Data exchange format and template

DELIVERABLE NO. 4.2

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 835896

Deliverable No.	D4.2	Work Package No.	WP4
Work Package Titel	Open modelling platform		
Status		Final	
Dissemination level	Public	PU	
Due date deliverable	2019.10.31	Submission date	2019.10.31
Deliverable version	1.0		

Deliverable Contributors:	Name	Organisation	Date
Deliverable Leader	Volker Krey	IIASA	2019.10.29
Work Package Leader	Daniel Huppmann	IIASA	
Contributing Author(s)	Sandrine Charousset	EDF	
	Luis Olmos Camacho	Commillas	
	Jed Cohen	JKU	
	Andrés Ramos Galán	Comillas	
	Paolo Pisciella	NTNU	
	Hettie Boonman	TNO	
	Theresia Perger	TU Wien	
	Philipp Haertel	Fraunhofer IEE	
	Ingeborg Graabak	SINTEF Energy Research	
Reviewer(s)	Pedro Crespo del Granado	NTNU	2019.10.30
Final review and approval	Ingeborg Graabak	SINTEF Energy Research	2019.10.30

History of Change

Release	Date	Reason for Change	Status
1.0	2019.10.29	-	First release

DISCLAIMER / ACKNOWLEDGMENT

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 835896.

List of acronyms used in this document

EMF	Energy Modelling Forum
EMP-E	Energy Modelling Platform for Europe
IAMC	Integrated Assessment Modelling Consortium
IPCC	Intergovernmental Panel on Climate Change
WP	Work Package

1. Introduction

Providing results from energy-economic modelling to policymakers and the public at large in a transparent and accessible format requires intuitive, web-based tools for visualization and analysis. In addition, making methods, structural information and detailed parametric model assumptions transparently available to the wider scientific community is important to live up to scientific standards and establish trust in the results of model-based studies.

In order to facilitate the linking of different energy models in an efficient way, on the one hand efficient processes and workflows are needed and on the other hand standardized ways of exchanging information among the models. This requires the development of both an improved modelling platform infrastructure as well as the standards and "handshake" definitions to exchange data between models within the consortium and disseminate results underpinning policy insights to the broader scientific community and all stakeholders of this project.

The openENTRANCE consortium has thus embarked on the creation of common data exchange formats to share model input data and quantitative results. Ideally such standards build on existing standards to enable faster uptake across the energy modelling community. IIASA has been involved in the process of developing of such standards in the context of the Integrated Assessment Modelling Consortium (IAMC), and drawing on the experiences and lessons learned there to ensure the development of standards that are compatible with norms used in other disciplines.

It is, however, important to recognize that in energy modelling which is in itself a multi-disciplinary field of research and therefore draws on a range of disciplines, multiple data exchange formats have been established in different parts of the community and will continue to co-exist for the foreseeable future. Therefore, beyond defining a standardized data exchange format within openENTRANCE, it is desirable to make it interoperable with those other efforts to agree on standardized data exchange formats to lower the threshold of collaboration across the wider community. This challenge will be taken up in the development of the open modelling platform by providing tools to convert between these formats.

The remainder of this document is organized as follows: Section 2 starts out with describing the IAMC data format as used to date. Section 3 presents the limitations of this data format vis-à-vis the use cases of the openENTRANCE project as well as extensions of the data format to overcome these limitations. Finally, Section 4 concludes and suggests a way forward for implementation as well as ensuring interoperability of the data format with other data standards used in different communities related to energy modeling.

2. IAMC data format

Over the past decade, the Integrated Assessment Modeling Consortium (IAMC) developed a standardized tabular timeseries format to exchange scenario data. Previous use cases include reports by the Intergovernmental Panel on Climate Change (IPCC), the model comparison exercises within the Energy Modeling Forum (EMF) hosted by Stanford University and a number of EU-funded projects (e.g., AMPERE, LIMITS, ADVANCE, CD-LINKS).

In addition, a number of tools have been developed around this data format which allows efficient data processing and analysis without having to write code from scratch (cf. IAMC GitHub organization[1]). For example, the Python-based pyam package and R-based iamc package are both geared for analysis and visualization of any scenario data provided in this format.

Data container (file format)

To guarantee a low entry barrier for the use of the IAMC data template, a format that allows both human and machine processing is preferable. In addition, to keep the data, meta data and further documentation together in a single file, a format that allows including multiple lists or tables is preferable. As a result of these basic requirements and continuity with previous data templates, a spreadsheet format has been identified to be the most appropriate container for the IAMC data template. More specifically, because of its widespread use, the Microsoft Excel file format is being used (both in its proprietary 2003 variant, xls, and in the more recent, non-proprietary 2007 variant, xlsx).

Figure 1 below shows a typical example of integrated-assessment scenario data following the IAMC format from the CD-LINKS project.

	A	B	C	D	E	F	G	H	I
1	**Model**	**Scenario**	**Region**	**Variable**	**Unit**	**2010**	**2015**	**2020**	**2025**
2	MESSAGEix-GLOBIOM_1.0	NPi2020_400_SSP1	WEU	Emissions|CO2	Mt CO2/yr	3838.271	2975.35	...	...
3	MESSAGEix-GLOBIOM_1.0	NPi2020_400_SSP1	WEU	Emissions|CO2|AFOLU	Mt CO2/yr	58.25322	60.129	...	...
4	MESSAGEix-GLOBIOM_1.0	NPi2020_400_SSP1	WEU	Emissions|CO2|Energy	Mt CO2/yr	3369.1	2522.043	...	...

Figure 1: Illustrative example of IAMC timeseries data format.

Data is reported for each model, region and scenario that is submitted. The columns included are Model, Scenario, Region, Variable, Unit, and any number of years. The names of the years need to be specified in the header, i.e. the first row of the, tabular format, but are flexible. The order of models,

[1] https://github.com/IAMconsortium

scenarios, regions and variables in the following rows is intended to be arbitrary, but it is essential to provide this information for each item on the sheet to allow automated data processing.

Variable definitions

In the process of putting together a new data template based on the IAMC format, compatibility with previously existing data templates is a major concern. This in particular includes using variable names that are identical with those variable names used before and consistency with the general approach of constructing new variable names. Two mechanisms are supposed to ensure standardization of variable names and backward compatibility of new templates:

1. Variable names and definitions of existing variables should be checked prior to defining new ones. New templates should build on this library of existing variable names and definitions and only construct new variable names in case these are not covered by the library of existing variable names.
2. A set of general rules that document that logic for constructing variable is published on the IAMC data template and database documentation[2]. These rules will only provide guidance for constructing new variable names, but do not define these in an unambiguous way.

The construction of variable names follows a few basic rules to ensure consistency in structure of new variable names with existing ones. These rules are published as part of the IAMC data template documentation[2] and are for convenience summarized below.

- Variables names should be kept as short as possible while at the same time being as self-explanatory as possible to minimize the possibility of misinterpretations without having to read the exact definitions.
- The first part of a variable name should define what the variable represents as opposed to being a structuring element (e.g., a sector) that combines several variables into a group of variables. In other words, the first part of the variable name should answer the question what type of indicator it is dealing with (e.g., an efficiency, energy use at a specific level, land cover of a certain type). There are exceptions to this rule, e.g., names of variables that are generated by post-processing tools (e.g., harmonization of variables to a historical inventory) so that these can be easily distinguished from variables natively reported by models.
- Different parts of variable names, including structural or hierarchical elements, are separated by a separator "|".

[2] https://data.ene.iiasa.ac.at/database/

- Semi-hierarchical structures in the variable names can be useful to add structure to variable names. On the other hand, too excessive use of such hierarchical structures should be avoided to keep variable names short. At the same time, there are often multiple hierarchies possible, i.e., multiple ways of structuring the variable tree exist. Therefore, in general the rule should be to use not more than two hierarchical levels in the construction of variable names (e.g., "Liquids|Petroleum Products" can be broken down into "Liquids|Diesel" and "Liquids|Gasoline" instead of "Liquids|Petroleum Products|Diesel" and "Liquids|Petroleum Products|Gasoline").
- In case different aggregates need to be constructed that cut across several categories of an existing hierarchy, the aggregates should be introduced at the same hierarchy level as the ones they are competing with (e.g., "Fossil" is the aggregate of "Coal", "Oil" and "Gas", "Solids" is the aggregate of "Biomass" and "Coal" and both would be introduced at the same level of the hierarchy).
- In particular, for aggregates of two groups of variables (e.g., sectors) a new aggregated group can be constructed by joining the original sector names with an "and" (e.g., the energy end-use sectors "Residential" and "Commercial" are often reported jointly as "Residential and Commercial", emissions for "Energy Supply" and "Energy Demand" can be aggregated into "Energy Supply and Demand"). These aggregates should also be placed at the same level of the hierarchy as the original groups to avoid proliferation of hierarchical levels.
- "Other" variables should be added for all variable categories to allow modelers to allocate variables that do not match any of the predefined categories. To document what is included in such "Other" variables, the comment sheet of the data template should be used.

3. Data format extension

While using an established data format instead of inventing a new one has the benefits of attracting a wider user community and benefitting from existing tools, it is important to ensure that all required use case are supported by that format. In openENTRANCE, this was done by testing to which degree modeling teams are able to report their in- and ouput data in the IAMC data format. In this process, two limitations of the IAMC format were identified. First, the original IAMC data format only covers annual data which limits its use for reporting information from energy models with high time resolution as it frequently occurs in electricity dispatch modeling or similar. Second, the format was thus far only used to report information at the region level, including aggregated flows into and out of a region, but not covering explicitly the source or destination region. This type of information typically occurs in trade models, but also energy system models and electricity dispatch models.

These two limitations therefore led to extensions of the IAMC data format to allow addressing the two important use case within the openENTRANCE project described above.

Temporal extension

In the process of extending the IAMC format to allow submission of subannual data, different variants were considered. An important feature identified in the consultation and testing process is allowing submission of data at different levels of temporal resolution (e.g., annual, monthly, weekly, hourly) in parallel. To achieve this, an optional 'Subannual' column after the 'Unit' column (Figure 2) was introduced. If this column is not provided or empty, it means that the spreadsheet contains yearly (average annual, total annual) values. If the template includes this column, the values in that column need to be specified when registering the model (similar to regions).

Examples for different sets of subannual timeslices:

- seasons: spring, summer, autumn, winter
- months: January, February, March, …
- more generic timeslices: summer-afternoon, winter-night
- hourly: h0000, h0001, h0002, …

	A	B	C	D	E	F	G	H	I	J
1	**Model**	**Scenario**	**Region**	**Variable**	**Unit**	**Subannual**	**2010**	**2015**	**2020**	**2025**
7	MESSAGEix-GLOBIOM_1.0	NPi2020_400_SSP1	WEU	Load	MW	h0001	351.231	333.669	…	…
8	MESSAGEix-GLOBIOM_1.0	NPi2020_400_SSP1	WEU	Load	MW	h0002	336.85	320.008	…	…
9	MESSAGEix-GLOBIOM_1.0	NPi2020_400_SSP1	WEU	Load	MW	h0003	342.976	325.827	…	…

Figure 2: Illustrative example of temporal extension of IAMC timeseries data format.

A limitation of implementing this extension using the different Excel data formats is that only 1,048,576 (xlsx) or 65,536 (xls) rows are supported which means that for hourly data (8760 subannual timeslices per year) only 119 and 7 model/scenario/region combinations, respectively, fit on a single workbook sheet. This limitation of Excel-based files can be overcome by either spreading data sets across multiple sheets of a single worksbook or implement the IAMC data format in csv format which does not have any such row limitations (cf. Section 4).

Spatial extension

In contrast to the temporal extension, the spatial extension required just an adjustment of the syntax used in the 'Region' column of the data format. Therefore, a new delimiting character '>' used for describing flows between different regions was introduced. The character, placed between two region identifiers 'region A' and 'region B' of a model indicates a flow from 'region A' to 'region B'. This extension is illustrated in the 'Region' column of the template in Figure 3.

	A	B	C	D	E	F	G	H	I
1	**Model**	**Scenario**	**Region**	**Variable**	**Unit**	**2010**	**2015**	**2020**	**2025**
5	MESSAGEix-GLOBIOM_1.0	NPi2020_400_SSP1	AFR>WEU	Trade|GHG Allowances	Mt CO2/yr	0	0	...	...
6	MESSAGEix-GLOBIOM_1.0	NPi2020_400_SSP1	AFR>WEU	Trade|Oil	EJ/yr	3.345	3.216	...	...

Figure 3: Illustrative example of spatial extension of IAMC timeseries data format.

4. Future extensions

As highlighted previously, within the energy modeling community multiple formats are used to report and exchange information. The so-called frictionless data specification[3] and associated software packages are used by some energy modeling projects. Frictionless data essentially combines a csv file for data storage with meta data specified in JSON. As a way of allowing easy access to data generated in openENTRANCE to others relying on frictionless data, the project will develop an implementation of the IAMC data format based on frictionless data, including a package to convert between the two formats.

Beyond the extension of the data format itself, the various case studies of openENTRANCE (WP6) will require additional variable definition that go beyond the set of variables defined previously. Therefore, as part of the case studies a dictionary of new variable definitions that are consistent with the variable naming convention of the existing IAMC data format will be developed and published.

Finally, a stakeholder dialogue with energy modeling groups beyond the openENTRANCE consortium has been initiated at the Energy Modeling Platform for Europe (EMP-E) 2019 conference[4] in Brussels to, among other things, elicit needs regarding data formats from the wider modeling community. This dialogue is planned to continue at the openmod conference 2020 in Berlin and EMP-E 2020 conference in Brussels to ensure that plans developed in openENTRANCE are consistent with needs of the broader modeling community.

[3] https://frictionlessdata.io/
[4] http://www.energymodellingplatform.eu/program-and-presentations.html