Report Industry Side Meeting at the 11th RDA plenary in Berlin

Peter Wittenburg, Ramin Yahyapour, Edit Herzcog, Sebastien Ziegler, George Strawn, Ana Garcia Robles

The Industrial Side Meeting was a continuation of our effort to foster the interaction between the Research Data Alliance and Data Industry, which started in September 2015 at the Plenary in Paris. The side event had different tracks such as (1) a joint IEEE-RDA hackathon/workshop on metadata management, (2) training type of sessions on machine learning and real time data[1], (3) a session with winners of the "Smart Data Use" competition, (4) a session with well-known experts about starting data businesses and (5) a day staring with a keynote from George Strawn followed up by 4 panels on different topics with representatives from RDA and industry. Sessions 1 to 3 will not be covered in this report.
In chapter 1 we present an executive summary of the meeting, in chapter 2 major observations, in chapter 3 short reports about the various sessions and in chapter 4 we discuss some terms being used in this report.

We would like to thank the following experts for their excellent contributions: Wolfgang Dorst, Georg Wittenburg, Nicolas Zimmer, Mark Hahnel, Ulrich Schwardmann, Zhou Jian, Christel Musset, Klaus Tochtermann, Milan Petkovic, Hanna Niemi-Hugaerts, Thomas Hahn, Hassan Chafi, Abdel Labbi, Larry Lannom

1. Executive Summary

This meeting was another moment to continue the interaction between RDA and industry about data related issues during the last 3 years when the third RDA Europe project was funded. The contributions gave an excellent overview about the topics where experts from science and industry widely agree and where they differ.

From both sides it was stated that the field of data is changing rapidly and that this dynamic development will continue. These dynamics result in huge inefficiencies and high data management and integration costs for the service users and in high development and maintenance costs for the technology providers, since they need to offer what customers are requesting. Therefore, it is not surprising that across all sectors there are attempts to reduce costs with differences in the approaches. In science only a bottom-up and cross-border (geographic, disciplines) approach is seen as successful as started in the Research Data Alliance which, however, need to be complemented by steps towards convergence as for example initiated with the focus on Digital Objects. In industry comprehensive reference architectures and integrating frameworks and platforms are being worked out to save investments. Both approaches are meant to identify and specify salient components as pillars of the emerging eco-system of infrastructures. Identifying and specifying components often beyond the work of the traditional standardisation organisations are the place to meet. Some suggestions for such salient components such as for global identification are already being discussed and implemented.

There is much talk about data economy and data markets and there is already an exchange of data in science and industry. But it is obvious that we are far away from a bazar like data market where data

[1] https://www.technologiestiftung-berlin.de/de/blog/echtzeitdaten-die-wirklichkeit-in-ihrer-ganzen-komplexitaet/

creators offer data products to facilitate the reuse of data in different contexts. Big industry relies mainly on bilateral interactions and projects, while in science there is a clear trend towards a more open scenario. It seems that also start-ups and SMEs would profit from a more open market. However, an open market would require that data is being associated information such as with persistent identifiers and metadata to enable data finding, accessibility and traceability as is currently pushed ahead in science by RDA and the FAIR principles[2].

Regulations as the new GDPR are a fact and need to be put in action. Experts seem to agree that at short-term these strict regulations will require investments and may hamper fast innovation. It needs to be shown whether this disadvantage will turn into an advantage in the long run. It is obvious that much more effort is required in Europe to overcome the many differences in legislation to make the required single digital market reality. In areas where sensitive data is being created such as in healthcare, too many barriers are still in place that prevent progress in exploiting the rich data.

A flourishing data market will depend on new mechanisms

- to improve trustworthiness of all actors and most probably to establish a layer of mediating brokers,
- to overcome gaps in data literacy to deepen the awareness about the new opportunities and to reduce the fears, and
- to improve data quality probably supported by increased certification regulations.

Summarising, we can state that this side meeting at the RDA plenary showed that there are many reasons to establish an interaction platform between interested experts from data science and data industry, since the common interests especially in the area of infrastructure building (components) and legislation became obvious. Such interaction needs to respect the differences as they can be seen for example with respect to licensing.

2. Observations from the Industrial Side Meeting

A number of observations seem to be widely agreed amongst the speakers, others will need more debates.

2.1 Field Dynamics

A wide agreement about the dynamics in the field of data could be sensed which one speaker expressed as "the universe of data is changing rapidly". These dynamics have a range of implications:

- There is an enormous proliferation of solutions in form of tools, formats, platforms, etc. and there is no evidence that this development will stop. In contrast, the transformations in nearly all scientific disciplines and also in industrial sectors are ongoing and the deployment of smart devices everywhere will also put the challenges to its extreme.
- As the keynote speaker expressed it, we are in a creolization phase with an increasing amount of interesting solutions optimized for specific purposes.
- The innovation lifecycle is very short, while many of our processes are rather slow in comparison. Thus, there is a major challenge to provide tangible answers and pragmatic solutions when needed. This poses a risk to companies but also to research. Solutions might be obsolete if they arrive too late.
- The consequences for customers are that it is difficult to take decisions about investments, that they continuously need to import/export data between tools/frameworks which is very inefficient, and that they are faced with interoperability problems also leading to huge inefficiencies[3].

[2] Different models are being discussed in industry.

[3] Various surveys from academia and industry clearly indicate that 80% of the time on data projects is lost with data wrangling which gives an impression of the huge inefficiencies we are confronted with.

- Also for technology providing companies enormous costs are the result since they need to include the heterogeneity of solutions by developing and maintaining all sorts of adapters and integration kits.
- However, companies need to act in a pragmatic way to offer what customers are requesting which at the end will drive the costs for data projects. Technology providing companies do not "see" the costs, but of course those who are dependent on smart analytics or want to do business with data.

2.2 Convergence

Despite the fact that there are so many different solutions, we can see the wish from all actors to take convergence steps. The differences are about the way this could be achieved.

- People from the science domain and RDA speak about the need to follow the FAIR principles of data, the need to use global resolvable persistent identifiers, the need to define a new interoperation layer comparable to what TCP/IP and HTTP achieved, and the need to identify salient components and specify their characteristics and interfaces. In science the FAIR principles and the RDA specifications around persistent identifiers are widely agreed now and a global PID service is available to support identification.
- People from industry speak about the need to develop comprehensive reference architectures to understand the complexity of the overall task and break it down, integrative frameworks that are flexible to be adapted and extended, and also integrating platforms for a selective community where companies define the rules of the game.
- Everyone speaks about the need for standards knowing that they are not sufficient and that in dynamic scenarios they need to be pushed forward by practical needs, i.e. relying on a bottom-up drive. Standardization is hitting on a moving target which does not allow lengthy decision processes. Since data integration is expensive, there is no alternative to finding more convergence, however.

In so far it is probably correct what the keynote speaker stated that we are close to steps of convergence to start a new level of exploitation, yet not knowing or agreeing at which level this convergence could be achieved.

2.3 Data Economy and Market

Speakers described various scenarios in which data is being exchanged or traded between creators and consumers; however, we are far away from a bazar-like data market.

- In science there are clear intentions to create a much more open data market where creators offer their metadata so that others (incl. industry) can search for what is out there. At the end this will lead to a bazar-like situation.
- In situations where much data is needed such as in digital health researchers would dream of such a situation, but regulations and legislations currently form severe hurdles. It is not obvious whether these hurdles also hold for descriptive metadata to inform others about useful sets which would be a minimal solution.
- Industry seems currently to often interact with partners at bilateral level to discuss opportunities. New models such as sectorial ecosystems, working with data incubators including industrial data sharing activities and platforms for data sharing are being investigated for example in the realm of BDVA.
- As we heard there still seems to be much anxiousness to deal data, in particular mission critical data, which points to a large trust gap. A mixture of education/advice, certification, introduction of new roles and more advanced technologies such as blockchain to trace re-usage based on smart contracts will be necessary to overcome the hesitations stepwise.

It should be noted that making bilateral deals has some **advantages**:

- Clear and resolvable identifications and metadata descriptions are not required, since the partners discuss all aspects of the exchanged data in detail, most probably transfer and

convert the data after the deal has been settled and take appropriate measures when the project is over.

- The contract partners are known and trust relationships can be established. There would be no need to rely on new technology for smart contracts (licenses) and data tracing as offered by blockchains for example.
- All juridical aspects are covered by the bilateral contract, i.e. no public statements about availability and reuse conditions have to be made which might be difficult in many cases.

On the other hand there are many **disadvantages**:

- There is no systematic approach to inform others about the existence of perhaps interesting data that can be reused in different contexts, i.e. lots of opportunities will be missed. Data is too often kept in silos which prevents the establishment of an interconnected data economy.
- We need to admit that at this stage the lack of data literacy and similar gaps would hamper data reuse. It will require new roles (brokers, translators, etc.) to give potential re-users an idea how data could be used for new businesses.
- Also in science data reuse is just starting and requires a cultural change and most probably a new generation that understands how to use data in new contexts.
- For Start-ups and SMEs the lack of a data bazar makes participation in data industry much more difficult and hampers innovation.

2.4 Legislation/Regulations

This will remain a topic of debates due to the ambiguous role of legislation and regulations: on the one hand, they are required for example to protect privacy, on the other hand they can hamper innovation. However, the GDPR is now reality and everyone needs to adapt strategies to be compliant with the law. Severe penalties can be given, but for industry it is also important to know whether the rules are effectively applied to create a fair market situation. Some argue that we should see the strict rules in Europe as a challenge to develop smart technologies which at the end could be seen as a competitive advantage since regulations will become stricter in other countries as well. A few more aspects seem to be obvious

- Ownership of and rights on data are not clarified and differences in culture complicate things in a global economy.
- In particular in Europe the national differences are roadblocks on creating an efficient single digital market.
- Investments in new organizational structures and smart technology are necessary to comply with the new regulations which will increase the costs.
- Considering ethical issues is important, but they are difficult to implement.

2.5 Trust and Data Literacy

There seems to be agreement across sectors that we still lack essential mechanisms to establish trust between data actors and that there is a great data literacy problem preventing realistic ideas about possible knowledge extraction from data. These two aspects are closely related, since without deeper knowledge about the opportunities given by data technology, it will be difficult to establish trust. Trust has many dimensions such as guaranteed data quality, long-term availability of data, clear identification of data, certification of the actors involved, availability of brokers etc.

With respect to improving literacy much is already being done in form of training, hackathons, joint developments etc. But we also need to accept that a young generation will be required to make the cultural change happen and this should start at early ages with gaming etc. Another path to cope with the problem is to install eScience centers in science or to found companies or centers of excellence[4] specializing in data and offering services to many.

[4] Also this option is being studied within the realm of the BVDA project.

2.6 Licensing

Defining and negotiating licenses seems to be non-trivial task in data industry. Some of the major reasons were expressed by the speakers:

- The value of data will change over time; therefore licenses are directed to specific reuse scenarios which are being defined very carefully.
- Much is still dependent on trust since we lack a system that allows tracing the reuse of data. This may lead to anxiousness about data miss-use which may lead to complex license constructions.
- Legislation/regulations are changing and it is not obvious which license formulations will be compliant with new situations.

2.7 Data Quality and Certification

It is widely agreed that data in general is in a bad state which partly has to do with the fact that the value of data for exchange and trade is not broadly recognized. For a bazar-like data market the quality of data and also metadata would have to be ensured which would require certification. In science certification methods have already been defined (CoreTrustSeal) and are increasingly often applied. We can also not ignore the need for curating useful legacy data which will require considerable funds.

2.8 Reference Architectures[5]

Here we can see a difference in approach between data science and data industry.

In **data science** all stakeholders have realized that the inefficiencies hamper progress in many areas and exclude many from participation.

- Data Scientists have started to work on flavors of advanced infrastructures (research, data, cyber) with the goal to come to harmonization which, however, is not easy to realize due to the special needs required by each scientific domain. Creating silos is still the dominant pattern.
- Funders have seen that the amount of funds spent on components that are reoccurring in all these different infrastructures in different flavors is huge and therefore push consolidation. But they are dependent on the steps taken by the data scientists.
- Research Organizations are adapting their strategies to cope with the inefficiencies.
- Founding the Research Data Alliance is a typical result of this bottom-up process of an increasing number of data scientists worldwide who see the need to overcome hurdles of all sorts. On purpose, the work is being done across-countries and across-disciplines. The RDA groups do not start from comprehensive reference architectures since they do not believe that there is one covering the "universe of data science". They rely on a bottom-up strategy analyzing in most cases existing use cases and extracting components by identifying reoccurring patterns. Some experts got together to cluster group results and rely on the principles of abstraction which led them to the concepts of Digital Objects and Persistent Identification. Others focus on defining the characteristics of repositories as salient pillars in the dynamic data universe.

In **industry** we see several companies and consortia developing comprehensive reference architectures (RAMI 4.0, BDVA Model, IIC, IDS,) integrative frameworks or platforms.

- Such reference architectures are created in general by large consortiums to have a solid basis for future investments and to allow the identification of the needed components in a process of stepwise refinement to break down complexity. Testbeds then are needed to achieve a maximum of interoperability between the various components.
- The development of larger frameworks and platforms is in general being done by large companies to reduce double work and inefficiencies, which cannot wait on lengthy

[5] In this paper we mean with "reference architecture" a document or set of documents that provides a template solution and a common vocabulary for system architects to identify and define components and their interfaces in a process of stepwise refinement.

agreement finding processes due to a competitive environment. The goals are basically similar.

Some groups define architectures for components of limited scope as for example the Web architecture which was basically defined by HTTP and HTML or the Digital Object Architecture which was defined around the notion of Digital Objects including a few essential and easy to understand sub-components such as metadata, PIDs, repositories/registries and a Digital Object Interface Protocol.

2.9 Components

The meeting points of all these approaches are components that will play an essential role in the evolving eco-system of infrastructures for the dynamically changing data universe. A number of such components have been identified which seem to become stable pillars in an emerging data bazar such as global identities of digital entities (Digital Objects), the necessity and availability of metadata, repositories and registries of different types, smart contracts where blockchain technology may play a role, etc.
In this area a closer collaboration across the different sectors was recommended by several speakers, i.e. the interaction started at a few meetings in the last 3 years should be continued. This would also include the participation in joint testbeds where applicable.

2.10 Automatic Processing

Another agreed need seems to be to make steps towards automatic processing of data management and data analytics tasks since this will be the only way to scale up at the end.

- Some argue in this respect about the advantages of the Digital Object concept with its inherent typing approach where standardized Data Type Registries could be used to link data types with a set of operations opening the door for efficient and effective automation.
- In industry new concepts are being developed that rely on integrative methods for run-time that have many bridges built in or that start from flexible reference architectures being extracted from reoccurring patterns.

Automatic processing requires, however, much more standardization allowing "stupid" machines to find their way. Big companies obviously have the power to build platforms or frameworks relying on their own definitions, however, these will again create hurdles for interoperability.

2.11 Time Frames and Incentives

With respect to time scales we can see different views and it is too early to make summarizing statements.
In China massive investments are being done to come to an integrated data economy by developing infrastructures, changing the culture of exchanging/trading data and in particular by making the young generation enthusiastic. One example is the deployment of a national infrastructure for identification based on the insights from pilot projects that have been carried out.
In science the expert understood that also a change is urgently required in these two dimensions: changing culture and infrastructure. Founding RDA can be seen as one result of this understanding.
In the production industry for example the insight in the coming transformations led to the creation of Industry 4.0 with many dimensions of standardizations such as expressed in RAMI 4.0.
In big IT industry there are yet no incentives to change their business models towards introducing new "standards" - the focus is clearly on coping with the heterogeneity as requested by the customers.

3. Session Reports

3.1 Session "Meet the Experts on Major Data Challenges"

Moderated by **Ramin Yahyapour** (director GWDG)

In this session 6 speakers from industry and industry related organisations presented their views about the challenges for starting data driven businesses.

Wolfgang Dorst (former Bitkom and now ROI management consultancy) focused on two topics: first, he explained the move in industry from pipeline to platform type of structures where customers and providers come together allowing agile interactions on new product developments. It allows collaboration and competition while maintaining a clear separation of concerns. Platforms are seen as viable ecosystems to foster faster adoption of technologies. Companies who do not make this step in particular in data driven markets will have problems to remain competitive. In addition, he pointed to the innovators dilemma in Europe, where the lack of venture capital is compensated partly by governmental funding, which, however, can only be used for pre-competitive developments and thus can hardly be used to yield market shares. He used the Industry 4.0 initiative to explain this mechanism. One can observe that after public funds are being involved an approach in consortia is mandated. This, however, creates a problem for companies in so far as they have difficulties to develop a competitive strategy.

Georg Wittenburg (co-founder of the Inspirient SME) reported about his key observations being responsible for a young company offering data services. He observed an increased willingness to share crucial data with trustworthy start-ups, a slowness with respect to corporate decisions about testing new ways and still a large gap between the collected raw data and the expectations with respect to possible results that could be extracted. This indicates a literacy problem in many companies about the actual potential of available technologies. This needs to be overcome to make the flourishing data economy happen. New roles such as "AI translators" need to be present to help overcoming this gap and to create a better understanding in industry in which cases AI can indeed improve decision finding. In addition, he made clear that the huge inefficiencies in data practices will only be overcome when more automatic processes are being introduced. And this is where his company is targeting on: understanding more business cases to increase the amount of available automatic processes.

Nicolas Zimmer (CEO of Technologiestiftung Berlin) explained why 90% of the start-up initiatives fail. Exemplified by case studies, he indicated that start-ups need to change their approach from an academic perspective where one first is looking what one can do and only then starting to think about possible customers who could be interested to an economic approach where the sequence should be in reverse order. Nevertheless, in some selective cases the academic approach may lead to success. He also spoke briefly about the role of disruptive innovation where in general entrepreneurs first tackle the needs of small customer groups by developing smarter solutions which then have the chance to become a mainstream selling point. However, we should not forget that more than 95% percent of the businesses until now are based on evolutionary and not on disruptive innovations. Another aspect preventing a flourishing data economy is the bad state of data (outdated, erroneous, poisoned, not described, etc.). This needs to improve, in particular, if we not only focus on the products for tomorrow, but on the products in 10 years.

Mark Hahnel (Founder of Figshare) stated that "data management" is a solvable issue which was the basis for starting his now globally acting company. From his experience it is obvious that researchers simply want to store and share all kinds of digital data be it files, slides, software code, electronic papers, etc., associated with some metadata and a DOI for citation purposes. What Figshare is offering in contrast to academic solutions is taking responsibility for code adaptions for example in the case of new regulations from funders which often change their policies. Here, his company can take profit from economy of scale effects since policies worldwide are becoming more similar. He also pointed out that pure storage is not a successful business case, but it is the added value of a software system with stable interfaces and processes and all the possibilities of linking various types of information that gives Figshare an added value.

Ulrich Schwardmann (Head of unit at GWDG, a company offering services to universities and research organisations) explained their service portfolio in the realm of persistent identifiers which are gaining more and more relevance in data science and data industry and which are the way to implement the FAIR principles for good data management. The Handle System is a global PID registry system guided by the Swiss DONA foundation which governs the international and redundant

network of root resolvers and where GWDG is one of the nodes. In addition, the GWDG is acting as registration authority to give prefixes to other organisations to run local Handle resolvers. It is also managing the ePIC collaboration in which some well-known European data centres established a redundant network of Handle services to guarantee a high service quality. In addition, GWDG is offering Handle related services such as Data Type Registries that allow to link data with procedures which can be seen as an extended MIME type registry and as door opener for increasing the degree of automation.

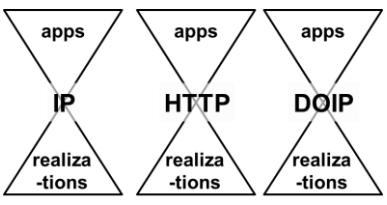
Finally, **Zhou Jian** (Director of the Electronic Technology Information Research Institute offering services to all kinds of industries in China) presented facts about the growing data economy in China. According to him, China is massively investing to manage the transformation towards a digital economy where much activity will be aligned towards an integrated eco-system of infrastructures and where there will be well-designed links with the domain of physical objects. One example for an infrastructure is the setup of a national identification system that everyone can use and which has the potential to become one of the core integration platforms. To achieve the needed level of innovation in particular the young generation needs to be engaged and drive the requirements based on active participation. China has ambitious goals for the coming decade and already now has the highest growth rate with respect to the digital economy. The data sharing mentality needs to be improved to facilitate the creation of the many new businesses which are expected to emerge.

The **discussion** indicated that it takes quite some time (10 years and more) to establish functioning platforms, that it is not always smart machine learning which is necessary to improve results, that we need to find strategies to not leave the many people behind us who are simply afraid of all these new technologies, that we lack the ethical dimension in discussing what is allowed to do with data, that disruption is overemphasized which may be caused by the fact that it does not require huge investments to come up with a disruptive web-service such as airbnb, etc. A special topic of concern was devoted to regulations. Politicians tend to overregulate areas which they do not fully understand. This may then lead to a lack of innovations and a gap in competitiveness. Only where data driven processes get out of boundaries such as in the Facebook case regulations are required.

3.2 Keynote on Common Patterns in Revolutionary Infrastructures

George Strawn (Chairman of the US BRDI committee and key player in the early development of the Internet) was invited to give the keynote in which he presented patterns found in comparing revolutionary infrastructure examples which may help to understand the developments in the data domain. He argued that large infrastructures in general evolve in a sequence of typical phases: Vision, Creolisation, Attraction, Convergence and Exploitation. This pattern can be found in the Telegraph networks (the first large scale electric infrastructure), the Railroad system (a major application of the steam engine), the Telephone system (another large scale electric infrastructure), the Electricity system (networks of power), the Internet (at first, just an application on the telephone network; now voice communication is just an application on the Internet) and the World Wide Web (at first just an Internet application; now an information infrastructure in its own right).

A detail analysis of the developments in the data domain indicates that we have seen much creolisation - there is a huge solution space for all dimensions of data. However, we can also see major attractors such as a global agreement on the FAIR principles and the use of globally resolvable Persistent Identifiers (PIDs) which can open the way to convergence by introducing concepts such as Digital Objects (DO) which could stimulate a phase of comprehensive exploitation again as indicated by terms such as "Open Science", as being offered by the automation of data wrangling or as described in the book "Reinventing Capitalism in the Age of Rich Data". Digital Objects with their inherent virtualisation indicate a natural path from Abstract Data Types and Object-Oriented Programming with all their advantages to build complex systems. A simple analogy with the famous hourglass metaphor used to describe the core idea behind the Internet (see diagram) led him to conclude that for building a DO-based data infrastructure the Digital Object Interface Protocol would be central and thus essential.



He finally concludes that (1) if the data infrastructure coalesces as the Internet and Web did, the data infrastructure could be revolutionary, (2) that the Internet of Things will require a better data infrastructure to reach its potential value, (3) that even where data remains in silos, a global FAIR data infrastructure could allow us to automate data wrangling (which currently takes an estimated 80% of data scientists' time), (4) that it could facilitate the emergence of a "data market" and (5) that it would facilitate the emergence of *Open Science*, which advocates believe will be as revolutionary as the 17th century science revolution.

3.3 Session Data Culture, Data Economy and Data Rights

Moderated by **Edit Herczog** (former MEP, now Director V&V)

Christel Musset (Director of Registration at European Chemicals Agency) explained the role of her agency, the information being collected, the importance of standards for data management and access, the services given to various stakeholders and the strategy for the coming years to maximize the use of all data in a growing interconnected data economy. Chemical industry is obliged to report on the usage of chemicals in their products and ECHA has the task to manage this huge amount of information going into hundreds of millions of structured and textual items and give access to it. A number of standards have been developed together with stakeholders and are crucial for all management and access activities and the exchange with large databases from other countries. The obligatory electronic submission of the Chemical Data Format (called IUCLID) developed together with OECD were the game changers. And the strong legislative REACH framework helped to achieve an appropriate level of data management. The strategy for 2023 requires maximizing the use of data and competences for the benefit of human health and the environment describing the path of being a strong pillar in the evolving data economy. However, a number of challenges (enforce the use of standards, investments in AI, removing legal barriers, improving data quality and richness, adapting the agency's mandate) need to be overcome.

Klaus Tochtermann (Director of the Leibniz Information Centre for Economics) reported about the global FAIR principles for research data which will help to overcome the huge fragmentation across various dimensions and about initiatives that will help implementing the transformation. The European Open Science Cloud initiative is one pillar since it is meant to open the path towards services for open science and it can be seen as a portal giving consolidated and integrated access to all the services offered by research and e-Infrastructures. The newly formed GO FAIR initiative is a bottom-up international approach for the practical implementation of the EOSC. Guided by a balanced governance structure, the work is carried out by implementation networks in three major dimensions: changing the data culture, training the many needed data managers and data scientists, and contributing to the implementation of an eco-system of infrastructures. In addition to a willingness of sharing data, the term "data culture" includes aspects such as ownership and copyright on data, as well as licenses for reusing data all being essential to make the data economy working. GO FAIR is an excellent place to achieve advancements towards open science and a flourishing data economy.

Georg Wittenburg (Co-founder and CEO of Inspirient) presented first his expectations about the moment where he would accept that we have a flourishing data market. As for other goods he would like to issue a request for a specific data set with a number of attributes (content, format, reuse options, quality and even validation, legal guarantees, etc.) and receive quotes from a number of providers. For many reasons we are far away from such a situation although in some areas such as fraud detection in banking people are already reusing a variety of data sets. When asking the question where we are now, we should rather compare data with uranium than with oil - many people in industry are simply afraid of exchanging data, since transportation channels are not secure, mission-critical data could be misused, etc. According to him, a number of specific challenges need to be addressed to change practices. (1) Standardization and identification needs to be improved including ways to integrate legacy data. (2) We need to have simple trusted data transfer mechanisms. (3) Data needs to be certified. (4) Obviously we need brokers as intermediaries

(usefulness, rights, etc.). (5) We may need some more regulations in future, but more important is a clear attitude with respect to enforcing the rules to create trust.
In the **discussion** the question was brought up in how far we need to build a sector crossing interaction platform with respect to data issues instead of discussing various aspects only in the three silos: academia, in industry and governmental services. Sector crossing discussions were seen as useful, but they need to take into account of different criteria such as with respect to data quality, the relevance of services on data and especially the relevance of licenses. Researchers traditionally do not think of licenses, but for industry licenses determining re-use is crucial. In this respect two completely different philosophies can be observed. The collected data from ECHA for example can be consulted, but not downloaded, since it is owned by the fee-paying companies. Handling this issue and improving re-usage conditions it the most complex task for ECHA. A complicating aspect is that re-usage will change over time, making it so difficult to agree on licenses.

3.4 Session Digital Markets and Data Protection Regulations

Moderated by **Sebastien Ziegler** (Director Mandat International, President of the Internet of Things Forum)
Milan Petkovic (Philips Head of the Data Science department, Vice President of the Big Data Value Association) reported on the huge relevance of data in the health care domain and the challenges to get access to data, since data is subject of strict regulations in Europe. In health care there are many areas where a deep AI-based analysis of data can lead to new insights and treatment methods. Just think of the increase of chronic diseases where new data-driven approaches are required. It is obvious that healthcare needs a transformation towards data driven methods and much data is already being generated. But strict regulations and also differences in regulations between the European countries make it hard to efficiently use such data. Philips is working on platforms and technologies that could help overcoming these restrictions. The HealthSuite Insights platform brings clinicians and data scientists together to share and analyze data sets. In the realm of the Big Data Value Association two projects are being carried out. BigMedilytics applies new methods such as sending algorithms to hospitals to make local calculations on protected data and only sharing the results. SODA brings mutually distrusting partners together (hospitals, insurers) such that they can agree on specific analytics to be carried out jointly without giving access to the data for other purposes. Obviously, these technologies are complex due to the high protection demands, but it is urgent to unlock the rich data sets to drive more effective and innovative care and to remain competitive.
Hanna Niemi-Hugaerts (Forum Virium Programme Director) describes the major tasks of a non-profit organisation owned by the Helsinki city as making Helsinki a leading smart city, fostering innovation and mediating between different stakeholders including industry. For any improvement data is in the core, but almost all data collected in smart cities is related to persons - so it is sensitive data. Here regulations such as the new GDPR need to be taken serious. She claims that we should not look at regulations as blocking innovation, but see them as challenges that need to be solved which in turn may give you a competitive advantage at the end. Since much of the created data in smart cities is getting stuck in different silos one of the core tasks is harmonisation of data formats and APIs. Another advanced concept is MYDATA which is a consent-based approach for the management of personal data in distributed scenarios. Basis for a more appropriate data management and for increasing trust is the separation of roles, i.e. a collector or operator is not per se the owner, etc. Based on these principles, Helsinki wants to create a trusted platform allowing managing rights and access to data and services.
Sebastian Ziegler (Director Mandat International, President of the Internet of Things Forum) focused on the various aspects that need to be considered in smart city activities to comply with the new GDPR regulations. He started by presenting the case of the H2020 European Large Scale Pilot on the Internet of Things for Smart Cities (Synchronicity) which brings together several international cities. Integrating data from various cities and sharing applications requires a lot of harmonisation not only in technical terms but also in meeting the various regulations which may differ slightly due to national regulations on top of the GDPR. GDPR is in various ways very strict and needs to be applied

by any company collecting data in Europe. When public data is being collected as in smart cities a Data Protection Officer (DPO) needs to be installed, which means that in collaborative projects various DPOs are being active and need to meet regularly to tightly synchronise data practices. For smart city activities also Data Impact Assessments need to be carried out. To understand the details tests with end users in realistic environments need to be carried out. In addition, he foresees that certification of data processes will become a duty and since current certification schemes have many gaps, new schemes need to be developed. He concluded that in order to meet the regulations in activities such as smart cities much effort needs to be done, but that there is no way out than taking the regulations serious.

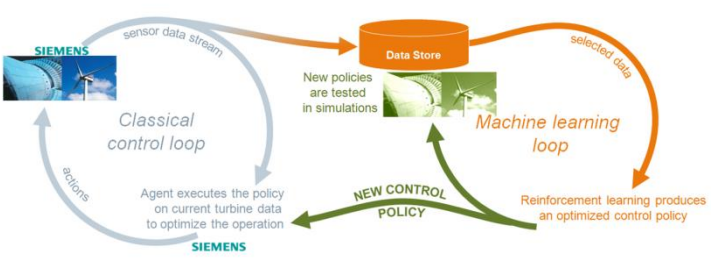
The **discussion** indicated the complexity inherent with regulations and the controversial nature of the debates. The term "ownership" seems to be not clarified and even misleading; we should better focus on "rights on data". While in Europe data sharing via an internet portal does not mean that you give up your rights, the situation in the US where most of the Internet giants reside is different: if you share your data, you also transfer the rights. Separating roles partly associated with certifications could be a way to improve trust. Identification of data would be ideal since it would also pave the way towards data tracing, however, it may take years to put mechanisms in place. The capability of tracing data usage would be a great step towards increasing trust and improving data sharing. If we take the view of individuals GDPR is very simple and this is the view we should take and not the one of companies. However, GDPR costs much overhead and training and it will cost effort and time to develop suitable technologies and organisational mechanisms to make it work in a way that it does not hamper innovation. What is seen as a burden in short run could lead to competitive advantages in the long run. What need to be overcome urgently are the differences in the national legislations in Europe since they are hampering access to the big amounts of data needed for data intensive science. Obviously we need to build teams with different skills (technical, juridical, etc.) to make progress in activities such as smart cities or in areas such as healthcare.

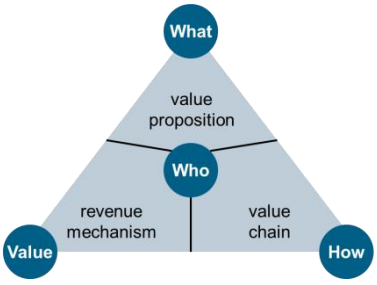
3.5 Session Architectures, Components and Standards

Moderated by **George Strawn** (Chairman of the US BRDI committee and key player in the early development of the Internet)

Thomas Hahn (Chief Expert Software at Siemens, BDVA Vice President) gave an overview about data related activities of Siemens and an impression of how the availability of data will influence the business of big companies. Sectors such as media, trade, mobility and health have already been subject of transformations due to digitisation, now even highly complex sectors such as discrete & process industries are at the tipping point of being subject of transformations. Indicators show that digitization will lead to market growth of about 8% which is determining the vision for the future and which guided Siemens to increase its research budget by 40% within the last 4 years. This includes data analytics and also studies how blockchain technology can be of use. This is also the reason to start core technology projects within the company and to collaborate closely with universities and startups. Companies' business cases are of course been driven by the needs and requirements of customers at the end, i.e. there are ongoing debates how the research and the business departments need to interact and how disruptions can be dealt with in a large company.

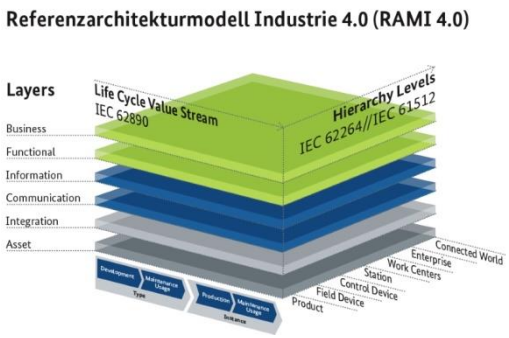
Two examples show how data is being used in Siemens. At CERN Siemens is responsible for the production of the elementary particles that are being accelerated. Continuously data from many





sensors is being collected that observe this production process. Specific patterns in these data

streams are being detected with the goal to optimize the production system. In wind parks the turbines are controlled by the classical control loops, however, now in addition aggregated data is being used to feed offline calculations to find even more optimal parameter sets that control the functioning of the turbines. A 1-3% increase in energy indicates the value of these optimizations.
He also gave concrete examples that show the different effects the use of data can have with respect to the so-called Magic Triangle pointing at the 4 dimensions of business models. When at least two dimensions are involved in a significant way, business model innovations can be distinguished from product or process innovations and this is important to define corporate strategies and design collaboration platforms.
Finally, he described the activities in the area of "Industry 4.0" which is a term describing the way towards smart factories where automation and rich data flows come together to optimize manufacturing. It is obvious that Industry 4.0 will lead to major changes and consortia in several industrialized countries are busy to define comprehensive reference architectures. The diagram shows the German RAMI 4.0 model which defines 3 dimensions: Functional Layers, Life Cycle Value Stream and Hierarchy Levels. Such reference architectures break down complex systems into easier to understand components including sensitive issues such as data privacy and IT security. They also ensure that the many actors involved in Industry 4.0 create a common language to understand each other. Different actors can work on components such as the Industrial Data Space working for example on the issue of smart contracts. To bring such complex systems into practice and to guide the integration work, broad testbed networks have been set up. Also standards need to be specified which is done in collaboration with well-known organizations such as ISO, IEC, DIN, W3C and others.



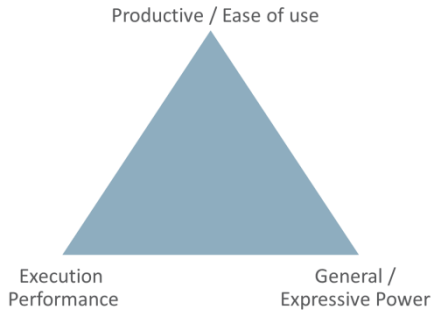
Peter Wittenburg (Senior Advisor MPCDF, RDA Europe Director) basically posed the questions whether industry and science talk about the same challenges and thus need to collaborate and whether there is a bridge between the huge efforts in industry to put, for example, Industry 4.0 in place and the more fundamental bottom-up approach followed in the Research Data Alliance. From various surveys in science and industry it is obvious that the inefficiencies in data-intensive science and industry are huge, about 80% of the time of data experts is lost with data wrangling, i.e. there seems to be a common core which we suffer from in industry and science. While reference architectures help to identify components in a "top-down" manner, RDA also talks about salient components which are extracted from many different use cases in a "bottom-up" fashion. Typical components are repositories storing and giving access to data and registries giving access to much information about data. Can we now, beyond the belief that we need culture changes to improved data sharing, share basic ideas that may help to overcome the huge inefficiencies? Together with an increasingly large group in RDA the concept of Digital Objects came now in focus. DOs are the atomic entities in the digital word which have some bit sequence storing the content, and which are associated with globally resolvable persistent identifiers and different types of metadata. Given the comprehensive reference architectures this concept of DOs seems to be a very small aspect, but in RDA many believe that agreeing on such a concept and implementing it systematically could help to master the data challenges of the future. Of course, the use of the DO concept can be systematically included in industrial reference architectures. He briefly referred to a new group of doers who want to turn DO-related specifications agreed upon in RDA groups to running code and to show their usefulness.
In the **discussion** it became obvious that companies such as Siemens of course make use of data sets from other companies based on bilateral interactions and restricted licenses. Yet we are far away from a market where companies offer data sets described by metadata and where different customers can see and evaluate the offers and then make contracts. Another question raised was what we need to do to overcome the huge knowledge gap with respect to state-of-the-art data management and data analytics. Many companies are far behind the level of the discussions

amongst experts. Siemens also sees this problem and provides a lot of help to start-ups and SMEs. Also in the Industry 4.0 activities much is done to help small groups to participate for example in testing new methods. A complementary way to overcome the knowledge gap and also the lack of experts is to create new types of services and here the mechanisms in science and industry are comparable. While the MPG, for example, extended a central group to offer professional knowledge about data issues to its many institutes and to help in solution finding, in industry SMEs such as Inspirient offer commercial services on solving data related tasks of other companies who do not have the required skills. It was agreed that a closer collaboration between industry and RDA should indeed be discussed to introduce the concept of Digital Objects into the conceptualisation of industrial reference architectures.

3.6 Session Components, Mechanisms and Services for Interoperability

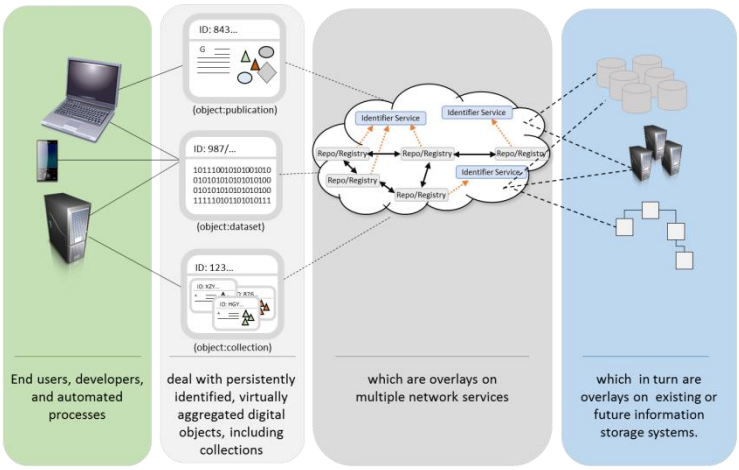
Moderated by **Ana Garcia Robles** (Secretary General of Big Data Value Association)

Hassan Chafi (Senior Director, Research and Advanced Development at Oracle Labs) first explained the enormous proliferation of tools for all kinds of problems with the Tradeoff Triangle. One can only improve performance on two dimensions at the cost of the third. This leaves space for new tools that are optimised for certain specific applications. Such proliferation, however, is dramatic for the customers, since it creates interoperability challenges, and for the software builders, since they cannot ignore these different tools. Customers' main task is to often transfer data between tools which is very inefficient. And there will be even more tools as the examples of the new workflow tools such as Jupyter etc. indicate. One of Oracles major question is now whether there are ways to achieve more interoperability between runtimes to increase workflow efficiency. The solution could be an Open Standard Intermediate Representation for analytical pipelines allowing to mix all kinds of analytics (relational, ML, graph, etc.), to easily connect different front-ends from various vendors and to include different kinds of compilers which generate driver code for specific endpoints such as relational databases, graph databases, spark databases, etc. A first component, the GraalVM will extend the existing Java developer kit. It is a polyglot compiler that can be embedded in various frameworks and that can be seen as a step to realising the integrative vision of Oracle.



Abdel Labbi (IBM Distinguished Engineer & Distinguished RSM, Cognitive Systems IBM Research) also addressed the heterogeneity of tools which will even get worse since in the data domain everything is changing rapidly. "The universe of big data is changing faster than tools can be built". Of course, we can introduce standards and they are necessary, but not sufficient. The way out is to start from architectures which are agnostic to tools and which pave the way to automatic solutions, since this automation will be the only solution to scale up. Architectural principles are derived from reoccurring patterns in the different types of tasks to solve, workflows being built and the types of data being used such as structured data, streams, etc. We need to see data as a new commodity where automation will be the key driver. Yet the degree of automation in the domain of data is very low. To make this work the development of suitable interfaces and micro services (adaptors) will be crucial. The IBM Cloud approach is based on such a flexible reference architecture which then should be easy to adapt to different application fields.

Larry Lannom (Vice president CNRI, RDA US Co-Chair) started with a few assertions such as the need for new approaches to information management, the usefulness of applying abstraction to tackle complex systems where the concept of digital objects could play a big role, the efficiency gain by treating all date entities in the same way and the need for a long-term perspective for interoperability solutions. He further explained the type of DO based abstractions he is thinking of and which are described in the diagram. The user only deals with logical representations of data, i.e. he operates in a domain of metadata and persistent identifiers. These representations are offered by a set of service providers such as repositories and registries which hide all details about storage systems and data organisations. There are a few major characteristics of DOs: they are associated



with persistent identifiers (PID), described by types and metadata, and have a set of operations. With respect to PIDs he stresses the importance of long-term accessibility of data and refers to existing stable resolution systems. With respect to types he points to the requirements of automation and the existence of a suggestion for a Data Type Registry the record structure of which is now being submitted to an ISO study group.

During the **discussion** a number of crucial points were raised. When asked about success stories for the DO concept, Larry referred to the 20 year long experience with the globally used Handle system which is also the basis for the DOIs. Although MIME types are known very well, the work with extended data types as needed for data science just started.

When the question was discussed whether relational databases will continue to exist, it became obvious that one needs to argue from the type of data and analysis to be dealt with, that there will always be applications where the relational model will be the most suitable one and that companies selling database technology already extended the model to include different types of data models.

Some time was spent on discussing whether the introduction of the DO concept which includes a common model of data organisation and which would allow to design a common Digital Object Interface Protocol to exchange data would have an effect. It was argued that such an approach may work in specific areas where there is a joint interest of the involved partners, but that industry does not see a need for another change at this moment, since there are no incentives yet to adopt another layer of abstraction. Industry primarily follows what customers want and a pragmatic approach in so far as it provides adaptors for all kinds of data types already.

The related question was raised how industry would see a realisation of what Georg Wittenburg was requesting: an open market where one can ask for data sets by specifying some metadata. The impression was that data exchange happens already massively based on direct interactions between companies. A need for an open market place was not obvious, although with blockchains there might be a technology now to securely agree on smart contracts and to trace re-usage.

4. Terminology Clarifications

It was not surprising that experts coming from so different backgrounds exposed some terminology differences. For this report we therefore believe that we should define a few core terms for the use in this report.

Digital Object: A DO is an atomic entity in the digital domain that has a meaning for scientific or industrial analytics or processing, i.e. stakeholders have an interest to exchange or trade it. It can be an electronic document, a media recording, a fragment of a stream of sensor measurements, a database or a table in a database, the result of a query on a database, a reusable piece of software, etc. The granularity depends on the usage intentions. A DO has thus some content represented by a bit sequence being stored in repositories, is identified by a unique and persistent identifier and associated with metadata of different types (descriptive, state/system, access rights, licenses, etc.).

Identification: Here the term "identification" describes the step of giving a digital entity a unique and persistent identifier that can be globally resolved to useful state information that allows finding the location where it can be accessed, assessing whether it is indeed the entity which is meant based on fingerprint information, finding metadata information to see what kind of entity it is and what the

license conditions are, etc[6]. Persistent in this context means that it can be resolved after many decades and that unauthorised actors cannot change the state information.

Data Market: The term "data market" describes a scenario where Digital Objects are being offered by data creators or brokers and where consumers can inform themselves about the offers with the help of metadata that will include descriptions of the content, of the accessibility and licenses conditions, the price etc. Let's call this the **bazar** scenario. The term "data market" also covers the scenario where two parties make a private deal to exchange a data set. Let's call this the **deal** scenario. Currently, it seems that the bazar for data does not exist and that all data is exchanged by deals. The deal scenario does not scale towards a flourishing data market.

Roles: In archaic markets the creator and the consumer directly exchange goods. In sophisticated markets there are different types of brokers that facilitate the exchange of goods. A bazar scenario does not work without brokers. A deal scenario does not require independent brokers. In science it is urgent to have repositories and registries to store and offer data. By establishing such entities new roles are being introduced and trustworthiness becomes an issue. There will be other roles as indicated already by the speakers.

FAIR: The FAIR principles (make data Findable, Accessible, Interoperable and Reusable) have now been accepted worldwide to increase the efficiency of data practices. A bazar scenario requires FAIR data, a deal scenario not, since the two partners discuss all details of the exchange operation.

Hourglass metaphor: The hourglass metaphor for the Internet describes a tiny bit of software in the whole stack when talking about complex applications. Why is the metaphor, nevertheless, of great relevance and why did the nucleus, the IP protocol, get so much attraction and is so important: it helped scaling up in message exchange on a global level without the need to maintain adaptors of all sorts and designing new types of network devices. It introduced a stable layer of commodity everyone could build on.

Reference Architectures: We need to distinguish between **comprehensive** reference architecture models which describe a whole eco-system of components from those that describe a specific component. While RAMI is such a comprehensive reference architecture describing the complexity of the Industry 4.0 production scenario, the Digital Object Architecture describes what Digital Objects are and refers to a few essential components around DOs such as the Digital Object Interface Protocol. All have the function to harmonise terminology and to initiate a process of specifying its components and their interfaces.

Data Organisation: We need to distinguish the term "data organisation" from the well-known term "data model". Although the latter term refers to an abstract model that describes elements of data and their relationships, it is mostly interpreted in the sense of structures of how to organise databases (flat, hierarchical, network, relational, object, etc.). The term data organisation is related with the concept of Digital Objects and refers to different entities such as its bit sequence, its PID, its metadata which is of different types such as descriptive, state (system), rights, context, embedding, provenance, etc. and the relationships between these entities. In the data warehouse approach all of this data is put into a large database following some database model.

Certification: When we speak about the certification of data we mean the certification of all actors involved in the management and processing of data until it is offered to customers. This includes the processes of creation, storing and managing, brokering, etc. For repositories a few suggestions have been made by ISO, DIN and RDA. The latter is now called CoreTrustSeal and is already broadly used in academia. It makes statements about the creation process and in particular the management process. It is a lightweight set of rules sufficient for academic use in general.

Repositories/Registries

[6] It should be noted here that, as Ulrich explained, there is one system that offers what is required: The Handle System is in operation for 20 years, is now governed by the Swiss DONA foundation and its services are guaranteed by now 10 globally distributed root resolvers.