

# Costs of FAIR Compliance and not being FAIR compliant

Peter Wittenburg, December 2017

This note is written from my own experience based on the data-driven research at the MPI for Psycholinguistics and infrastructure work mainly in DOBES, CLARIN and EUDAT. The note grew over time and is meant to react on a request to describe the costs of being FAIR and not being FAIR.

## Summary

- Ground-breaking research wants to use as much data as possible also from other disciplines. Efficiency of access to and re-combination of data is crucial for maintaining a competitive advantage. The scientific workflows incl. the data needed are not always predictable. Researchers are not interested in details as long as they can get a competitive advantage.
- Data practices are far from being FAIR and thus are highly inefficient, about 80% of the time can be wasted with data "wrangling". Max Planck Institutes can pay student assistants and have technical support so that they can afford data-driven research even when data needs to be integrated from other silos. However, even at MPI many research questions could not be tackled, since accessibility of suitable data was not guaranteed to be solved in predictable times at reasonable efforts.
- A proper data organisation, a high quality of data and FAIR compliance have shown that they can offer new opportunities to tackle new advanced data-driven research. But this does not come for free - running a larger and proper online archive with convenient access methods costs about 400 k€ per year enabling economy of scale factors.
- The closer infrastructures are to the research questions the higher are their values for the researchers, however, at risk of silo solutions. In almost all disciplines we have seen an enormous growth of silo solutions to data problems resulting in a huge solutions space and many software components which cannot be maintained. A turn from this "creolisation phase" to a "convergence phase" is urgently required<sup>1</sup>.
- In DOBES, encoding and metadata standards were jointly developed and the request from the funder to offer proper data to the archive worked very well. Only very few teams had problems to adapt their way of doing. So carrots and sticks both were applied to achieve high quality.
- In CLARIN and EUDAT it turned out that different or unclear data organisations and lack of quality were the biggest roadblocks for more efficient dealing with data. Mapping metadata semantics were not so problematic, although often the lack of explicitness of structure and category definitions are hampering efficient work.
- Bridging between semantic classifications and annotations will remain a challenge due to the dynamic nature of research, the dynamics of classifications and unhandy tools. Large ontologies seem to be unhandy and underused in daily practice.
- Currently, automatic workflow procedures are not used very often by researchers although they will be the only guarantee to improve the practices given the increasing volumes and complexity.
- Establishing intensive interaction platforms for the different actors involved in data work is crucial for improvements.
- A "level of commodity on top of which we can grow again" (J. Hendler) is urgently required to synchronise minds and to achieve momentum.

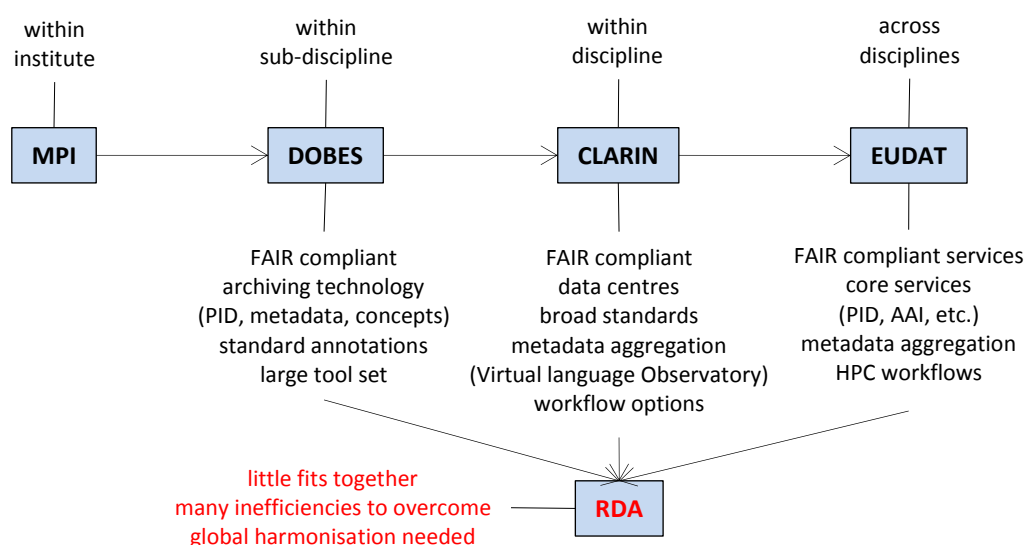
---

<sup>1</sup> The concept of separate infrastructures seems to be misleading. What is needed are common standards which all infrastructures adhere to and then automatically different types of services will establish - some closer to one discipline, others hosting services for different disciplines.

## Background

In the 1990s the Technical Group at the MPI for Psycholinguistics decided to switch to an all-digital domain, i.e. we started digitising all audio and video material from our field linguists, also started building our first annotation tools for PCs and tested first metadata descriptions anticipating a rapidly increasing number of digital files that will populate the PCs and servers. This change and new tools developed by the team motivated researchers to do new type of studies such as for example gesture research. Within two years researchers switched from an analogue style of working with audio/video equipment to a digital type of working with PCs. Due to this early work the MPI team became very well-known within the community for its ground-breaking work revolutionising traditional workflows. Due to this advanced role the team was very successful in applying for grants to foster its work which continued until 2014 when I officially retired.

In 2000 we were selected by the VW-Foundation to lead the technological/archiving work of the international DOBES project on documenting endangered languages. Also in 2000 we started looking for funds from the EC and other funders such as BMBF in DE and NWO in NL. Most influential wrt. our choices was the unique DOBES project where we had to help preserving a unique part of human heritage (languages & cultures) requiring a fundamental revision of our strategies and using our experience to design proper methods for data creation, management, archiving<sup>2</sup> and re-use. These methods found their way later into infrastructure projects such as CLARIN where I was responsible for implementing the technical infrastructure and EUDAT where I acted as scientific director. These methods were also being discussed in groups of the Research Data Alliance to become widely accepted now in many research disciplines.



My personal view of own developments of the MPI team is illustrated in the diagram above. These developments were driven by the enormous drive of psycholinguistics research to get access to as much data from within the discipline, but later also across disciplines when non-linguistic factors on human language behaviour and processing by the human brain became of high relevance for new findings.

The step towards RDA was done since in particular in EUDAT we realised the many differences in data solutions at all layers between the disciplines and within the disciplines often without a scientific reason for them. Solutions simply evolved within certain contexts.

<sup>2</sup> We did not make a deep difference between proper data management and stewardship (use of open standards etc.) on the one hand and archiving on the other hand except for creating "lots of copies" of our data to prevent data loss. We relied on proper and widely tested software components.

It should be noted here that in this document I do not speak about all the other data at MPI resulting from experimental studies. Processing of digital-born data such as speech, eye tracking, gesture tracking, virtual reality streams, brain images (eeg, meg, mrt), etc. were part of the daily work of researchers and the technical team. Until recently they were captured in file hierarchies without being FAIR compliant since their re-usage potential seemed to be small. This view changed during the last few years.

## Project Overview

Here, I want to indicate roughly the funding base for the work which the Technical Group did over the years. I will only mention those projects that helped to pave the way towards a FAIR<sup>3</sup> compliant data domain within the discipline.

### **2000 - 2012 MPI funding basis over all years: 3 FTE per year, i.e. 180 k€ per year**

- Here only those persons are mentioned that were directly involved to push ahead the "data management/archiving" work.

### **2000 - 2012 funded by VW Foundation (total funding for MPI: 1.5 Mio €)**

<http://dobes.mpi.nl/>

- start of DOBES project
- installation of Handle server - 5000 € (current costs) and running Handle service (little maintenance costs)
- building LAMUS archiving system (<https://tla.mpi.nl/tools/tla-tools/lamus/>)
- building XML-based IMDI metadata scheme (<https://tla.mpi.nl/imdi-metadata/>)
- building IMDI metadata tools (editor, browser, search etc.)
- building ELAN media annotation tool (open XML based annotation format) (<https://tla.mpi.nl/tools/tla-tools/elan/>)<sup>4</sup>
- building conversion + digitisation lines for audio & video material (open standard formats)
- building a variety of other tools (lexicon tool, joint content & metadata search, web-based annotation tool, etc.) (<https://tla.mpi.nl/tools/tla-tools/>)
- uploading data/metadata and data/metadata curation
- setting up remote archives (Argentina, Brazil, Mexico, Cameron, Moscow, Canberra, etc.)
- DOBES Archive had about 80 TB data of data
- total Archive at MPI had about 200 TB of data, not all being described by metadata

### **2000-2001 ISLE Project funded by EC (funding MPI part: 67 k€)**

<http://www.mpi.nl/ISLE/>

- building IMDI metadata set and tools

### **2000-2003 MUMIS Project funded by EC (funding MPI part: 273 k€)**

<https://tla.mpi.nl/projects/past-projects/mumis/>

- improving the annotation framework

### **2002 - 2004 ECHO Project funded by the EC (funding MPI part: 202 k€)**

<http://www.mpi.nl/echo/>

- further developing and improving ELAN annotation tool

---

<sup>3</sup> We should note here that "FAIR" was not a term used in the discipline, but from 1998 on we had broad and continuous discussions about standards in our field (LREC conferences) and later we accepted the DSA and WDS requirements for proper repositories.

<sup>4</sup> ELAN is the most widely used media annotation tool worldwide.

**2003 - 2005 INTERA Project funded by EC (funding MPI part: 130 k€)**

<http://www.mpi.nl/intera/>

- building the IMDI community
- improving IMDI tools
- developing perspectives for metadata for language resources

**2006 - 2007 DAMLR Project funded by EC (funding MPI part: 130 k€)**

<http://www.mpi.nl/DAM-LR/>

- unifying metadata, forming a jointly searchable metadata domain
- improve and unify archiving
- unify PID registration (install and use Handle resolvers)
- unify access based on distributed authentication and authorisation

**2008 - 2012 CLARIN Research Infrastructure funded by EC (funding MPI part: 609 k€)**

<https://www.clarin.eu/>

- CMDI metadata concept and CMDI tool implementation
- Virtual language Observatory (metadata portal based on metadata harvested worldwide)
- data curation
- new language tools
- work on standards
- support, help, etc.

**2011 - 2013 EUDAT Data Infrastructure funded by EC**

<https://eudat.eu/>

I only mention this project here since we continued to convince colleagues to accept generic standards, i.e. use of PIDs, use of schema based metadata, proper repositories etc. All EUDAT services are based on FAIR principles as far as possible.

## **Funding and Achievements**

The total amount of funding for 12 years of language oriented infrastructure building was about 5.182 Mio € with about 40% own MPI funds and 60% from external projects. In average we had about 431 k€ per year which is about 6 fte persons working on this.

### **DOBES Period**

In the **DOBES period** the following was achieved by the MPI team as the following diagram indicates:

- (1) The LAMUS archiving system was built including its IMDI related tools and a running PID service taking care of a proper data organisation including all relevant relationships (collections).
- (2) A complete and professional annotation tool was developed for annotating time series data (audio, video, eeg, eye tracking, etc.).
- (3) Archive access tools were added such as IMDI search function, a GIS integration for geo-browsing, the ANNEX tool to do web-based streaming and annotation analysis and the TROVA search tool that includes metadata and annotations.
- (4) A lexical component was added, but changed character over time.
- (5) Some other tools were developed but where not so widely applied.
- (6) The ISOcat<sup>5</sup> category registry was added to register domain concepts which was later turned into a SKOS based registry.

---

<sup>5</sup> <https://tla.mpi.nl/tools/tla-tools/older-tools/isocat/>



(7) For the archive dynamic synchronisation software was built that replicated every incoming file to 4 large German data centres and that copied selected collections to about 14 data centres in 12 different countries worldwide (based on LAMUS/IMDI).

All tools were built following open standards and explicitness of structure and semantics as far as possible. For the repository we participated in the Data Seal of Approval and World Data Systems evaluation procedures. Several of these tools are still working although the available funds dropped drastically after my official retirement in 2014.

In 2008<sup>6</sup> I presented the following numbers<sup>7</sup> for an MPI-like archive which in 2012 had about 200 TB of data of which were 80 TB in the online available archive, i.e. a FAIR compliant and accessible store. For the MPI the basic IT infrastructure (CPU, storage, network, etc.) had to be maintained anyhow to serve its researchers, adding some other TB of capacity up to 2 PB (disc space, tape robot) to cover the

online archive was not that expensive. Also system management had to be available and economy of scale factors allowed running the online-archive with the same basic system experts. In peak times of data pre-processing and ingest we had about 4 trained student assistants to help uploading data, curating some metadata, processing analogue devices for digitisation, etc. The cost indication in the following table does not cover such peak times. We had an "open archiving offer" which was used by many external people in so far as serious researchers could upload their data (and metadata) into our archive via the online tools making it immediately visible based on a simple request. Due to many online checks on formats and metadata conformance we could give such an offer. Due to the usual network bandwidth limitations there was no risk to overload our storage system.

type	k€/y	comment
basic IT infrastructure	80	4-8 years innovation cycle (up to 2 PB capacity in 2012)
digitization and workflow	10	new recorders, capturing, student assistants
copies at large centers	<5	4 external copies of about 80 TB in 2012
system management	60	shared for different activities
archive management	80	advice, curation, consistency, format checks, etc.
repository software maintenance	60	without new functionality
utilisation software maintenance	>120	dependent on spectrum of tools
building, energy, etc.	?	not counted here
<b>total</b>	<b>415</b>	<b>rough sum to run an existing archive</b>

How do these numbers in the table relate to the available funds not counting basic IT infrastructure, system management and peak load times for data ingest? From the reported roughly 6 fte persons we spent in average 1.2 fte on archive management, 1.8 fte on repository software maintenance and functional updates, 1 fte on the ELAN Tool maintenance and functional extension, and 2.0 on the maintenance and extension of the remaining software stack.

### CLARIN Period (first 3 years when I was involved)

Together with other collaborators the CLARIN time was used to extend the above mentioned and other methods to more generic, discipline wide solutions. CLARIN worked on major issues such as

<sup>6</sup> <http://www.alliancepermanentaccess.org/index.php/community/conferences/apa-conferences/apa-2008-conference/>

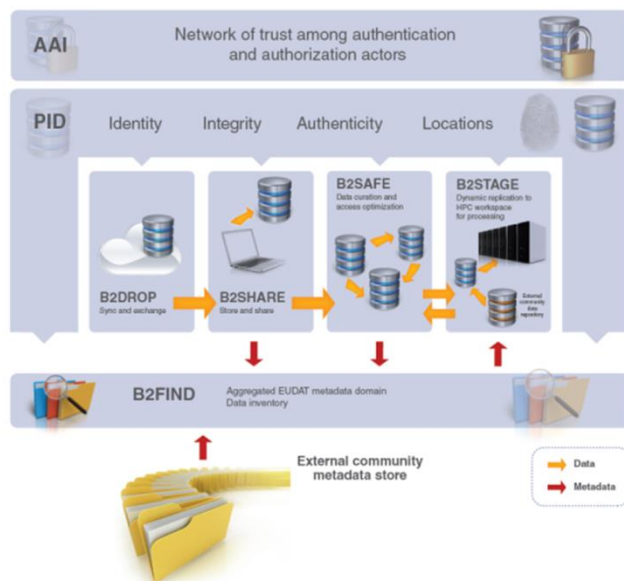
<sup>7</sup> At that time I could base the average costs for a person-year on 60 k€, since building, energy etc. was available. After 2008 the costs for trained personnel were increasing.

- making language resources visible and accessible
- developing discipline standards for data and support conversion
- developing standards and requirements for data repositories to participate in the CLARIN network (DSA, WDS)
- developing a component based metadata system to cope with the varying requirements of the sub-disciplines
- developing the ISOcat concept registry as basis for semantic explicitness and interoperability in CMDI
- aggregating all metadata worldwide about language resources and building the Virtual Language Observatory portal with about 800.000 records
- working on workflow machines available via the web<sup>8</sup>
- giving help and support

Also these components assist in making the CLARIN infrastructure FAIR compliant as far as possible. The major tools are still maintained and are being upgraded and extended. CLARIN has received the status of an ERIC at its 3<sup>rd</sup> year of existence and was the second ERIC.

### EUDAT Period (first 3 years when I was involved)

EUDAT developed a number of services that had been selected and specified by the representatives of the 5 participating and driving disciplines (biophysics, plate tectonics, climate modelling, languages, biodiversity). In addition, a few



core services (PID service, AAI service, etc.) were developed enabling the scientific services. While the B2DROP service is similar to a dropbox service where FAIR principles cannot be applied, all other services were designed as FAIR compliant services. B2SHARE was meant to allow researchers to upload data of limited size; at upload time a PID and metadata were associated. All metadata in the EUDAT data domain and beyond were aggregated and made available via the B2FIND portal. Difficult to realise were the B2SAFE service to replicate large collections and the B2STAGE service to include HPC pipelines in the domain of registered and thus FAIR compliant data. All data organisations<sup>9</sup> found in the different

disciplines and also between the repositories in the disciplines were widely differing requiring the development of special adaptors for each replication task. This did not scale and no one can maintain a lot of adaptors. In a later chapter I will discuss one concrete example for a problematic data organisation. From this work it became obvious that there is an urgent need of a unified API.

For the B2STAGE service the big problem was that once data is being computed to new data in HPC pipelines the history of the data is being lost and needs to be re-built afterwards which in general

<sup>8</sup> [https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main\\_Page](https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page)

<sup>9</sup> With "data organisation" we mean the way the various relationships between related entities such as data, metadata, provenance, PIDs, rights information, collection information (relationships between data) and structural/semantic information. How do I find the metadata of some data? All repositories and data models used (file systems, cloud objects, SQL databases, No-SQL databases, etc.) have different ways to express these relationships and often they are not explicit at all.



does not happen. Data made available via a B2SAFE service was pushed into the HPC workspace to carry out some computation and the new data being created by the computations should be returned from the workspace to the registered data domain with proper metadata including provenance and PID registration. Methods were implemented that showed that FAIR compliant workflows were possible, but it is a long way to get this into the daily practices at HPC centres.

The conclusion from the CLARIN and EUDAT work was that only little fits together in a way so that data work becomes efficient.

## Statements on Costs and Inefficiencies

In 2016 **RDA Europe** published an analysis called "RDA Europe: Data Practices Analysis<sup>10</sup>" which covered in total 50 interviews and 80 intensive discussions with experts from various disciplines and types of organisations carried out by the German RADIESCHEN and the two EC-funded projects EUDAT and RDA. I will describe the main results of this analysis:

- Data Management and Processing is too time consuming and costly due to the heterogeneity in how data is organized in particular with respect to logical information. Researchers see the need to change habits, but do not have agreed suggestions for solutions.  
*many researchers stick to file systems – often at its limits*
- Federating data including logical layer information which is relevant for tracing provenance, for understanding creation context, for checking identity and integrity, etc. is so costly that in practice it is not done, although most data professionals understand that this practice cannot be continued like this.
- Data Management and Processing is not ready for Big Data due to the lack of usage of automated procedures incorporating proper data organization mechanisms.  
*too many ad hoc scripts without proper documentation are used*
- Due to a lack of software that is supporting proper data organizations we continue to create legacy data that cannot be integrated easily into our growing domain of accessible data.  
*most disciplines have massive problem with legacy, we still create legacy*
- When we only consider that for example one of the key biologists in a large research institute is spending 75% of his time for manual data management, we can estimate how much money and human capital is wasted by the way we are doing.  
*huge waste of money and skills*
- When we look at all this it is obvious that we need a change of data organizations and data sharing/management procedures.
- We need to start with training our young people in how to overcome this situation.
- Most of the software developed in communities is not maintainable.

**Michael Brodie** from MIT<sup>11</sup> reported at the DAMDID conference in Obninsk<sup>12</sup> in 2015 about a study in the US where it turned out that 80% of the scientist time in data driven projects is wasted with data wrangling<sup>13</sup>. The difference to the RDA finding can be neglected. At the Big Data Summit 2017<sup>14</sup> industry experts reported about 60% of the costs of data driven projects on data wrangling work. Since data work in industry is still more streamlined in many cases as in science the difference is understandable. The change from the term "data warehouse" to the term "data lake" in industry,

---

<sup>10</sup> <http://hdl.handle.net/11304/6e1424cc-8927-11e4-ac7e-860aa0063d1f>

<sup>11</sup> <http://michaelbrodie.com/>

<sup>12</sup> <https://10times.com/damdid-conference>

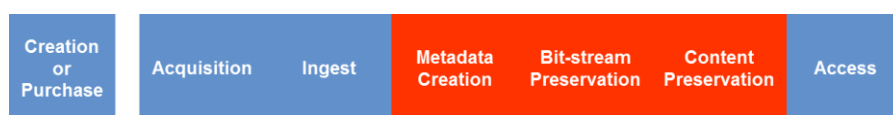
<sup>13</sup> The term "data wrangling" is used for all types of data aspects and work that needs to be done before the real analysis can start.

<sup>14</sup> <https://www.big-data.ai/de> (reference to the 2017 conference could not be found)

however, indicates that also in industry things are changing dramatically. For data warehouses all data available is integrated into one big database to enable query based analytics. It's the complexity and the dynamics of the data with many different types and from different silos that makes data integration a challenge requiring new methods.

To give an impression about the costs we can compare approaches about **cancer research** at a well-known and professional German institute with a similar one in the US. At the German institute researchers did not move to semi-automatic workflows since there are too many exceptions and parameter variations to be considered, thus they rely on ad hoc scripts and manual management steps with the consequence that 75% of time is spent on data management. In the US institute 2 IT experts were hired for 3 years to work closely with the researchers and develop a flexible workflow system. Thus costs in order of about 600 k\$ or more were invested to reduce the wasted time on data management. I do not know whether the workflow concept has been a success and how much money is spent on maintaining and updating the workflow system.

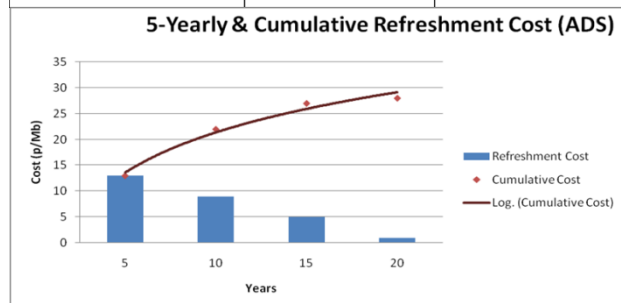
At the APA conference in 2008 in Budapest<sup>15</sup> **Neil Beagrie and Keith Jeffery** reported about various cost aspects. Beagrie differentiated between three phases in data management: the first three steps making up for 42% of the effort/costs, the three red marked tasks for 23 % and the access for 35%. The last number very much depends on the level of access given, here only basic access (disseminating the files) can be meant.



Institutional Repository (e-publications):	Staff	Equipment (capital depreciated over 3 years)
Annual recurrent costs	1 FTE	£1,300 pa
Federated Institutional Repository (data): Annual recurrent costs	Staff	Equipment (capital depreciated over 3 years)
Cambridge	4 FTE	£58,764 pa
KCL	2.5 FTE	£27,546 pa

Beagrie also compared the costs of a few repositories in UK. It required 1 FTE person to run an e-publication repository. Running two data repositories required much more effort and the difference between the required staff indicates that the effort very much depends on the ambitions and the size of the repository.

Beagrie also gave a good indication that it would require about 30 times more effort to create metadata after 10 years compared to creating metadata immediately.



Not surprising are his numbers about refreshment costs, since the costs per byte are decreasing rapidly. But as has been indicated earlier for MPI the replication of 80 TB at large scientific data centres in Germany that are used to deal with capacities of hundreds of PB are comparatively low.

In this realm we can report about typical storage **costs in industry**. Cloud storage costs at Amazon/Google/Microsoft are about 7 k\$ per 5 TB per 5 years which includes just a basic service. For the MPI archive with its 200 TB in 2014 this would mean 280 k\$ for 5 years just for the storage

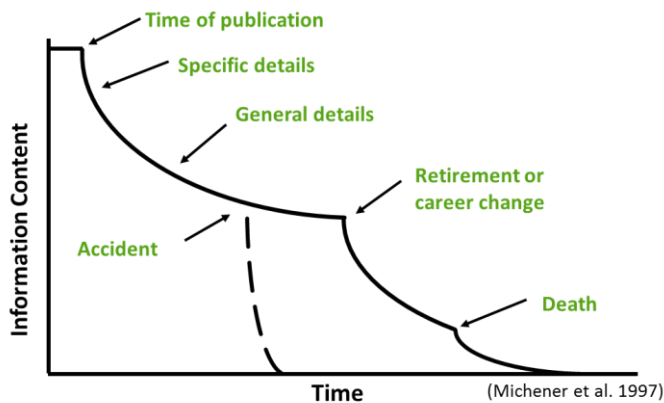
<sup>15</sup> <http://www.alliancepermanentaccess.org/index.php/community/conferences/apa-conferences/apa-2008-conference/>



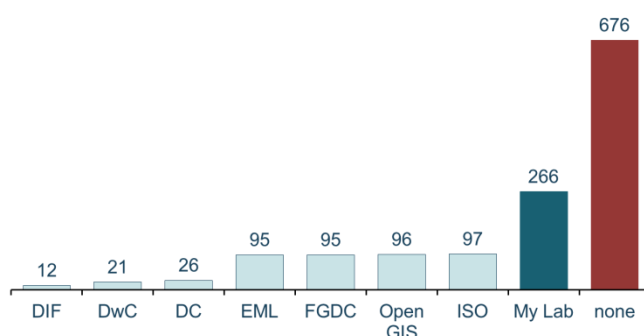
without advanced services as requested by the researchers. In the MPI case where a server/storage system has to be maintained anyhow to store the various forms of data (experiments, brain images, etc.) the 80 TB of the online archive just meant to add disc and tape space.

A data repository as Dryad is requesting about 25 k\$ for 5 TB per year offering its archiving and access services including metadata support which also indicates that the developments at MPI were justified.

**Bill Michener** from the US-based DataONE project reported about two interesting phenomena. The



first diagram about "data entropy" indicates the rapid loss of information about data over time which is well-known for almost all scientific institutes. In particular when data creators (PhDs, researchers) leave an institute, there is hardly any information left about their data. Even if the researchers created some metadata mostly in form of spreadsheets it is hard to find them and if they are found, it is even harder to interpret them.



The second diagram reflects about the practices in metadata creation. Only in about 30% of the cases researchers used well-defined metadata semantics to describe their data. More than double as many use some self-defined metadata (often in spreadsheets) or even no description at all. For data that is being used in a small research group this already may create inefficiencies. Re-using such

data beyond the silos becomes very inefficient.

One **researcher** leading one of the DOBES teams which had between 2 and 4 international collaborators reported about a change of practice in his team when they discovered that they could not manage anymore their data set existing of more than 6000 files with different types of data, with different versions and partly fragments of other data. They started to use the DOBES archive more actively as a reference platform, i.e. relevant versions were uploaded at an early stage including the necessary metadata.

I should at least present one example from **EUDAT** that explains that different organisations of data lead to non-scalable solutions. One one-line database with excellent information and about 11 TB had to be replicated into the EUDAT domain. This database evolved as so many over the years and was developed by excellent and engaged researchers. However, data was scattered in file and content management systems and metadata was partly even hidden in scripts. We had to invest roughly in 3 weeks programmer time to replicate most of the data according to the norms of EUDAT and to make them FAIR compliant. This was just one of the thousands of databases which are out there. How to replicate or federate such databases and how to maintain all the different adaptor scripts? This example can be compared to the FAIR compliant and properly organised DOBES archive where we simply needed to synchronise all files from a certain root with the help of rsync and/or Andrew File System. A simple IMDI XML file crawler at the replication site would be sufficient to recreate the browsable and searchable archive - an almost neglectable effort once tested.

## Major Reasons for Inefficiencies

From my experiences in almost 15 years of data infrastructure building the biggest factors for inefficiencies were as follows:

- unclear and non-explicit data organisation
- unclear rights
- lack of quality
- lack of structural and semantic explicitness
- semantic mapping
- knowledge extraction from data

The relevance of a proper **data organisation** is often underestimated and has to do with our tradition to work with files that have names, exist in directory structures and are self-supporting. As explained earlier it is about the relationships between related entities such as data, metadata, provenance, PIDs, rights information, collection information (relationships between data) and structural/semantic information. Currently many like to speak of Digital Objects that are assigned a PID that leads us to a resolution system that can resolve a PID into information pointing to these related entities. There are may be other ways doing it, but it is of great importance for the complex data landscape with its varying solutions that this problem will be solved in a systematic and systemic fashion. When federating data, i.e. integrating data from different sources, it is crucial to find all these information entities in particular when we apply automatic workflows. Currently, nothing fits together and in many cases even the creators do not know where to find the various information types.

In many cases the **rights on data** are unclear which is a barrier for re-use in particular in industry. Many interesting projects are simply left un-tackled since no one wants to take risks of being accused. Often it takes time and requires expensive advice from lawyers to find out what the rights are and how to find the rights holders to negotiate about usage. This area has even a number of additional dimensions such as coming to usage agreements, defining licenses, taking care of liability issues, etc.

Often there is no **explicitness of structure and semantics** applied in data and metadata, i.e. there are no schemas and the semantic categories are not defined or the specifications are hidden somewhere on a personal storage system in some form. In general it requires much effort to find out what the underlying structures and meanings are. Often projects of relevance will simply not be done since the effort is estimated as too high. In many cases semantics have been defined, but it is difficult to find the definitions. Using generic web mechanisms may work for humans, but not machines.

**Lack of quality** is another obstacle for efficient re-use. Even in case that structure and semantics are being specified and can be accessed, it is often the case that data and metadata require a considerable curation effort to make it usable and to allow a formal mapping to other structures and semantic spaces to enable joint analytics.

FAIR principles request rich metadata which are indeed important when data is going to be shared and re-used. Poor metadata is the reality and can also be categorised under the header "lack of quality".

The "**semantic gap**" has been mentioned frequently and does not need much comment. I would like to point to a difference between semantics in metadata and data. In my experience we can manage the mapping between different metadata sets quite well, however mapping semantic spaces embedded in the data is a big challenge and will remain a big challenge. The semantic scope of metadata is restricted and researchers use metadata as a tool to do selections, searches etc., i.e. metadata concepts are not crucial to underpin a certain theory. Most metadata concepts are easy to

understand and even in the case of specific semantics describing discipline categorisations of data robust mappings can be done to support searches or selections.

**Semantic categories** describing phenomena such as annotations of data streams are inherently bound to the research question and the theory a researcher is developing. Simple mappings between category systems are often problematic and require fine-tuning by the researcher who wants to do data integration to carry out a certain research task. Yet we do not have flexible enough facilities to enable this kind of research. Ontologies that have been created by excellent and specialised teams are often complex and inflexible and thus heavily underused. In our domain it was the separation of category definitions (ISOcat) and relationships between the categories that allowed us to make steps, since relationships are often dependent on the task to be carried out.

The special case of **automatic processing** should be mentioned here. Metadata describing a digital object in such a way that a procedure can decide if it can operate on some data needs to be very specific. Therefore, there is a trend to associate "types" with each digital object and register the operations that are possible to process specific type. The existing MIME type concept does exactly this, for scientific data this concept needs to be extended.

Another big issue closely related with the semantic categories describing phenomena is how to **extract knowledge** from data streams and how to present that knowledge. There is so much data that there is an increasing need to (automatically) extract knowledge in form of assertions and to analyse (statistics, reasoning) these assertions manually or automatically. This task will keep us busy due to the dynamics in the semantic descriptions which are relevant for both the subject of analysis and the result of the analysis.

### Summary

Most of the barriers towards efficient data processing with the exception of the deep semantic challenges described at the end can be dealt with by a new systematic and systemic approach. Tackling these problems consistently would allow us to focus on the huge deep semantic tasks. It is time to define methods and to develop software that help removing these many barriers in a systematic way. It is time to look at the whole system and identify the weak points. To give an example: there are many schema and vocabulary registries out there, but how to find them and how to know which one is the right one, etc. A systemic approach is required.

## Value of the Changes

As Keith Jeffery pointed out convincingly in his conclusions, it is possible to estimate costs for creating and maintaining data/information, but it is more or less impossible to estimate the value. We can measure the impact of certain databases for improving our global understanding of certain phenomena such as for the Protein Databank or the Voices of the World database. In some disciplines it may even be possible to calculate the value of data in terms of benefit in economy who may have used the PDB or one of the many other databases around.

In science we can speak about impact based on citations or examples. Let me here make a few statements of the impact of the DOBES database and being FAIR compliant:

- DOBES was a blueprint of how to change the field globally and how FAIRness has an immediate impact on research, many other archives have been set up since then following similar principles
- Due to the global availability and in particular the ease of accessing data new types of research questions could be addressed which were impossible beforehand
  - comparative studies between languages for example comparing intonation patterns based on data could be carried out

- the development of languages throughout thousands of years in certain areas could now be based on large feature matrices extracted from data
- Offering FAIR compliant data allowed many more researchers globally to work on endangered languages in particular also in the countries where the languages are spoken.

Such examples do not answer the question about the value of data since this is very much dependent on societal processes. As is known from current US politics for example the value of data on climate change is rated very low despite the huge costs of the increasing amount of nature catastrophes. Certainly the value of data capturing cultural heritage phenomena is rated much lower as data about climate change in many countries.

If we could reduce researchers' involvement in data wrangling from 75 % to for example 30% much time would be gained to do other work. For sure more studies could be done in the same time. We could, however, not say if this would speed up insights about cancer and reduce the costs spent by societies on cancer treatment etc.

We could also argue that efficient treatment of brain diseases such as dementia which seems to be the number one source of costs in health would not be possible without the access to huge amounts of data. The worldwide costs of dementia were estimated by a US survey to \$818 billion in 2015, an increase of 35% since 2010; 86% of the costs occur in high-income countries. We know the costs of dementia treatment globally (\$818 billion in 2015<sup>16</sup>) and it is increasing. Therefore, if projects such as the Human Brain Project, for example would help to decrease this number substantially due to new insights which need to be based on processing huge amounts of data, we could calculate the cost/benefit ratio. What would be the impact of creating FAIR data in this case? We cannot identify this yet precisely, but we know that efficiency would be increased enormously and we also know that research would be democratised, i.e. more researchers and even citizen scientists could participate.

## Some Side Remarks/Questions

Finding solutions to what has been described above mainly as technical issues is actually much more of a **social and also economic challenge** than a technical one. This can best be explained by a number of exemplary questions:

- How can we establish new trust relationships when changing current mechanisms and workflows?
- How can we define brokering frameworks to sort out legal, usage, etc. issues and who is willing to pay for support?
- How much does it cost to build and maintain adaptors to generic interfaces that will emerge and to stepwise adapt the organisational structures of repositories? Who is willing to spend money on this transition and what will convince decision takers?
- Do we have the trained personnel to manage the changes?
- How dynamic do data management/stewardship plans need to be to make sense in a dynamic scientific lab situation? Who is willing to make the effort and who is going to pay for specialists, since it will have to be taken from the research funds?
- Who is going to do the work to make data FAIR compliant - the scientists need to publish new results and can't invest even more work than the 75% mentioned?
- Should we not better turn over to support ready-made code snippets that can be integrated into workflows and event to adhoc scripts? How can we convince people to change practice?
- How to come to a global eco-system of data infrastructures tackling the need for systematic and systemic approaches evolving from the many suggestions already existing?

---

<sup>16</sup> <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5232417/>

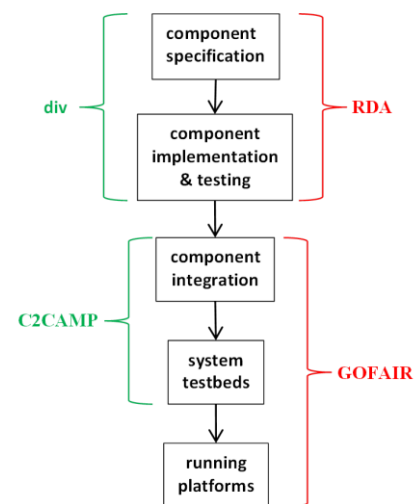
This note will not give answers to these questions. But when discussing cost aspects we need to understand that the way towards an efficient eco-system of data infrastructures will be expensive. The question is not whether we should start spending more money, but rather why we do not start now to invest more money, since the data volumes and complexity are increasing and without a solid basis we will end up in a "data tower of Babel".

Now that we tried so many different solutions for scientifically driven and IT driven data infrastructures it is time to accept that the primary task is to create a fundament for the future which is primarily an advanced organisational/logistic challenge<sup>17</sup> with all its communication and engineering dimensions: putting the FAIR principles into practice and let "data come out of pipes seamlessly as water" is a costly global enterprise. Defining a solid basis will allow us to tackle deep scientific aspects such as advanced semantic processing and knowledge extraction at a professional level. Also in the early days of Internet, researchers were also busy to design their own computer networks instead of spending all their time on research issues since there was nothing they could use. This changed after a few standards were simply accepted as common basis such as Ethernet and TCP/IP. Networking Digital Objects in suitable ways that are stored in distributed data repositories is the challenge for future management and processing.

## EOSC Funding Aspects

Here I will mention a few conclusions for EOSC funding which are based on the assumption that we are in a phase of dynamic changes where we cannot predict the characteristics of the next more stable phase. EOSC funding needs to have the following characteristics:

- it will have to be a long term funding and not one big amount of funds spent once
- this will include hard- and software which needs to be rethought based on digital object virtualisation
- technological developments will have to follow an agile process and also funding needs to respond to this agile way
- component developments and flexible testbeds will have to be funded and evaluated since they will be crucial to tackle the big challenges
- discussion processes need to be organised at least at three levels beyond the usual lobbyisms with global and technologically neutral characteristics as depicted in the following diagram<sup>18</sup>
  - Specification of components and their characteristics – here a.o. RDA has been set up as a bottom-up driven initiative and needs to be used more efficiently, as a bottom-up initiative it will find efficient ways due to self-organisation
  - For the implementation level we do not yet have a global interaction platform, but GOFAIR could be extended to be accepted broadly - it also has the bottom-up characteristics that are needed to bring the implementers together



<sup>17</sup> We can compare the challenges with earlier big infrastructure plans starting with the Roman water supply infrastructure.

<sup>18</sup> This diagram was created as a response to an excellent workshop in the US. C2CAMP is a global initiative to build an extensible testbed that was presented at the US symposium:

[http://sites.nationalacademies.org/pga/brdi/pga\\_181009?utm\\_source=BRDI+Mailing+List&utm\\_campaign=a9d7d34e0e-EMAIL\\_CAMPAIGN\\_2017\\_10\\_16&utm\\_medium=email&utm\\_term=0\\_5b187d867a-a9d7d34e0e-129095337](http://sites.nationalacademies.org/pga/brdi/pga_181009?utm_source=BRDI+Mailing+List&utm_campaign=a9d7d34e0e-EMAIL_CAMPAIGN_2017_10_16&utm_medium=email&utm_term=0_5b187d867a-a9d7d34e0e-129095337)

- Nevertheless a third layer will be needed that exists of experts from various disciplines, sectors and background that have a good oversight about the components and their state, the test and testbed results to make evaluations and recommendations - this will be the needed platform to give critical comments to the self-organising processes, to push convergence and to indicate what could be seen as the appropriate "level of commodity on top of which we can grow again" as Jim Hendler (RPI) stated it<sup>19</sup>.

## Acronyms

MPI	Max Planck Institute
DOBES	Dokumentation von Bedrohten Sprachen (Endangered Languages)
EUDAT	European Data Infrastructure
CLARIN	Common Language and Technology Infrastructure
VW-Foundation	Volkswagen-Foundation
BMBF	German Ministry for Education and Science
NWO	Dutch Science Organisation
RDA	Research Data Alliance
LAT Tools	Language Archiving Technology (some are still maintained)
LAMUS	Complete Archive Software (similar to scope of D-SPACE)
IMDI	complete Metadata Framework (schema, categories, search, browser, etc.)
CMDI	new component based metadata solution
ELAN	most widely used data series (video, audio, EEG, etc.) annotation tool
LEXUS	Lexicon tool
ARBIL	data organiser
ANNEX	web-based media annotation visualiser
TROVA	integrated metadata and annotation search tool
VICOS	building ontologies and relating concepts to lexica and annotations
VLO	Virtual Language Observatory - a big metadata aggregation portal
ISOCAT	ISO compliant category registry
SKOS	Simple Knowledge Organisation System (W3C)
DSA/WDS	Data Seal of Approval, World Data System
ERIC	Legal Entity Framework for European Research Infrastructures
B2xxx	Service Family developed in EUDAT
B2DROP	dropbox like service
B2SAVE	upload of files into repository with access via PID or metadata
B2SHARE	replication of large collections via special adaptors
B2STAGE	getting data into HPC workspace and carry out computation
B2FIND	aggregation of all metadata in EUDAT domain and search
B2HANDLE	registration and resolution of PIDs
B2ACCESS	distributed authentication service
APA	Alliance for Permanent Access (EU project)
FTE	Full-time employee
Rsync	Internet protocol to synchronise data
C2CAMP	global testbed project to foster FAIR compliant data infrastructure work
EOSC	European Open Science Cloud initiative
GOFAIR	an initiative to foster science to create FAIR compliant data
RPI	Rensselaer Polytechnic Institute

<sup>19</sup> Industry take up would change the game again as in the case of Internet and the WWW, but yet we cannot see industry relying on global best practices.