Deliverable D1.3

Data Management plan

Authors :	Vincent Minier (CEA) David Barrado (INTA) Pierre-Olivier Lagage (CEA)	Date : 18/06/2018
Reviewed by	Executive Board	Date :
Release by	Pierre-Olivier Lagage (CEA)	Date : 29/06/2018

DISSEMINATION LEVEL		
PU	Public	X
CO	Only for the consortium members (including Commission Services)	

DOCUMENT STATUS

Issue	Date	Description
0.9	25/06/2018	First Draft
1	29/06/2018	Initial version

DISTRIBUTION LIST

Members of the Executive Board
Members of the Steering Committee
EC

TABLE OF CONTENTS

1 INTRODUCTION

In the framework of the ExoplANETS-A project, archival data from ESA Space Science archives (HST) combined with NASA Space Archives (Spitzer, Kepler) will be exploited with novel data calibration and spectral extraction tools, novel retrieval tools, to produce a homogeneous and reliable catalog of exoplanet atmosphere properties. In parallel, a coherent and uniform database of the relevant properties of host stars will be developed from ESA Space Science archives (XMM, Gaia, Herschel), combined with data from international space missions and ground-based telescopes. These exoplanet and host star catalogues will be accompanied/interpreted with models to assess the importance of star – planet interactions.

The project gathers the expertise of seven laboratories:

No	Name	Short name	Country	Project entry month[8]	Project exit month
1	COMMISSARIAT A L ENERGIE ATOMIQUE ET AUX ENERGIES ALTERNATIVES	CEA	France	1	39
2	INSTITUTO NACIONAL DE TECNICA AEROESPACIAL ESTEBAN TERRADAS	INTA	Spain	1	39
3	UNIVERSITY OF LEICESTER	ULEIC	United Kingdom	1	39
4	MAX-PLANCK-GESELLSCHAFT ZUR FORDERUNG DER WISSENSCHAFTEN EV	MPG	Germany	1	39
5	UNIVERSITY COLLEGE LONDON	UCL	United Kingdom	1	39
6	UNIVERSITAT WIEN	UNIVIE	Austria	1	39
7	STICHTING SRON NETHERLANDS INSTITUTE FOR SPACE RESEARCH	STICHTING SRON	Netherlands	1	39

This document presents an initial version of the data management plan (DMP) of the project, the deliverable number 3 of the management workpackage (WP1) due 6 months after the start of the project. It follows the template given in Reference Document 1 (see below). It is a living document which will be updated as the implementation of the project progresses.

2 APPLICABLE DOCUMENTS (AD)

AD-1	ExoplANETS-A Grant Agreement	N° 776403
AD-2	ExoplANETS-A Consortium Agreement	Version 3, 2017-12-22; DRF 0647_X30423

3 REFERENCE DOCUMENTS (RD)

RD-1	http://ec.europa.eu/research/participants/data/ref/h2020/gm/reporting/h2020-tpl-oa-data-mgt-plan_en.docx	Version 1.0 ; 13 October 2016

4 DATA SUMMARY

4.1 PURPOSE OF THE DATA COLLECTION/GENERATION AND ITS RELATION TO THE OBJECTIVES OF THE PROJECT?

The objectives of the project are :

- To establish new knowledge on the atmosphere of exoplanets by exploiting archived space data (HST, Spitzer, Kepler) using novel data reduction methods, as well as improved techniques to retrieve atmospheric parameters from data.
- To establish new insight on the influence of the star on the planet atmosphere by exploiting archived space data (GAIA, XMM, Chandra, Herschel, IUE, HST) on the host stars, as well as complementary ground-based data.
- To disseminate new knowledge.

Data are thus essential to the ExoplANETS-A project. Archival data are at the start of the project and new data sets with added scientific value are generated by the project and are made available to the community via a knowledge server. The global concept is shown below.

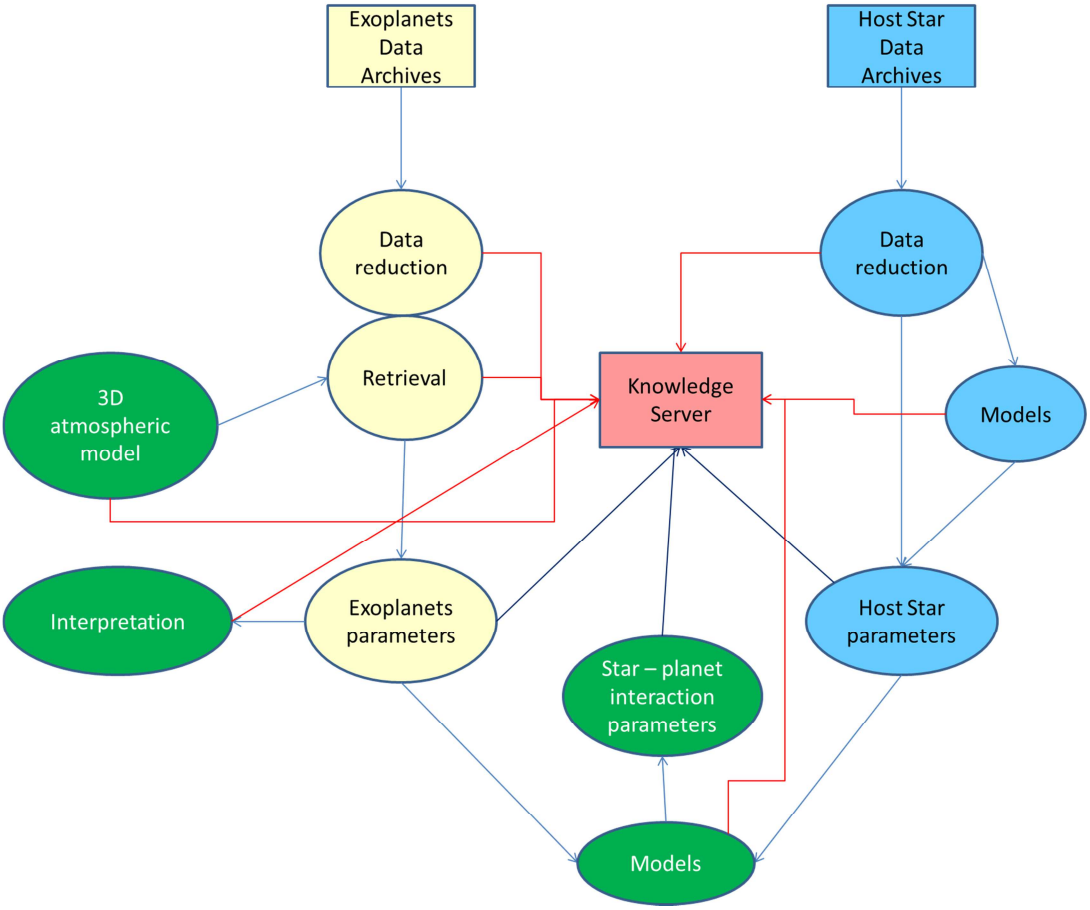

Figure 1 : starting from archive data on exoplanets and their host star, the project will develop novel data reduction and retrieval techniques to get an homogeneous catalogues of a hundred of exoplanets; the data will be available from a knowledge server.

4.2 WHAT TYPES AND FORMATS OF DATA WILL THE PROJECT GENERATE/COLLECT?

The science products of the project consist of spectra of exoplanet atmospheres (see Figure 2), exoplanets and host-star parameters (such as molecular content of exoplanet atmosphere see Figure 3), modelling and retrieval algorithms, tools to analyze the data, ab-initio models of sources…

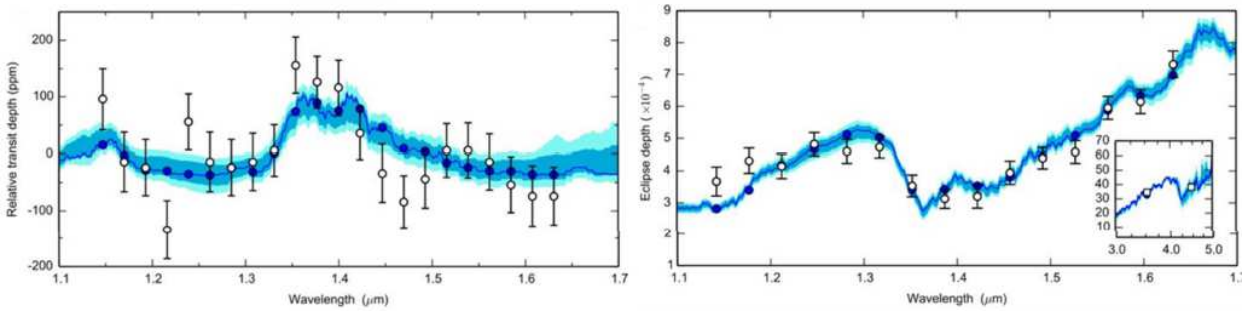

*Figure 2. Example of transmission spectrum (left) and emission spectrum (right) of an exoplanet (WASP43 b), as observed with the WFC3 instrument on board of HST (white circle); the dark blue line show the best fit models from retrieval analysis. The feature observed around 1.4 microns is a water feature. The insert on the right image shows the photometric points from observations with the Spitzer Space Telescope. (L. Kreidberg et al. ApJL **793**, Issue 2, L27-32; arXiv:1410.2255)*

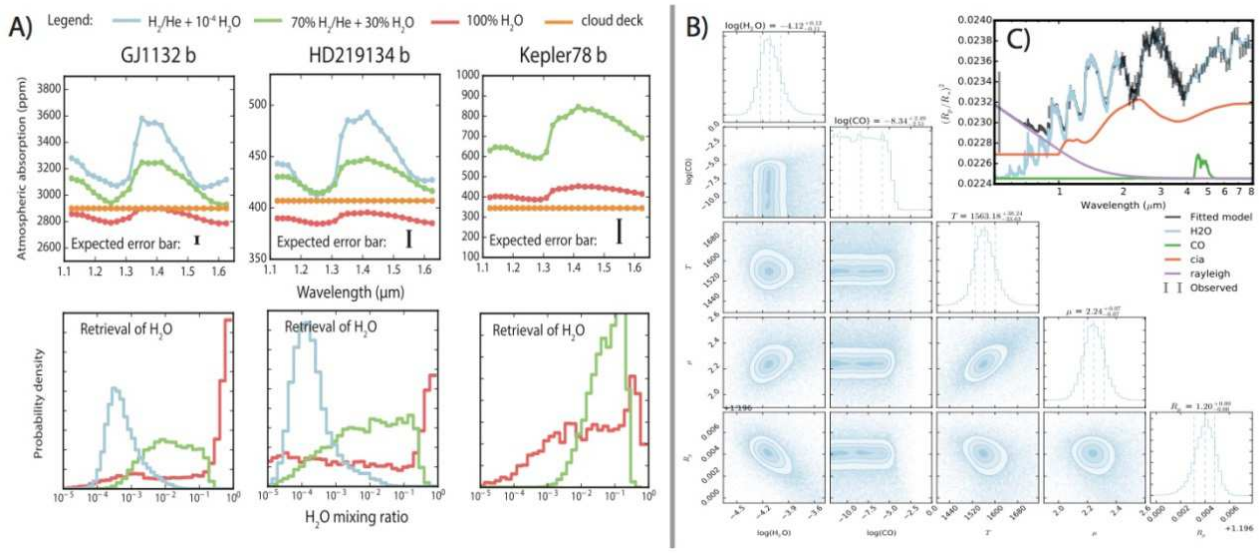

Figure 3: A) Examples of Tau-REx results for three simulated super-Earths HD 219134 b, GJ 1132 b and Kepler 78 b (planetary parameters from exoplanet.eu). Top: atmospheric spectra for varying compositions at Hubble/WFC3 wavelengths. Expected error bars for observations are also shown. Bottom: Tau-REx retrieved constraints of H_2O abundance for the spectra shown above. B) Posterior distributions of complex likelihood functions encountered in spectral retrievals. Parameter spaces are often highly dimensional (>20D) with non-linear inter-parameter correlations. We will fully map these correlations and, using manifold learning, identify model degeneracies. C) Simulated observational data analysed by the Tau-REx framework. Multiple atmospheric components are shown visually.

Our aim is to use, as much as possible, the Virtual Observatory standard or one of the standard formats in the astronomical community, i.e. the fits format.

4.3 WHAT IS THE ORIGIN OF THE DATA?

The data will be of various origins:

- We will use archival data from observations of exoplanet atmospheres with space observatories, as well as ground-based observatories.
- Thanks to the development of novel data reduction, we will produce from the archival data, new calibrated data sets.
- From this new set of data and thanks to the development of new retrieval techniques, we will derive parameters of exoplanets atmospheres, such as its molecular contain.
- Data will also be generated from the modelling the atmosphere of exoplanets.
- We will use archival data from observations of exoplanet host stars with space observatories, as well as ground-based observatories. When needed, we will submit observing proposal to complete information on some of the host stars of our target list.
- From those data, we will derive, either directly or thanks to models, the star parameters: effective temperature, luminosity, gravity (also as an age estimate), metallicity, rotational period, variability, proper motion, multiplicity, magnetic field, topology of the field, wind …
- From the parameters and thanks to star – planet interaction models, we will determine the importance of such interactions.

4.4 WHAT IS THE EXPECTED SIZE OF THE DATA?

To be determined precisely; but not big; in the gigaoctets range. One of the end products will be a catalogue with the properties of the atmosphere of about 100 targets.

Work Package number	Deliverable type	Data format	Data size
WP1	Documents	PDF, excel	At maximum in the 100 Moctets range
WP2	Calibrated spectra of about 100 exoplanets Data reduction Codes Documents, Scientific papers	Fits, PDF, excel	At maximum in the 10 gigaoctets range
WP3	Retrieved parameters for the atmosphere of about 100 exoplanets Retrieval codes Documents, Scientific papers	Fits, PDF, Excel	At maximum in the 10 gigaoctets range
WP4	Catalogues of host stars Codes Documents, Scientific papers	TBD, FITS, PDF, Excel	At maximum in the 10 gigaoctets range
WP5	Models Documents, Scientific papers	TBD, FITS, PDF, Excel	At maximum in the 10 gigaoctets range
WP6	Science products Web site Videos MOOC – SpoC Documents	TBD, FITS, PDF, Excel, Virtual Observatory standards, MP4, TXT, PNG, JPEG, EPS	In the few 100 gigaoctets range

4.5 TO WHOM MIGHT IT BE USEFUL ('DATA UTILITY')?

The data will be useful in first place to the scientific community working on exoplanets. It will also be of interest to students, as well as general public.

5 MAKING DATA FINABLE ACCESSIBLE INTEROPERABLE AND RE-USABLE (FAIR)

The dissemination of knowledge is a key aspect of the project and a WorkPackage, WP6, is dedicated to this aspect (see Figure 4).

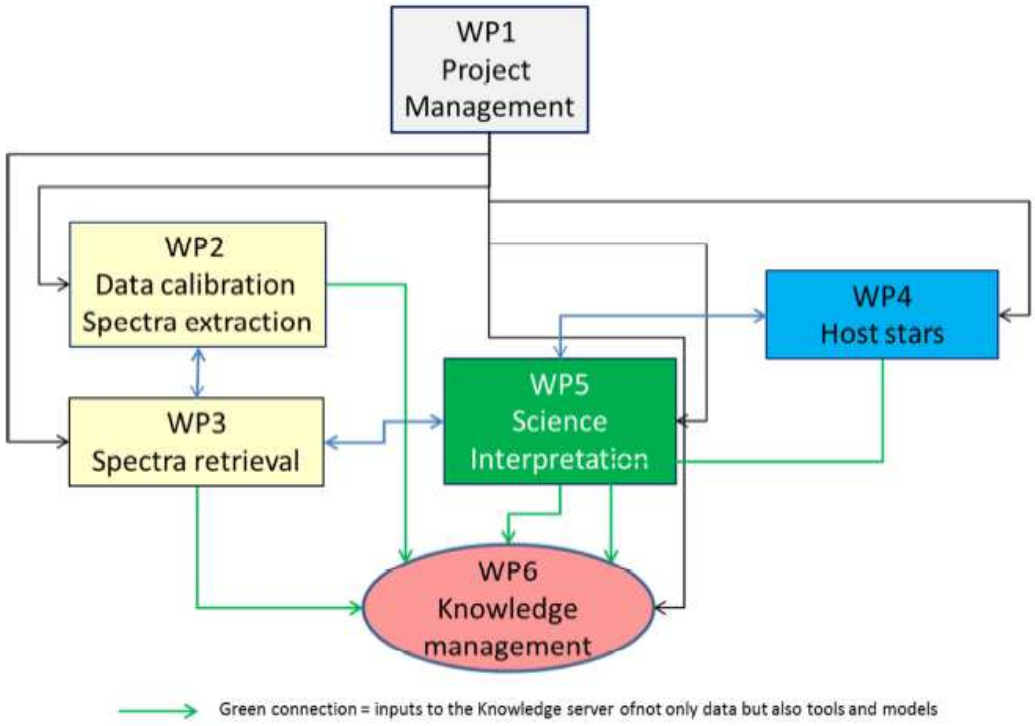

Figure 4 : The data generated by the various WPs will be integrated to a knowledge server.

The Knowledge Management WP aims at

- Capturing knowledge produced by the other WPs (see Figure 4) within a **knowledge base** including all scientific products (data, models, tools and interpretation).
- Providing open access to them through a **knowledge displayer** with two interfaces, one for the scientific community with a direct access to the science products, the second for the general public with educational resources based on the science products.
- Getting feedback from the users

Figure 5 describes the various aspects of knowledge management.

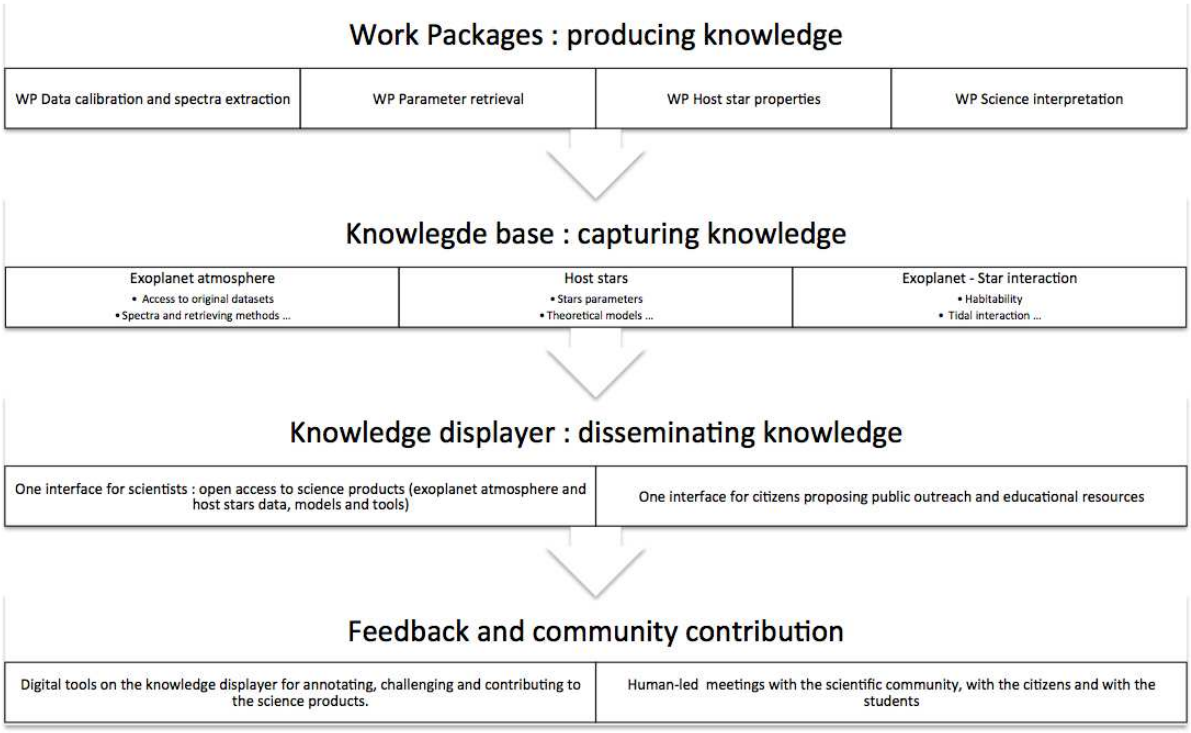

Figure 5: Overview of the knowledge management of the science products

5.1 MAKING DATA FINDABLE AND OPENLY ACCESSIBLE

The Definition, design and production of a Knowledge Base is scheduled from Month 6 to Month 36. The data produced in the framework of the project will be made findable by means of a standard identification mechanism. The data will be encapsulated with metadata. Metadata categories will be defined to follow the different types of science products (ex: Observation, Code…) or educational resources (ex: Video, Image…). For science products, we plan to follow the naming convention used for the Virtual Observatory.

The architecture of the knowledge base (KB) that will capture, record and format the science products for dissemination will be defined first. Once defined collectively with the WorkPackages (WP) 2, 3, 4, 5, the KB will be produced by deploying a server. The data will be stored under a relational database. Indexes, Data Base (DB) engines and cache mechanism will be set up to ensure the scalability and efficiency of the overall data access. Regarding binary files, an indexed and versioning file system will be deployed to keep track of modifications without losing content. The KB will be hosted on a server (NodeJS or PHP according to the actual requirements identified during the specification phase) providing REST (REpresentational State Transfer) access to the data (i.e. a set of standardised Uniform Resource Identifier (URI) allowing to obtain the information in a JSON format). Specific routes will be defined to query the KB according to several criteria: hierarchy, semantic, historic, syntactic... This server will also provide a standardised export mechanism to export data to a user-friendly product (e.g. PDF files or JPG images). We expect a hundred of exoplanet atmospheres at the beginning for the Beta version; but the project will be dimensioned to be able to handle thousands of entries. The Definition and production of a Knowledge Displayer (KD) will start on Month 12. Once recorded in the knowledge base, the scientific community and general public will be granted open access to our science products using two dedicated interfaces. The Knowledge displayer will be in charge of displaying the KB to the end-user by accessing the Knowledge Base Server REST API's (Application Programming Interfaces). All the data will be made openly available at the latest at the end of the project.

5.2 MAKING DATA INTEROPERABLE

The data produced in the project will be interoperable throughout virtual Observatory standard for science products and Learning Tool Interoperability (LTI) standard for education resources.

5.3 INCREASE DATA RE-USE

The data will be made available for re-use through the knowledge server following the deliverables plan of the project. There is no restriction in the re-use of the data generated by the project.

The idea is to keep on feeding the data base after the end of the ExoplANETS-A project, for example in the framework of ARIEL; Ariel is the ESA space mission just selected for the M4 slot of the ESA cosmic vision 2015 – 2025 program and whose adoption is scheduled end of 2020. In any case, we can guarantee that the database will be designed to remain operational for at least 5 years after the project end (for example putting the data on the CERN ZENODO data base).

6 ALLOCATION OF RESOURCES

The data management is one of the work packages of the project; its cost is included in the project cost; it is covered partly by EC and partly by the institutes participating in the project. 46 person.months have been attributed to the workpackage. There is a lead from CEA and a co-lead from INTA.

7 DATA SECURITY

The data security (including data recovery as well as secure storage) will be taken into account in the design of the data base and of the knowledge server.