groups/data-foundation-and-terminology-wg.html
[25] RDA Data Type Registry WG, https://rd-alliance.org/groups/data-type-registries-wg.html
[26] RDA PID Information Type WG, https://rd-alliance.org/groups/pid-information-types-wg.html
[27] RDA Practical Policy WG, https://rd-alliance.org/groups/practical-policy-wg.html
[28] RDA Metadata Standards Directory WG, https://rd-alliance.org/groups/metadata-standards-directory-working-group.html
[29] RDA Data Citation WG, https://rd-alliance.org/groups/data-citation-wg.html
[30] RDA Repository Audit and Certification WG, https://rd-alliance.org/groups/repository-audit-and-certification-dsa%E2%80%93wds-partnership-wg.html
[31] Data Seal of Approval, http://datasealofapproval.org/en/
[32] World Data Systems, https://www.icsu-wds.org/
[33] RDA Adoption Day, San Diego, 2015, https://www.rd-alliance.org/plenary-meetings/fifth-plenary/programme/adoption-day.html
[34] RDA Data Fabric IG, https://www.rd-alliance.org/group/data-fabric-ig.html
[35] Bridget Almas et.al., Data Management Trends, Principles and Components - What Needs to be Done Next?, 2015, http://hdl.handle.net/11304/33430f2e-f598-11e4-ac7e-860aa0063d1f
[36] ESFRI Roadmap 2006, http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri-roadmap&section=roadmap-2006
[37] CLARIN Research Infrastructure, http://www.clarin.eu/
[38] CLARIN Virtual Language Observatory, http://clarin.eu/content/virtual-language-observatory
[39] PRACE e-Infrastructure, http://www.prace-ri.eu/
[40] EC Open Science and Innovation, http://ec.europa.eu/research/conferences/2015/era-of-innovation/index.cfm
[41] National Center for Supercomputer Applications, http://www.ncsa.illinois.edu/
[42] National Data Service, http://www.nationaldataservice.org/
[43] Australian National Data Service, http://www.ands.org.au/
[44] NoMaD - Novel Materials Discovery Project, http://nomad-repository.eu/cms/
[45] Data Archiving and Networked Service, http://www.dans.knaw.nl/nl
[46] 3TU Data Centrum, http://datacentrum.3tu.nl/home/
[47] SURF Sara, https://surfsara.nl/
[48] Netherlands eScience Center, https://www.esciencecenter.nl/