Roadmap for a European open science alliance for ATM research

Tatjana Bolić, Andrew Cook
Centre for ATM Research
University of Westminster
London, UK
t.bolic@westminster.ac.uk

Rainer Koelle, Enrico Spinielli, Quinten Goens
Aviation Intelligence Unit
EUROCONTROL
Brussels, Belgium

Martin Strohmeier
OpenSky Network
Burgdorf, Switzerland

***Abstract*—We here propose an open science alliance for ATM, advocating for open data in such a framework. The benefits of adopting an open science approach are to be found, *inter alia*, through the independent verification and validation of reported impacts/results and achieved performance levels, i.e., through reproducibility. We consider that this can only be achieved through: (1) open access to scientific methods and data utilised; (2) open access to (analytical) code and methods; (3) open review of reported analyses/research. The proper application of such practices will reduce the innovation cycles in ATM, which is much needed by industry and society. Steps for forming an open science alliance for ATM are described. We propose further initial, specific recommendations for supporting open data and improved access for research.**

***Keywords*—Open science, open data, open code, open access, reproducibility, research community.**

I. Introduction

The European Commission (EC), through the Horizon Europe programme, strives for open science, which is now a mandatory part of the funding actions: "Open science is an approach to research based on open cooperative work - that emphasizes the sharing of knowledge, results and tools as early and widely as possible. It is mandatory under Horizon Europe, and it operates on the principle of being 'as open as possible, as closed as necessary' " [1].The open science is based on the FAIR principles: findable, accessible, interoperable, and reusable. It provides a wide range of benefits, intended to lead to:

- efficient academic research due to having more access to data, methods, results and publications;
- knowledge transfer and innovation;
- greater impact, visibility, and citation rates; and,
- eventually providing increased transparency and promoting higher levels of trust between research/governmental institutions and citizens.

Open science practices involve: open access (to publications), open data (sharing research data), open source/code (sharing software), open methodology (sharing models, methods, etc.), open peer review and open educational resources.

When these practices are suitably implemented, they foster greater efficiency in research through increased collaboration, higher levels of verification/validation, and reduced duplication. Furthermore, collaborations broaden the user community, improve methods and code testing, and enable more reproducible research [2]. Reproducibility is an important principle of the research process, as we can have more confidence in results that are confirmed through multiple findings, increasing the transparency of research. The ability of interested and independent researchers and practitioners to validate findings requires a spectrum of open science practices, including open access to the underlying data analytical code.

Each field of science has slightly different characteristics and requirements when it comes to research and open science. Here, we are focusing on research performed within the Single European Sky Air Traffic Management (ATM) Research (SESAR) project. SESAR research is based on public, EC funding[1] and, as such, it should be, and is generally freely available to the research community and industrial or institutional stakeholders. In general, the public deliverables of the funded projects are available through the EC's CORDIS platform[2], and much of the research deliverables from the first SESAR programme can be found in the EngageWiki[3]. Research papers from SESAR 'exploratory research' (ER; i.e. research mostly at low technology readiness levels (TRLs)), is published through the SESAR Innovation Days conference website[4], and other related (peer-reviewed) ATM/aviation conferences, and through open access articles in academic journals (as requested by Horizon Europe rules). The more mature research, the results of SESAR 'industrial research' (IR), is published in the form of SESAR 'Solutions'[5] [3], which contain the specifications to enable the industrialisation of the Solutions.

Even though research *results* are generally accessible, over the years of running the SESAR programme, *requests for open data* access became increasingly articulated. In particular, this

[1] The research programme is managed and executed through the SESAR 3 Joint Undertaking, previously the SESAR Joint Undertaking.
[2] https://cordis.europa.eu/en
[3] https://wikiengagektn.com/EngageWiki
[4] https://www.sesarju.eu/sesarinnovationdays
[5] A Solution represents a change in the way ATM is performed. ATM Solutions are new operational concepts, procedures and relevant technologies.

goes in hand with the emergence of data-driven methods, especially artificial intelligence / machine learning, which require much larger (e.g. longer-term) data sets. In this paper we focus on the open data, the issues surrounding the access to data and benefits of opening it. Furthermore, we propose a roadmap towards the *Open science alliance for ATM*, a community dedicated to practising open science, for which the open and accessible data is the foundation.

II. BACKGROUND

The majority of ATM research in Europe centres on the SESAR project, which was established by European Council Regulation No. 219/2007 [4] with the ambition "to integrate and coordinate research and development activities which were previously undertaken in a dispersed and uncoordinated manner". The development and validation of SESAR Solutions is managed by the SESAR 3 Joint Undertaking (JU), which was first instituded in 2007, and extended in 2014 [5], and in 2021 [6]. The SESAR is built on a consortium of industry and research organisations.

The programme covers different aspects of ATM, incorporating new topics, such as: improving the environmental impact of ATM; enabling multimodality and new types of air vehicles (e.g., drones, unmanned aircraft systems, air taxis, super-high altitude aircraft, hydrogen and electrically propelled aircraft); integrating new (for ATM) techniques and technologies (e.g., machine learning, AI). This, and the necessity for a wide-ranging transformation of ATM to provide cost-efficient air navigation services that support sustainable aviation [7] are also linked with the great need to shorten the time from idea to deployment in ATM, which often takes decades.

Furthermore, the Covid-19 impact "shows that we need an ATM system that is sustainable, scalable, and resilient, which does call for the envisioned transformation" [5] and faster innovation cycles. This is also a requirement recognised by the international community. For example ICAO's Global Air Navigation Plan[6] - currently in the process of being updated - identifies higher levels of agility and flexibility as key features of the future air navigation system. The Performance Review Commission (PRC) also identified the ATM transformation in its flagship approach/transformation support strategy [8].

Taking all the above into account, the openness, transparency and reproducibility of research under open science practices **can enable faster innovation cycles** and will be a key ingredient for driving the transformation over the next years. Over the course of SESAR programme, a large number of Solutions have been delivered, up to TRL6 (which is still not a full system, as per TRL8 - see next section). Many still require additional data required by the regulators, as ATM is a safety-critical industry and thus subject to regulations. The input side of the innovation pipeline – the ER (low TRLs), are often hampered in their research efforts by data access, as has been documented by the Engage Knowledge Transfer Network (KTN) project [9], and discussed later. To sum up, the entire ATM innovation pipeline is dependent on data, for research, development, industrialisation and even after deployment, for post-ops assessments. The easier it is to access needed data, the shorter the innovation cycle becomes. In the next sections we expand on Solutions, data access and documented data issues, which are followed by the description of the OpenSky Network, as an example of open data, code and community that grew around the 'open data' principles.

[6]https://www4.icao.int/ganpportal

A. SESAR Solutions

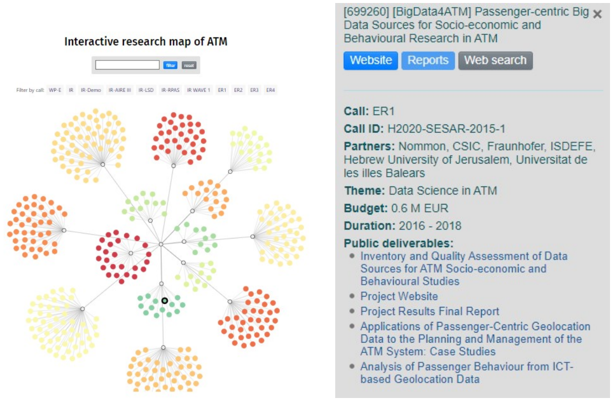
To date, the SESAR programme has 60 Solutions in the pipeline (between the ER and IR), and it has delivered 127 Solutions, 70 of which are deployed [3]. Each Solution is delivered with a collection of safety, business, performance, human performance cases, and the material needed to support standardisation and regulatory requirements. To reach TRL 6, the Solutions need to pass the validation and verification steps. However, this process does not allow for independent verification of established claims (e.g., performance levels, benefits), partly due to the funding mechanism, partly due to culture (that we expand on in section III-B) However, there is a stepping stone between the Solution developed by SESAR (delivered at TRL6) and a Solution to be deployed requiring TRL8. This gap includes activities such as standardisation, development and the collection of evidence for certification. These activities are often data intensive and expensive. They might also be performed by different (industrial) organisations. The respective data may be restricted to the specific requirements of the assurance case, comprise proprietary methods, or are confidential for other reasons (e.g., safety).

B. Engage

The 'Engage' (1) KTN ran from January 2018 to June 2022, and now continues as 'Engage 2', launched in June 2023[7]. It was tasked with promoting and *facilitating* the development of ATM research in Europe, through the development of multiple, parallel activities. One of these was the EngageWiki [10], which collected most of the research outputs since SESAR1, clustered by topic similarity (using semantic similarity analysis) in an interactive research map of ATM (see Figure 1). This opening of the access to European ATM research also served to inform the ATM future concepts roadmap, which can also be found on the EngageWiki.

The KTN consolidated views on the most important research barriers and enablers, with the associated lessons learned [9]. One of the most cited issues was data availability, that we will term *"open data"*: in particular, we focus on its accessibility and licence-/regulation-free use. This has been recognised by various stakeholders as a bottleneck in SESAR's ER, as vital data is often difficult to acquire, and often subject to non-disclosure agreements and/or other limitations, such as restricted use for one specific project. As such, this lack

[7]https://sesarju.eu/news/engage-2-fostering-knowledge-transfer-air-traffic-management-research-and-innovation



(a) Research clusters (b) Project details

Figure 1. EngageWiki research clusters.

of open data, presents a "barrier to improving experimental comparability across projects" [9]. The KTN collected opinions of many projects and PhD students. A common denominator was that they spent approximately 6 to 12 months (sometimes more) in acquiring, consolidating and cleaning data. This observation was a recurring theme in the workshops organised by the KTN. The data required for ATM research is composed of different types, and the exact requirements vary across the projects. Some data can be obtained freely (e.g. EUROCONTROL R&D data, ADS-B data from the OpenSky Network), some need to be paid for (e.g. schedule data, passenger itineraries and fares), and some need to be acquired from multiple sources if a greater geographical area is being researched (e.g. MET lightning or radar observations, public transit schedules) [9]. Since, frequently, some sort of licensing and non-disclosure agreement is required, the input data cannot be shared. Depending on the specificities of the licensing/agreements, sometimes the results of the research can be shared. However, if the input data cannot be similarly shared, achieving the desired comparability and reproducibility across projects and between research groups becomes close to impossible. A couple of solutions for better data access for ER projects were proposed by the KTN, and were the inspiration for the alliance proposed in this paper:

- "Creation and application of non-disclosure agreements regarding the acceptable *form* of sharing of confidential (or subject to GDPR) information by the data owners, having in mind that for the research results to be validated and/or reproducible, the sharing of input data is required";
- "Creation of a framework to share ATM-relevant data (including MET and multimodal data), to afford easier access without having multiple agreements in place, which would require the provision of centralised licencing for certain ATM and commercial data (and/or the creation of synthetic datasets for the ATM community)." [9]

C. *OpenSky Network*

The OpenSky Network functions as a collaborative sensor network, collecting surveillance data for air traffic control purposes. Its core mission is to grant the general public access to real-world air traffic control (ATC) data and *promote the advancement of ATC technologies and processes*. Since 2013, the network has continuously gathered air traffic surveillance data, distinguishing itself from commercial flight tracking networks such as Flightradar24, or FlightAware, by preserving the original Mode S replies in an extensive historical database accessible to researchers and scientists across various fields. Initially, the non-profit network began with eight sensors in Switzerland and Germany but has expanded to include over 6500 registered receivers worldwide. As of 2023, OpenSky's dataset spans over a decade of aircraft broadcast data. While its initial focus was on ADS-B, it broadened its data range to include the complete Mode S downlink channel in March 2017. More recently, it integrated other technologies such as FLARM and VHF. The dataset comprises over 35 trillion Mode S replies, experiencing a peak influx of over 20 billion messages per day. The global data reception of the OpenSky Network relies on a crowd-sourced network of receivers, primarily comprised of enthusiasts, academics, and supporting institutions. Each sensor's coverage is limited by the range of antennas' line of sight, typically around 400–500 km for the best-performing antennas reaching the radio horizon. The organic growth areas of such a crowd-sourced network effectively represent densely populated and wealthier regions worldwide. Between 2018 and 2023, the network's global coverage (see Fig. 2) reached saturation, with many new sensors mostly enhancing reception at lower altitudes in areas already covered in Europe, the US, and other developed countries. However, notable coverage expansions are still observed in the Middle East, South Asia, and New Zealand. Geographical regions such as deserts and oceans lack ground-based coverage due to physical constraints, prompting commercial ADS-B providers to partially rely on space-based ADS-B or ADS-C data [11]. OpenSky, too, will add ADS-C data in the future.

OpenSky offers several interfaces to access the live and historical data (using different underlying technologies). The OpenSky community has developed extensive tooling for easy data access, processing and visualisation most notably *traffic* [12] and *pyModeS* [13].

For further details on the history, architecture, and use cases of OpenSky, please refer to [14], [15]. OpenSky data has been used in more than 470 publications as of September 2023. While the trend is growing (see Fig. 3), open data is clearly only used in a fraction of the air traffic data-based research conducted every year, as many other publications are using restricted-access data, for a variety of reasons, which are explained in the following section.

III. Barriers and Opportunities

All ATM research projects need data. As the ATM systems are complex, interconnected and heavy data consumers and producers, anything innovative also needs to be based on appropriate data, often from multiple sources. As already explained, it often takes between 6 and 12 months to obtain such data. Taking into account that the duration of most of

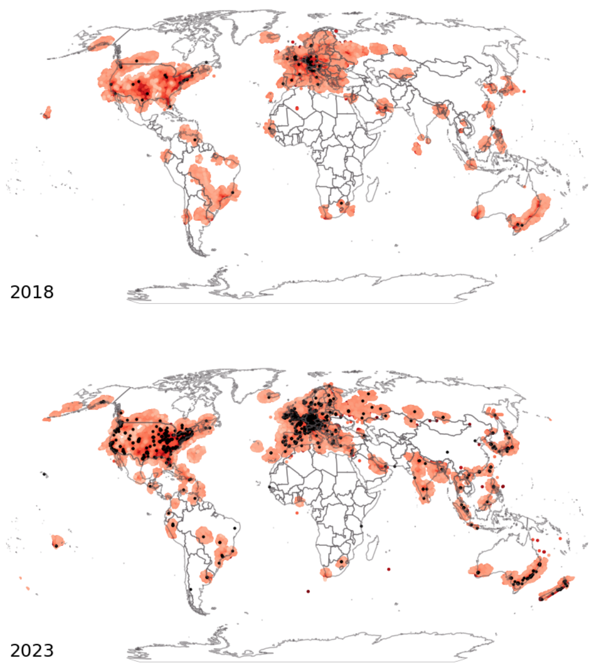


Figure 2. OpenSky's global coverage in 2018 and 2023

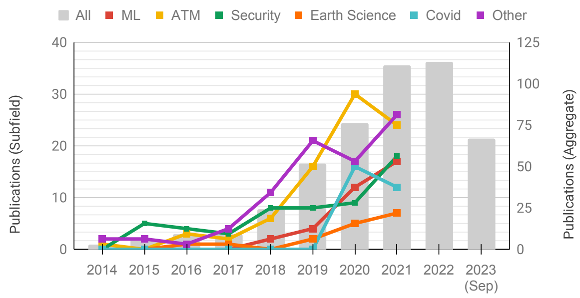


Figure 3. Publications using OpenSky Network data from 2014 - 2023.

the SESAR (and Horizon Europe) projects is 24 months, and that modelling, testing, and validation cannot be performed without data that arrives only half-way through the project duration, the possibilities of ultimate progress within a single research project are very often restricted by data availability. Furthermore, most of the data in ATM is treated as confidential by the data providers, which, in turn, requires various licensing or non-disclosure agreements. The impact of this is twofold – the time required to complete the appropriate paperwork and, more often than not, the limitations on what the data can be used for: very frequently linked to a specific project only, thus precluding any input data sharing and dictating the shape or format of results' sharing.

Such a state of affairs severely impairs the reproducibility of ATM research, or even precludes it. Many fields are moving towards requiring data and code to be shared for acceptance in journals/conferences. Provocatively speaking, a paper without openly available data is not trustworthy, and might not be even considered science.

A. General availability of data for ATM research

In this section we describe the availability of some of the most frequently used datasets in ATM research, and the impacts on research. The list and usages are not exhaustive, but serve as an illustration of the ever growing need for open data.

ATM manages traffic and as such, almost all research needs some form of **flight data**. EUROCONTROL provides free access under license to archive data for R&D[8] that is two years old, for March, June, September and December. Two-year old data are not highly appreciated by scientific publications. The lag can also cause other issues, such as delays to post-Covid data release, where researchers await more 'normal' traffic data. The flight data cover commercial flights (no military, state and general aviation) trajectories, last-filed and actual flight plans. The trajectories contain crossings into flight information regions (FIRs) and ATC unit airspaces. Theses data are useful for research into strategic (i.e., from planning phase) and some tactical (i.e., actual operations) innovations in ATM that deal with ATM performance – airborne and at airports. For example, as the data contain aircraft registration numbers, one can investigate reactionary delays and turnarounds. Further, originally-filed flight plans are not included, thus the original intention of the airspace user is not known, nor is it possible to (accurately) model the situation *before* responses to flow-management regulations. Furthermore, the trajectories here are not exactly 'true' trajectories, but approximations between the filed and actual trajectories as they diverge only When the difference between the two is large enough[9]. Thus, such data are not suitable for more precise analyses, such as those on operations around airports.

More precise actually flown trajectories are available from the OpenSky Network. The ADS-B/Mode S messages broadcast by aircraft contain the positioning, altitude, timestamp and other information. The frequency of certain messages can be quite high (e.g., twice per second for ADS-B) enabling creation of precise trajectories, at the expense of high memory and data preparation requirements. Free access to OpenSky Network's historical data (for research purposes), combined with shared tools make data acquisition and preparation faster, shortening the preparatory phase of research projects. Moreover, the significant amount of data available is conducive to the application of data science, machine learning and artificial intelligence techniques, that require large sets of data for model training.

Airspace and airport capacity data. 'Airspace data'[10] refers to the division of airspace into sectors. Each air navigation service provider (ANSP) divides its airspace into sectors, portions of airspace controlled by a controller team. Depending on the traffic, the supervisor decides which opening scheme (i.e., which sectors) to activate, to provide the best service

[8]https://www.eurocontrol.int/dashboard/rnd-data-archive

[9]For example, 5NM lateral difference or 10 minute delay.

[10]Also referred to as 'environmental data' – the environment where operations take place.

for the foreseen traffic, given the staffing levels available. The airspace data contains the indications of declared capacities of sectors, under different operational conditions, and the declared airport capacities. Currently, the airspace data is not available for academic research as it is not available in the R&D archive. Even though air traffic still has not reached the 2019 levels[11], delays have reached the levels of the capacity crisis in 2018/2019 [16], pointing to the need for further developments in demand-capacity balancing. This requires understanding the airspace and airport environments, their capacity characteristics and traffic characteristics.

Investigating the impact of delay on the network and, more importantly, on passengers, requires information on **schedules and fare data**. In Europe, schedule data are available to purchase from providers such as Cirium or OAG Aviation. Airline passenger and fare data are available from global distribution system providers and IATA, for example, and are often extremely expensive. Airport passenger data are available at the airport level from a variety of sources, including civil aviation authorities and individual airports (usually for purchase), or freely from Eurostat[12] at more aggregated levels. Commercial providers impose limitations on data usage: no sharing of the purchased data is a standard requirement (understandable from the business point of view), but this can also extend to the usage of such data (which can sometimes be limited to the research project for which the buyer enters into contract).

In the USA, the Bureau of Transportation Statistics (BTS) makes scheduled times available for the 17 airlines that "have at least 0.5 percent of total domestic scheduled-service passenger revenues" [17], where this reporting is required by Federal Aviation Administration regulations. This has led to a significant volume of European research (including PhDs) using these datasets, when schedule information is needed. The data include a sample of passenger and fare data.

Recent years have seen the introduction of multimodality topics into ATM research, as the EU is striving for improved mobility. To address multimodality challenges, additional data are needed from other modes of transport (especially rail, e.g., as an appropriate aviation short-haul substitute), which is often even harder to obtain than aviation data. The Digital Europe Programme[13] that focuses on bringing digital technology to the EU, is funding project aimed at creating European mobility data spaces[14]. The initial goal is to map existing mobility data-related ecosystems and initiatives, and to propose a common European mobility data space. These initiatives demonstrate that mobility data are recognised as important, but also that they are not easy to find and access, for both academia and business. As an example, train schedule (historic) data is not usually freely available (it differs across states), but may be obtained from rail associations. Train capacity and occupancy data are often collected only at the single company level.

[11] https://www.eurocontrol.int/Economics/DailyTrafficVariation-States.html

[12] https://ec.europa.eu/eurostat

[13] https://digital-strategy.ec.europa.eu/en/activities/digital-programme

[14] E.g., https://mobilitydataspace-csa.eu/

Weather impacts are different across the flight phases (e.g., take-off, climb, cruise, etc.), for which **meteorological or 'MET'** data are needed. The provision of MET data is freely available for aviation purposes, it is standardised and regulated [18], [19]. These provisions cover certain types of observations and reports, such as METAR[15] and TAF[16], and these ('live' ones) can be easily accessed by any interested party. Problems can arise when historical data are needed, as not every provider of this information (usually each state covers its own territory) can provide archiving services. An example of a collaborative platform that collects and shares data is Iowa Environmental Mesonet (which collects environmental data) where METAR and TAF historical information can be found[17]. The European Centre for Medium-Range Weather Forecasts[18] provides different products for research and public access, such as the ERA5 atmospheric re-analysis of the global climate, and many others are available from the Copernicus programme. These datasets contain vast amounts of atmospheric data, but might be cumbersome to prepare and analyse for transportation purposes, as these are not standard data set used.

For aircraft performance data - EUROCONTROL's BADA4 data has become an industry standard. The downside (for expediting the research process) is that it can take a long time to get access approval, which is given per project only.

Most ATM problems require multiple data sources. Sourcing and preparation of data from any single database requires time, which can grow significantly if information from different datasets is needed.

B. Barriers

The introduction highlighted the benefits of an open science (data) approach. An underlying requirement is the cultural change embracing transparency, reproducibility, and a public discourse. This change will require removing barriers to opening data. Such barriers can be broadly categorised as: reputational repercussions; privacy concerns (both at individual or organisational levels); business considerations (comprising, *inter alia*, the confidentiality of business-critical data, intellectual property protection, legal liability or restrictions); national security and defence issues; costs (associated both with the collection, processing, and storage of 'massive' data in general, and the provision/maintenance of open data).

A recurrent theme when discussing open data is data ownership (and, implicitly, the associated authority controlling its use). Air transport is a public transportation domain. As a public service, involving various stakeholders and organisations, their processes and operations are governed by a variety of rules and laws. Traditionally, air transportation data collection, processing, and distribution was dependent on dedicated technologies (e.g. surveillance data required radar installations). Interference of the safe execution of flight is regulated in all national laws, prohibiting non-aviation

[15] Meteorological Aerodrome Report

[16] Terminal Area Forecast

[17] https://mesonet.agron.iastate.edu/request/download.phtml

[18] http://www.ecmwf.int/

stakeholders from interacting with associated aviation data. In consequence, state organisations or service providers involved in air traffic management, including MET and control or airlines are considered 'data owners'. In light of the emerging debate on climate change – including the role and impact of air transportation – a wider interpretation constrains the exclusive private authority over data ownership. The continual pressure of the climate impact will likely tailor exemptions related to privacy: individual rights to support the broader societal well-being. Accordingly, a prerequisite for open science/data is a proper separation of the legitimate use of data from the concerns of individuals, organisations, and legal/regulatory regimes.

While of a different nature, (concerns regarding) loss of reputation, confidentiality, and national security/defence fall into a similar category. The concern is the disclosure of information, or behaviour or level of capability that may result in 'unfair' assessment/comparison or leaked secrets. Similar to operating a vehicle on the streets, the operation of aircraft follows certain rules and requires, *inter alia*, the display/dissemination of unique flight identification, to ensure safe operations in the public space. Investigative journalism or use of such information by environmental interest groups (e.g. Bernard Arnault's selling of his private jet [20], Elon Musk tracking, and recent NATS system disruption [21]) have gained high visibility in the press. National security and defence operations may apply specific rules, which might ultimately entail limiting the identification of aircraft or their interaction with other sensors. The principle of open data *per se* does not infringe these considerations, as it relates to the operation of flights within the existing rules.

The emergence of 'big data' and recent data processing and storage techniques are not cost-free. Next to the pure technological operating cost (e.g., server hardware infrastructure, electricity), the collection of the data, developing data interfaces, converting received data into a format to be stored, and potential data cleaning efforts present additional costs. Community-based approaches help to share the burden, increase coverage, and strengthen the collection of data. Challenges of maintaining and sharing collected and (partly) pre-processed datasets, could be overcome through different established repositories, such as the OpenSky Network or Zenodo[19]. Web-based networks of enthusiasts, researchers, and institutions could help facilitate the effort of data collection. Further example can be found in the Open Aviation Data Github repository[20].

Aviation is a data-rich environment, and while certain operational data are readily processable, other data sets require a substantial suite of sensors. MET data, particularly on a wide-area level and high granularity are based on cost-intensive sensors (e.g. satellite imagery, air mass related sensor measurements). Particularly in these fields, public funding will be an essential enabler for open access data in the future. As of today, ESA and NASA are subscribing to open science and are providing free access to a vast amount of MET and atmospheric information.

C. Opportunities

Today's environment is characterised by an increasing number of isolated developments that can serve as the foundation for the envisioned open science/data approach put forward in this paper. The last decade or so has seen a general increase in data literacy by individuals, modern data processing and analysis techniques, and a general increase in available data sets. At the political level, the regulatory players (e.g., the European Commission, national governments) have embraced the benefits of open data, and across Europe it is a common feature to couple public funding with open access to both results and underlying data for research, policy analysis, etc.

Within the aviation domain, a continuously growing number of initiatives to crowd-collect/acquire data and share it exist - at least within the respective community. This phenomenon is supported by web-based platforms for data storage and processing. These community-based efforts range from aircraft inventories, aeronautical information, to air transport movement data [10], [14], [22]. These interest groups may have different backgrounds (e.g. technology discussion groups, aviation enthusiasts, aircraft tracking, private flight planning, researchers). They share the same goal: to establish, complement, clean, and provide insights based on shared open data sets. The political ambitions and goals surrounding the future of air transportation and the discussion of aviation's impact on climate change, call for higher levels of transparency, increased levels of operational efficiency as near-term measures, and a faster adoption of novel operational concepts. Open data will also make it easier to apply the Performance Asssessment Framework within SESAR, ensure transparency and reproducibility, and enable a higher research ranking on the global stage.

More broadly, open data goes hand-in-hand with open code and open publications, which we next address.

1) Open code: 'Open code' refers to sharing pieces of software and code (including functions) within a community, which may help with common data manipulation or calculational tasks, sometimes as libraries. Open code can be "openly accessed, inspected, modified, and enhanced by anyone" [23]. A good example can be found in the OpenSky Network community, which shares extensive tooling for easy data access, processing and visualisation. There are many such communities around Github, even one dedicated to ATM[21].

2) Open aviation journal: The Journal of Open Aviation Science (JOAS)[22] is a novel initiative started in 2022 and supported by TU Delft. JOAS is dedicated to advancing high-quality open science, open data, and reproducible scientific research within the aviation domain. Recognising a deficiency of such practices in current aviation research, the principal objectives are to foster reproducible research in aviation through

[19]https://zenodo.org/
[20]https://atmdata.github.io/sources/
[21]https://atmdata.github.io/
[22]https://journals.open.tudelft.nl/joas

the utilisation of open-source tools and open data, and to promote the replication of existing research within the aviation research domain. Submissions must include data, accompanied by publicly available code, and explanations, ensuring that the code is tested for reusability by other researchers.

IV. Roadmap towards an 'open science alliance' for ATM

The preceding text describes open science, focusing on open data, the benefits it offers and where we are regarding open data in ATM research. It also describes some of the key challenges and opportunities. We now turn to proposing a roadmap towards an 'open science alliance' by describing the first, among (we hope) many intitiatives to come, followed by our vision of the future for such an alliance.

A. Open performance data initiative

Ambitious political goals and strategic objectives ultimately require a 'democratisation' of access and use of ATM data. This would allow political decision-makers, strategic planners, practitioners and researchers to tap into a harmonised open data environment. As an initial step towards such an environment and level playing field, the Performance Review Commission (PRC) of EUROCONTROL launched the open performance data initiative (OPDI) [23], aiming at closing the gap towards data accessibility and openness in European aviation research - supporting research validation and reproducibility.

The OPDI data sets are based on OpenSky Network crowdsourced (ADS-B) data for the period 2019 to present. It comprises: (1) a flight list (typically, flight ID, aerodrome of departure/destination (ADEP/ADES), (UTC) day of flight, aircraft type); (2) a list of milestones for each flight (see examples); (3) A 10-second state vector representation of the relevant flight trajectory as recorded by the Network - a sequence of 3D positions and relevant timestamps). Milestones are a way to summarise/synthetise a (typically 1-second samples) state vectors for a flight to a set of 'significant' events encoded as a 4D position (flight ID, 3D + timestamp), milestone type and contextual info. Selected examples of milestones are: off-block (with parking position ID); top of climb; start/end of holding; start/end of level segment; landing.

The milestones data set, for example, can be used to extract departure, arrival, internal, overflight statistics by using FIR crossing milestones, or in the estimation of CO_2 and NO_x emissions by modelling a flight via the milestones collected for the ground, departure, en-route and arrival data.

The availability of an open milestones data set linked to a flight list and state vector flight profile is useful for research as a common reference upon which to build analysis and models. Milestone types are not fixed and new ones can be easily accommodated, allowing the data set to grow with the needs of the research community. The work of cleanup, milestones extraction, and flight list identification is already available and can lift the burden of initial data acquisition and raw data preparation, which can require long lead times. Such data sets, freely available for research purposes make research reproducible, increasing research quality. The data set could be used as a *de facto* standard for actually-flown trajectories.

B. Next steps - roadmap towards an open science alliance

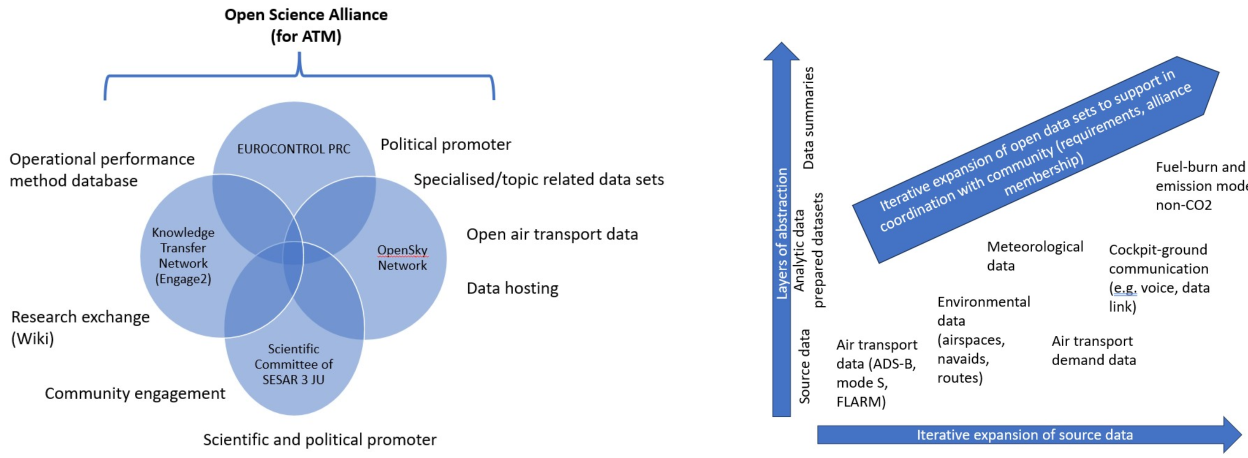
Mindful of the challenges and opportunities set out above, we propose the establishment of an open science alliance, which fully encompasses the broader 'open' principles described above, not least open data. Such an alliance could build on the existing collaborations of Figure 4a, leveraging the existing progress and dialogue, and, moreover, commitment to the principles laid out herein. EUROCONTROL PRC may promote further the values of opening data and offer their support in building the operational performance methods, and the definition of specialised datasets. The Scientific Committee (SC) of the SESAR 3 JU[23] also identified open data as an issue requiring support and could help in identifying opportunities and promoting the need for open data. The OPDI is a substantial first step, showcasing the alliance between EUROCONTROL PRC and the OpenSky Network. The latter also has broadly and historically supported the collection and sharing of open air transport data. Engage 2, as a SESAR coordination and support action, is committed to open science principles, further development of the EngageWiki, and, in particular, community building and outreach. Key data provisions for the community include 'source' data sets that support disaggregated / fundamental work (e.g. analytics, AI/ML, algorithmic work, data signal processing, etc.) and aggregated / prepared data, including existing summaries (thus also supporting 'first mile' accessibility for those wishing to avoid data extraction/preparation).

Figure 4b depicts a possible expansion path. This core alliance could first be elaborated through an initial terms of reference (ToR), and then extended to invite further specific membership including executive bodies (such as the European Research Executive Agency) with the power to support the funding of specific recommendations (where required), and the wider scientific community and associations, further elaborating the requirements and ToR, in an iterative process.

V. Conclusions and recommendations

We strongly believe in, and subscribe to, the practices of open science. Openness and reproducibility are two sides of the same coin and will ultimately enable trust, flexibility, and ensure the positive developments identified through this paper. Individual, organisational, and institutional sensitivies can range from privacy concerns to issues of liability. However, the benefit of adapting an open science approach will ultimately pay dividends. The challenges faced by air transport and ATM require higher levels of transparency. A fundamental enabler will be the independent verification and validation of reported impacts/results and achieved performance levels. We consider that this can only be achieved with: (1) open access to the scientific methods and data utilised; (2) open access to

[23]https://www.sesarju.eu/about-sesar/scientific-committee



(a) First step

(b) Possible expansion path

Figure 4. Open science alliance for ATM

the analytical code and implementation of methods; (3) open review of reported analyses/research. This paper has identified an initial roadmap to establish an open science alliance for ATM research in Europe. Such an open partnership would represent a stepping stone towards leveraging the benefits of open data, *inter alia*, to increase transparency and accountability, accelerate the validation and outreach of research, and promote faster innovation cycles. The cultural change required is to move away from considering data as the 'new gold', and make it an asset. Open data is a particular ingredient, which will allow us to leverage the benefits highlighted throughout the paper. Europe is in an excellent position to come together and realise such an alliance, which can be achieved by bringing together activities within different domains through a concerted action on open science in the widest sense.

For specific consideration within the framework of this broad alliance, we further recommend, in no implied order of priority, to:

- together with regulators and policy makers, identify future data requirements as they pertain to the Green Deal and Digital Europe programmes, in the ATM domain;
- support and finance a common licencing programme for commercial data required by multiple research projects to obviate the loss of finance and human effort of multiple spends on the same, or very similar, data (such as for airline schedules and passenger itineraries);
- review the requirements for domain-specific (e.g., environment, multimodality, MET, cybersecurity) data with the current SESAR projects, and collate these to be considered for common licences and/or for improved open access (e.g., by relaxing project-specific approvals for broader activities within SESAR ER and IR, such as for BADA 4 data) - this could be initially reviewed at the SESAR flagship level;
- consider the open science survey results, currently being developed by Engage 2 to specifically investigate the data needs and access challenges of the wider ATM research community.

Acknowledgment

The authors would like to thank all colleagues who contributed to shaping these thoughts over recent collaborations. Parts of the article originate from material from the Engage (1) KTN, Grant agreement No. 783287, funded by the SESAR Joint Undertaking.

References

[1] European Research Executive Agency. Open science. [Online]. Available: https://rea.ec.europa.eu/open-science_en
[2] National Academies of Sciences, Engineering, and Medicine, *Open Science by Design*. National Academies Press, 2018.
[3] SESAR 3 JU. Digital sesar solutions catalogue. [Online]. Available: https://www.sesarju.eu/catalogue
[4] Council of the European Union, "Council regulation (ec) no 219/2007 of 27 february 2007 on the establishment of a joint undertaking to develop the new generation european air traffic management system (sesar)," 2007. [Online]. Available: http://data.europa.eu/eli/reg/2007/219/oj
[5] T. Bolić and P. Ravenhill, "SESAR: The past, present, and future of european air traffic management research," *Engineering*, vol. 7, no. 4, pp. 448–451, 2021.
[6] Council of the European Union, "Council regulation (eu) 2021/2085 of 19 november 2021 establishing the joint undertakings under horizon europe," 2021. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2021/2085/oj
[7] SESAR Joint Undertaking, *European ATM master plan:digitalising Europe's aviation infrastructure: executive view: 2020 edition.* Publications Office, 2020.
[8] Performance Review Commission. (2023) Prc transformation support strategy. [Online]. Available: https://ansperformance.eu/transformation/
[9] A. Cook, D. Schaefer, M. Bourgois, P. Hernandez, G. Tanner, and T. Bolić, "D3.10 research and innovation insights," Engage KTN, Tech. Rep., 2022.
[10] Engage KTN. Engagewiki. [Online]. Available: https://wikiengagektn.com/EngageWiki
[11] B. I. Woodrow, "FAA Chooses Enhanced ADS-C Over Space-Based ADS-B for Oceanic Airspace," 2019. [Online]. Available: https://interactive.aviationtoday.com/avionicsmagazine/august-2019/faa-chooses-enhanced-ads-c-over-space-based-ads-b-for-oceanic-airspace/
[12] X. Olive, "Traffic, a toolbox for processing and analysing air traffic data," *Journal of Open Source Software*, vol. 4, no. 39, pp. 1518–1, 2019.

[13] J. Sun, H. Vû, J. Ellerbroek, and J. M. Hoekstra, “pymodes: Decoding mode-s surveillance data for open air transportation research,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 7, pp. 2777–2786, 2019.

[14] M. Schäfer, M. Strohmeier, V. Lenders, I. Martinovic, and M. Wilhelm, “Bringing up OpenSky: A large-scale ADS-B sensor network for research,” in *Proceedings of the 13th International Symposium on Information Processing in Sensor Networks*, 2014, pp. 83–94.

[15] M. Strohmeier, “Research usage and social impact of crowdsourced air traffic data,” in *Proceedings*, vol. 59, no. 1. MDPI, 2020, p. 1.

[16] Performance Review Commission, “Prr 2022 performance review report, an assessment of air traffic management in europe,” EUROCONTROL, Tech. Rep., 2023.

[17] Bureau of Transportation Statistics. Airline on-time statistics. [Online]. Available: https://www.transtats.bts.gov/ONTIME/

[18] ICAO, “Annex 3 - meteorological service for international air navigation,” ICAO, Tech. Rep., 2018.

[19] European Commission, Directorate-General for Mobility and Transport, “Commission implementing regulation (eu) 2017/373 of 1 march 2017 laying down common requirements for providers of air traffic management/air navigation services and other air traffic management network functions and their oversight (text with eea relevance.),” 2017. [Online]. Available: http://data.europa.eu/eli/reg_impl/2017/373/oj

[20] Insider. The world’s 2nd-richest man, louis vuitton’s ceo, sold his private jet after people started tracking it on twitter: ’no one can see where i go’. [Online]. Available: https://www.businessinsider.com/louis-vuitton-bernard-arnault-sold-private-jet-people-tracking-twitter-2022-10

[21] James Haydon. Uk air traffic control meltdown. [Online]. Available: https://jameshaydon.github.io/nats-fail/

[22] OpenAIP. (2023) Open aip, free worldwide aviation database. [Online]. Available: https://www.openaip.net/

[23] Performance Review Commission, EUROCONTROL. Open performance data - an enabler to further performance enhancements? [Online]. Available: https://www.eurocontrol.int/publication/performance-insight-5-open-performance-data