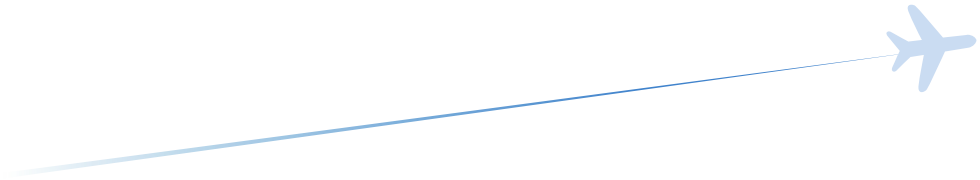

D2.1 Data Management Plan

Deliverable ID:	**D2.1**
Dissemination Level:	**PU**
Project Acronym:	**BEACON**
Grant:	**893100**
Call:	**H2020-SESAR-2019-2**
Topic:	**SESAR-ER4-08-2019**
Consortium Coordinator:	**University of Westminster**
Edition Date:	**03 February 2021**
Edition:	**01.00.01**
Template Edition:	**02.00.02**

Founding Members
EUROPEAN UNION EUROCONTROL
SESAR JOINT UNDERTAKING

Authoring & Approval

Authors of the document

Name/Beneficiary	Position/Title	Date
Paola Bassi / UNITS	Project member	23 December 2020
Lorenzo Castelli / UNITS	Project member	23 December 2020

Reviewers internal to the project

Name/Beneficiary	Position/Title	Date
Gerald Gurtner / University of Westminster	Project coordinator	23 December 2020
Tatjana Bolic / University of Westminster	Project member	23 December 2020

Approved for submission to the SJU By - Representatives of beneficiaries involved in the project

Name/Beneficiary	Position/Title	Date
Gerald Gurtner / University of Westminster	Project coordinator	23 December 2020

Rejected By - Representatives of beneficiaries involved in the project

Name/Beneficiary	Position/Title	Date
N/A		

Document History

Edition	Date	Status	Author	Justification
01.00.00	23 December 2020	Release	BEACON Consortium	New document for review by the SJU
01.00.01	03 February 2021	Second Release	BEACON Consortium	Modifications required by the SJU implemented

Copyright Statement

BEACON

BEHAVIOURAL ECONOMICS FOR ATM CONCEPTS

This Data Management Plan is part of a project that has received funding from the SESAR Joint Undertaking under grant agreement No [893100] under European Union's Horizon 2020 research and innovation programme.

Abstract

This deliverable presents the approach of BEACON for the data management, and details the data sources considered.

Table of Contents

List of tables

Executive summary

BEACON's general goal is to explore the role of airspace users' complex behaviour, such as bounded rationality in the design of new management procedures. Having sufficient data to be able to model the different interactions is paramount. BEACON defines models to capture the AUs' decision process exploiting the Behavioural economics theory, with the aim of reducing the cost impact caused by Air Traffic Flow Management (ATFM) delays. In particular, BEACON tries to expand the ATFM slots exchange concept and to improve the current under studying procedures, such as the User-Driven Prioritization Process (UDPP), via the design of new market mechanisms.

BEACON wants firstly to create a strategic model with long-term planning capabilities for the agents and then a tactical model to capture network effects and compute various key performance indicators and metrics to assess the impact these models have on the AUs' and the entire ATM network. The data regarding the system infrastructure, traffic (flights), aircraft performance data and passenger itineraries are required to be stored and prepared for the use in these models. Finally, BEACON will use statistical analysis on the results of the model, thus the management of the outcome of the modelling is equally important as the management of the input data.

The data BEACON will acquire, elaborate, and manage, are based on the data requirements, coming from the modelling needs. The data that BEACON will manage can be categorised as follows:

1. Traffic and delay data (e.g., trajectories, causes and amounts of delay, flight plans, flight attributes, Air Traffic Flow and Capacity Management (ATFCM) measures);
2. Aircraft performance data;
3. Airspace environment data (e.g., airspace structures, procedures, routes and flying restrictions);
4. Airline specific data (e.g., detailed schedules, crew rotations, flight priorities);
5. Passenger flows data;
6. Other data (e.g., ADS-B messages of flights in Europe).

Most of the datasets have already been acquired, and just need to be loaded into a bespoke database. Another database to support output data analysis will be established.

Both databases will be hosted by a cloud service, Oracle cloud or Amazon AWS, that will guarantee secure access to the BEACON partners.

Next steps involve the population of the input database with the data relevant to the BEACON project, mainly data from September 2018. As the project progresses, new tables will be added to store intermediate data input and the modelling results, applying a versioning scheme, in order to keep track of the project progress.

1 Introduction

BEACON's general goal is to explore the role of airspace users' complex behaviour, such as bounded rationality in the design of new management procedures. Having sufficient data to be able to model the different interactions is paramount. BEACON defines models to capture the AUs' decision process exploiting the Behavioural economics theory, with the aim of reducing the cost impact caused by ATFM delays. In particular, BEACON tries to expand the ATFM slots exchange concept and to improve the current under studying procedures, such as the User-Driven Prioritization Process (UDPP), via the design of new market mechanisms.

BEACON wants firstly to create a strategic model with long-term planning capabilities for the agents and then a tactical model to capture network effects and compute various key performance indicators and metrics to assess the impact these models have on the AUs' and the entire ATM network. The data regarding the system infrastructure, traffic (flights) and passenger itineraries are required to be stored and prepared for the use in these models. Finally, BEACON will use complexity science tools to analyse the results of the model, thus the management of the outcome of the modelling is equally important as the management of the input data.

Moreover, the BEACON consortium is formed by six institutions (partners) located in five different countries. Different partners are responsible for the development of different parts of the model and the analysis of the outcome data. This presents a particular set of requirements in terms of data accessibility, security and reliability that will be detailed in the following sections.

The acquisition and management of data used in BEACON is performed within the WP2 of the project.

2 Data requirements

The consortium agreed on the list of data requirements needed for the modelling and the results analysis. Some of the data have already been collected, even if further cleaning is required, as explained in Section 3, while other data are still in the acquisition phase.

BEACON will investigate several mechanisms that will be modelled to create the different case studies. The detailed description of the mechanisms developed and studied during the project will be given in D3.1. Still, regarding the data required for their simulation and analysis some preliminary and basic considerations, common for all of these frameworks, can be made.

All the mechanisms that BEACON is going to test aim to reduce the cost impact caused by airport and airspace regulations. Moreover, they all involve different actors, such as airlines, airports, air navigation service providers (ANSPs) and the Network Manager, which can be interpreted as agents who make decisions based on the current situation.

2.1 Airlines' data

Airlines aim to reduce the delay impact on their fleet. In each of the mechanism under exam they are asked to perform an evaluation of their current situation and determine a set of actions (prioritisation, bidding, trading, etc.). Their decision will be made by considering:

- the specific characteristics of each of their aircraft involved
- the passenger load and connections
- crew rotations
- current flights' schedule delays and delays cost forecast
- aircraft turnarounds.

In order to reproduce this decision process and study how it is applied to the different mechanisms, a detailed flight delay cost model is required. This cost model will also be needed to evaluate the overall performance of the different mechanisms in terms of cost impact. The consortium has access to CODA data (EUROCONTROL, 2018) through EUROCONTROL's OneSky portal, and these data contain the basic information on the delay causes as reported by airlines.

2.2 Network data

The Network Manager has to guarantee that any rescheduling operation, resulting from the application of a certain mechanism in a particular hotspot, does not lead to undesired ripple effects on the network. This represents the second goal of the project. In order to develop a realistic simulation that considers the possible network knock-on effects caused by some local action and ensure reliable and robust results, it is required to have at disposal the following data referring to the chosen day of the operations sourced from the R&D data archive and Demand Data Repository (DDR2) data:

- air traffic demand including flight trajectories (R&D archive);
- airports and flight sectors capacity (DDR2 environment data);
- airports and flight sectors load (intersection of the previous two).

If tackling this problem on a global scale might result in an unfeasible task, in BEACON a first exploratory attempt will be performed considering just a small set of hubs and their connections.

2.3 Modelling the behavioural effects

Another goal of BEACON is to include behavioural economics into the analysis of the different mechanisms. An initial high-level formulation of the potential biases that might affect the agents' decision process will be developed in WP4. A sufficient amount of data to analyse these effects is not available at the moment, however, a first calibration of these behavioural models will be performed in WP4 collaborating with different stakeholders, such as duty managers, flight dispatchers etc. The calibration phase will be conducted through a series of questionnaires. The chosen stakeholders will be asked to fill the questionnaires designed in WP4 in order to discover and highlight potential non-rational behavioural patterns for the evaluation of the proposed market mechanisms developed in WP3. To further characterise and measure the effects of these non-rational biases, in WP5, the same actors will be involved in a series of "human in the loop" live simulations, in which they will be asked to make decisions on a series of realistic scenarios. The data collected during this phase will be used to test the performance of the different mechanisms even in the case of non-rational agents.

2.4 Data confidentiality concerns

Personal data acquired during the project will be protected following the procedures described in D8.1 and D8.2. These deliverables include procedures on the storage and use of personal data, including anonymisation and pseudo-anonymisation techniques.

All proprietary data acquired during the project will be protected, first by using proper data protection measures (encryption etc) as described in this deliverable and second by ensuring that the results published or communicated are sufficiently aggregated.

2.5 Other

Putting a price on various delays incurred by the flights is also needed in order to properly derive the behaviour of the agents via their cost function. This will be done using the results of task T3.4, which will update for 2018-2019 the reference study on costs (Cook & Tanner, 2015) (the reference's baseline was 2014). This document and its predecessor, created by UoW, has become a standard in the industry, for instance for the Performance Review Body (PRB) to estimate the total cost of delays in Europe.

Various other data might be required to properly calibrate the model. Among them, the Central Route Charges Office (CRCO) charges for the day of operation for every ANSP will likely be required to properly assess the operational costs. Details on different alliances and partnerships between airlines will also probably be used in the model, in particular for delay management strategies and for passenger connections. WP2 runs for most of the project to monitor the needs of the project, acquire more data if needed and manage the model output data.

3 Data sources, acquisition and elaboration

Based on the data requirements detailed in the previous section, here we describe the data sources, acquisition strategy and elaboration of data in order to prepare the input data for the models. The following data categories have been identified:

1. Traffic and delay;
2. Aircraft performance data;
3. Airspace environment;
4. Passengers;
5. Airline specific data;
6. Other.

The sources, acquisition status and elaboration needs are further detailed for each of the categories in the following text. In order to elaborate the data, we will first load the data in their raw form into the input database. Its structure is described in Section 4.

3.1 Traffic and delay

Different sources will be used and consulted in order to prepare the traffic data for BEACON's models:

- EUROCONTROL'S R&D data archive: R&D archive offers filed and actual flight trajectories, more specifically flight points, sequence of national regions and sequence of airspace blocks. The data from September 2018 are available and will be used for everything concerning the trajectories, or more generally, the traffic.
- EUROCONTROL's DDR2: DDR2 also contains trajectory data. Older data from a number of other AIRACs (Aeronautical Information Regulation and Control) are available to consortium members, so these will be used to perform statistical analyses when required (these analyses cannot be done with R&D because data of only 4 months per year are available).
- Daily ATFCM summary data: contain detailed information on the regulations that were applied on the day of study. These are obtained from Network Manager (NM) ATFCM statistics.
- CODA summary delay data: will be used to analyse delay and enable realistic delay generation (for all the delays not caused by the ATFM actions).

- CODA taxi times: standard taxi times, published by CODA, are useful to model the time between gate and runway and from the runway to the stand as these values are required to accurately estimate the arrival and departure delay, and passenger connections.
- Commercial data sources: BEACON will purchase commercial data to obtain information such as schedules.

3.2 Aircraft performance data

- BADA performance models: BEACON will use BADA 4.2 performances from EUROCONTROL especially with regard to aircraft performance. Some consortium members have access to BADA 4.2.

3.3 Airspace environment

- R&D archive: it contains data for both filed and actual trajectories, and information about Flight Information Regions (FIR) and Air Traffic Control unit Airspace (AUA) crossed. More specifically: minimum and maximum vertical boundary of the airspace volume, FIR's shapes expressed as a set of points (latitude, longitude).
- EUROCONTROL's DDR2: the DDR2 repository also contains information regarding the airspace environment in terms of sector shapes, sector activations, sector and airport capacities, and basic ATFM regulations.

3.4 Passengers

- Passenger itineraries: passenger itineraries will be based on previous datasets developed by the University of Westminster for 2010 and 2014, which will be used as long as the 2018 data are not acquired. Data for 2018 are being acquired from Global Distribution System (GDS) providers and will be available to BEACON during the project.
- Airline load factors: passenger load factors reported by airlines.
- Airport connectivity information: airport reports on passenger connectivity from ACI EUROPE and individual airports.

3.5 Airline specific data

- Thanks to the presence of SWISS International Air Lines among the project partners, BEACON has some specific data of the airline at its disposal. The data that may be provided to the consortium are: detailed schedules, crew rotations and flight priorities.

3.6 Other

- Cost of delay: cost of delay models developed in-house by the University of Westminster (Cook & Tanner, 2015) will be revised for 2018.
- CRCO unit rates: unit rates in effect on the selected test day for every Member State will be required to properly assess the operational costs of modelled flights.

3.7 Summary of data sources

Table 1. Summary of data sources

Category	Datasets	Acquisition status	Elaboration status
Traffic and delay	• R&D archive • DDR2 • ATFCM summary data • CODA summary delay data • CODA taxi times • BADA performance models • Schedule data	• Traffic data for September 2018[1] • DDR2 traffic data of several AIRACs available to consortium • ATFCM summary data available for 2018 • CODA summary delay data yet to be acquired for September 2018 • CODA taxi times available for the summer season 2018 • BADA – BEACON partners in possession of BADA licenses • Schedule commercial data will be acquired from OAG or Innovata (Cirium).	Need to load the data into the input database, with the exception of schedule commercial data which has yet to be acquired
Passengers	• Previous itineraries 2010 and 2014 • 2018 itineraries • Airline load factors • Airport connectivity information	• Available in-house (University of Westminster) • being acquired from suppliers of GDS data such as IATA PaxIS or OAG • to be acquired from airline sources or other projects (e.g. Domino) • Connectivity information yet to be acquired for 2018	Need to load the data into the input database, with the exception of 2018 itineraries which has yet to be acquired
Airspace environment	• DDR2 files containing airport and airspace capacity	• DDR2 airspace environment data of selected AIRACs available to consortium members	Need to load the data into the input database

[1] We use 2018 data because some 2019 data are not available, in particular the R&D archive.

	• R&D files containing information about Flight Information Regions (FIR) crossed	• R&D FIR data available for September 2018	
Other	• Cost of delay	• Available in-house (University of Westminster)	Cost of delay needs an update.
	• CRCO unit rates	• Unit rates available for 2016, 2017, 2018	Unit rate data needs to be uploaded into input database
	• Airline alliances	• Available in-house (University of Westminster), though yet to be updated with 2018 alliance changes	

3.8 Selection of the test day

A day belonging to September 2018 was selected because, as can be seen from table 2, it is one of the months with the highest traffic of 2018 but quite low ATFM delay. For example, June has almost the same traffic, but an average ATFM delay of 1.44 min/flight higher than September.

Table 2. Traffic and ATFM delays in 2018

Month	All traffic	En-Route ATFM Delay per Flight	Airport ATFM Delay per Flight	Total Delay per Flight
2018-01	766.330	0.17 min	0.47 min	0.63
2018-02	718.297	0.32 min	0.47 min	0.79
2018-03	835.547	0.64 min	0.47 min	1.11
2018-04	896.037	1.08 min	0.44 min	1.52
2018-05	975.303	2.61 min	0.59 min	3.20
2018-06	1.024.707	3.25 min	0.70 min	3.95
2018-07	1.083.577	3.87 min	0.72 min	4.59
2018-08	1.076.263	2.87 min	0.67 min	3.54
2018-09	1.030.926	1.81 min	0.70 min	2.51
2018-10	985.102	1.22 min	0.59 min	1.81
2018-11	819.399	0.49 min	0.57 min	1.06
2018-12	799.100	0.71 min	0.65 min	1.36

We have targeted a busy, but not unduly disrupted day[2] in September 2018. Note that Fridays have been selected as they are usually the busiest day of the week.

[2] We restrict ourselves to one day of data, due to the high effort required to align different sources of data for a single day.

In summary the choice is Friday 14 September 2018:

- Ranked #2 by traffic count in September 2018;
- Ranked #21 by delay in September 2018;
- Total ATFM delay quite low

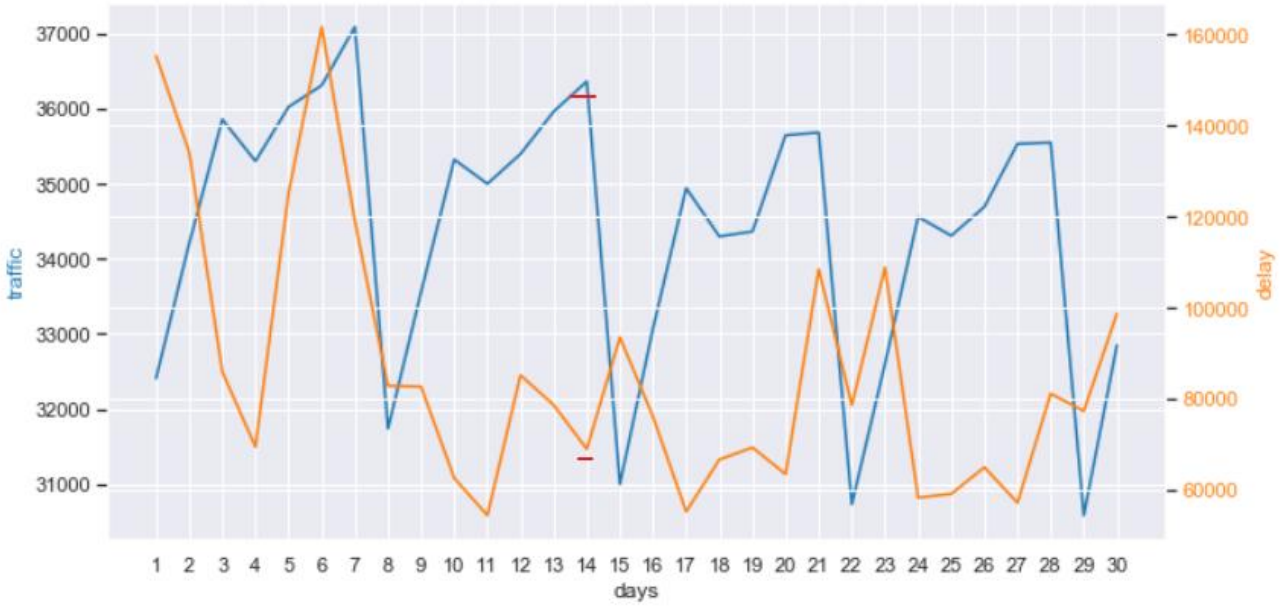

Figure 1. Traffic count and ATFM delay (minutes) for September 2018 source: EUROCONTROL's ATFM daily summary)

Table 3. Daily summary of 14 September 2018

All Traffic[1]	Regulated Traffic[2]	Delayed Traffic[3]	% of Delayed Traffic[4]	ATFM Delay[5]	Airport ATFM Delay[6]	Avg ATFM Delay per Delayed Flight[7]	Avg ATFM Delay per Regulated Flight[8]	Avg ATFM Delay per Flight[9]
36.362	8.967	4.751	13,1%	68.971	24.872	14,5	7,7	1,9

[1]: the number of flights entering daily in the NM area

[2]: part of All Traffic passing through one or more regulations

[3]: Part of Regulated Traffic delayed by a regulation, i.e. for which CTOT - ETOT > 0.

[4]: ratio between columns 3 and 1

[5]: the sum of the delays calculated from CASA regulations, assigned to the traffic demand (expressed in minutes)

[6]: part of ATFM Delays induced by Airport protecting regulations (expressed in minutes)

[7]: the ATFM Delay[5] divided by the Number of Delayed Flights[3] (expressed in minutes)

[8]: the ATFM Delay[5] divided by the Number of Regulated Flights[2] (expressed in minutes)

[9]: the ATFM Delay[5] divided by the All Traffic[1] (expressed in minutes)

4 Database infrastructure

All data used in the BEACON project will be stored into two secure databases hosted by a cloud service, one dedicated to the input data and another one to the output data.

4.1 Database access

Due to the various Non-Disclosure Agreements (NDAs) signed by UoW and the other partners for the data, as well as the need to protect the results of the model, access to the database needs to be properly secured. The access to databases will be password-protected.

Once logged-in, partners have permission to use the database resources for testing and production.

Access to data by different partners is limited considering different data requirements by the different institutions and subject to having the adequate licencing agreements. The control of data access ensures that possible data corruption is minimised. For instance, UNITS and UoW have full writing and reading access to the data, since they are managing the content of the databases in BEACON while other partners involved in the modelling have read-only access, or can create new tables but not erase any.

4.2 Database structure

In BEACON, due to different data structure of the input and the output, different technologies and databases will be used to save the data as described in sections 4.2.1 and 4.2.2.

BEACON uses the databases for two purposes:

- To have standard input data with easy access.
- To store the results of the model(s) in an efficient way.

The structure of the databases should be compatible with the following requirements of the models:

- Reproducibility: getting the same output from the same input with the same code.
- Reliability: making sure that the input data has not changed between two runs of the model.
- Consistency: making sure that the input in particular is self-consistent.
- Traceability: making sure that the output data can be linked unambiguously to a given input dataset.

4.2.1 Input data

All input data are structured to be inserted in a relational database, so we decided to store all the input data in a MySQL database that will be called *beaconInputEnvironment*. MySQL is an open source standard for relational databases. It is well documented, reliable, and well suited for mid-range databases.

Building-up on data management experience from past projects, BEACON will thus use different types of tables/schemas:

- Some schemas for the primary data, which should never be modified. This includes the R&D data for instance and other sourced data (see previous section).
- Some tables/schemas for the secondary data, which are built 'off-line' by some pre-processing codes of the models. These data change with the maturity of the models, and should be versioned.

By versioning the secondary data, the project ensures the traceability of the results.

4.2.2 Output Data

Unlike the input data, the format and structure of the output data are not yet established. Based on experience from past projects, the output will consist of unstructured data, so it will not be saved in a relational database. The output data will change during the project and the changes will be tracked with a versioning method.

To save the unstructured output data, two possible systems are considered:

- Non-SQL database (such as Apache Cassandra)
- Data lake (such as Amazon S3 or Apache Hadoop)

The choice of one of these services will be strictly related to the type of output obtained.

Apache Cassandra is an example of NoSQL database management system (DBMS), free and open source. This type of DBMS offers flexibility (the flexible data model makes NoSQL databases the ideal solution for semi-structured and unstructured data), scalability and high performance (higher than the results obtained by trying to achieve similar functionality with relational databases).

Amazon S3 is an example of data lake i.e. A Data Lake is a type of shared data repository that can store large and varied raw data sets in their native format. This service offers the opportunity to archive data with very different formats without the need to standardize and "normalize" them, useful when the data format is not a priori known.

The services chosen and a more detailed description of the database structures will be reported in D2.2.

5 Next steps and look ahead

In terms of data, most of the datasets are already available to the consortium, excluding the data concerning passengers and schedules, which still have to be acquired (as mentioned in sections 3.1 and 3.4).

Regarding the preparation of data, the team still needs to load the available AIRACs into the input database. This will be straightforward for DDR2 data, and only minor changes to the code used to upload the same type of data in the past are needed, usually caused by the data version changes within DDR2. A similar argument applies to R&D Archive data: even if the data structure is different than in DDR2, the content is similar. Passenger itineraries also need to be updated and loaded for September 2018. When the new data are loaded into the input database, as these will be in the same format, the update process will be relatively seamless. In this way, data availability will not slow down the development of the rest of the project.

A MySQL instance for the input data will be created on the cloud service and access to the database will be given to all the partners. The team will also need to manage the new personnel joining the project and any potentially leaving it.

Finally, the team will actively monitor the needs for data, should new ones arise. WP2 runs almost until the end of the project to make sure that there is no bottleneck due to data availability and that new data can be obtained and prepared, if required.

6 References

Cook, A., & Tanner, G. (2015). *European airline delay cost reference values - updated and extended values (Version 4.1).* Retrieved from https://www.eurocontrol.int/sites/default/files/publication/files/european-airline-delay-cost-reference-values-final-report-4-1.pdf

EUROCONTROL. (2018, June 8). *Central Office for Delay Analysis (CODA)*. Retrieved from https://www.eurocontrol.int/articles/central-office-delay-analysis-coda

7 Acronyms

ACI EUROPE: Airport Council International Europe

AIRAC: Aeronautical Information Regulation and Control

ANSP: Air Navigation Service Provider

ATC: Air Traffic Control

ATFCM: Air Traffic Flow and Capacity Management

ATFM: Air Traffic Flow Management

ATM: Air traffic management

AU: Airspace user

AUA: ATC Unit Airspace

BADA: Base of Aircraft Data

CASA: computer-assisted slot allocation

CODA: Central Office for Delay Analysis

CRCO: Central Route Charges Office

DBMS: database management system

DDR2: Demand Data Repository

GDS: Global Distribution System

NDA: Non-Disclosure Agreement

NM: Network Manager

PRB: Performance Review Body

UDPP: User Driven Prioritisation Process

UNITS: Short name of BEACON partner: Università degli Studi di Trieste

UoW: Short name of BEACON coordinator: University of Westminster

-END OF DOCUMENT-