https://doi.org/10.1016/j.envsoft.2019.07.005

The METACLIP semantic provenance framework for climate products

J. Bedia[a,b,*], D. San-Martín[a], M. Iturbide[c], S. Herrera[b], R. Manzanas[c], J.M. Gutiérrez[c]

[a] *Predictia Intelligent Data Solutions SL. C/ Benidorm 8 Bajo. 39005 Santander, Spain*
[b] *Grupo de Meteorología. Dpto. de Matemática Aplicada y Ciencias de la Computación. Universidad de Cantabria. Avda. de los Castros, s/n. 39005. Santander. Spain*
[c] *Grupo de Meteorología. Insituto de Física de Cantabria (IFCA). CSIC - Universidad de Cantabria. Avda. de los Castros, s/n. 39005. Santander. Spain*

Abstract

Having an effective way of dealing with data provenance is a necessary condition to ensure reproducibility, helping to build trust and credibility in research outcomes and the data products delivered. METACLIP (METAdata for CLImate Products) is a language-independent framework envisaged to tackle the problem of climate product provenance description. The solution is based on semantics exploiting the web standard Resource Description Framework (RDF), building on domain-specific extensions of standard vocabularies (e.g., PROV-O) describing the different aspects involved in climate product generation. We illustrate METACLIP through an example application within the open source R computing environment, generating a climate product for which full provenance information is recorded. Finally, the METACLIP Interpreter, a web-based interactive front-end for metadata visualization is presented, helping a diversity of users with different levels of expertise to trace and understand the provenance of a wide variety of climate data products, and to fully reproduce them.

Keywords: Metadata, climate services, Resource Description Framework, Ontologies, Earth Science Information

*Corresponding author
Email address: bediaj@unican.es (J. Bedia)

https://doi.org/10.1016/j.envsoft.2019.07.005

1. Introduction

Provenance is defined as a *"record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing"* (Provenance Working Group, 2013). This information can be used to form assessments about data quality, reliability or trustworthiness. In the context of climate science, having an effective way of dealing with data provenance is a necessary condition to ensure the reproducibility of results. Keeping track of data provenance is expected to increase the value of both the final products and the original data as this would make evident and traceable the impact one has on the other. Furthermore, an effective and officially approved metadata schema is a must-have in order to allow for the certification of a specific data workflow. Therefore, an effective provenance tracking system becomes crucial in order to allow users to perform an in-depth scrutiny of the data generation, fostering a best scientific practice that favours open science, transparent peer-review and double-checking of both products and raw data.

In an era of rapid development of data services, and climate services in particular (see e.g., Hewitt et al., 2012), there is a compelling need by the different data users (and producers) for a comprehensive provenance description of the products generated (Hills et al., 2015), that has motivated the development of several transnational initiatives aimed at fostering international standards for data processing and sharing (e.g., the Research Data Alliance –RDA–, `https://rd-alliance.org`).

More specific to the climate science and related geoscientific disciplines are the Climate and Forecast Metadata Conventions (CF, Eaton et al., 2011, ; `http://cfconventions.org`), that are nowadays a widely adopted international standard for metadata encoding. The CF conventions provide a description of the physical meaning of data and of their spatial and temporal properties, although they are dependent on a specific data format (the Network Common Data Form –netCDF–, Rew and Davis, 1990), thus restricting the access and the scope of metadata information to a technical context (see how-

https://doi.org/10.1016/j.envsoft.2019.07.005

ever recent efforts to abstract the CF conventions from netCDF, e.g., Hassell et al., 2017). Recent initiatives towards the improvement of the traceability of geoscientific products have been developed in some EU-funded projects. For instance the QA4ECV Project (Scanlon et al., 2015) was aimed at recording the traceability chain of several essential climate variables from remote sensing products (for instance Leaf Area Index, Peng et al., 2017), providing for each dataset comprehensive details of the processing algorithm and the estimation of uncertainties. Besides, the CHARMe Project (Clifford et al., 2015) was focused on linking climate datasets with publications, user feedback and other items (collectively designated as "commentary metadata"), using a Resource Description Framework annotation approach (RDF, see Sec. 2). The resulting metadata accompany datasets, helping end-users to choose the products best suited to their particular research aims. Likewise, a RDF-based approach has been followed by the U.S. Global Change Research Program, with the development of the "Global Change Information System Ontology" (see `https://data.globalchange.gov`), in order to document provenance in global change research (Ma et al., 2014a). The provenance tracking system developed allows linking to research papers, datasets, models, analyses etc. related to key global change research findings, that in words of their authors *"improve the visibility into the assessment process, increase understanding and possibility of reproducibility . . ."* (Tilmes et al., 2013). Another relevant example is the DataONE Network, `https://www.dataone.org/`, a project providing access to Earth and environmental data, supporting enhanced search and discovery of environmental data through the coordinated use of an inter-operable metadata model, conceived with the expectation of broader community adoption.

As shown by these previous initiatives, Semantic Web technologies are gaining an increasing importance among data scientists. In this context, METACLIP is intended as a solution for identifying, extracting, linking and assembling the pieces of information needed to fully describe the provenance of a climate product, also providing a tool for effective visualization. Thus, METACLIP

https://doi.org/10.1016/j.envsoft.2019.07.005

somehow widens the scope of the different initiatives above enumerated, being focused both on the low-level scientific details of specific products or file formats (e.g., it benefits from the CF conventions for climate variable naming or cell method definitions, for instance), and at the same time provides a more general, user-oriented provenance information for example by linking datasets with their DOIs (Digital Object Identifier, `https://www.doi.org`), or with relevant technical documentation, as well as describing the agents involved in their production and distribution. Furthermore, it provides a full description of the code, thus enabling full reproducibility of the products.

The METACLIP approach was initially developed in the project QA4Seas (Manubens et al., 2017), framed in the Copernicus Climate Change Service (C3S, `https://climate.copernicus.eu`). Since the beginning, it proved as an effective way of dealing with seasonal forecast product provenance description (there are seasonal forecast examples in the METACLIP gallery at `http://www.metaclip.org`), responding to specific project's needs such as linking low-level processing steps with calibration/verification activities and with known community datasets and organizations. These needs are common to other projects and research contexts requiring a complete set of tools for provenance tracking, from product generation to visualization. As a result, METACLIP is currently being used in different initiatives broadening its initial scope, for example by providing semantic metadata for downscaling products in the VALUE initiative (Gutiérrez et al., 2018, `http://www.value-cost.eu/`).

METACLIP is based on RDF and focused on the semantic description of *climate products* (i.e., maps, plots or any other climate research outcome stored in a file), so that each product and its provenance information are inseparable and jointly delivered (although provenance information can be easily detached if needed). Its aim is to ease metadata discovery and understanding by a wide range of users, from experts requiring a complex technical provenance description to other users interested in a higher level representation. The metadata can be explored through a specific application (the METACLIP interpreter), that thanks to the granularity of the schema provides the level of detail best tailored

https://doi.org/10.1016/j.envsoft.2019.07.005

to the user needs. The resulting provenance information is not only human-readable, permitting an intuitive navigation through a semantic description of the product at hand, but also "machine-readable", allowing data mining and the deployment of search and discovery tools.

The main objective of this paper is to provide an overview of the METACLIP framework. It is illustrated through a worked example in the open source R language and environment for statistical computing (R Core Team, 2019), given its popularity among a growing number of members of the climate data user community. To this aim, we introduce the METACLIP extension for R (the `metaclipR` package), as a new component of the `climate4R` framework for climate data analysis (Iturbide et al., 2019). It must be noted that METACLIP is a semantic framework based on RDF, and as such it is not dependent on any particular language/environment. Its extension to R via the package `metaclipR` is used here for illustration, but it can be applied within any other computing environment(s). This paper also introduces the METACLIP Interpreter, the interactive web-based tool conceived as a user-oriented front-end to explore and visualize the provenance information attached to products.

This paper is structured as follows: First, in Sec. 2 a description of the METACLIP vocabularies is presented; these vocabularies have been designed as a domain-specific extension of widely used "domain-agnostic" data models (like PROV) to describe the different activities, agents and transformations involved in climate data product generation. Secondly, the `climate4R` framework is briefly presented (Sec. 3) to set the scene for the R package `metaclipR` (Sec. 3.2). Next, in Sec. 4 a worked example of METACLIP encoding using `metaclipR` is presented through a simple, real-world case study. Finally, the main components of the METACLIP Interpreter are described in Sec. 5. In addition, further worked examples covering other specific aspects of METACLIP (e.g., description of validation of climate products, bias correction etc.) are provided as supplementary information in a companion paper notebook, as indicated in Sec. 6.

https://doi.org/10.1016/j.envsoft.2019.07.005

2. METACLIP: A RDF-based approach for provenance description

RDF is a family of World Wide Web Consortium (W3C) specifications originally designed as a metadata model (W3C, 2004; Candan et al., 2001). It is an abstract model that has become a general method for conceptual description or modelling of information that is implemented in web resources, using a variety of syntax notations and data *serialization* formats. Serialization is understood in this context as the process of translating data structures or object state into a format that can be stored, i.e. a particular file format (see the link in the caption of Fig. 1 for an example). RDF extends the linking structure of the Web to use URIs[1] to name the relationship between things as well as the two ends of the link (this is usually referred to as a *triple*). As a result, an RDF representation consists of a collection of triples in which each triple is represented as a node-arc-node link (hence the term "graph"). An example is given in Fig. 1.

[1]URI is the acronym for Uniform Resource Identifier, a character string that unambiguously identifies a particular resource (URI Planning Interest Group, W3C/IETF, 2001)

https://doi.org/10.1016/j.envsoft.2019.07.005

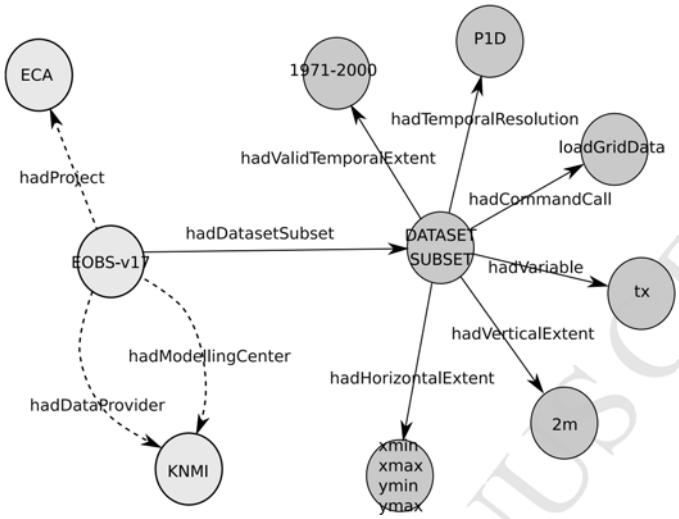

Figure 1: An RDF graph showing the Data source description (the EOBS-v17 dataset), and the first step within the data workflow (a dataset subset). Entities (nodes) are linked by properties (arrows) following the so called *triples* of the form Subject–Predicate–Object. Predicates are often referred to as *object properties* (labelled with blue fonts). Thus, this particular example of a RDF graph (see Sec. 2) is formed by a set of ten triples. For instance, a particular Dataset-class entity ("EOBS-v17") is associated with a specific activity (the "ECA" Project), and with two agents: i) A data provider (the agent distributing the data) and ii) A modelling center (the agent generating the data). These nodes are filled in grey and the arcs linking them represented by dashed lines. The grey part of the graph is represented in more detail in Fig. 2. The entity classes, their possible individual instances, and the relationships between them (object properties) are not free, but controlled by the rules given by the vocabularies (see Sec. 2.1). In METACLIP, the RDF representation is serialized into JSON-LD format. A JSON-LD representation of this graph can be viewed in the following link: `http://metaclip.org/fig1.json`

.

Designed to provide a framework that ensures interoperability between metadata frameworks, RDF allows for structured and semi-structured data to be mixed, exposed, and shared across different applications (RDF Working Group, 2014). Although the definition of the mechanism is domain-neutral, it is suitable for describing information about any specific domain (RDF Working Group, 2014) thus being a very extensible schema suitable for the design of specialized

https://doi.org/10.1016/j.envsoft.2019.07.005

metadata models. As a result, RDF has been widely adopted in many different fields and there are hundreds of vocabularies to describe many different things from geospatial features (Jiang et al., 2018) to music products (Raimond and Sandler, 2012) or geological map features (Ma et al., 2014b). To this aim, specific *vocabularies* (a 'near-synonym' for ontologies[2]) are written in RDF using the fully-featured Web Ontology Language (OWL, W3C OWL Working Group, 2012). As a result, ontologies contain a conceptual model of a particular (more or less broad) domain of knowledge, listing the types of objects, the relationships that connect them and constraints on the ways that these objects and relationships can be combined, being used for description, classification and reasoning.

In this context, METACLIP ontologies are designed in order to describe complex features and processes specific of the climate science domain (Fig. 3). Since the term ontology isn't well-known outside the semantic web community, we use here the more colloquial term *vocabulary* to refer to the METACLIP ontologies. METACLIP has a broad scope within the climate science, and is envisaged to adequately serve to the description of scientific outcomes in various related disciplines (short, medium and long range predictions, climate change projections, observational studies . . .). The METACLIP vocabularies are briefly described in the following section.

2.1. METACLIP vocabularies

The METACLIP framework is conceived as a domain-specific extension of the more general ("domain-agnostic") PROV Data Model (PROV-DM; Provenance Working Group, 2013), next briefly introduced. PROV-DM is the conceptual framework which defines the general types and relationships among features forming a basis for the W3C provenance (PROV) family of specifica-

[2]This thread may help the non-expert reader to clarify both terms and in which contexts ontology and vocabulary may be used interchangeably: `https://stackoverflow.com/questions/20200270/ontology-vs-vocabulary`

https://doi.org/10.1016/j.envsoft.2019.07.005

tions. PROV-DM is organized in six components, respectively dealing with: (1) entities and activities, and the time at which they were created, used, or ended; (2) derivations of entities from entities; (3) agents bearing responsibility for entities that were generated and activities that happened; (4) a notion of bundle, a mechanism to support provenance of provenance; (5) properties to link entities that refer to the same thing; and, (6) collections forming a logical structure for its members (Provenance Working Group, 2013). Even though PROV-DM is domain-agnostic, it is equipped with extensibility points that allow domain-specific information to be included. As a result, PROV-DM has currently a wide international adoption in many different fields.

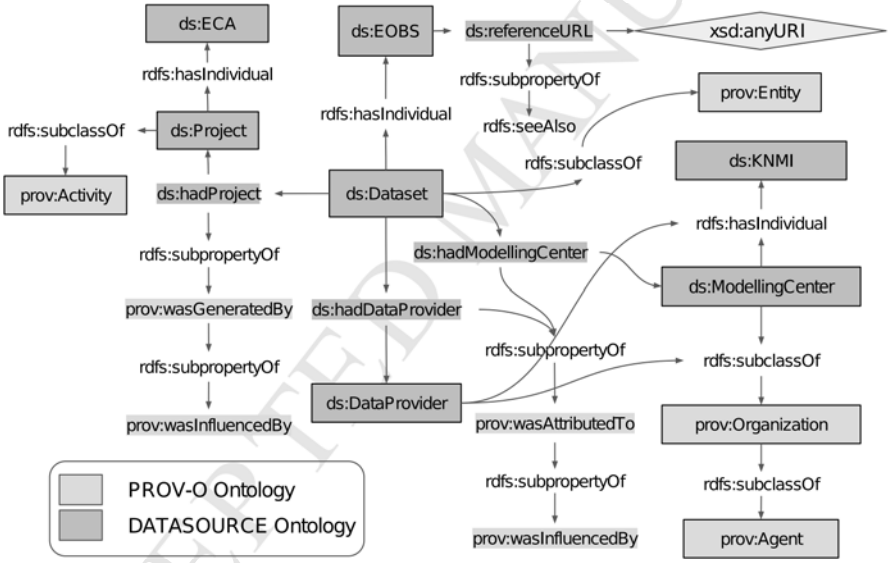

Figure 2: Schematic example showing the re-use of the PROV-O ontology by the METACLIP's datasource ontology. The graphical representation corresponds to the RDF graph nodes highlighted in Fig. 1. The two ontologies are differentiated by color, and their namespace prefixes are also indicated. See Table 1 for details about the *rdfs* and *xsd* schemas appearing in the figure.

The PROV Ontology (PROV-O; PROV Working Group, 2013; Moreau et al., 2015) allows the mapping of the PROV data model to RDF, providing a set of

https://doi.org/10.1016/j.envsoft.2019.07.005

Vocabulary name	prefix	Namespace URI	Reference
Dublin Core Metadata Element Set	dc	`http://purl.org/dc/elements/1.1/`	Dublincore.org (2012)
GeoSPARQL	geosparql	`http://www.opengis.net/ont/geosparql#`	Cox et al. (2002)
PROV Ontology (PROV-O)	prov	`http://www.w3.org/ns/prov#`	PROV Working Group (2013)
RDF Schema	rdfs	`http://www.w3.org/2000/01/rdf-schema#`	Brickley and Guha (2014)
Simple Knowledge Organization System (SKOS)	skos	`http://www.w3.org/2004/02/skos/core#`	Miles et al. (2009)
Dublin Core Metadata Initiative Metadata Terms	terms	`http://purl.org/dc/terms/`	Powell et al. (2005)
eXtensible Markup Language Schema	xsd	`http://www.w3.org/2001/XMLSchema#`	W3C XML Core Working Group (2001)

Table 1: Summary of pre-existing ontologies and schemas used by METACLIP.

classes, properties, and restrictions that can be used to represent and interchange provenance information generated in different systems and under different contexts. In order to achieve domain-specific extensibility of the PROV data model to the climate science context, METACLIP has developed specific vocabularies extending some basic PROV-O classes and properties, enriched with the necessary annotations to provide meaningful representations of the different steps involved in the generation of a specific data product. Furthermore, METACLIP extends other widely used data models apart from PROV-DM when relevant. For instance, spatially explicit features are described as an extension of the OGC GeoSPARQL standard, through its vocabulary (*geosparql*) for representing geospatial data in RDF (Cox et al., 2002). Also, different concepts involved in the quality assessment of forecasting systems have been introduced in the *verification* vocabulary as SKOS concepts (Miles et al., 2009). Another widely used metadata schema adopted by METACLIP is the Dublin Core (Powell et al., 2005), providing a number of elements to describe data sources (title, creator, description etc.). The use of these previously existing vocabularies reduce duplicity and promote interoperability, provided these are based on formal recommendations of working groups widely adopted by the Semantic Web community. A summary of these vocabulary imports are summarized in Table 1. An example of how the PROV-O is reused by the METACLIP ontologies is presented in Fig. 2.

The METACLIP vocabularies are under current development and evolution, as METACLIP is used in the context of new projects and initiatives, some of

https://doi.org/10.1016/j.envsoft.2019.07.005

them already indicated in Sec. 1. In these projects, domain experts discuss the need for defining new classes and properties and their naming during meetings held for this purpose, and their possible mapping to other pre-existing vocabularies is carefully analysed by the METACLIP core development team (authors of this article). Several technical aspects are considered in order to avoid errors compromising the validity of the vocabularies. One mechanism to ensure integrity is through *reasoning*, a process that can use the semantics of the different classes and properties to interpret the data and to infer new knowledge. Some errors in RDF only reveal themselves after reasoning (Hogan et al., 2010), and therefore the METACLIP vocabularies are tested using a reasoning engine to check for potential problems (HermiT OWL reasoner, Glimm et al., 2010, `http://www.hermit-reasoner.com/`). In addition, in creating the METACLIP vocabularies, we leveraged existing vocabularies (Table 1) by creating sub-classes/properties from their terms and avoided creating super-classes/properties of their terms, which would imply changes to the core semantics of those ontologies, a problem known as "ontology hijacking" (Hogan et al., 2009) that causes issues when combining provenance from multiple sources for reasoning.

As a result, METACLIP is currently composed of the four core vocabularies next described (the prefix of each ontology is indicated between parenthesis after its name). Both stable and development versions of the METACLIP vocabularies can be reached in their public GitHub repository (`https://github.com/metaclip/vocabularies`). A high-level representation of a typical climate data workflow, and the vocabularies involved at each stage of the climate product generation is presented in Fig. 3. The vocabularies were written using the open-source Protégé software (Musen, 2015).

- *datasource (ds:)* This vocabulary describes the origin of the input data (dataset description), and data transformations (subsetting, aggregation, anomalies, PCA, climate indices etc.). It also establishes the links between the different transformation commands and arguments in each step (source

https://doi.org/10.1016/j.envsoft.2019.07.005

code). The ontology version IRI[3] is `http://metaclip.org/datasource.owl`.

- *calibration (cal:)* The Calibration vocabulary encodes the metadata describing bias correction, downscaling and other forms of statistical adjustment (variance inflation, ensemble recalibration etc.). The development of this specific ontology is partially aligned with the conceptual framework designed in the COST Action VALUE (Maraun et al., 2015; Gutiérrez et al., 2018), providing a European Network for a comprehensive validation and development of statistical downscaling methods. The ontology URI is `http://metaclip.org/calibration.owl`.

- *verification (veri:)* The forecast verification vocabulary encodes the metadata related with the verification of seasonal forecast products, providing a description of the verification measures applied as well as a description of the verification aspect addressed by each measure. Furthermore, the vocabulary also provides a conceptual scheme for the definition of other more general forms of climate validation. The ontology URI is `http://metaclip.org/verification.owl`. There is also an ongoing initiative to develop a vocabulary for climate model validation, following the framework developed in the above-mentioned VALUE initiative (Gutiérrez et al., 2018).

- *graphical_output (go:)* This vocabulary is aimed at graphical product description (charts, maps), including a characterization of uncertainty types represented and how these are communicated. It has two main components: i) The graphical product description and ii) the description of the uncertainty types communicated by the different graphical elements of the product. The ontology URI is `http://metaclip.org/graphical_output.owl`.

[3]Internationalized Resource Identifier

https://doi.org/10.1016/j.envsoft.2019.07.005

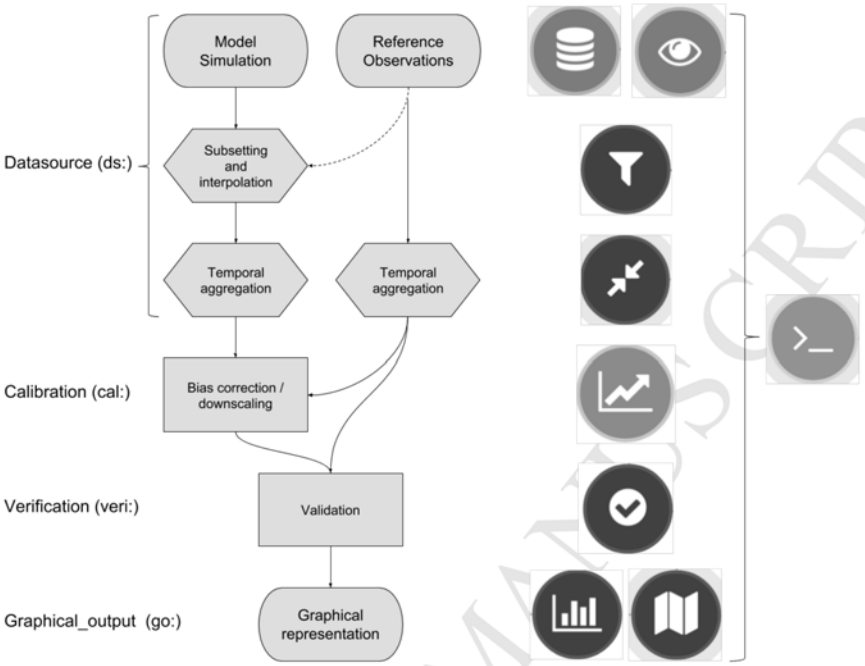

Figure 3: Schematic representation of a typical climate product generation workflow, from database description, subsetting and data transformation to final graphical product generation (a map, chart etc.). METACLIP specifically considers the different intermediate steps consisting of input data transformations, statistical adjustment/downscaling and model data validation. The different vocabularies describing each stage are indicated in the left (the vocabulary prefixes are indicated in parenthesis followed by ":"). In the right column, some icons used to visually represent each step in the METACLIP Interpreter (Sec. 5) are presented. Note that all steps include a detailed description of the command calls generating them (represented by the "> _" symbol).

In order to illustrate the scope of the vocabularies and the way the PROV data model has been extended, a brief explanation on how a primary climate data source is described by the *datasource* vocabulary is next given as example. This is in general the first step for provenance tracking (i.e., the first boxes of the high-level representation in Fig. 3). In METACLIP, the input data descrip-

https://doi.org/10.1016/j.envsoft.2019.07.005

tion is achieved by the class *Dataset* of the *datasource* vocabulary (it will be referred to by indicating the vocabulary prefix followed by the class name as in *ds:Dataset* hereafter), that extends *prov:Entity*, and splits into 6 different subclasses attending to the nature of the dataset (e.g., *ds:MultiDecadalSimulation*, *ds:Observations*, *ds:Reanalysis*, etc.). Further classes are linked to *ds:Dataset* via object properties that provide further provenance details such as the *ds:ModellingCenter* producing the data and the *ds:DataProvider* distributing the data, both defined as subclasses of *prov:Agent*), the *ds:Project* in which the data can be framed (e.g., CMIP5, CORDEX etc.) or experiments (e.g., evaluation experiments or historical/future emission or radiative forcing scenarios etc.), defined as subclasses of *prov:Activity*. Other details are also given that are specific for each subclass of *ds:Dataset* (e.g., the driving model *ds:GCM* for a given *ds:RCM* regional simulation –which extends *prov:SoftwareAgent*–, the URLs[4] serving as entry points for the data etc.). A schematic overview of the climate data source description by the *datasource* vocabulary is given in Fig. 4.

[4]URL is the Uniform Resource Locator, a type of URI that identifies a resource via a representation of its primary access mechanism (e.g., its network "location" URI Planning Interest Group, W3C/IETF, 2001)

https://doi.org/10.1016/j.envsoft.2019.07.005

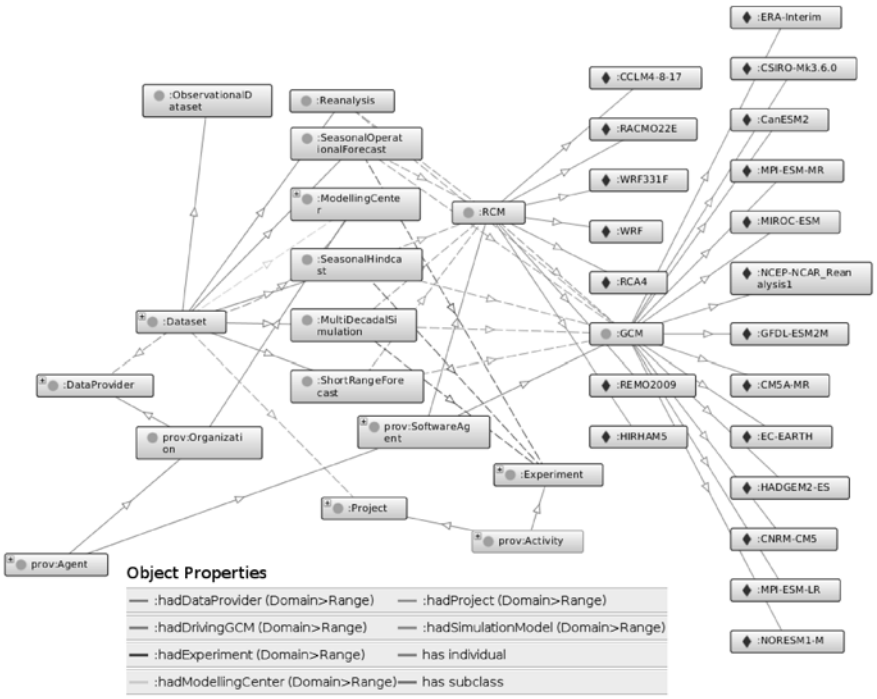

Figure 4: Schematic overview of the main classes (yellow circles), object properties (arrows) and individuals (purple diamonds) of the *datasource* ontology used to describe a climate dataset. To avoid a congested graph, only the individuals defined for *ds:GCM* and *ds:RCM* are shown (there are other individuals describing *:ModellingCenter*, *:Project* and *:DataProvider* classes). The PROV-O classes extended by the datasource ontology are indicated by the *:prov* prefix (the *ds:* prefix is omitted in the figure for conciseness).

After recording the provenance of the input data sources, the different transformations experienced by the data are encoded as instances of the *ds:Step* class, which is a specific METACLIP subclass of the more general PROV *prov:Derivation* class, defined as *"a transformation of an entity into another, an update of an entity resulting in a new one, or the construction of a new entity based on a pre-existing entity"* (PROV Working Group, 2013). *ds:Step* class itself is a general transformation with different specific subclasses such as *ds:DatasetSubset*, *ds:Aggregation*, *veri:ForecastVerification*, *cal:BiasAdjustment*,

https://doi.org/10.1016/j.envsoft.2019.07.005

go:GraphicalRepresentation and many others belonging to the different METACLIP vocabularies, aimed at the description of the different stages involved in the climate product generation (Fig. 3). Similarly, the construction of multi-model ensembles is described by *ds:Ensemble*, as a subclass of *prov:Collection*, and a specific class (*ds:CombinationMethod*) is aimed at the technical description of the method for ensemble member combination. A schematic overview of the main *ds:Step* classes is shown in Fig. 5.

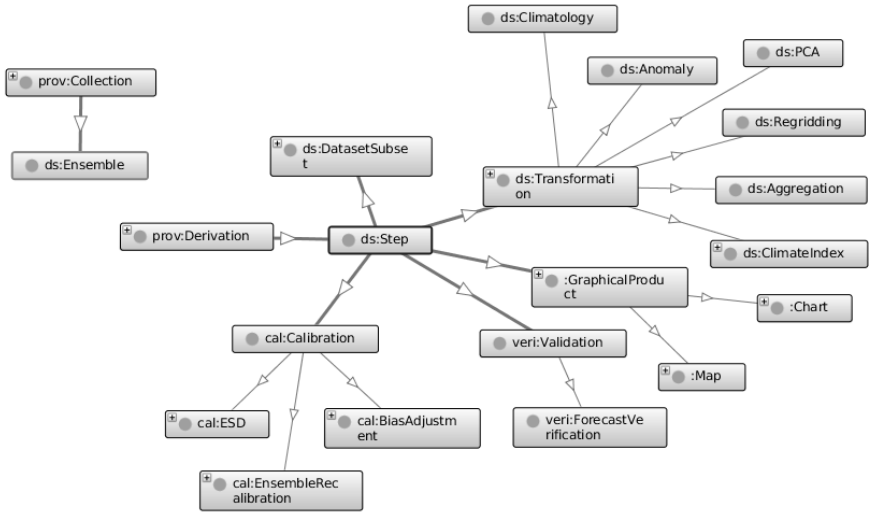

Figure 5: Schematic overview of the main classes (yellow circles, blue arrows represent subclasses) in the different METACLIP vocabularies used to describe climate data transformations (*ds:Transformation* superclass). The vocabularies defining each class are indicated by their prefixes, as indicated in Sec. 2.1 (see also Fig. 3).

Furthermore the software used and the command calls associated with each *ds:Step* can also be recorded. Specific software applications used to perform different climate data manipulation tasks (*ds:Package*) extend the *prov:SofwareAgent* class, while the specific commands invoked (*ds:Command*) extend the *prov:Activity* class. The recording of all the command calls triggering the different transformations ensures the full reproducibility of the final

https://doi.org/10.1016/j.envsoft.2019.07.005

outcome.

3. `metaclipR`: a *climate4R* extension for using METACLIP in R

In this article, the METACLIP concept is illustrated through an applied example using `climate4R` (Iturbide et al., 2019, see also `http://www.meteo.unican.es/climate4R`), a R-based framework for climate data access, postprocessing (including bias correction and downscaling) and visualization. `metaclipR` is a package implementing the METACLIP framework in R, specialized in (although not restricted to) the particular context of `climate4R`, in order to efficiently deal with the specificities of this group of packages, allowing for an easier abstraction of command calls and data structures to the entities defined in the METACLIP vocabularies. Both `climate4R` and `metaclipR` are briefly presented in this section.

3.1. The climate4R framework

`climate4R` builds on two main data structures (grid and station, including metadata) to deal with gridded and point data from observations, reanalysis, seasonal forecasts and climate projections. Data access is primarily obtained by package `loadeR`, enabling transparent, "user-friendly" access to (possibly harmonized) data from R through the NetCDF-Java API (Unidata, 2018), and a wide range of common transformation operations through package `transformeR`. `climate4R` considers ensemble members as a basic dimension of the data structures. Also, compatibility with some external packages has been achieved by either two-way bridging functions or wrapper packages, thus enhancing the `climate4R` core packages with extended functionalities addressing specific community needs or project developments such as the calculation of extreme climate indices (`climate4R.climdex`, Bedia, 2018), model validation (package `VALUE`, Gutiérrez et al., 2018), seasonal forecast visualization (`visualizeR`, Frías et al., 2018), species distribution modelling (`mopa`, Iturbide et al., 2018), fire danger applications (`fireDangeR`, Bedia et al., 2018) or the provision of remote access to harmonized seasonal forecast datasets (`loadeR.ECOMS`,

https://doi.org/10.1016/j.envsoft.2019.07.005

Cofiño et al., 2018), among others. Furthermore, the *climate4R Hub* is a cloud-based computing facility that allows to run `\code{climate4R}` on the cloud using docker and jupyter-notebook (`https://github.com/SantanderMetGroup/climate4R/tree/master/docker`).

3.2. Package `metaclipR`

`climate4R` attains full compatibility with the METACLIP framework with `metaclipR`, envisaged to keep track of the different operations undertaken during the data workflow and translate them into the RDF description provided by METACLIP. This is achieved through two main mechanisms:

i.) On the side of the vocabularies, by introducing in the METACLIP vocabularies individuals describing particular features of `climate4R` in detail (e.g., the UDG data access layer of `climate4R` is an individual instance of the *ds:DataProvider* class –subclass of *prov:Organization*–, `loadeR` is an individual instance of *ds:Package* –subclass of *prov:SoftwareAgent*–, etc.). Every time `metaclipR` is loaded, it will automatically perform a check to connect to the METACLIP vocabularies. In case this connection fails (either the remote server is temporarily unavailable or there is not internet connection), there will be a warning message. `metaclipR` can still be used, but there are several automated metadata recording steps that the package will not be able to undertake, leading to a less detailed, more generic metadata representation because the individual class instances (see e.g., Fig 4) won't be read from the vocabulary.

ii.) On the side of `metaclipR`, by including specific functions referring to known `climate4R` functions when these are used (e.g., the function `metaclipR.loadGridData` handles *ds:DatasetSubset* entities when computed with function `loadGridData` from package `loadeR`, as next shown in the example of Sec. 4).

We emphasize here that in spite of the specialization of `metaclipR` on `climate4R` packages, it can handle any other functions (even outside the R

https://doi.org/10.1016/j.envsoft.2019.07.005

environment), as the different entities and transformations can be recorded without assuming a particular function call to generate them. For instance, following with the example in the above paragraph *ii.)*, the generic function to encode a dataset subset is `metaclipR.DatasetSubset`, although optionally `metaclipR.loadGridData` can be used if the latter function is being used to generate the subset. As the mapping of specific command calls to METACLIP classes is highly sensitive to code modifications, METACLIP controls the `climate4R` package versions when these are used. For this reason, the package version is an input argument in all `metaclipR` functions, and internally controls the valid versions that apply for each `metaclipR` package.

Thus, in essence `metaclipR` generates metadata for the products generated by mapping the specific function calls or input arguments received onto the METACLIP vocabularies (Sec. 2.1), creating a RDF representation of the metadata (a RDF graph, see example in Fig. 1). The directed graph structure is internally constructed using the `igraph` package (Csardi and Nepusz, 2006). The sequence of the data workflow is achieved by recording, after each step the *terminal node* of the resulting RDF graph. After each call to the `metaclipR` functions, the RDF graph is updated (and therefore the terminal node, which is updated with the latest operation undertaken).

As data workflows are often non-linear, different graphs can be initiated and grown in parallel, until a certain operation binds them to create a unique, larger graph structure. For instance, one may want to describe different primary data sources that undergo subsetting and interpolation onto a common grid prior to ensemble generation. The last step entails a merge of several previous graphs for each ensemble member into one single collection (there are specific examples of this in the companion paper notebook, see Sec. 6).

https://doi.org/10.1016/j.envsoft.2019.07.005

4. Worked Example: Maps of summer day biases for a EURO-CORDEX sub-ensemble

The main functionalities of `metaclipR` are showcased in this section, describing the complete workflow to compute the climatological map of an ETCCDI core climate index (`http://etccdi.pacificclimate.org/list_27_indices.shtml`) relevant for the assessment of climate change impacts (Karl et al., 1999). In particular, in this example we consider summer days (SU, defined as the number of days with maximum temperature $> 25°C$) over a Mediterranean domain, using observational data the E-OBS interpolated grid (Haylock et al., 2008), in order to compute and visualize the resulting index climatology. This example is an extremely simple case study used to illustrate the steps involved in the generation of a climate product (a map) with the corresponding provenance information. More advanced examples providing a more comprehensive overview of the METACLIP functionalities are illustrated in the companion notebook of this paper (see Sec. 6) and in the gallery of the METACLIP interpreter (Sec. 5).

Throughout this section, the R code is interwoven within the text in order to show how `metaclipR` operates. All lines of code are identified by the R prompt symbol "`>`". Furthermore, in order to differentiate the calls to the `climate4R` functions performing the climate data operations from the `metaclipR` functions that record the corresponding provenance information, the latter code chunks are written in blue.

4.1. Working with the observations (E-OBS)

E-OBS is a daily gridded observational dataset of reference in Europe (Haylock et al., 2008). It is available through a public OPeNDAP server maintained by the KNMI (the Royal Dutch Meteorological Institute). In this example, the 0.25-degree regular grid of maximum temperature will be used. E-OBS is defined by the METACLIP's *datasource* ontology as an Individual instance of the *ds:ObservationalDataset* class. Hence, all the required metadata associated to

https://doi.org/10.1016/j.envsoft.2019.07.005

the individual will be automatically annotated by the function (every time that there is an open connection with the vocabulary, as explained in Sec. 3.2).

```
> library(metaclipR)
> showUDGDatasources()[7, ]
```

Similarly, KNMI is also a known institution, for which a specific individual exists:

```
> "KNMI" %in% knownClassIndividuals("ModellingCenter")
```

When the data comes from the UDG data provider (see `knownClassIndividuals("DataProvider")`), the URL pointing to the data is automatically recorded. However, in this case we are using an alternative data provider (the KNMI server), and its URL can be optionally included. This URL will be internally encoded in the provenance info as a data property (note that the URL string in the command below has been trimmed to allow readability in a single line; the full URL is indicated in the foot note[5])

```
> eobs.url <- "http://opendap.knmi.nl/knmi/.../tx_0.25deg_reg_v17.0.nc"
```

```
> metadata <- metaclipR.Dataset(Dataset.name = "E-OBS_v17_0.25regular",
                                DataProvider = "KNMI",
                                DataProvider.URL = eobs.url,
                                Dataset.subclass = "ObservationalDataset",
                                Project = "ECA",
                                ModellingCenter = "KNMI")
```

To give an idea of the operation just undertaken, the function has started an RDF graph with the dataset information, partially displayed in Fig. 1. There is much more information inside the graph encoded as *data properties* (relevant

[5] `http://opendap.knmi.nl/knmi/thredds/dodsC/e-obs_0.25regular/tx_0.25deg_reg_v17.0.nc`

https://doi.org/10.1016/j.envsoft.2019.07.005

URLs, class belonging and other annotations), but this requires the METACLIP Interpreter (described in Sec. 5) in order to conveniently display it in an interactive way. Also note that in this particular example, the *ds:DataProvider* and the *ds:ModellingCenter* correspond to the same individual (KNMI, Fig. 1). This is not necessarily so in other cases, and often the data produced by modelling centers is distributed by other providers (e.g., the User Data Gateway, an Earth System Grid Federation –ESGF– node, etc.). There is also an associated *ds:Project* generating the data (the European Climate Assessment –ECA–, Klein Tank et al., 2002).

4.1.1. Data subsetting

As shown in Iturbide et al. (2019), the function `loadGridData` from package `loadeR` was used to access the specific data slice used in the study. Note that `loadGridData` performs several steps in one single command call, depending on the different arguments used. Thus, it is possible to undertake dimensional subsetting + index.calculation + aggregation on-the-fly when using this function. For this reason, a specific `metaclipR` function was designed to account for this characteristic (see example in Sec. 3.2). This allows for a more accurate description of the different transformations experienced by the original data following the METACLIP schema when using the `climate4R` data loading functions.

In this example, the function `loadGridData` performs subsetting according to the arguments specified. Climate index calculation and temporal aggregation will be performed afterwards using other command calls, so in this case the original daily data of maximum temperature from E-OBS is retrieved without further aggregation. The data collocation parameters used for subsetting are indicated by the different specific arguments (`var`, `lonLim`, `latLim`, `season` and `years`, in this case). Note that package `loadeR` is first loaded:

```
> library(loadeR)
> lon <- c(-10, 20)
> lat <- c(35, 46)
> tasmax <- loadGridData(dataset = eobs.url,
```

https://doi.org/10.1016/j.envsoft.2019.07.005

```
                    var = "tx",
                    season = 1:12,
                    years = 1971:2000,
                    lonLim = lon,
                    latLim = lat)
```

Note that no arguments indicating temporal aggregation (`time`, `aggr.m`, etc.) are being used. This means that we are loading the data in its native temporal resolution, that in this case is daily. Next, a call to the corresponding `metaclipR` function is done in order to record the subsetting step just undertaken:

```
> metadata <- metaclipR.loadeR(package = "loadeR",
                               version = "1.4.0",
                               graph = metadata,
                               output = "tasmax",
                               fun = "loadGridData",
                               arg.list = list(dataset = eobs.url,
                                               var = "tx",
                                               season = 1:12,
                                               years = 1971:2000,
                                               lonLim = lon,
                                               latLim = lat))
```

4.1.2. Climate index calculation

The function `climdexGrid` is the workhorse for the calculation of all the ETCCDI core indices in the `climate4R.climdex` package (Bedia, 2018), building upon the original code of package `climdex.pcic` (Bronaugh, 2015). The specific index is indicated by the `index.code` argument. Additional specific arguments can be passed to this function (these are detailed in the help menu of `climdexGrid`). Here, we apply the default configuration of the SU index (Summer Days, i.e., the annual number of days recording a maximum temperature above 25°C).

https://doi.org/10.1016/j.envsoft.2019.07.005

```
> library(climate4R.climdex)
> SU <- climdexGrid(index.code = "SU", tx = tasmax)
```

A specific `metaclipR` function (`metaclipR.etccdi`) was designed for the characteristics of the ETCDDI indices (these are defined as individual instances of the *ds:ClimateIndex* class):

```
> metadata <- metaclipR.etccdi(graph = metadata,
                               output = "SU",
                               arg.list = list(index.code = "SU"))
```

Note that the temporal resolution (as well as other relevant metadata) is updated after climate index calculation. In this case, the original daily maximum temperature (daily resolution) has been aggregated to annual after the index calculation. This characteristic is taken into account by `metaclipR`, and the temporal resolution of the transformed data is automatically updated accordingly.

4.1.3. Climatology calculation

The climatological mean field is next calculated. In this case, the function `climatology` from package `transformeR` is used, By default, this function will apply the mean over the time dimension. However, for clarity, this option is explicitly indicated here in the argument `clim.fun`.

```
> library(transformeR)
> SU.clim <- climatology(SU, clim.fun = list(FUN = "mean", na.rm = TRUE))
```

And the corresponding metadata of this step is annotated. Note that by default, the function will assume that the `mean` cell method is being used (i.e., the climatological mean).

```
> metadata <- metaclipR.Climatology(graph = metadata,
                                    arg.list = list(clim.fun =
                                                    list(FUN = "mean",
                                                    na.rm = TRUE)))
```

https://doi.org/10.1016/j.envsoft.2019.07.005

4.1.4. Climatology map

The default behaviour of `spatialPlot` (from package `visualizeR`) will produce a basic map without text and with a basic (color-blind friendly) color palette. Thus, in order to reproduce the Fig. 2a in Iturbide et al. (2019), we add further customization options, indicating the same color palette and lines delimiting political boundaries. As a result, the following code generates Fig. 6:

```
> library(RColorBrewer)
> library(visualizeR)
> SU.colors <- colorRampPalette(colors = rev(brewer.pal(11, "RdYlBu")))
> main <- "Climatology of Summer Days (ETCCDI-SU) 1971-2000"
> spatialPlot(SU.clim, col.regions = SU.colors(61),
              at = seq(0,260,10),
              backdrop.theme = "countries",
              main = main)
```

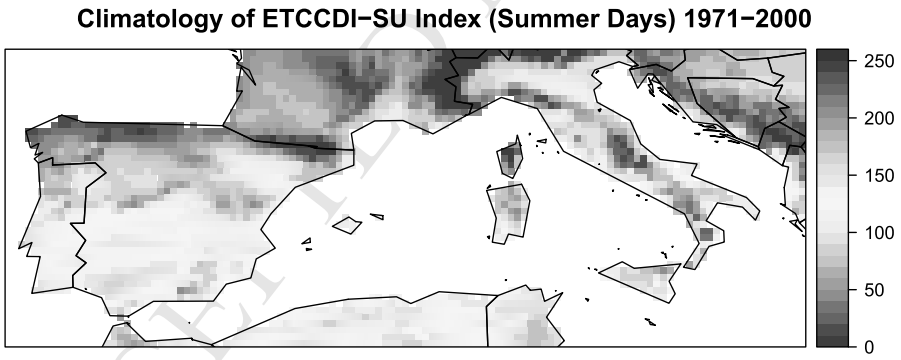

Figure 6: Mean number of summer days (SU index from ETCCDI) for the 30-year climatological period 1971-2000. The figure reproduces Fig. 2a from Iturbide et al. (2019).

The metadata is next updated with the step generating the graphical output. The function `metaclipR.SpatialPlot` has been specifically designed to describe the provenance of graphical products generated with this function:

https://doi.org/10.1016/j.envsoft.2019.07.005

```
> metadata <- metaclipR.SpatialPlot(graph = metadata,
                                    input.grid = SU.clim,
                                    arg.list = list(grid = SU.clim,
                                                    col.regions = SU.colors(61),
                                                    at = seq(0, 260, 10),
                                                    backdrop.theme = "countries",
                                                    main = main))
```

4.1.5. Figure file creation and metadata embedding

Finally, both the file containing the map and the embedded provenance information can be produced using the function `embedFig`, that undertakes all operations (metadata encoding of the graphical product + graphical product generation + metadata compression + metadata embedding) in a single function call:

```
> embedFig(plot.fun = "spatialPlot",
           arg.list = list(grid = SU.clim,
                           col.regions = SU.colors(61),
                           at = seq(0,260,10),
                           backdrop.theme = "countries",
                           main = main),
           full.metadata = metadata.EOBS$graph,
           format = "png",
           filename = "EOBS_SU_climatology.png",
           width = 950, height = 800, res = 150)
```

The final figure file with full provenance description is available in the following link: `http://metaclip.org/EOBS_SU_climatology.png`. The image is also available in the example gallery of the METACLIP Interpreter (see Sec. 5, `http://metaclip.org`), and included as Supplementary Material in this article.

https://doi.org/10.1016/j.envsoft.2019.07.005

5. Metadata exploration and visualization

Once the climate product has been produced and all the metadata conveniently recorded and embedded into the product, as illustrated in Sec. 4, it is ready for delivery. The provenance information embed into the figure can be now explored using the METACLIP Interpreter.

5.1. Interpreter overview

The METACLIP Interpreter has been designed as an interactive provenance visualization tool to navigate through complex data workflows, and obtain for each step a semantic description of the operations undertaken (as provided by the specialized vocabularies). This makes the METACLIP Interpreter a unique visualization tool that allows an easy interpretation of the provenance information to users with different levels of expertise. The visualization interface allows provenance description at different levels of granularity, in such a way that the most technical details (e.g., command calls, only relevant for expert users) remain hidden unless explicitly queried. At first sight, only a high-level description of the provenance information will be displayed, that will suffice in most occasions for end-users to have an overview on how the product was generated, without entering in unnecessary technical descriptions, often meaningless for non-experts. However, the user can query further details by clicking in the corresponding nodes or using a topic selector on the left panel, so an expanded level of detail is shown. As a result, technical details regarding downscaling, validation, command calls etc., can be easily obtained, including a full description of the source code if needed.

5.2. Technical aspects about the Interpreter implementation

As commented in Sec. 2, RDF representations must be serialized into a specific data format. In particular, `metaclipR` writes the final igraph-class object in JSON-LD format (function `graph2json`), although many other formats are possible. The interpreter is able to handle all of these RDF serialization formats to ease the usage of METACLIP, although the examples presented in

https://doi.org/10.1016/j.envsoft.2019.07.005

this paper and the companion material are coded in JSON-LD (see this link[6] for an example showing the JSON-LD representation of the RDF graph of Fig. 1).

The support for parsing the different serialization formats in JavaScript is quite limited and very heterogeneous. Therefore, the interpreter has been designed following a two-component architecture where there is (i) a back-end service to extract and parse the METACLIP provenance information and (ii) a front-end component that handles the interactive visualization. The back-end is based on Java and uses *Apache Jena* (`https://jena.apache.org`), a widely used library to deal with linked data. This library has an API that abstracts the user to the underlying format of linked data files. It supports many different format files, has a large community of users and is actively maintained. On the other side, the front-end is based on *d3js* (`https://d3js.org`), a JavaScript library to implement interactive visualization components.

In addition, the METACLIP Interpreter has the ability to extract and decompress the metadata information when this is embedded in the product (compression is applied in this case to minimize the file size overhead). This is the case in the examples provided in this paper. As a result, any graphical product file can be loaded into the METACLIP Interpreter (the option exists to use a drag-and-drop area to ease usability), and metadata visualization (including decompression) is automatically performed. The METACLIP Interpreter can be accessed through the following link: `http://metaclip.org`. A gallery of examples is available, including the example of Sec. 4.

6. Conclusions and future work

In the context of climate data information, METACLIP is conceived as a new solution for data provenance tracking, bringing together semantic web technologies, visualization and domain-specific knowledge in the field of climate science,

[6] `http://metaclip.org/fig1.json`

https://doi.org/10.1016/j.envsoft.2019.07.005

to provide a user-oriented, product-based solution to data provenance. It must be noted that METACLIP is not a closed solution, but a recent development likely to experience changes in the mid-term as a result of new requirements or user needs arisen in future projects and initiatives, some of them currently ongoing. Moreover, it is possible that METACLIP will adopt new or pre-existing ontologies to "better" describe certain features. The extensibility and reusability of the RDF schema ensures its maximum flexibility to adapt to the specific requirements of each situation.

The main METACLIP features are illustrated in this work through a simple example in which a climate product (an image file containing a climatological map of a specific climate index) is generated in parallel with its provenance information. A web-based front-end (the METACLIP Interpreter) completes the METACLIP framework, facilitating the navigation through the provenance information to users with different levels of expertise. Therefore, end-users ultimately receive a product and a metadata exploration tool readily available. Nevertheless, this approach does not preclude the separation of the raw provenance information (JSON-LD files) from the products in order to develop independent product databases oriented to content management systems or interactive search-and-discovery tools, exposing the provenance information in a way that linked open data services could read/interpret, thus allowing for advanced provenance analytics.

All the steps followed to generate the climate index map presented in this paper (with extended details and additional information), are available in the companion Paper Notebook:

- pdf file: `https://github.com/SantanderMetGroup/notebooks/blob/v0.1.0/2018_metaclip_EMS.pdf`
- source code (R markdown): `https://github.com/SantanderMetGroup/notebooks/blob/v0.1.0/2018_metaclip_EMS.Rmd`

Additional examples of anomaly calculation, bias correction and future cli-

https://doi.org/10.1016/j.envsoft.2019.07.005

mate index projections are also provided. The R software and all the packages required to reproduce the results are freely available as indicated in the paper notebook, where more specific details for installation are given.

Software availability

Name of the software:

- `metaclipR` (paper version: 1.1.0): `https://github.com/metaclip/metaclipR/releases/tag/v1.1.0`.
- METACLIP Web Interpreter: (paper version 1.0): `https://github.com/metaclip/interpreter/releases/tag/v1.0`

The ontology version IRIs used in this paper are next indicated. Note that these versions are already archived, so no further changes are done to the linked files. The current stable vocabulary version IRIs are indicated in Sec. 2.1.

- datasource (paper version 0.11): `https://github.com/metaclip/vocabularies/blob/master/archive/datasource/0.11/datasource.owl`
- graphical_output (version 0.1): `https://github.com/metaclip/vocabularies/blob/master/archive/graphical_output/0.1/graphical_output.owl`

Year First Available: 2017
Developers: Bedia, J., San-Martín, D.
E-mail: bediaj@predictia.es, daniel@predictia.es
Website: `https://github.com/metaclip`
Hardware Requirement: General-purpose computer
Programming Languages: R, Java, Javascript
Software Requirements: R version 3.1.0 or later. A generic web browser and an internet connection to use the METACLIP Interpreter.

https://doi.org/10.1016/j.envsoft.2019.07.005

Licensing

This software is made freely available under the terms and conditions of the GNU General Public License Version 3.

Data availability and Supplementary Materials

All the code (and data) used to produce this paper is available in a companion paper notebook (see Sec. 6). The full code of the different METACLIP components (including the vocabularies) are available in the pubic GitHub repo `https://github.com/metaclip`.

Figure 6 with full provenance information embedded is included as supplementary material.

Acknowledgements

We are grateful to the three anonymous reviewers and to the Editor for their insightful comments, helping to improve the original manuscript. Our colleague Prof. Antonio Cofiño (University of Cantabria) has helped with valuable comments to the development of the METACLIP framework. Dr. Jonas Bhend (MeteoSwiss) has contributed to the design of the vocabulary for seasonal forecast verification. This work has been partially funded by the European Union's Earth Observation Programme COPERNICUS, through the ITT C3S-51-Lot3 (QA4Seas Project, "Quality Assurance for Seasonal Forecast Products"), the Spain's Research and Innovation Programme under project INSIGNIA (co-funded by MINECO and FEDER, grant no. CGL2016-79210-R) and the European Union Research and Innovation Programme under projects IS-ENES2 (FP7 grant agreement no. 312979) and IS-ENES3 (H2020 grant agreement no. 824084).

https://doi.org/10.1016/j.envsoft.2019.07.005

References

Bedia, J., 2018. climate4R.climdex: Climate Change Index calculation for climate4R data. R package version 0.1.3.
URL http://meteo.unican.es/climate4R

Bedia, J., Golding, N., Casanueva, A., Iturbide, M., Buontempo, C., Gutiérrez, J. M., 2018. Seasonal predictions of Fire Weather Index: Paving the way for their operational applicability in Mediterranean Europe. Climate Services.

Brickley, D., Guha, R. E., 2014. RDF Schema 1.1. W3C Recommendation, World Wide Web Consortium.
URL https://www.w3.org/TR/rdf-schema/

Bronaugh, D., 2015. climdex.pcic: PCIC Implementation of Climdex Routines. R package version 1.1-6.
URL https://CRAN.R-project.org/package=climdex.pcic

Candan, K. S., Liu, H., Suvarna, R., 2001. Resource description framework: metadata and its applications. ACM SIGKDD Explorations Newsletter 3 (1), 6–19.

Clifford, D., Alegre, R., Bennett, V., Blower, J., Deluca, C., Kershaw, P., Lynnes, C., Mattmann, C., Phipps, R., Rozum, I., 2015. Capturing and Sharing Our Collective Expertise on Climate Data: The CHARMe Project. Bulletin of the American Meteorological Society 97 (4), 531–539.

Cofiño, A., Bedia, J., Iturbide, M., Vega, M., Herrera, S., Fernández, J., Frías, M., Manzanas, R., Gutiérrez, J., 2018. The ECOMS User Data Gateway: Towards seasonal forecast data provision and research reproducibility in the era of Climate Services. Climate Services.

Cox, S., Cuthbert, A., Daisey, P., Davidson, J., Johnson, S., Keighan, E., Lake, R., Mabrouk, M., Margoulies, S., Martell, R., 2002. Opengis geography markup language (gml) implementation specification, version. OGC

https://doi.org/10.1016/j.envsoft.2019.07.005

Standard. Citeseer.
URL http://www.opengis.net/doc/IS/geosparql/1.0

Csardi, G., Nepusz, T., 2006. The igraph software package for complex network research. InterJournal Complex Systems, 1695.
URL http://igraph.org

Dublincore.org, Jun. 2012. Dublin Core Metadata Element Set, Version 1.1: Reference Description. DCMI Recommendation, W3C.
URL http://dublincore.org/documents/dces

Eaton, B., Gregory, J., Drach, B., Taylor, K., Hankin, S., Caron, J., Signell, R., Bentley, P., Rappa, G., Höck, H., Pamment, A., Juckes, M., 2011. NetCDF Climate and Forecast (CF) Metadata Conventions V1.6. Last access: 23 May 2018.
URL http://cfconventions.org/

Frías, M. D., Iturbide, M., Manzanas, R., Bedia, J., Fernández, J., Herrera, S., Cofiño, A. S., Gutiérrez, J. M., Jan. 2018. An R package to visualize and communicate uncertainty in seasonal climate prediction. Environmental Modelling & Software 99, 101–110.

Glimm, B., Horrocks, I., Motik, B., Stoilos, G., 2010. Optimising Ontology Classification. In: Patel-Schneider, P. F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J. Z., Horrocks, I., Glimm, B. (Eds.), The Semantic Web – ISWC 2010. Vol. 6496. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 225–240.
URL http://link.springer.com/10.1007/978-3-642-17746-0_15

Gutiérrez, J. M., Maraun, D., Widmann, M., Huth, R., Hertig, E., Benestad, R., Roessler, O., Wibig, J., Wilcke, R., Kotlarski, S., San Martín, D., Herrera, S., Bedia, J., Casanueva, A., Manzanas, R., Iturbide, M., Vrac, M., Dubrovsky, M., Ribalaygua, J., Pórtoles, J., Räty, O., Räisänen, J., Hingray, B., Raynaud, D., Casado, M. J., Ramos, P., Zerenner, T., Turco, M., Bosshard, T., Štěpánek, P., Bartholy, J., Pongracz, R., Keller, D. E., Fischer, A. M.,

https://doi.org/10.1016/j.envsoft.2019.07.005

Cardoso, R. M., Soares, P. M. M., Czernecki, B., Pagé, C., Mar. 2018. An intercomparison of a large ensemble of statistical downscaling methods over Europe: Results from the VALUE perfect predictor cross-validation experiment. International Journal of Climatology.
URL http://doi.wiley.com/10.1002/joc.5462

Hassell, D., Gregory, J., Blower, J., Lawrence, B. N., Taylor, K. E., 2017. A data model of the climate and forecast metadata conventions (cf-1.6) with a software implementation (cf-python v2.1). Geoscientific Model Development 10 (12), 4619–4646.

Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P. D., New, M., Oct. 2008. A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. Journal of Geophysical Research 113 (D20).

Hewitt, C., Mason, S., Walland, D., 2012. The global framework for climate services. Nature Climate Change 2 (12), 831–832.

Hills, D. J., Downs, R. R., Duerr, R., Goldstein, J. C., Parsons, M. A., Ramapriyan, H. K., 2015. The importance of data set provenance for science. Eos 96.

Hogan, A., Harth, A., Passant, A., Decker, S., Polleres, A., 2010. Weaving the Pedantic Web. In: CEUR Workshop Proceedings. Vol. 628. Raleigh, NC, USA, p. 10.
URL http://ceur-ws.org/Vol-628/ldow2010_paper04.pdf

Hogan, A., Harth, A., Polleres, A., 2009. Scalable Authoritative OWL Reasoning For the Web. nternational Journal on Semantic Web and Information Systems 5 (2), 49–90.

Iturbide, M., Bedia, J., Gutiérrez, J. M., 2018. Tackling Uncertainties of Species Distribution Model Projections with Package mopa. The R Journal 10, 18.

https://doi.org/10.1016/j.envsoft.2019.07.005

URL https://journal.r-project.org/archive/2018/RJ-2018-019/index.html

Iturbide, M., Bedia, J., Herrera, S., Baño-Medina, J., Fernández, J., Frías, M., Manzanas, R., San-Martín, D., Cimadevilla, E., Cofiño, A., Gutiérrez, J., 2019. The R-based climate4r open framework for reproducible climate data access and post-processing. Environmental Modelling & Software 111, 42–54.

Jiang, L., Yue, P., Kuhn, W., Zhang, C., Yu, C., Guo, X., 2018. Advancing interoperability of geospatial data provenance on the web: Gap analysis and strategies. Computers & Geosciences 117, 21–31.

Karl, T. R., Nicholls, N., Ghazi, A., 1999. CLIVAR/GCOS/WMO Workshop On Indices And Indicators For Climate Extremes. Workshop Summary. Climatic Change 42, 3–7.
URL http://danida.vnu.edu.vn/cpis/files/Papers_on_CC/CC/CLIVAR-GCOS-WMO%20Workshop%20On%20Indices%20And%20Indicators%20For%20Climate%20Extremes%20-%20Workshop%20Summary.pdf

Klein Tank, A. M. G., Wijngaard, J. B., Können, G. P., Böhm, R., Demarée, G., Gocheva, A., Mileta, M., Pashiardis, S., Hejkrlik, L., Kern-Hansen, C., Heino, R., Bessemoulin, P., Müller-Westermeier, G., Tzanakou, M., Szalai, S., Pálsdóttir, T., Fitzgerald, D., Rubin, S., Capaldo, M., Maugeri, M., Leitass, A., Bukantis, A., Aberfeld, R., van Engelen, A. F. V., Forland, E., Mietus, M., Coelho, F., Mares, C., Razuvaev, V., Nieplova, E., Cegnar, T., Antonio López, J., Dahlström, B., Moberg, A., Kirchhofer, W., Ceylan, A., Pachaliuk, O., Alexander, L. V., Petrovic, P., Oct. 2002. Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. International Journal of Climatology 22 (12), 1441–1453.

Ma, X., Fox, P., Tilmes, C., Jacobs, K., Waple, A., 2014a. Capturing provenance of global change information. Nature Climate Change 4 (6), 409–413.

Ma, X., Zheng, J. G., Goldstein, J. C., Zednik, S., Fu, L., Duggan, B., Aulenbach, S. M., West, P., Tilmes, C., Fox, P., 2014b. Ontology engineering in

https://doi.org/10.1016/j.envsoft.2019.07.005

provenance enablement for the National Climate Assessment. Environmental Modelling & Software 61, 191–205.

Manubens, N., Hunter, A., Bedia, J., Bretonnière, P. A., Bhend, J., Doblas-Reyes, F. J., Dec. 2017. Evaluation and Quality Control for the Copernicus Seasonal Forecast Systems. AGU Fall Meeting Abstracts.
URL http://adsabs.harvard.edu/abs/2017AGUFMIN33D..05M

Maraun, D., Widmann, M., Gutiérrez, J. M., Kotlarski, S., Chandler, R. E., Hertig, E., Wibig, J., Huth, R., Wilcke, R. A., 2015. VALUE: A framework to validate downscaling approaches for climate change studies. Earth's Future 3 (1), 2014EF000259.

Miles, A., Bechhofer, S., (Eds.), 2009. SKOS Simple Knowledge Organization System Reference. W3C Recommendation, World Wide Web Consortium.
URL https://www.w3.org/TR/skos-reference

Moreau, L., Groth, P., Cheney, J., Lebo, T., Miles, S., Dec. 2015. The rationale of PROV. Web Semantics: Science, Services and Agents on the World Wide Web 35, 235–257.

Musen, M. A., Jun. 2015. The protégé project: a look back and a look forward. AI Matters 1 (4), 4–12.
URL http://dl.acm.org/citation.cfm?doid=2757001.2757003

Peng, J., Blessing, S., Giering, R., Müller, B., Gobron, N., Nightingale, J., Boersma, F., Muller, J.-P., 2017. Quality-assured long-term satellite-based leaf area index product. Global Change Biology 23 (12), 5027–5028.

Powell, A., Nilsson, M., Naeve, A., Johnston, P., 2005. Dublin core metadata initiative - abstract model. White Paper.
URL http://dublincore.org/documents/abstract-model

PROV Working Group, Apr. 2013. PROV-O: The PROV Ontology. W3C Recommendation, W3C.
URL https://www.w3.org/TR/2013/REC-prov-o-20130430/

https://doi.org/10.1016/j.envsoft.2019.07.005

Provenance Working Group, Apr. 2013. PROV-DM: The PROV Data Model. W3C Recommendation, W3C.
URL https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-provenance

R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL https://www.R-project.org/

Raimond, Y., Sandler, M., 2012. Evaluation of the Music Ontology Framework. In: Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Simperl, E., Cimiano, P., Polleres, A., Corcho, O., Presutti, V. (Eds.), The Semantic Web: Research and Applications. Vol. 7295. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 255–269.
URL http://link.springer.com/10.1007/978-3-642-30284-8_24

RDF Working Group, Feb. 2014. RDF - Semantic Web Standards.
URL https://www.w3.org/RDF/

Rew, R., Davis, G., 1990. NetCDF: an interface for scientific data access. IEEE Computer Graphics and Applications 10 (4), 76–82.

Scanlon, T., Nightingale, J., Muller, J.-P., Boersma, F., De, A., Lambert, J.-C., 2015. QA4ECV: A robust quality assurance service for terrestrial and atmospheric ECVs and ECV precursors. In: Proceedings of the RSPSoc and CEOI Joint Conference. The University of Southhampton, UK, pp. 1–4.

Tilmes, C., Fox, P., Ma, X., McGuinness, D. L., Privette, A. P., Smith, A., Waple, A., Zednik, S., Zheng, J. G., 2013. Provenance Representation for the National Climate Assessment in the Global Change Information System. IEEE Transactions on Geoscience and Remote Sensing 51 (11), 5160–5168.
URL http://ieeexplore.ieee.org/document/6558476/

https://doi.org/10.1016/j.envsoft.2019.07.005

Unidata, 2018. Network Common Data Format (NetCDF). [software].
URL https://doi.org/10.5065/D6H70CW6

URI Planning Interest Group, W3C/IETF, 2001. URIs, URLs, and URNs: Clarifications and Recommendations 1.0. Report from the joint W3c/IETF URI Planning Interest Group, World Wide Web Consortium.
URL https://www.w3.org/TR/uri-clarification/

W3C, 2004. Resource Description Framework (RDF): Concepts and Abstract Syntax.
URL https://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

W3C OWL Working Group, 2012. OWL2 Web Ontology Language Document Overview (Second Edition). W3C Recommendation, World Wide Web Consortium.
URL https://www.w3.org/TR/owl2-overview

W3C XML Core Working Group, May 2001. W3C XML Schema. W3C Recommendation, W3C.
URL https://www.w3.org/XML/Schema

https://doi.org/10.1016/j.envsoft.2019.07.005

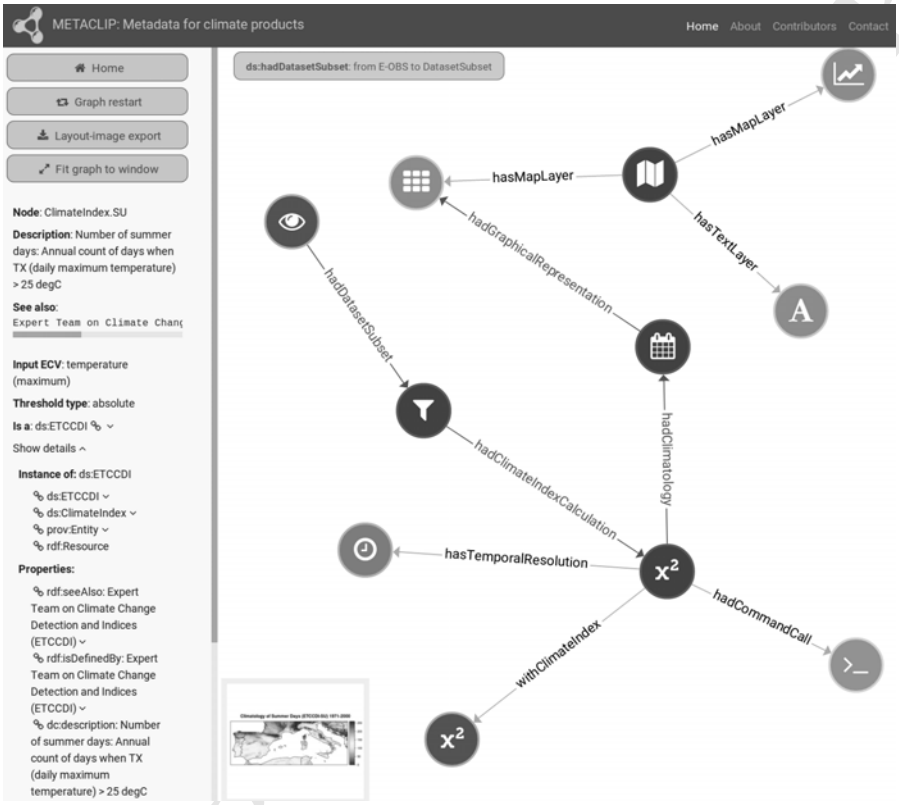

Figure 7: A screenshot of the METACLIP Interpreter (`http://metaclip.org/interpreter`), displaying the provenance representation of Fig. 6. Here, the metadata of the Climate Index node (the grey node labelled "X^2") is being displayed on the left panel). The information of the different nodes can be interactively queried by the user. Double-clicking each node will expand it to further nodes displaying other sub-properties and their corresponding annotations, until the lowest representation level is reached.

https://doi.org/10.1016/j.envsoft.2019.07.005

Highlights

- We introduce METACLIP, a new user-oriented provenance framework exploiting semantic web technologies for describing climate products
- It allows for the exploration of provenance representations of climate products using an interactive, web-based front-end.
- METACLIP is based on RDF, and develops a set of vocabularies designed as domain-specific extensions of PROV-O and other international standards.