



**HAL**  
open science

## PARSEC Data and Digital Output Management Plan and Workbook

Shelley Stall, Corrêa Pedro Luiz Pizzigatti, Alison Specht, Romain David,  
Rorie Edmunds, Laurence Mabile, Jeaneth Machicao, Margaret O'Brien,  
Lesley Wyborn

► **To cite this version:**

Shelley Stall, Corrêa Pedro Luiz Pizzigatti, Alison Specht, Romain David, Rorie Edmunds, et al..  
PARSEC Data and Digital Output Management Plan and Workbook. 2020. hal-02885254

**HAL Id: hal-02885254**

**<https://hal.science/hal-02885254>**

Preprint submitted on 30 Jun 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

## Data and Digital Output Management Plan and Workbook for the Belmont Forum

Collaborative Research Action (CRA) Science-driven e-Infrastructure Innovation (SEI) for the Enhancement of Transnational, Interdisciplinary and Transdisciplinary Data Use in Environmental Change



### **Project Building New Tools for Data Sharing and Reuse through a Transnational Investigation of the Socioeconomic Impacts of Protected Areas (PARSEC)**

June 2020

#### Introduction

This Data and Digital Output Management Plan and Workbook (DDOMP) are used by the PARSEC team to establish the policies the team will follow, document the operational procedures necessary to comply with those policies, and the planned activities necessary to manage PARSEC data, software, and other digital outputs. This is a working document.

This document describes activities necessary during the PARSEC research lifecycle as well as those necessary to preserve all digital outputs for use into the future. It is the intent of the PARSEC team to make our digital outputs as open as possible, discoverable, accessible, well-documented to promote the broadest reuse. These elements are both recommended and required by the **Belmont Forum Open Data Policy and Principles**<sup>1</sup> and the **FAIR Guiding Principles**<sup>2</sup>.

The Belmont Forum Open Data Policy and Principles state that:

Datasets should be:

- Discoverable through catalogues and search engines
- Accessible as open data by default, and made available with minimum time delay
- Understandable in a way that allows researchers—including those outside the discipline of origin—to use them
- Manageable and protected from loss for future use in sustainable, trustworthy repositories

The FAIR Guiding Principles state that data should be Findable, Accessible, Interoperable, and Reusable.

<sup>1</sup> <http://www.belmontforum.org/about/open-data-policy-and-principles/>

<sup>2</sup> <https://www.nature.com/articles/sdata201618>.



The **Table of Contents** (page 3) lists the operational procedures addressed. In each procedure we provide reference to the specific elements requested by the Belmont Forum to support the review process.

For an **understanding of repository process steps for data preservation** referenced in this DDOMP, we recommend the one preferred by our selected preservation repository, the Environmental Data Initiative. Specifically the Data Management Overview and Processing Pipeline as defined by the Santa Barbara Channel Marine Biodiversity Observation Network (SBC MBON): <https://sbc.marinebon.org/data/overview/>  
This aligns with the metadata standards used by the Environmental Data Initiative, EML 2.2.

**Glossary of terms:** A well-managed glossary that partners with the Research Data Alliance, a PARSEC partner, is the [CASRAI Research Data Management Glossary](#).



## Table of Contents

Introduction	1
PARSEC Team Members Roles	4
PARSEC - Operational Procedures for selecting datasets for preservation.	5
PARSEC - Operational Procedures for protecting digital objects during the life of the project.	6
PARSEC - Operational Procedures for ensuring quality of digital objects during the life of the project and prior to preservation.	8
PARSEC - Operational Procedures for ensuring data security, privacy, and intellectual property restrictions associated with datasets and digital outputs are honored and preserved in derivative products.	9
PARSEC - Operational Procedures for assigning usage licences to data and digital outputs.	9
PARSEC - Operational Procedures for preparing supporting documentation (i.e., metadata) for data and digital outputs that supports public access, discovery, and long-term reuse.	11
PARSEC - Operational Procedures for selection of a digital object Preservation Repository	12
PARSEC - Sensitive or Restricted Data Operational Procedures	14
PARSEC Data Sets (to include Data Types, Transition to Preservation Repository, etc.)	15
PARSEC Software (to include Models, Workflow, etc.)	15
PARSEC other Digital Outputs (to include Posters, Presentations, Training Materials, Workshop Materials, etc.)	15
PARSEC - Operational Procedures for determining costs associated with long-term data management.	16
Appendix 1 - WHO Multi-Country Studies Data Archive, Data Access Agreement for the dataset "France - World Health Survey 2003"	17

## PARSEC Team Members Roles

*Belmont Forum Rubric Reference: 3.0: Team Members responsible for developing, implementing, overseeing, and updating the Data and Digital Outputs Management Plan*

The Principal Investigators (PI) for each of the PARSEC strands are responsible for the development, implementation, oversight, and maintenance of the Data and Digital Output Management Plan and Workbook. The type of outputs are different for each strand. This collaboration allows for all types of digital objects to be addressed.

- Responsible for Synthesis Science Digital Content Used and Developed  
David Mouillot, Synthesis Science Strand PI  
david.mouillot@umontpellier.fr  
<https://orcid.org/0000-0003-0402-2605>
- Responsible for Data Science Strand Digital Content Used and Developed  
Shelley Stall, Data Science Strand PI  
sstall@agu.org  
<https://orcid.org/0000-0003-2926-8353>
- Responsible for Complying with Belmont Forum Open Data Policy and Principles  
All team members.

**All PARSEC team members have signed a Code of Conduct document** concerning data and digital output management that states “Derived data and digital outputs generated during PARSEC activity will be documented and released at the time of publication where possible into the public domain in compliance with Belmont Forum requirements. Policies for broad access and sharing, including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements, metadata description, intended repositories, and so on will be clearly described in the Project’s Data and Digital Outputs Management Plan.”

Signed PARSEC Code of Conduct forms from all team members are kept on file and available for review.

Responsibility following the end of the project for data and digital objects created by the team is entrusted to the selected preservation repositories. Selected repositories have indicated a commitment to maintaining the preserved objects and their persistent identifiers in support of the scientific record.

All PARSEC team members are responsible to have and manage an ORCID profile that links to their digital products. This promotes credit and automated attribution.

## PARSEC - Operational Procedures for selecting datasets for preservation.

*Belmont Forum Rubric Reference: 1.1 Datasets and other digital objects of long-term value are identified, including data type and encoding, 1.2 How the data and other digital objects will be collected, captured, or created.*

A fundamental principle of the PARSEC project is to promote data reuse by NOT creating any new earth observation data. All raw data needed for our work already exists and includes household survey datasets and satellite imagery. Datasets significant to our research are logged in our [Dataset Tracking](#) tool as well as any processed or derived dataset.

As an overview, PARSEC is using existing raw data that includes satellite imagery available at various resolutions, time-series, and costs from instruments associated with our country partners: US, Brazil, France, and Japan. To perform ground truth analysis with our models we are using local household surveys for towns and villages associated with our selected sites. These are sourced from primarily global operations such as WHO World Health Survey with more precise data available (if needed) from the Demographic and Health Surveys (DHS), carried out by ORC Macro for the United States Agency for International Development (USAID) and the Living Standards Measurement Study (LSMS) surveys, conducted with technical assistance from the World Bank. Depending on the selection of protected area sites, we likely will need additional surveys.

Much of this data is available in both detailed and aggregated forms. For the detailed data we must submit formal requests and comply with the terms of use, and necessary protections for this sensitive information.

PARSEC output data will include global gridded socio-economic information and socio-economic dynamics specific to our study sites.

Our primary data and digital products planned for preservation include:

- Cleaned observation records and measurements
- Aggregated datasets subjected to quality control protocols, used for machine-learning and artificial intelligence models. This includes the provenance of what data was used to develop the aggregated dataset.
- Documentation of the quality control protocols used
- Algorithms developed to prepare data (cleaning scripts) and perform analyses
- Software supporting machine learning and artificial intelligence models
- Results of our analysis
- Training and workshop material

**Dataset Preservation:** If a dataset is used for analysis, it is to be preserved and made available prior to the end of the project or if more time is needed then before the final report is submitted to the Belmont Forum. We track this in the table where all datasets are logged. For

# PARSEC - Data and Digital Output Management Plan and Workbook



any papers written during the PARSEC project, datasets used to support that research must be preserved and available at the time of publication.

**Software Preservation:** All software developed is to be preserved at each major release.

**Training and workshop materials:** All materials are to be preserved on or near the time of the workshop.

## PARSEC - Operational Procedures for protecting digital objects during the life of the project.

*Belmont Forum Rubric Reference: 4.1 Describe security measures for the data and other digital outputs to prevent unauthorized access with specific examples.*

The PARSEC project has selected a number of tools to use during the project. We have elements of our project that we want to keep confidential during the development, but other objectives that benefit from a very broad community, much larger than our team, supporting review and feedback. The PARSEC Data Strand provides support and expertise for the data management needs of the entire team.

For material development and temporary storage: **Google Drive**

We have organized the content allowing different permissions for leadership, team members, and other sub-groups. **Security measures:** minimum protocols are in place. Links to the materials are easy to share across the team, partners, and other stakeholders.

For team communications and information dissemination: **Email, Slack, Zotero**

**Security measures:** minimum protocols are in place. The email list for the project is managed by the leadership. The Slack channel is available only to project members, partners, and other invited stakeholders. [PARSEC's group in Zotero](#) is open to those interested in the references gathered.

For dataset storage during the project: **[Université de Montpellier/CINES](#), [Amazon Web Services \(AWS\)](#), [Open Science Framework \(OSF\)](#)**

We are currently working through the best way to integrate these platforms and apply the terms of use for the datasets we plan to use. Each of these platforms supports access/controls for our team members to protect the project space. Montpellier's repository supports protections necessary for sensitive data. The team will have access to AWS in June 2020. OSF can integrate with AWS, GitHub, Google Drive, and other platforms.

**Security measures CINES:** [Le C.I.N.E.S.](#) (Centre Informatique National de l'Enseignement Supérieur) is a national public institution, based in Montpellier (France) and placed under the Ministry of Higher Education, Research and Innovation (MESRI). The C.I.N.E.S. has three national strategic missions: Supercomputing, Long-term preservation and Hosting computer platforms nationwide.

# PARSEC - Data and Digital Output Management Plan and Workbook



Details on security can be found here: <https://www.cines.fr/en/long-term-preservation/expertises/quality/security/>

**Security measures AWS:** AWS provides robust security measures for project teams. The full set of best practices can be found here: <https://aws.amazon.com/blogs/security/tag/best-practices/>

**Security measures OSF:** Each member of PARSEC must have an OSF login in order to manage access/controls. The project space can be kept private or public. Project files are backed up each night. More information here: <https://help.osf.io/hc/en-us/articles/360019930993-FAQs-Security>

For dataset and supporting object preservation (except for software and training/workshop materials): **Environmental Data Initiative**

The **Environmental Data Initiative** (EDI, independently funded by the U.S. National Science Foundation) is a partner on the PARSEC project, and a preferred domain repository. EDI promotes and enables curation and re-use of environmental data for large ecosystem-level research projects including the U.S. Long Term Ecological Research Program (LTER) and many field stations, individual laboratories and synthesis research projects. The EDI repository fully supports metadata and data immutability and strong versioning for auditing and reproducibility. Each dataset version receives a Digital Object Identifier (DOI), registered and resolved through DataCite. EDI is working toward CoreTrustSeal certification, and collaborates with other data professionals to operationalize FAIR principles. EDI is a signatory of the Enabling FAIR Data Commitment Statement.

**Security measures:** The EDI repository ensures stable access via servers hosted on the campus of the University of New Mexico (UNM) and connected via UNM's Science DMZ research network, DOE Energy Sciences Network (ESNet) and the Western Regional Network. EDI's data storage capacity is currently at 30TB (expandable), with an equivalent mirror storage for near-time backups. EDI also uses the AWS Glacier storage as a long-term "cold data" archive.

EDI's authentication procedures use organizational LDAP systems (Lightweight Directory Access Protocol), with mapping to 3rd-party identity service providers (Google, GitHub and ORCID), using the OAuth/OpenID Connect protocols. EDI metadata employs Ecological Metadata Language (EML), which includes access rules at the data package and data entity levels to support embargo when necessary. EDI advocates for open and unfettered access, but does not forbid providers from declaring more restrictive licensing agreements for use of their data packages. If providers do not include a statement of Intellectual Rights in EDI adds a default declaration based on the Creative Commons CC0 "No Rights Reserved" waiver. Details are posted: <https://environmentaldatainitiative.org/edi-data-policy-2/>

For software development: **GitHub**

GitHub allows a team of software developers to work together with version control, collaboration, and easy access. During the project the team site will be kept private. At the end of the project or if referenced in a paper, the site will be made open. The PARSEC GitHub repository is located at <https://github.com/PARSECworld>.

**Security measures:** GitHub is GDPR compliant and adheres to the [Privacy Shield Framework](#), certified January 26, 2017. More information on GitHub security can be found here: <https://github.com/security/trust>

For software preservation: **Zenodo** (using integrated link with GitHub)

Zenodo is a well known general repository that has specialized in software preservation through their partnership with GitHub.

**Security measures:** Our digital objects preserved in Zenodo are open and licensed CC BY 4.0. More information here: <https://about.zenodo.org/policies/>

For training and workshop materials preservation: Zenodo, [PARSEC Community](#)

**Security measures:** Our digital objects preserved in Zenodo are open and licensed CC BY 4.0. More information here: <https://about.zenodo.org/policies/>

For credit and automated attribution: PARSEC team members will have active and up-to-date ORCID profiles that allow for automated updates from Crossref and DataCite.

PARSEC - Operational Procedures for ensuring quality of digital objects during the life of the project and prior to preservation.

*Belmont Forum Rubric Reference: none.*

The PARSEC project is developing a set of processes to ensure that the quality of datasets generated by the team through aggregation or analysis meets all the quality requirements and are consistent across the country teams. We plan to report on these processes in future reports.

Processes to ensure Quality and consistency will include:

1. Methods for handling missing data
2. Methods for handling data at different granularities.
3. Methods for handling data using different time series.
4. Use of file naming standards.
5. Completion and accuracy of metadata.
6. Use of standard vocabularies.
7. Adequate capture of changes made to digital objects to maintain transparency.
8. Tracking of those involved to ensure an accurate list of creators/authors with their ORCIDs.
9. Correct levels of access/controls for sensitive/restricted data.
10. Compliance with all elements defined in the DDOMP.

PARSEC - Operational Procedures for ensuring data security, privacy, and intellectual property restrictions associated with datasets and digital outputs are honored and preserved in derivative products.

*Belmont Forum Rubric Reference: 7.1 Describe **intellectual property rights** to the data and other digital outputs.*

**All PARSEC team members have signed a Code of Conduct document, as mentioned previously,** that along with expectations on data and digital output management, also includes the following statements about how intellectual property rights will be handled:

Investigators will adhere to the appropriate standards for reporting the results of their scientific activities including respecting the intellectual property rights of others consistent with the European Code of Conduct for Research Integrity (2011) downloadable from: <http://www.esf.org/coordinating-research/mo-fora/research-integrity.html>.

Derived data and digital outputs generated during PARSEC activity will be documented and released at the time of publication where possible into the public domain in compliance with Belmont Forum requirements. Policies for broad access and sharing, including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements, metadata description, intended repositories, and so on will be clearly described in the Project's Data and Digital Outputs Management Plan

The Belmont Forum claims no intellectual property rights (including copyright) in any reports or data produced by the funded party except for summary reports on grant outcomes required by the Belmont Foundation or other bodies to whom the Belmont Foundation is required to provide reports.

Signed PARSEC Code of Conduct forms from all team members are kept on file and available for review.

PARSEC - Operational Procedures for assigning usage licences to data and digital outputs.

*Belmont Forum Rubric Reference: 7.2 Describe **licensing** of data and other digital outputs.*

PARSEC digital outputs are to be as **open as possible and closed as necessary**. PARSEC team members will follow these steps to ensure reuse of datasets and other digital objects or for created digital objects.

**Existing Datasets or Digital Objects:**

# PARSEC - Data and Digital Output Management Plan and Workbook



Step 1 - determine the terms of use and licensing of datasets (of digital objects) being used by the team. Determine if there are additional country or regional laws that must be considered, or indigenous sovereignty governance required of the data.

PARSEC team members will:

- Give credit to the dataset authors
- Be responsible with the data
- Share what is learned
- Respect the dataset license or waiver
- Understand and follow any restrictions or regulations

Step 2 - determine how those existing requirements should be inherited to aggregated/quality-controlled datasets that result in derived data products. Derived data products (to which the project has added intellectual content) will cite the original source(s) in the metadata.

## **PARSEC Created Datasets/Digital Objects:**

Step 1 - determine if any requirement must be followed. Otherwise, the dataset is as open as possible with the ability to create derived data products. Creative Commons CC BY 4.0 is the default for the project, but CC0 for US government agencies as required.

Step 2 - confirm that the selected repository supports the desired access/controls. The digital object must have the most open license possible. Creative Commons CC BY 4.0 or CC0

**Software:** the license PARSEC will use is: **GNU GPLv3** which allows others to do almost anything they want with the software, *except* distributing closed source versions.

**Training and Workshop Materials:** the license used is Creative Commons CC BY 4.0.

## **Note:**

**CC0** is a waiver and allows any type of usage of the object is assigned. This includes access, mining, redistribution & reuse.

**CC BY 4.0** applies internationally and requires all derivative works to acknowledge all contributors to all works from which they are derived. Digital objects can be accessed, mined, redistributed, and reused as long as attribution is indicated.

References to be used by the team making these determinations:

Creative Commons Licenses: <https://creativecommons.org/about/cclicenses/>

RDA-CODATA Legal Interoperability Interest Group: LEGAL INTEROPERABILITY OF RESEARCH DATA: PRINCIPLES AND IMPLEMENTATION GUIDELINES: [https://rd-alliance.org/sites/default/files/attachment/Legal%20Interoperability%20Principles%20and%20Implementation%20Guidelines\\_Final.pdf](https://rd-alliance.org/sites/default/files/attachment/Legal%20Interoperability%20Principles%20and%20Implementation%20Guidelines_Final.pdf)

CARE Principles for Indigenous Data Governance: <https://www.gida-global.org/care>

Legal Side of Open Source: <https://opensource.guide/legal/>

How to License Research Data: <http://www.dcc.ac.uk/resources/how-guides/license-research-data>

PARSEC - Operational Procedures for preparing supporting documentation (i.e., metadata) for data and digital outputs that supports public access, discovery, and long-term reuse.

*Belmont Forum Rubric Reference: 8 Describe **supporting documentation and metadata** that will be created to make data and digital outputs publicly accessible, discoverable, and reusable.*

PARSEC will follow the community standards for each product type.

**For data and supporting products preserved in the Environment Data Initiative repository**, each data package will include metadata in the Ecological Metadata Language (EML, <https://eml.ecoinformatics.org/>).

**For software preserved in Zenodo** we will use the Force11 recommendation of [CodeMeta](#), with support from the online tool to [generate CodeMeta Files](#).

**For training and workshop material preserved in Zenodo** we will use metadata criteria defined in the Data Management Training Clearing House: <https://dmtclearinghouse.esipfed.org>. We will also register the materials in the clearinghouse.

**For any digital product not included above** we will use metadata criteria defined by the selected repository, specifically Zenodo, enhanced as needed depending on product type.

All digital products, no matter the repository, will include the following core metadata:

**Provided by submitter:**

- Creator/Author(s) (with given name, family name, ORCID, and affiliation)
- Title
- Resource Type (e.g. paper, pre-print, dataset, software, etc.)
- Contributors (with given name, family name, ORCID and affiliation)
- Date of submission
- Related identifier and relationship (if it applies)
- Version (if it applies)
- Description/Abstract
- Geolocation information (if it applies)
- Key parameters and units used
- Size of file (if it applies)

## **Controlled by repository:**

- Persistent Identifier (e.g. Digital Object Identifier)
- Publisher or Repository Name
- Publication Date
- Recommended citation

To enable automated attribution and credit, ORCID IDs for authors will be included with all products preserved in repositories.

## PARSEC - Operational Procedures for selection of a digital object Preservation Repository

*Belmont Forum Rubric Reference: 2.3 Name of the Repository; 5.1 Repository T&C and/or sustainability plan / cost, 5.2 Entity responsible for managing long-term accessibility*

The PARSEC team used the following guidelines when selecting repositories for our digital output preservation.

### **Desirable Repository Characteristics:**

1. Supportive of the [Belmont Forum Open Data Policy and Principles](#) - clipped here:
  - a. Discoverable through catalogues and search engines
  - b. Accessible as open data by default, and made available with minimum time delay
  - c. Understandable in a way that allows researchers—including those outside the discipline of origin—to use them
  - d. Manageable and protected from loss for future use in sustainable, trustworthy repositories
2. Compliance with the [Enabling FAIR Data Commitment Statement](#) for Repositories - clipped here:
  - a. Ensure that research outputs (e.g., data, software, technology, and physical samples) curated by repositories are open and FAIR, have essential documentation, and include human-readable and machine-readable metadata (e.g., on landing pages) in standard formats that are exposed and publicly discoverable.
  - b. Ingest and expose data to promote interoperability and reuse.
  - c. Ensure that unique, persistent identifiers are used for authors (e.g., [ORCID](#)), research objects (e.g., [Digital Object Identifier](#)), and physical samples (e.g., [IGSN](#)).
  - d. Create associations among the research outputs that they manage and other related entities.
  - e. Ensure that data and software have licenses that are as open as possible, and as protected as necessary.

- f. Support peer-review of related manuscripts by enabling access to the research outputs prior to publication.
    - g. Gain third-party validation of trustworthy and sustainable practices and capabilities.
  3. Dataset or file size limits can support PARSEC requirements.
  4. Costs for preservation during the project are within the PARSEC budget
    - a. PARSEC country members have accounted for data management in their budgets. Each country-lead monitors their budget and usage.
  5. Access and controls necessary to protect sensitive data.
  6. Indicates in the terms of use (or similar) their commitment to long-term preservation and persistence of the digital objects in their care as part of the scientific record.
  7. For domain repositories, has a third-party certification of services such as CoreTrustSeal (for domain repositories only) or are a [World Data System Regular Member](#) compliant with their [Data Sharing Principles](#).
  8. Uses English as a primary language, or has an English version that allows for searching, and other discovery activity in English. [Note: Journals prefer that preserved datasets to be in a repository that supports the English language.]
  9. Working towards compliance with the [TRUST Principles for digital repositories](#)<sup>3</sup>.
  10. Complies will all PARSEC funder requirements

## PARSEC Funder Repository Requirements – Reference Information:

**São Paulo Research Foundation, FAPESP (São Paulo, Brazil):** This workbook follows the FAPESP guidelines that recommends to deposit in repositories where the data will be maintained and preserved according to the standards of each scientific domain, under the ethical and security constraints defined by the project and its associated domains. FAPESP has been working since 2017 in the open research data repositories that will be considered as soon as available (<http://www.fapesp.br/openscience/en>).

<http://www.fapesp.br/openscience/en>

**Agence Nationale de la Recherche, ANR (France):** This workbook aligns with ANR's [Open Science Policy](#). In it they reference "[Practical Guide to the International Alignment of Research Data Management](#)" which has an extensive list of guidance for selecting trustworthy repositories (pp. 26-30).

**National Science Foundation, NSF (US) -** The NSF currently does not have guidance for selecting repositories. In NSF 19-069, [Dear Colleague Letter: Effective Practices for Data](#) they indicated the importance of data having persistent identifiers. This workbook supports that best practice.

<sup>3</sup> Lin, D., Crabtree, J., Dillo, I. *et al.* The TRUST Principles for digital repositories. *Sci Data* 7, 144 (2020). <https://doi.org/10.1038/s41597-020-0486-7>

**Japan Science and Technology Agency, JST (Japan)** - This workbook is aligned with [Implementation Guidelines: JST Policy on Open Access to Research Publications and Research Data Management](#). Here it is indicated that “It is preferable that research data be made openly available by depositing them in an existing public database or common disciplinary repositories established by academic societies to facilitate reuse of research results. If there is no appropriate public database, institutional repositories could also be used. In some cases, the academic journals to which research publications are submitted suggest designated repositories to deposit data.”

## PARSEC - Sensitive or Restricted Data Operational Procedures

*Belmont Forum Rubric Reference: 6.1 - Describe how sensitive data and other digital outputs are made available beyond the project team, 6.2 - Describe any limitations on the ability to share data and other digital outputs.*

PARSEC is using socio-economic data sourced from primarily global operations such as WHO World Health Survey with more precise data available (if needed) from the Demographic and Health Surveys (DHS), carried out by ORC Macro for the United States Agency for International Development (USAID) and the Living Standards Measurement Study (LSMS) surveys, conducted with technical assistance from the World Bank. Depending on the selection of protected area sites, we likely will need additional surveys.

These datasets include details about communities that are defined as ***sensitive***.

To date, only one dataset has been requested from the WHO. The data access agreement is listed in [Appendix 1](#). All elements defined will be complied with by the PARSEC team. Those elements specific to this DDOMP are in bold.

For data that does not have a specified data access agreement or terms of use:

Before sharing data, PARSEC team members practice includes:

- Understanding the Belmont Forum’s policies on data sharing, and any additional institution or country requirements.
- Only share data that has been aggregated and can no longer be used to identify individuals.

Before using existing datasets previously created, PARSEC team members practice includes:

- Giving credit in publications, presentations, and other materials
- Respecting the assigned license
- Protecting the datasets appropriately

PARSEC may be using data associated with indigenous peoples. Should that occur, the team will consult the [CARE Principles](#) for Indigenous Data Governance. At the time of this report there is not implementation guidance, but we do expect that to be developed during the time of this project.

## PARSEC Data Sets (to include Data Types, Transition to Preservation Repository, etc.)

*Belmont Forum Rubric Reference: 1.1 - Types of data and other digital outputs of long-term value, 1.3 - How much data is anticipated, 2.1 - Metadata standards or formats for the data and other digital objects, 2.2 - When data and other digital outputs will be made available outside and within the project team, 2.3 - How data and other digital outputs will be made available beyond the project team, 6.2 - Limitations on the ability to share data and other digital outputs, 7.2 - licensing of the data and other digital outputs*

**Datasets** developed by the PARSEC project to include the date of identification, name, metadata standard used, available formats, name of repository, date of initial submission.

### PARSEC Dataset Tracking Tool

<https://docs.google.com/spreadsheets/d/1XQuGBRJKrgoZsxim9SKzi8cv0P-6tQOOpuxrLzGdEmY/edit?usp=sharing>

## PARSEC Software (to include Models, Workflow, etc.)

*Belmont Forum Rubric Reference: 1.1 - Types of data and other digital outputs of long-term value, 1.3 - How much data is anticipated, 2.1 - Metadata standards or formats for the data and other digital objects, 2.2 - When data and other digital outputs will be made available outside and within the project team, 2.3 - How data and other digital outputs will be made available beyond the project team, 6.2 - Limitations on the ability to share data and other digital outputs, 7.2 - licensing of the data and other digital outputs*

**Software** developed by the PARSEC project to include the date of identification, name, description, metadata standard used (CodeMeta), language, name of repository, date of initial submission.

### PARSEC Software Tracking Tool

<https://docs.google.com/spreadsheets/d/10sLE0CcvHwM3vIAyilby7nuH-e6iyffY13veBLzg4WU/edit?usp=sharing>

## PARSEC other Digital Outputs (to include Posters, Presentations, Training Materials, Workshop Materials, etc.)

*Belmont Forum Rubric Reference: 1.1 - Types of data and other digital outputs of long-term value, 1.3 - How much data is anticipated, 2.1 - Metadata standards or formats for the data and other digital objects, 2.2 - When data and other digital outputs will be made*

*available outside and within the project team, 2.3 - How data and other digital outputs will be made available beyond the project team, 6.2 - Limitations on the ability to share data and other digital outputs, 7.2 - licensing of the data and other digital outputs*

**Other Digital Objects** developed by the PARSEC project that includes the date of identification, name, metadata standard used, available formats, name of repository, date of initial submission. These objects include posters, presentations, training materials, and workshop materials organized by event where they were presented.

<https://docs.google.com/spreadsheets/d/1dH-uvcim-kIC590HKJOLNNHOL8UaRTmhY7iGLBVL5Bw/edit?usp=sharing>

PARSEC - Operational Procedures for determining costs associated with long-term data management.

*Belmont Forum Rubric Reference: 9 Describe costs or estimated costs associated with long-term data management or an assigned data manager.*

## **Cost Description:**

**Acquiring External Data Sources:** For the PARSEC project this primarily includes satellite images.

**Quality Check Data:** Develop and use common methods for mitigating quality challenges.

**Prepare Metadata:** Develop and use common templates for data and digital objects.

**Data Storage and Backup During Project Duration:** Ensure data is protected throughout the project. This includes cloud usage and storage costs.

**Data Access and Security:** All members of the team, from six countries, need access to the project data and digital objects.

**Data Preservation:** Includes conversion costs for preparing data.

**Data Sharing and Reuse:** Costs for copyright clearance or legal advice. Costs for curation of datasets.

The PARSEC team members are providing the above data management services as in-kind contributions except for the following items:

- Fees for satellite imagery - the team from France is managing this budget item.
- Access to AWS - the team from Brazil has allotted \$11,000 USD per year to support the PARSEC team. This access starts in June 2020.

Appendix 1 - WHO Multi-Country Studies Data Archive, Data Access Agreement for the dataset “France - World Health Survey 2003”

1. **Access to the restricted data will be limited to the Lead Researcher and other members of the research team listed in this request.**
2. Copies of the restricted data or any **data created on the basis of the original data will not be copied or made available to anyone other than those mentioned in this Data Access Agreement, unless formally authorized by the Data Archive.**
3. The data will only be processed for the stated statistical and research purpose. **They will be used for solely for reporting of aggregated information, and not for investigation of specific individuals or organizations.** Data will not in any way be used for any administrative, proprietary or law enforcement purposes.
4. The Lead Researcher must state if it is their intention to match the restricted microdata with any other micro-dataset. If any matching is to take place, details must be provided of the datasets to be matched and of the reasons for the matching. **Any datasets created as a result of matching will be considered to be restricted and must comply with the terms of this Data Access Agreement.**
5. The Lead Researcher undertakes that no attempt will be made to identify any individual person, family, business, enterprise or organization. If such a unique disclosure is made inadvertently, no use will be made of the identity of any person or establishment discovered and full details will be reported to the Data Archive. The identification will not be revealed to any other person not included in the Data Access Agreement.
6. The Lead Researcher will implement **security measures** to prevent unauthorized access to licensed microdata acquired from the Data Archive. The microdata must be destroyed upon the completion of this research, unless the Data Archive obtains satisfactory guarantee that the data can be secured and provides written authorization to the Receiving Organization to retain them. Destruction of the microdata will be confirmed in writing by the Lead Researcher to the Data Archive.
7. Any books, articles, conference papers, theses, dissertations, reports, or other publications that employ data obtained from the Data Archive will **cite the source of data in accordance with the citation requirement provided with the dataset.**
8. An electronic copy of all reports and publications based on the requested data will be sent to the Data Archive.
9. The original collector of the data, the Data Archive, and the relevant funding agencies bear no responsibility for use of the data or for interpretations or inferences based upon such uses.



10. This agreement will come into force on the date that approval is given for access to the restricted dataset and remain in force until the completion date of the project or an earlier date if the project is completed ahead of time.

11. If there are any changes to the project specification, security arrangements, personnel or organization detailed in this application form, it is the responsibility of the Lead Researcher to seek the agreement of the Data Archive to these changes. Where there is a change to the employer organization of the Lead Researcher this will involve a new application being made and termination of the original project.

12. Breaches of the agreement will be taken seriously and the Data Archive will take action against those responsible for the lapse if willful or accidental. Failure to comply with the directions of the Data Archive will be deemed to be a major breach of the agreement and may involve recourse to legal proceedings. The Data Archive will maintain and share with partner data archives a register of those individuals and organizations which are responsible for breaching the terms of the Data Access Agreement and will impose sanctions on release of future data to these parties.