Rethinking transport
27–30 April 2020

Proceedings of 8th Transport Research Arena TRA 2020, April 27-30, 2020, Helsinki, Finland

Exploring the Establishment of a European Transport Research Cloud

J. Rod Franklin[a], Martin Böhm[b], Sarah Jones[c], Tatiana Kovacikova[d], Katarzyna Nowicka[e], Rafal Rowinski[f], Katerina Folla[g]*, George Yannis[g]

[a] *Kühne Logistics University, Großer Grasbrook 17, 20457 Hamburg, Germany*
[b] *AustriaTech, Raimundgasse 1/6, 1020 Vienna, Austria*
[c] *Digital Curation Centre, Argyle House 3 Lady Lawson Street, Edinburgh EH3 9DR, UK*
[d] *University of Žilina, Univerzitná 8215/1, 010 26 Žilina, Slovakia*
[e] *SGH Warsaw School of Economics, al. Niepodległości 162, 02-554 Warsaw, Poland*
[f] *European Commission, Directorate-General for Research & Innovation H.1., 1049 Bruxelles/Brussel, Belgium*
[g] *National Technical University of Athens, 5 Iroon Polytechniou str., Athens GR-15773, Greece*

Abstract

The aim of this research is to explore the potential of a European Transport Research Cloud (TRC) as a subset of the European Open Science Cloud (EOSC). On that purpose, a group of six experts was created to perform a preliminary analysis concerning the current practices, the main barriers, needs, and possible benefits of establishing a TRC. The experts collected data based on their personal experiences and contacts, prior research, as well as a survey was carried out with the participation of researchers from the transport sector. The results of this research led to ten recommendations, grouped into five thematic areas, which are considered essential for the development of a viable European Transport Research Cloud.

Keywords: open science; open data; research cloud; transport

1. Introduction

During the last two decades, a growing number of governments and funding agencies have begun to promote the sharing of research data in order to make research results available to a wider public, including other researchers, education and business sectors. Many policies for promoting open access data have been developed mainly by publicly funded authorities and agencies, while many private foundations have also begun to release data from their research. While data sharing policies vary widely by research domain, country and agency, they have many goals in common.

The European Commission (EC) has shown great interest in the areas of open science and open data, since it has adopted the philosophy that the money spent by the Commission on research should be leveraged in a manner that generates returns (scientific, economic, environmental and social) well after the originally funded research has concluded (Council of the EU, 2018). Within this context, the European Commission has been committed to open up science and has launched many actions in recent years to develop the required policies and infrastructure. More specifically, a vision has been proposed for the establishment of the European Open Science Cloud (EOSC), which aims to link existing infrastructures from research sectors and Member States in order to achieve the sharing of

research data. The upper aim of the EOSC is to provide a single point of access to all European research data and data services, world-class tools and standards required for their re-use.

The transport research community is consisted by a diverse group of researchers who collect and produce large amounts of data, either by tracking actual movements of goods or persons, recording sensor data from vehicles and infrastructures or capturing video of various transport-related phenomena. Unfortunately, most of the data collected by researchers are used once and then stored at points that are inaccessible to other researchers. In order to enhance the use of the existing transport research data, there is a need to establish a European Transport Research Cloud (TRC) in line with the EU's open science efforts and the EOSC. A primary aim for a TRC will be to provide researchers in the transport and logistics domain with access to open data sets relative to their research interests.

The objective of the current research is the exploration of the state of the art, the barriers, the opportunities and the needs for setting up a European Transport Research Cloud. On that purpose, a group of six experts was created, with aim to record current practices, the main needs, obstacles and the opportunities for data sharing in the area of transport research. The experts collected data and information based on their personal experience, their contacts and the available literature findings. Additionally, a survey was conducted aiming to identify the existing data documentation and sharing practices of transport researchers. The synthesis of the results of the overall work led to the formulation of ten recommendations grouped into five thematic areas, which are considered essential for the development of a sustainable Transport Research Cloud.

This paper is based on the results of the research conducted for the development of the report on "Analysis of the state of the art, Barriers, Needs and Opportunities for Setting up a Transport Research Cloud" for the European Commission.

2. Survey findngs

In the context of the current research, a survey was undertaken in the summer of 2018 to determine existing data documentation and sharing practices of transport researchers, and what the transport research community would expect from a cloud service. Within this survey, researchers from academic institutions and representatives of public authorities and the commercial sector participated. The survey included four thematic areas of questions: transport research data, cloud service requirements, opportunities and barriers and funding mechanism. A total of 87 responses were collected. Respondents came from 29 countries – primarily from the United Kingdom (11%), Slovakia (9%), Greece (9%), Germany (7%), Austria (7%), Israel (6%), Poland (5%) and Spain (5%).

The largest group of respondents were researchers from academic institutions (85%), followed by representatives of commercial sector (6%), public bodies (4%) and others. In terms of relationship with transport data, in most cases, participants responded that "analyse transport data" (28%), 25% of the participants responded that they "use transport data" and 20% of respondents "process data".

The participants in the survey were asked if some data should not be accessible via an open data model. A consensus between most respondents was reached that data should be available primarily for research purposes, however, some survey data might be more sensitive due to privacy issues. At the same time, 95% of respondents expect data to be described (supported with documentation and metadata information). In terms of expectations to these transport metadata and documentation, respondents would like to get information concerning descriptive metadata to aid discovery (16%), study design and methodological information (15%), and data dictionaries explaining abbreviations, lab notes etc. (14%).

In the second part of the survey respondents' opinions on requirements and expectations for a TRC service potentially offered by the European Commission were examined. More than half of the participants responded that they would use the TRC, while the remaining participants responded that this depends on the quality of the service provided. In terms of functionality of the TRC, respondents would expect to have access to existing transport data collections (18%), advanced search options to filter and find relevant content (17%) and open data sharing (11%).

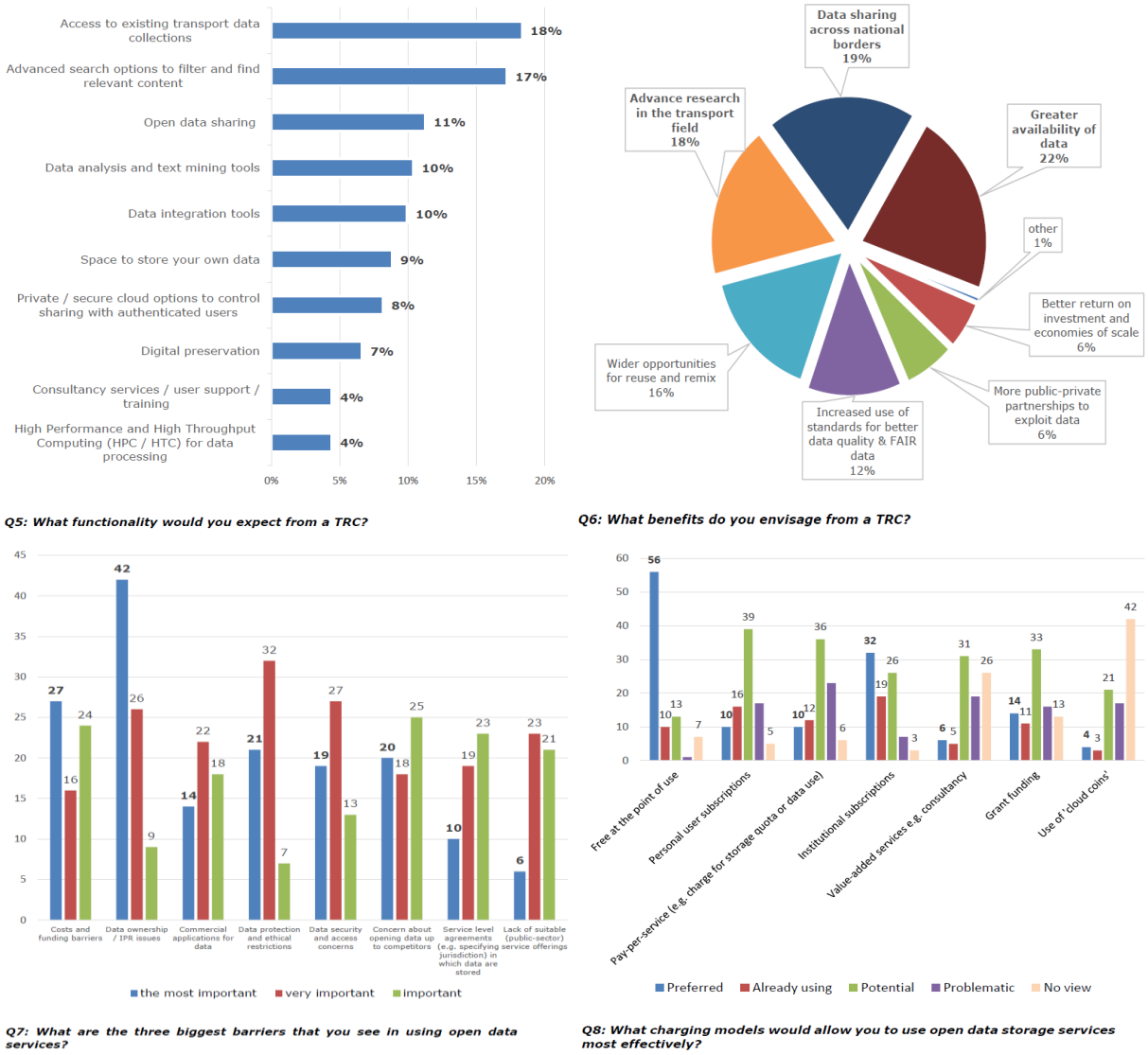

Figure 1: Indicative survey results on TRC needs, obstacles and expectations (Source: Bohm et al., 2018)

In the third part of the survey, participants were asked to underline benefits and barriers in using open access data. The most important benefit from TRC considered is the “greater availability of data", followed by “data sharing across national borders” and “advance research in the transport field”. As the three most significant barriers that can be foreseen in using open data services, respondents pointed out: (a) “data ownership / IPR issues”, (b) “cost and funding barriers” and (c) “concern about opening data out to competitors”.

In the last part of the survey participants were requested to evaluate potential funding mechanisms for the TRC. Currently, 76% of respondents do not pay for storing data in an open data service, while 21% do so. In terms of charging models that would allow to use open data storage services most effectively, respondents marked in the first place “free at the point of use”, next “institutional subscriptions”, and then “grant funding”.

3. Synthesis of results

3.1. Characteristics and scope of transport research data

In order to identify the scope of the Transport Research Cloud, the dimensions of the transport sector should be defined and well understood. The transport sector is consisted of different types of transport modes (e.g. road, rail, marine etc.), "intermodals" (e.g. inter-modal, multi-modal transport etc.), transport sectors (e.g. passenger, freight

transport), vehicles, infrastructure, evaluation perspectives, policy aspects, technology, applications and data (Rodrigue, 2017). The definition of these dimensions will also help to better explain the complexity of the transport research.

Concerning the existing transport related databases, that could be included in the TRC, some of them contain data from governmental entities, such as cities, prefectures, states and federated communities, which collect regularly data on transit, traffic, safety and other operational data. Commercial data could also be included, if commercial transport companies, service providers, vehicle manufacturers etc. gained interest in the TRC through their contribution. Additionally, data from research programmes within the transport sector, which are funded by governmental authorities, could also be shared with the TRC.

However, an important issue for the development of the cloud is to identify what data are considered as "transport research data" in contrast to transport data in general, which could be achieved through a more detailed analysis of the research problems that researchers are working on. The potential outline of transport research data classification should refer not only to the different modes and types of transport, but also to areas of transport research that enable and define the efficiency, safety, cost/value, environmental impacts, and security of transport operations. Additionally, all phases of the lifecycle of a transport project should be taken into account (planning, design, implementation, operation and management), as well as all types of content (raw and processed data, research outputs and publications).

Consequently, three main categories of research data are suggested to be included in the TRC:

- Original research data, e.g. data from Field Operational Tests (FOTs), Naturalistic Driving Studies (NDS), research results and research models from published papers.
- Operational data directly related to research, such as accident data, transport volumes data, etc. This category of data consists mainly of data from public authorities, either national or European/ international.
- Data from published transport research appearing in scientific journals, delivered at conferences, workshops, etc.

The main target of the cloud will be the collection and re-distribution of the transport research data. On that purpose, the datasets should be provided to the researchers for reuse in a curated and open manner. More specifically, a detailed description of the available data should be provided, catalogues (e.g., for datasets, services, standards) based on machine readable metadata and identifiable through a common identification mechanism available should exist. Data re-use requires also knowledge about the data itself. Consequently, high quality metadata should also be provided, describing for instance the conditions under which data have been collected, for which purpose, how they have been stored, processed, etc. (EOSC Declaration, 2017). These standards should be evaluated in light of the needs of the transport research community and adopted, modified or replaced depending on whether they are determined to be acceptable for the broad requirements of this community.

An important issue for consideration, related to both research data and operational data, is the size and complexity of datasets. “Big data” require ample storage space and powerful computers in order to be processed. Curating, storing and handling these large, and often unstructured datasets will also require specialized databases, data management systems and infrastructure to ensure that access and reuse are made as simple as possible. Finally, the development of new software tools will be required in order to enable researchers to search, browse, review and access available data. Software and search tools will make use of the metadata that annotates and identifies the underlying data. Ideally, all research data will be available through web APIs, so that they can be identified and accessed by search engines and automated systems.

Besides the need for specific transport data formats and the ability to have machines process the data, a clear understanding of the benefit of linking cloud platforms together in a federated manner forms the basis for providing access to the data that is stored on those platforms. Linking cloud platforms together will require standardizing communications protocols between the various clouds, search approaches for accessing metadata concerning the data stored on the platforms, and upload/download mechanisms, so that researchers can easily upload data and metadata to their preferred platforms and access cloud hosted datasets that they believe can be used in their research (Hey & Trefethen, 2005).

3.2. Current approaches to support data sharing

There are practically as many operational models for open data platforms as there are platforms (OECD, 2017), e.g. domain focused platforms, governmental platforms, cross-domain platforms, research society platforms etc. Current international open data platforms have generally formed around a specific "big science" project in which data requirements are not confined to the borders of nations. Many of these existing platforms have been in existence for a number of years, driven by the needs of the particular domain and research topic (e.g., CERN or SSRN). With respect to their applicability as frameworks for the TRC, they all have potential for implementation in managing open transport data.

Each platform has evolved operating models that address their particular user community and funding agencies. Examples vary from comprehensively curated data that is searchable with strong query and segmentation tools to simple data repositories that provide little by way of curation or tools. The key to which of the several models will be successful for the operation of any of the federated transport research platforms is that the platforms in question have well defined business models, a clear understanding of their stakeholder value propositions, and sufficient start-up funding to ensure that they gain traction with their stakeholders so that they can become relatively self-sufficient.

When it comes to the management aspects of the various open data platforms, it can be stated that platform management varies by how and for what purpose the platforms were established. Platforms that were formed via lead universities tend to be managed through the library function within the lead university. Domain specific platforms, such as the CERN platform, have dedicated management structures that are funded through the platform partners. Governmental platforms and platforms developed based on ad hoc governmental requirements, are managed through either pure governmental departments or public/private partnerships. Finally, platform sustainability requires a business model that generates value for stakeholders of the platform and funding sources that, recognizing this value, provide long term funding for the platform. However, very few platforms that have been studied to date, save for those that are funded purely by governments or supra-governmental bodies, have a truly sustainable business model and, therefore, funding sources (Duch-Brown et al., 2017).

Current recommendations on how the EOSC might become sustainable envision various "for fee" business models including "cloud coins" (Science Business, 2018), subscriptions and pay-as-you-go (EC, 2018) and continued government funding. Each of these ideas has the potential to act as a means of sustainable funding for the EOSC, however they do not define a truly sustainable business model for the EOSC.

3.3. Opportunities and barriers to transport data sharing

Enabled by recent developments in Information and Communication Technologies (ICT), cloud computing, artificial intelligence, machine learning and the Internet of Things (IoT), a large amount of data is generated, collected, processed and used in research. By the end of 2020 it is estimated that there will be over 50 billion connected devices globally collecting over 2.3 zettabytes of data each year (Statistica, 2018; Cisco, 2018). Open and easily accessible data will facilitate research across communities and countries, advancing the state-of-the-art in the field more quickly. It could also facilitate more public-private partnerships as commercial companies are encouraged to make their data available and research teams do not have to approach data "owners" individually and make separate agreements for reuse.

Some examples of where greater access to data could lead to advances in the area of transport research come for the increased use of FOT and NDS. These approaches to understanding how vehicles and people interact create extremely large datasets that contain much more data than those that are used for the original purpose of any single study. These data can be reused by researchers, giving them the opportunity to reduce the funding and effort requirements. Another example where sharing transport data provides a significant opportunity are advanced mobility solutions, such as journey planners and control systems that can save businesses money, increase safety and reduce congestion.

One final area where significant research benefit could be realized through data sharing is the integration of cross modal, multi-modal and synchro-modal transport operations. The ability of researchers to obtain, analyse and integrate data from multiple modes of transport operation will facilitate their efforts to develop realistic models of how transport actually occurs and lead to improvements in the actual delivery of transport services across Europe.

These benefits were also supported by the individual researchers that responded to the TRC survey.

However, numerous challenges still hinder the reuse of transport related data (Janssen et al., 2012). Transport data are often stored in distributed data silos, which makes data analysis difficult and causes problems with many analytical models. Another obstacle to the efficient exploitation of these data assets is the fragmentation of data ownership and a lack of interoperability between datasets and platforms. Data ownership varies by who generates and collects the data. Transportation system operators (both public and private), various transport focused agencies local governments, transportation researchers and other generators of transport data may not be willing to share their data due to privacy, legal liability, IP, competition, or cost related issues. Additionally, the existence of different stakeholders in the transport sector leads to great variations in their interests and thus, to variations in their requirements for data access.

When working with commercial partners, the sensitivity of the data may preclude the use of cloud services, particularly when it represents trade secrets, evidence central to forthcoming patents and other data in commercial interests. The data owners may be unwilling to use cloud services for fear of data breaches or unauthorised access. Privacy issues, particularly associated with data sets that are granular in nature and contain time and date information on individual trips, cause many data owners to restrict access to their data sets. In addition, issues with how data is used by external researchers can create legal problems if the data set is employed in a manner that places individuals at risk or that yields results that are not valid. Transport data are often ethically or commercially sensitive and thus, tight controls are required concerning access to the data and ensuring that they are stored within geographic areas where the legislative frameworks match European data requirements. Transport research may include human participants, so controls are needed concerning the access to the data and how they can be reused.

Another barrier for reusing the open data is the data quality. As a result of the diversity of data sources, all developed with a particular problem definition in mind, numerous data types and data sets of differing quality are produced in the transport domain. This creates problems for those wishing to integrate the data for further analysis. Moreover, the data that are collected without consideration for any potential reuse usually lack any standards-based collection structure. Variations in hardware and software used for collecting the data also create problems for reuse, as hardware and software vendors generally utilize proprietary standards in their products. Finally, a critical challenge for Europe is ensuring the availability of skilled experts in the data ecosystem, including data scientists and engineers with expertise in analytics, statistics, machine learning, data mining, and data management.

3.4. Needs for transport data and the TRC

The types of cloud services required in a TRC reflect the needs of researchers observed in other domains, including access to datasets, search tools, data analysis, storage, data sharing and preservation. What emerged most strongly from the survey was the need for access to datasets and advanced search tools to help research communities assess the relevance of content for their work. Since research is global and transport data inherently crosses geographic boundaries, having a single point of access to discover relevant datasets from different countries is significant. The quality of the datasets that a researcher finds in the TRC is also critical. Data sharing is not yet common practice in transport research so increasing the amount of high quality data made available is a priority to not only engage researchers, but also build the reputation of the TRC.

Discovery and access mechanisms should provide a coherent catalogue with advanced search options to enable filtering based on criteria appropriate to the particular transport research question being asked. Data should be automatically indexed so shared datasets are discoverable via external catalogues and search engines, not just via the TRC. To encourage the broadest reuse of the stored data, the TRC should provide facilities for collaborative data access so that consortia can access the complex data available applying their combined skills to analyse the data.

In order to make the transport research cloud a viable long term project, it should provide value to the various stakeholders that will use, supply, operate and fund it. Each of these stakeholders, due to their different needs, will have a different view of value based on their role (data suppliers, data users, platform operators, TRC funders, society). These primary value areas will need fine tuning for the various platform business models, so that all stakeholders perceive the value they receive from participating in the cloud.

4. Recommendations

Based on the above results, ten recommendations to the EC for the establishment of a TRC as a subset of the EOSC were developed, presented in the following five thematic areas:

4.1. Reusable research data

Since not all transport data are of value to the research community, the first step for the development of the TRC is to define what research data is. The EC is recommended to gather researchers and various data users in order to define which types of data should be included in the TRC. Moreover, the researchers should be motivated to reuse existing datasets. For this reason, the objections of researchers behind the limited use of data collected by others should be defined through a detailed study and recommendations on how to overcome these objections should be developed. Researchers should also be informed about best practices in the collection and reuse of data so that their concerns can be overcome.

4.2. Data as a public good

It should be made clear to the research community that data collected within projects and processes paid by public money needs to be made available to the public as an asset that can be reused by others. It is, therefore, recommended that any data collected under contracts that are paid by tax payer funds to be classified as public data and be available in the EOSC and/or its subsidiary cloud infrastructures (e.g., the TRC). However, it is recognized that certain data, due to privacy or secrecy requirements would need to be excepted from this requirement. Furthermore, in order to encourage the researchers to release the public data, the intellectual property that the researcher has developed should be protected. On that purpose, clear guidelines that separate the public asset from the private asset should be developed.

4.3. Standards

Research data should be available in a standardised format in order to be easily findable and reusable by other data users. On that purpose, the EC should gather members of the transport research community and governmental authorities that produce public data in order to define the necessary standards for the collection of transport data by public institutions, the data formats these data should adhere to, the metadata that must be used to describe the data, and formats of this metadata so that automated search engines can easily find and characterize the data.

4.4. Infrastructure

Transport research datasets can be extremely large, composed of unstructured data requiring sophisticated data management and analysis technologies to store, curate and add value to. The TRC must be able to handle these complex datasets in a seamless manner if it is to be viewed as a valuable tool for transport researchers. The infrastructure to handle this requirement, along with the personnel needed to operate the infrastructure, will not be inexpensive. The EC is therefore recommended to define the infrastructure and operating requirements for a TRC, in order to ensure that an appropriate level of service can be provided at a cost that is understood by all stakeholders.

In contrast with the passive data warehouses, the TRC must actively work to create value for its users and stakeholders. Besides services such as easy upload/download, quick search and query, rapid data extraction, etc., the TRC should provide services that will excite data users and encourage them to continually visit and use the platform. Thus, a sustainable business model should be built and the requirements for making the TRC a "go to" place for researchers should be defined though dedicated studies and pilot implementations. It should also be taken into account that the TRC will be a supporting pillar of the EOSC and as such must conform to processes and procedures established for that overarching cloud infrastructure in order to avoid confusion and conflict in the future.

4.5. Incentives, education, and training

Researchers need to be motivated to publish not only the results of their research, but also to make the data used available to other researchers. EU policies for academic promotion, training, publication, and knowledge

generation at public universities should be examined and harmonized to ensure that researchers are uniformly trained in the process of placing their research data into the TRC. Additionally, universities are recommended to provide proper incentives to their faculty and researchers to ensure that their research data are made openly available, and that proper credit for the generation of data that is reused is given to the individuals who originally collect the data.

Finally, training/education programs are recommended to be developed for existing and future researchers, libraries and librarians, data curators, and other individuals who will be needed to carry out the development of a data cloud.

5. Conclusions

The present research focuses on the requirements for sharing data within the transport research and more specifically explores the potential of the establishment of a Transport Research Cloud as a subset of the European Open Science Cloud (EOSC). Within this context, the needs of a TRC in research data, software, analysis methods, ways of storing large data and infrastructures, the possible barriers to be overcome are highlighted, as well as potential models for a viable Cloud. Based on the above analysis, ten recommendations to the European Commission grouped into five thematic areas were developed concerning the development of a TRC.

It should be emphasized that the recommendations made for the development of the TRC are, by their nature, broad. This is due to the fact that much work needs to be done in setting standards, understanding the proposals and needs of related stakeholders, identifying the infrastructure that could be used and its requirements for the proper operation of a TRC. This means that additional research is needed before the creation of a TRC, while important information can be drawn from the development of the EOSC, so that the establishment of the TRC will lead to a growing service and not an interesting but unused service.

Finally, it is essential that the TRC will be considered as a sustainable long-term project. On that purpose, the TRC must focus on understanding how it can provide value to its users, educate the users and continue to add new value in order to become it becomes the go to place for doing research. This could be achieved by linking the TRC with open science and transport research projects. If this final point is not taken into account, the TRC, no matter how well is constructed, will not achieve its goals.

Acknowledgements

This paper is based on the results of the research conducted for the development of the report on "Analysis of the state of the art, Barriers, Needs and Opportunities for Setting up a Transport Research Cloud" for the European Commission.

References

Böhm M., Franklin J. R., Jones S., Kovacikova T., Nowicka K., Yannis G., (2018). Analysis of the State of the Art, Barriers, Needs and Opportunities for Setting up a Transport Research Cloud. European Commission, Brussels.

Council of the European Union, Draft Council Conclusions on the European Open Science Cloud, 9029/18.

Duch-Brown N. B., Martens, B., Mueller-Langer, F., (2017). The economics of ownership, access and trade in digital data. JRC Digital Economy Working Paper 2017-01.

EC, (2016). Guidelines on FAIR Data Management in Horizon 2020. Version 3.0. European Commission. Directorate-General for Research and Innovation. Brussels.

EC, (2018). Prompting an EOSC in Practice. Interim report and recommendations of the Commission 2nd High Level Expert Group on the European Open Science Cloud. Brussels.

EOSC Declaration, (2017). EOSC Declaration. European Open Science Cloud. New Research & Innovation Opportunities. Brussels.

Hey, T. Trefethen A.E., (2005), "Cyberinfrastructure for e-Science," Science, vol. 308 no. 5723, pp.817-821; Research Data Alliance (2015), "Sustainable Business Models for Brokering Middleware to support Research Interoperability," Sustainable Business Models Team Report to the Brokering Governance Working Group, Research Data Alliance.

https://home.cern/

Janssen M., Y. Charalabidis, A. Zuiderwijk, (2012). "Benefits, Adoption Barriers and Myths of Open Data and Open Government, Information Systems and Management, vol. 29, no. 4, pp. 258-268.

OECD (2017), "Business models for sustainable research data repositories", OECD Science, Technology and Industry Policy Papers, No. 47, OECD Publishing, Paris, https://doi.org/10.1787/302b12bb-en.

Rodrigue J. P. (2017), The Geography of Transport Systems, Fourth edition, New York: Routledge. ISBN 978-1138669574

Science Business (2018), “How the Science Cloud could pay its way,” Science Business Network’s Cloud Consultation Group, Brussels.
Statistica (2018), https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/
Cisco (2018), http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html
Wilkinson, M.D., M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman (2016), “The FAIR Guiding Principles for scientific data management and stewardship,” Scientific data, vol.3.
www.fot-net.eu