Update of the European Data Market
SMART 2016/0063

Story 8 – Accelerating the Impact of Data Commons

29th February 2020
<table>
<thead>
<tr>
<th><strong>Author(s)</strong></th>
<th>Laia Pujol Priego, David Osimo (The Lisbon Council);</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Deliverable</strong></td>
<td>D 3.8 Quarterly Stories – Story 8</td>
</tr>
<tr>
<td><strong>Date of delivery</strong></td>
<td>29th February 2020</td>
</tr>
<tr>
<td><strong>Version</strong></td>
<td>1.0</td>
</tr>
</tbody>
</table>
| **Addressee officer** | Katalin IMREI  
Policy Officer  
European Commission - DG CONNECT  
Unit G1 – Data Policy & Innovation  
EUFO 1/178, L-2557 Luxembourg/Gasperich  
katalin.imrei@ec.europa.eu |
| **Contract ref.**   | N. 30-CE-0835309/00-96                              |
# TABLE OF CONTENTS

EXECUTIVE SUMMARY .................................................................................................................. 4

1 INTRODUCTION ......................................................................................................................... 6

2 WHAT ARE 'DATA COMMONS': SCOPE AND TYPOLOGY ...................................................... 8

3 TWO CASES OF DATA COMMONS ............................................................................................. 11
   3.1 Open Targets: A data commons case in the pharmaceutical industry ............................. 11
   3.2 Open Subsurface Data Universe (OSDU): A data commons case in the oil and gas industry ........................................................................................................................................... 15

4 THE TWO SIDES OF DATA COMMONS: COSTS AND BENEFITS ......................................... 19

5 MECHANISMS TO LEVERAGE THE IMPACT OF DATA COMMONS ........................................ 20

6 CONCLUSION .............................................................................................................................. 22

7 REFERENCES .............................................................................................................................. 23
Executive Summary

With the explosion of data availability and resulting increase of data-related challenges pervasively affecting companies in a variety of industries, a third generation of systems emerge, so-called data commons, to integrate data from heterogeneous sources, computing and storage infrastructure with software services required for analysing and working with the data (Grossman et al. 2016). Data commons are currently managed by multiorganizational collaborations, usually a consortium or a group of organizations including competitors, that come together to share costs and resources to build a common infrastructure that supports data analysis and helps them extract value from the integration and re-use of the data shared.

From the lesser or higher degree of sharing across companies, such efforts of building common data pools need to address some challenges:

- To generate the appropriate incentives for organizations to share some data
- To provide the conditions under which they are willing to do it, that is a governance approach that defines rights and responsibilities across data owners and users
- To agree in a set of data formats, data structure and quality in which the data and its contextual information needs to be shared, which imply an agreement on data standards and metadata to allow interoperability across systems
- A sustainability model that guarantees that such infrastructure and efforts are not only maintained over time but eventually scale.

In brief, the development of such data commons displays the need for some technical and organizational creativity to make them happen and sustain over time. The present report brings some insights from two inspiring examples of data commons, which have achieved a governance approach that overcome the challenges at the base of developing and scaling data commons.

The first example, Open Targets, is an inspiring case that comes from biomedical research and the pharmaceutical and biotech ecosystem. Life science and, as a result, the biopharma industry have been very active in developing data commons, especially since the data deluge resulting from the completion of the human genome in 2003. The Open Targets case is valuable since relevant industry players such as Biogen, GSK, Sanofi, Takeda, and Celgene have come together with academic partners such as EMBL-EBI and Sanger to agree and developed an infrastructure that support early stages of R&D process in the drug discovery, by sharing data, tools, and knowledge related to targets.

The second example, OSDU is an exemplar case coming from the oil and gas industry where leading companies came together in 2018 to jointly face the challenges brought by the data deluge coming from the expansion of the Internet of Things (IoT) and sensor devices combined with seismic and historical data, amongst others. Following the same rationale that the OT case, companies followed the strategy to share investments and costs regarding data sharing and compete in the software applications that provide relevant insights from such data.

Both examples have in common to have successfully achieved to make compatible both the cooperation of organizations and data sharing generating positive externalities while preventing the competitive interests of their contributors. Both provide insights on the organizational characteristics and technical architecture of the infrastructure that allow companies to dynamically move from sharing and restricting access to their data and knowledge, and a governance model that preserve trust amongst partners and manage to successfully stay and scale.

Some of the mechanisms that the present report describes, and which were put in place by the two cases so that companies got incentives to collaborate with competitors, include:
1) ex-ante safeguards;
2) arbitration;
3) a stratified and layered infrastructure;
4) user-centric approach;
5) interaction with data generators or data holders to accelerate re-use; and
6) a federated approach towards data-sharing.

The combination of all these mechanisms created, in the inspiring cases of OT and OSDU, a favourable environment for the organizations to share costs and realize scale in the collaboration to accelerate their innovation processes.

As data intensity and computationally intensive innovation methods are predicted to conquer different industries, we can expect an increase in required investment levels, rendering them prohibitive to single organizations. As data commons emerge in response, the two cases shared in the present report, can be inspirational for the different governance modalities, how they relied in open source and open standards strategies (vendor neutral), and in general the mechanisms that need to be put in place to successfully navigate the competitive and cooperative dynamics inherent in such data commons.
1 Introduction

This study addresses the topic of data commons, the idea to pool data between different entities (notably commercial entities) to provide data that can feed innovation. The idea is prominent in the most recent European Data Strategy, notably through the notion of data spaces and data pools that would allow Europe to federate industrial data repository and enable the next generation of data-driven innovation. But what exactly are data commons? Are they effective? How do they work, and what are the success factors?

Data commons co-locate data, storage and computing infrastructures with commonly used services and tools for analyzing and sharing data to create interoperable resources for a heterogeneous base of users (Grossman et al. 2016). Originally devoted to R&D domains where the growth of data, and in particular in big science, has been significant since the last decade, data commons have conquered pharmaceutical, oil and gas, environmental and a heterogeneous number of sectors. Examples expand from collaborative arrangements between corporations to agree on data standards, ontologies and metadata to potentially share data amongst organizations to different and heterogeneous data infrastructures where companies share their data, which they combine with public or open data, to reuse it as part of different R&D and innovation processes.

From the lesser or higher degree of sharing across companies, such efforts of building common data pools to extract value from the integration and re-use of such data shared, need to address some difficulties, which are common across such initiatives. Overall the challenges of developing such data commons display some patterns across these different efforts, being the most important one to generate the appropriate incentives for organizations to share some data; to provide the conditions under which they are willing to do it, that is a governance approach that defines rights and responsibilities across data owners and users; to agree in a set of data formats, data structure and quality in which the data and its contextual information needs to be shared, which imply an agreement on data standards and metadata to allow interoperability across systems; and a sustainability model that guarantee that such infrastructure and efforts are not only maintained over time but eventually scale. In brief, the development of such data commons displays the need for some technical and organizational creativity to make them happen and sustain over time.

The present report seeks to bring some insights from two inspiring examples of data commons, which have achieved a governance approach that overcome the challenges at the base of developing and scaling data commons. Both examples have in common to have successfully achieved to make compatible both the cooperation of organizations and data sharing generating positive externalities while preventing the competitive interests of their contributors.

The first example, Open Targets, is an inspiring case that comes from biomedical research and the pharmaceutical and biotech ecosystem. Life science and as a result the biopharma industry have been very active in developing data commons, especially since the data deluge resulting from the completion of the human genome in 2003. This sector has offered different inspiring examples of how organizations have managed to successfully allocate rights and responsibilities amongst contributors. Some examples include Pistoia Alliance, Yoda, amongst different pre-competitive consortia funded in the framework of IMI who have tried to develop a common pool of collective resources to be shared amongst organizations (see (Pujol Priego and Wareham 2018) for further examples). In particular, the Open Targets case is valuable because it relevant industry players such as Biogen, GSK, Sanofi, Takeda, and Celgene have come together with academic partners such as EMBL-EBI and Sanger to agree and developed an infrastructure, so-called the open targets platforms and genetics platform, that support early stages of R&D process in the drug discovery, by sharing data, tools, and knowledge related to targets (i.e the term “targets” typically refers to proteins that have three-dimensional structures to which specific molecules can bind to provoke some physiological effect). This initiative provides insights on the technical architecture of the infrastructure that allow companies to dynamically move from sharing and restricting access to their data and knowledge, and a governance model that preserve trust amongst partners and manage to successfully stay and scale since 2015 (e.g. one company per year has been incorporated in the open targets initiative and the number of
success stories and benefits from open targets outcomes has populated in particular since the last two years).

The second example, OSDU is an exemplar case coming from the oil and gas industry where leading companies came together in 2018 to jointly face the challenges brought by the data deluge coming from the expansion of the Internet of Things (IoT) and sensor devices combined with seismic and historical data, amongst others. Following the same rationale that the OT case, companies followed the strategy to share investments and costs regarding data sharing and compete in the software applications that provide relevant insights from such data.

Finally, the present report will try to provide some understanding of how these two examples not only are inspiring for other data commons initiatives, but also how they inform the policy space seeking to promote data commons to foster innovation and economic growth. We acknowledge that it has been a particular intensive policy agenda since the last decade around the open science leitmotiv and the promotion of data economy as a relevant pillar to foster a digital single market, yet data sharing is still incipient or have not achieved the desired level. As such, the present report seeks to provide some lessons from the two cases that can provide some insights about the barriers that policy institutions can help to overcome while promoting some of the drivers to unblock the growth of such data commons.

The remainder of the report is structured as follows: first, we provide some clarification about what data commons are and their trajectory. We provide a description of the two cases and from their presentation, we extract the ‘two sides’ of such data infrastructures; that is a general understanding of their costs and challenges and benefits or positive externalities. We extract some governance models of data commons and identify some mechanisms that contributed to leverage the impact in our cases. Finally, we offer some concluding remarks.
2 What are ‘data commons’: Scope and typology

Data commons also referred as data collaborative, data spaces, open data partnerships, common data infrastructures (Mishra et al. 2016; Perkmann and Schildt 2015; Susha et al. 2017), have been defined as platforms that openly share data and knowledge with a computational infrastructure that supports data sharing across an heterogeneous base of users supporting different services on top of the data (Contreras 2010).

As described by Robert Grossman (2018) (see figure 1) the first generation of platforms operated databases in which datasets were deposited. "As the web became the dominant infrastructure for collaboration, data portals emerged as applications that made the data in the underlying databases readily available to researchers" (Grossman, 2018 p. 11). Data portals were websites that provided interactive access to data in the underlying database.

With the explosion of data availability (in the case of biomedical research after the completion of human genome, but in the oil and gas industry with the growing increase of granular data thanks IoT, datasets began to grow and soon large scale computing infrastructures were needed to process, manage and distribute data. Several systems were put in place and developed to process datasets. This second generation of systems (so-called data clouds) collocates computing with the data to enable users to compute over the data. Over the past decade, cloud computing has supported the collocation of "on-demand large scale computing infrastructure that has created new opportunities for the largescale analysis ..." (Grossman, 2018, p.12).

The third generation of systems emerges with the so-called data commons, to integrate data from heterogeneous sources (through data standards), computing and storage infrastructure with software services required for working with the data. Data commons are currently managed by multiorganizational collaborations (usually a consortium or a group of organizations) that come together to share costs and resources to build a common infrastructure that supports data analysis and helps extract value from data.

![Figure 1. Evolution towards data commons (Source: Robert Grossman, 2018)](https://www.slideshare.net/rgrossman/crossing-the-analytics-chasm-and-getting-the-models-you-developed-deployed)

The major differences between data commons and traditional R&D consortia and collaborations in the past are: First, the extreme data intensity, being their primary focus on large data pools and related processes to generate, analyze redistribute highly valuable, often proprietary, data. Second, data commons are characterized by the members’ commitment to reveal results developed in the framework of the collaboration, which is the data, methods, technologies that they share (Mishra et al. 2016). While traditional R&D consortia and collaborations used to operate under traditional-IP agreements, data-commons promote free revelation of the data and related outputs and eventual

---

1 Source: Crossing the Analytics Chasm and Getting the Models You Developed Dep...LinkedIn SlideShare(2018), [https://www.slideshare.net/rgrossman/crossing-the-analytics-chasm-and-getting-the-models-you-developed-deployed?next_slideshow=1](https://www.slideshare.net/rgrossman/crossing-the-analytics-chasm-and-getting-the-models-you-developed-deployed?next_slideshow=1)
collaboration with the public. Within such data commons, ‘sharing’ poses benefits and challenges: companies need to control that such collaboration in developing the data commons do not lead to unintended knowledge spillovers that damage the firm’s competitive position and interests; while certain degree of sharing is needed to work jointly towards an infrastructure and tools around data sharing that effectively accelerates their innovation processes and generates positive externalities for the organizations contributing and the public. Finally, a third characteristic is the relatively high uncertainty of the potential value that can be extracted from pooling data, resources, and capabilities.

The strong focus on (1) data, (2) sharing practices, and (3) uncertainty shifts the emphasis of the collaboration from a strictly defined transactional model towards governance that allows greater adaptability to negotiate the best paths towards effective collaboration amongst the organizations. This entails understanding the governance process within the ecosystem of organizations developing and contributing to the data commons, which implies spanning the boundaries of participating organizations.

Data commons emphasize the value created through large-scale and diverse data. Their primary goal is to generate technological standards, infrastructures (including data platforms) and, in the particular context of R&D, novel scientific knowledge while pursuing their corporate interests (De Vruhe and Crommelin, 2017). In that sense, data commons can be classified according to the primary focus of their collaboration (Altschuler et al., 2010). Although the main focus relies always upon ‘data’, the different intentions and goals shape a different ethos of data commons. As such, data commons differ depending on if their focus is:

i. Generation of mutual standards to foster scientific data sharing;

ii. Data generation and aggregation, where partnerships pool together financial resources and expertise to create, maintain and aggregate data;

iii. Knowledge creation, where collaborations leverage existing standards and shared tech infrastructures to create new collective knowledge in early drug discovery phases.

We can also identify different types of data commons depending on their origin (i.e. who initiated it), which affects how they work:

i. Industry-driven: when the collaboration around the development of data commons have been originated by companies

ii. Policy-driven: when the cooperation has originated as a policy instrument to support multi-stakeholder long-term collaboration to implementation platforms supporting data sharing in specific technology areas (e.g., IMI).

Following such characterization, and going back to the examples shared in our introductory lines, Pistoia Alliance is an industry-driven collaboration created in 2009 by large pharmaceutical companies, such as GlaxoSmithKline (GSK), AstraZeneca, Pfizer, and Novartis, without any government intervention to collaborate and share collective resources in a pre-competitive space. Pistoia started with the goal to generate mutual standards in industry, ontologies and web services that are made available under an open-source framework to academic institutions, vendors and companies, to facilitate data sharing, data representation, text-mining activities and improve the R&D efficiency of the drug discovery and development process. Pistoia has been scaling since its origins accepting applications from both private and public organizations that share the ideology of sharing data and resources of drug discovery research pipeline. Today some of the partners including, but not only: Lundbeck, Accelrys, ChemAxon, ChemITment, BioXPR, EdgeConsultancy, GGA, Rescentris, and UPCO. With many initiatives underway since its foundation, spanning from ‘ontologies mapping’, semantic standards and translation of FAIR\(^2\) principles to the organizations data to foster data

\(^2\) FAIR is an acronym that refers to data findability, accessibility, interoperability, and reusability
integration, the alliance has been pooling resources from many of their company contributors and openly sharing the output of its work with the outside community.

The second example shared in our introduction is Open Targets. It is also an industry-driven collaboration started with the collaboration of different pharmaceutical and biotech companies in collaboration with two leading academic partners such as EMBL-EBI and sanger, and the goal of Open Targets was not only to agree on some mutual standards related to data but to generate data through joint experimental projects and openly release it to be integrated with other public data sources in the open targets platform. As part of the goals, such as data commons or collaboratives foresaw also to generate knowledge together by extracting value from the data generated and aggregated in the framework of such collaboration.

OSDU, the example from oil and gas, is also an industry-driven collaboration that started in 2018 by key players in the industry (see section 3.2.). While the development of data standards is part of the work that OSDU needs to implement, the scope goes beyond the agreement on standards towards building a platform where organizations share data and software applications and services are developed on top.

Finally, there are different types of stakeholders involved in data commons. Stakeholders include (based on Grossman, 2017):

- **Data commons service provider** (DCSP), which refers to the organization operating the data commons;
- **Data contributor** (DC), which labels the organization or individual providing the data to the data commons service provider.
- **Data user** (DU), which is the organization or individual accessing and using the data.
3 Two cases of data commons

3.1 Open Targets: A data commons case in the pharmaceutical industry

Open Targets (OT) is an innovative data commons infrastructure developed by a large-scale, public-private collaboration on pre-competitive research. It was initiated by the Europe flagship laboratory EMBL-EBI in collaboration with leading global pharmaceutical companies to provide comprehensive and up to date data for drug target identification and prioritization\(^3\), which is the first phase in the drug discovery process. OT integrates publicly available information and data relevant to targets and diseases in the OT Platform\(^4\); while it also performs high throughput experimental projects that generate highly valuable target-centred data in physiologically relevant systems to shed light on the causal relationships between targets and diseases. Open Target's experimental projects are focused in three therapeutic areas: Oncology, Immunology, and Neurodegeneration (Pujol Priego and Wareham, 2019).

The foundations of OT infrastructure are the result of a major paradigm shift that the industry has experienced since the completion of the human genome. In a context where the costs of bringing a new drug to market have never been higher - 2.6 billion- more than double the estimate of just a decade ago according to Tufts Center for the Study of Drug Development, and the time longer, companies have decided to partner with competitors and academia to leverage the potential value from the new biological data coming from sequencing the genome and transform it into valuable knowledge about disease mechanisms and systems biology. The data deluge in biomedical research has led to an increasing granularity of information that requires new computational approaches, tools, and techniques to accelerate drug discovery. Considering all the duplication of efforts in research devoted to providing the basic biological knowledge required for successful drug design (Altshuler et al., 2010), some companies have followed the strategy of partnering and opening up their boundaries in heterogeneous and dynamic multiorganizational collaborations that seek to build common pool of data resources and technology solutions that leverage their capabilities in the early phases of drug discovery.

Following such rationale, OT was born to bring together the skills, expertise, technologies and diverse data types that R&D managers at pharma and scientists in academic labs need to establish the genetic links between targets and disease development. Target validation is an exercise that defines the role that a biological process plays in disease and is the first step in drug discovery. "Currently, an estimated 90 percent of compounds entering clinical trials fail to demonstrate the necessary efficacy and safety requirements, never reaching patients as medicines. This is often because the biological target chosen is not well understood", explains the Biogen team (EMBL-EBI, 2014).

In brief, OT employs advances in cutting-edge genetic methods to support researchers in the first step of exploring new drugs; concretely, helping them to prioritize and identify “where to start”. Figure 2 visualizes how OT infrastructures both generate and integrates data that is openly shared with the research community in the OT platform.

---

\(^3\) OpenTargets website: https://www.opentargets.org/
\(^4\) Open Targets Platform: https://www.targetvalidation.org/
Some of the technologies that OT employs include:

1. Gene editing (CRISPR)
2. Induced pluripotent stem cells
3. Single-cell genomics
4. Organoid and tissue culture
5. Large-scale genomics and epigenomics (e.g. ENCODE)
6. Genome-wide association studies
7. Next-generation sequencing
8. Bioinformatics
9. High-performance computing

OT methods include a combination of large-scale genomic experiments with objective statistical and computational techniques to identify and validate causality between targets, pathways, and diseases (Open Targets outreach, 2016).

At present, OT infrastructure contains more than:

- 27,069 targets
- 6,336,307 associations
- 13,579 diseases
- 20 data sources.

OT collective work has resulted in:

- The identification of 2,540 potential new indications for 791 current drug targets.
- The discovering of “thousands of genes essential for cancer’s survival and ranked which ones show the most promise as drug targets for developing new treatments” by employing a novel computational framework that integrates “multiple lines of evidence to assign each gene a target priority score” (Behan et al. 2019 p. 511).
The number of users of Open Targets Platform also reflects the interest and value offered to the community, which exploits the data shared to accelerate their R&D and innovation processes. According to one of the early impact studies available in 2016\(^5\), 900 unique IP addresses access the Open Targets Platform every week. Some metrics available about Open Target platform usage, from April 2016- March 2017, reveal that the platform is used substantially (see figure below).
The metrics are also aligned with the qualitative feedback reported by Open Targets team from platform users. For instance, as one drug discover scientist said: “Powerful resource, clear links and easy to use without training, especially for a non-bioinformatician!” (Kafkas et al., 2017)

Several academic publications have reported already successful results from mining Open Targets data and its target diseases associations with different computational approaches. Such publications uncover the on-going opportunities of the re-use of such data for drug discovery processes. As such, Khaldakar et al. (2017), for instance, describe different computational workflow implemented, which systematically uses data from OT to identify potential repositioning opportunities. Concretely, it uncovers 2540 potential new indications for 791 existing drug targets, and it further categorizes them based on diverse and multiple types of evidence. From these new opportunities uncovered, 1366 are for Orphanet rare diseases where the target is a known drug target for a common disease.

Rare diseases are defined as diseases affecting less than 1-in-2000 people in the EU or less than 200,000 patients in the USA (Tambuyzer, 2010.) There are today severe unmet needs for the population suffering from rare disease. The potential of repositioning an existing drug or clinical compound for a rare disease is an attractive approach being implemented for orphan drug development. OT has been able to suggest potential drug-repositioning opportunities for 14 rare diseases. In addition, taking into account the recent importance of genetics support for having higher success rates on drugs for targets, Khaldakar et al. (2017) highlighted that 628 (24%) of the 2540 new indications are supported by genetic evidence.

Consulting with the commercial source Pharmaprojects6, which gathers worldwide drug development pipeline data, 6% of all new target-disease pairs uncovered in OT are in the drug discovery pipelines of companies, which is considered a conservative estimate due to the fact that only indications that exactly match were considered. Overall, the computational workflows implemented show substantive potential opportunities that will need to be evaluated in detail, combining the evidence from the OT platform with endogenous evidence about the disease. Many of the opportunities already identified by OT are expected to eventually lead to effective drugs.

Finally, the analysis of OT case provides valuable insights about:

i) The costs and benefits of data commons (section 4);

ii) Different modalities of data commons, that is the diverse governance approaches and models that allocate rights and responsibilities across users and contributors and orchestrate the development and sustainability of the data commons; and the mechanisms that underlie in such a governance approach that provide the right incentives for contributors to share the data while preserving their interests (section 5).

---

6 Pharmaprojects website : http://www.pharmaprojects.com
3.2 **Open Subsurface Data Universe (OSDU): A data commons case in the oil and gas industry**

From seismic analysis to reservoir modeling, drilling services, production reporting and a large number of multiple activities, the oil and gas industry has been a data-intensive sector where companies have had to face already several challenges. With the evolution of digital oilfields, where sensors and recording devices are increasingly generating millions of data points each day, one of the critical challenges is to transform such data into valuable information for the companies (Mohammadpoor and Torabi 2018). Just like any other sector where optimization is relevant, but more importantly in those with high cost of error and large capital investment, the potential benefits of predictive analytics and data science combined with the trends on computer processing, storage and technological advances in digital imaging, are driving oil and gas companies towards data sharing models to steer innovation creating an interesting movement in their ecosystem of suppliers.

The oil and gas industry has been managing large amounts of data for a while, leading some scholars to refer to it as the ‘original big data industry’ (O’Reilly, 2015). The growing number, resolution, and frequency of seismic data, advances in IoT, combined with large amounts of heterogeneous historical data, has created severe pressure to the companies raising data science capabilities that have required partnering with other stakeholders.

As an answer to some of such problems, Open Subsurface Data Universe (OSDU) initiative started in 2018 as a collaboration towards the provision of common infrastructure (also called platform) for the oil and gas industry to reduce data silos and offer efficiency gains to the companies. OSDU was founded to facilitate secure, and reliable access to subsurface and wells data, support disruptive workflows in the organizations, and foster the implementation of new digital solutions that support decision-making processes. This requires reaching a consensus and creating an open standards-based ecosystem.

Similar to the previous case on OT for the pharmaceutical industry, OSDU follows a platform-complementors logic where the main business rationale of companies is to separate the core- that is the data that they share- from the potential applications, where their competitive value comes from. To do so the basic idea of OSDU was to develop a commons data infrastructure with standard public APIs supported by global cloud hosting vendors that allow companies to reduce resources and cost burden in independently managing subsurface data platforms while accelerating key reference components at the foundation of upstream digital transformation (OSDU, 2020)

The commons approach of OSDU, that means making it openly available to different stakeholders, makes the initiative attractive for the oil and gas operators, but also cloud service companies, technology providers, software suppliers, academic organizations, amongst others.
The main milestones of OSDU include the publication of a reference architecture, which is a cloud-native subsurface and wells data platform reference architecture with committed initial implementations by Microsoft Azure, Amazon Web Services and Google GCP, as of today. As part of its endeavors, OSDU needs to define application standards that guarantee services developed by different organizations on top of OSDU data. Finally, OSDU will require to leverage industry data standards for frictionless integration and data access.

The major business impact predicted by OSDU organizations by reducing the data silos is that an ecosystem of organizations with different services develops around the data of the OSDU platform. OSDU integrates a heterogeneous base of stakeholders from oil and gas operators, service companies, technology providers, software companies, academic partners, amongst others.

The founding organizations of OSDU are the following (see figure below):

While there are not yet data to describe the impact of OSDU, due to its recent launch, it is worth noting the ability of the initiative to convince and engage such a significant number of companies in the industry. The value offered by OSDU case is translated in the first place in the current number of members, which include:
- **Operators**: Anadarko, BHP Billiton Deepwater, BP, Cepsa, Chevron, Concho Resources, ConocoPhillips, Devon Energy, Equinor, ExxonMobil, Hess, Marathon Oil, Noble Energy, Oil and Gas Authority, Pandion Energy, PETROBRAS, PETRONAS, Reliance Industries, Shell, Total, and Woodside


The OSDU is a trademark of The Open Group, which is a global consortium that supports the development of technology standards. The Open Group has designed a process that they have put in place to reach an agreement amongst partners about the different standards that need to be implemented so that the data platform can operate. The main process (see Figure 4 below) seeks to comply with a set of 7 principles: Openness; Consensus; Timely and Deterministic Process; Public Availability of Published Specifications; No Legal Impediment to Implementation or Adoption; Confidentiality; and Executable Standards. In brief, the procedure seeks to reach stable decisions (this implies that silence is not interpreted as agreement), which are inclusive (as many organizations involved as possible), participatory (the process actively solicit input from the participants) and cooperative (participants need to discuss and strive to reach the best possible collective decision for all members).

**Figure 7.** Main procedures for standards agreement (Source: The Open Group, 2020)
Finally, for OSDU to operate, and as in the case of OT in the pharmaceutical industry, a set of procedures have been put in place so that the open group behaves as a trusted third-party amongst the different competitors and organizations in OSDU, especially when handling with confidential material that members do not want to disclose. In brief, the open group membership agreement includes provisions that define the general obligations of The Open Group and members so that through individual non-disclosure agreements that may also apply to third-party information, the confidentiality of some of the information shared is preserved. Additionally, guidelines have been put in place with practical steps to be followed to identify and protect confidential information.
4 The two sides of data commons: Costs and benefits

One of the biggest challenges for data commons, in particular for large-scale data commons as in the case of OT and OSDU, is generating the business case for companies to be willing to share their data with a pool of organizations via a common data infrastructure; and developing a long-term sustainability model that affords the different activities behind the operations that support such infrastructure.

Regarding the first challenge, several industries have realized the critical value created by sharing and re-using digital data (in short data) for different purposes unforeseen when data was collected in the first place. Data is considered a non-rival good, meaning that it can be reused and recombined to generate positive spillover effects (OECD 2015). Although the re-use of data is assumed to foster innovation and contribute to economic efficiency avoiding duplicative investments, amongst other positive externalities (Drexl, 2018), data is still kept in silos and little re-used is reported (Pujol Priego et al. 2019). As of today, data is not subject to property rights, such as knowledge, but other legal bounds can inhibit data to be reused, such as trade secret protection, right to data portability in GDPR, amongst others. However, most importantly, the decision of a company to share its data to be reused depends on its factual interest in sharing it.

There are different costs associated. Depending on the sector, liability concerns may arise for companies preventing them to share some potentially sensitive data (e.g. clinical trial data) with the public, even when there are potential single and collective benefits behind it. Competitive concerns are related to how to share information without damaging competitive interests and advantage of for profits. Besides such concerns behind the companies’ decision of engaging in the development of data commons, there are different costs associated with doing it.

Worth noting are some operational costs related to the services provided by such data commons, which include data curation, periodic updating, and monitoring of datasets, the agreement and implementation of data standards to ensure interoperability, data storage services and all activities related to guarantee data security and privacy standards. Interoperability is an important requirement of data commons to permit the aggregation and integration of datasets through a variety of tools and support its re-usability. Data commons must be also secure to preserve permissions, guarantee data protection and prevent corruption of data. This aspect is critical where data involves sensitive personal information (e.g. health-related data, financial or legal related data).

On top of such operational activities, additional tasks, depending on the scope of the data commons, may be related to the provision of data visualization tools that foster data re-used, operational support for users seeking to re-use the data and asking for additional contextual information about the data or clarification, amongst others. See Grossman, 2018 for more exhaustive detail.

Potential benefits are related to efficiency gains and the generation of new products and services afforded by the re-use and recombination of other’s data, fostering an ecosystem that co-creates value around data. As such, data is a non-rivalrous resource allowing the same data to support the generation of heterogeneous products and services. Any company can engage with the same data in different data-sharing arrangements, being unlimited the potential value that can be extracted from the same data. For instance, as some of the data commons cases reveals, including the OT case, properly designed data commons can serve to R&D processes as an active and accessible repository for research data; as a platform for reproducing research results; to support discoveries by adding data and new algorithms developed and implemented around the commons and as new software applications and tools are integrated with the common pool of data. As the OSDU case shows, benefits include to reduce data silos from different organizations and integrate workflows in the cloud taking advantage of a common cloud solution; and supporting software vendors, oil companies, academia and an open community of projects to be developed (competitive dynamics) on top of a common data platform developed (cooperation).
5 Mechanisms to leverage the impact of data commons

In developing data commons, the case of OT introduced in section 3 and reinforced by the OSDU example, uncovers some mechanisms put in place so that companies got incentives to collaborate with competitors. Mechanisms include:

1. *Ex-ante safeguards:* Carefully selecting the organizations that can join the data commons contributes mitigating opportunistic behavior and provides some guarantees to sustain the collaboration over time. By qualifying participants with complementary assets, making sure there is a firm commitment to cooperate and share data, and guaranteeing some rules of the game before joining the data commons-pool of organizations, is relevant in particular for its sustainability. Both cases show how an agreement with the governing party (i.e. an operational team or an external organization supporting the organization of the data commons) is required to define the rights and responsibilities of members while establishing some protocols and processes to follow to allow the collaboration.

2. *Arbitration:* Human expertise and intervention are required to orchestrate the integration of the heterogeneous data by organizing the pooling of the heterogeneous resources from the different contributors while coordinating the data and knowledge flows so that stakeholders feel comfortable in their contribution to the development of the data commons. In brief, a trusted intermediary allows the coordination needed for companies to cooperate and work together towards a common goal, while dissipates potential disputes and concerns. Whether an internal team selected by members, as in the case of OT or an external organization providing support to the processes, arbitration and orchestration activities is required.

3. *Stratified and layered infrastructure:* It is critical to define explicit domains of cooperation and set clear boundaries where the race starts. The case of OT revealed an explicit agreement between contributors of different visibility levels of data according to where data fell under the pre-competitive label, or where it was considered competitive. This demarcation between data domains made it possible to share open data about targets (considered pre-competitive) while restricting any access to data about the following and sequential steps in the drug discovery process, considered competitive (i.e. lead optimization and onwards). The stratification in layers promoted a ‘smart openness’ towards accelerating drug discovery and in general innovation. It moved away from the ideological debate that confronts the pro-open with the pro-IP advocates or the binary option between opening-closing, by providing an appropriate framework that foster disclosure of data towards scientific progress, while not threatening innovation or industrial collaboration from the research labs and commercial entities.

Additionally, besides providing the right incentives for data holders to join the data commons, the analysis of OT reveals also that the publication or disclosure of data is not enough to foster its re-use, but it constitutes only a small part of it. As a result, the case reveals different mechanisms that need to be put in place to foster data re-use by any external and potential data user:

4. *The user-centric design of the infrastructure:* There are significant barriers to the re-use of data. The case of OT reveals that the collaboration sought critical to developing an infrastructure that supported the two basic workflows that scientists and R&D managers in the private or public labs follow when investigating about targets and
their association with diseases. As such, the interface was designed as a Google-type engine where the user can search for a disease (one path or workflow) or a target (second path or workflow). As a result, OT will display all targets associations with a disease (in the first path) or the other way around for the second. Different visualization tools (in some cases third-party visualizations) support the task of such user, by displaying the data in a way that generates value and useful insights. In sum, it is worth noting that a usable data commons requires significant investment both in backend and front-end systems, not only in the design phase but also during its life requires clear policies and protocols related to data flows, including access and usage.

In the case of OSDU, the architecture of the platform is currently being developed through different cooperative activities amongst members to reach some agreements. Building a platform that is usable and offers the value expected by its members requires understanding the different processes that the platform needs to support. As such a participatory process shows mandatory.

5. **Interaction with data generators or data holders:** Sharing data in the public domain is necessary but, in some cases, an insufficient requirement for data re-use or other forms of commercialization. In some cases, depending on the type of data and contextual conditions around how data was collected and for what purposes, a close interaction from data generators or who generated the data in the first place, and those that seek its re-use it is still needed to understand what data tells, how it was generated, and how to interpret it. Following that rationale, the OT case, for instance, stresses the importance of having access to the data generators by the users and the importance of being in OT consortium to be closed to those who generate the data to foster its re-use.

6. **A federated approach towards data-sharing:** Some examples of data commons, including one of the cases shared in the present report, relied on federated approaches to effectively (and with some guarantees of quality) facilitate access to the increasingly available repositories of data. In the case of OT, the aim was not to store all the data or evidence objects related to targets but the organizations opted for a meta-data approach where the platform provides summary information about the heterogeneous and granular sources. Since extant databases are uniquely tuned to deal with many of the specialized data sources, developing meta-data or summary information of the available data proved to be an effective solution to deal with data heterogeneity. Additionally, the very user-friendly search engines (see point 4), APIs, analytical and visualization tools provided on top, and integration with important statistical applications further supports data users’ utility.

An additional observation that is worth noting, for its role in the successful governance amongst companies - which is present in both OT and OSDU cases- is the role of **open source and open standards** strategies, which are vendor neutral and provide the common ground required for companies to successfully agree and jointly develop such data commons. Being open source, the tools commonly built around such data commons architectures and being open the standards required to make them interoperable, provides the commercial neutrality and transparency required for companies to collaborate.

In sum, the combination of all these mechanisms created, in the inspiring case of OT and OSDU, a favorable environment for the organizations participating in it to share costs and realize scale in the collaboration to accelerate their R&D processes by allowing them to protect their strategic assets and competitive options while maximizing the positive externalities towards any potential data user.
6 Conclusion

Over the last few years, there has been a growing interest in how to design and operate data commons that are scalable and offer increased value to organizations. As data intensity and computationally intensive innovation methods conquer different industries, we expect an increase in required investment levels, rendering them prohibitive to single organizations. As data commons emerge in response, the two cases shared in the present report can be inspirational for the different governance modalities and mechanisms that need to be put in place to successfully navigate the competitive and cooperative dynamics inherent to such data commons.

As the types of data commons grow and diversify, we can expect to see a variety of higher-order services on top of such infrastructures with different offerings operating within and across data commons while displaying different data value-chain configurations - including third-party contributions - to maximize the value of the data.

European policy in this domain is certainly necessary to facilitate the creation of sufficient critical mass. But it should draw on existing experience for building such “data spaces”. In particular, what emerges clearly is that there are great challenges to data pooling that can only be addressed by sophisticated and carefully designed governance mechanisms. Data pooling, to be effective, needs to be framed in a wider context of precompetitive collaboration between companies, and needs to have a compelling business case: in other words, it has to be demand driven.

The cases presented in this report are pioneering and offer strong precedential value in understanding the challenges, but more importantly, in providing the toolbox to policy-makers for designing successful data commons.
7 References


