An Extensible Conceptual Model for Tabular Scientific Datasets

Javad Chamanara*, Michael Owonibi*, Alsayed Algergawy*, and Roman Gerlach†

Friedrich Schiller University of Jena
*Institute for Computer Science
†Institute for Geography
Jena, Germany
Email: firstname.lastname@uni-jena.de

***Abstract*—There is a proliferation of datasets generated by various scientists of different scientific disciplines. Therefore, there is a growing need to construct and develop platforms that enable scientists to capture, exchange, process, and interpret data for immediate use, as well as to store and manage data to support future reuse. Modeling and organizing data within such platforms are key challenges. To this end, in this paper, we introduce the dataset model of the *BExIS 2* platform and how data can be organized inside the model. In particular, we describe the anatomy of a general purpose tabular dataset, which consists of data tuples to represent the table rows and data cells that are compound objects holding the obtained values and their auxiliary information. The structure of datasets is defined and applied separately in order to factor out shared concepts such as unit of measurement, methodology, data type, valid and missing values, processing functions and so on. The datasets are extensible in multiple ways and can be annotated on various levels utilizing taxonomies, ontologies, and custom metadata structures.**

Keywords–Scientific data; Dataset structure; Biodiversity data.

I. INTRODUCTION

In research data management, one of the utmost goals is to support data sharing, as this facilitates the reproduction and evaluation of scientific results as well as the reuse of the data for other purposes. Traditionally, researchers focused on collecting, processing and analyzing data and then published their findings in the scientific literature. Preparing and publishing research data was not part of the general scientific workflow. This has been changing. Publishing data is becoming a standard in most disciplines thanks to the advent of dedicated data repositories (e.g., Dryad [1], Pangaea [2]), data journals (e.g., Natures Scientific Data [3], Earth Science Data Journal [4], Biodiversity Data Journal [5]) and funding organizations requesting data publication. With such publications data becomes persistently available, documented, citable, and to some extent validated [6].

Many of these data repositories follow a rather generic approach to data management and accept a broad range of data models, data formats, and data types. They provide facilities to store data as files, together with a description of the content, structure, and administrative information in metadata documents. For some repositories (e.g., Pangaea) data submission is a curated process, which improves data quality in terms of consistency, completeness, and reusability. But the primary focus is still to make data discoverable by humans and allow them to download data files. When thinking about reuse in the sense of (automatic) data integration additional requirements need to be satisfied, e.g., flexible access patterns (selection and projection), data change/update/provenance management, integrated analysis, human and machine interpretability, flexible security and access management, data context provisioning, and semantic enablement. For instance, it will be really difficult to automatically integrate datasets which have not been parsed in the first instance (the dump table files), and in instances where they are dynamically parsed, the question of determining equivalent variables in different datasets and unit conversion comes into play.

A study conducted by Rexer [7] has shown that more than 90% of datasets contain less than 100 million records and are mostly managed/ processed by tools such as RDBMSs, Excel, or R. Another study done recently by O'Reilly indicated tabular datasets are among the most used forms of data [8]. This is due to the popularity in usage of spreadsheets for handling (storing and analyzing) data by the data providers, which is in turn due to the fact that spreadsheets are relatively easy to use, flexible, and compatible with a lot of applications across several disciplines. Also many of data acquisition tools simply generate raw data in tabular form, mostly comma separated flat files.

In this paper we focus on the domain of biodiversity, where spreadsheets, relational databases and statistical tools like R are widely used for managing data [9]. Biodiversity data is highly heterogeneous, including information about species distribution and abundance, genetic sequences, trait measurements, organisms, their morphology and genetics, life history and habitats, and geographical ranges. These data is mostly linked to spatial, temporal, and environmental data [10][11][12]. These heterogeneities can be broadly classified into five categories: technical, syntactic, structural, semantic, and data models [13]. Data model heterogeneity is the problem that systems and tools employ different data models, such as relational, XML, or semantic-based data models. A recent study shows that most existing biodiversity repositories are based on relational database models [12]. In contrast, structural heterogeneity focuses on the problem that information can be represented in multiple ways for a given data model.

Therefore, in order to effectively manage tabular data in a data repository, there is a need to model the composition of tabular datasets such that it satisfies the manifold data management needs outlined above. The current paper is an effort to extend the conceptual model presented in [14] and provide more details on the concept of a generic dataset. Although the model was developed for this particular domain we expect it to be applicable to others as well.

Copyright (c) IARIA, 2015. ISBN: 978-1-61208-415-2

The rest of the paper is organized as follows: a brief survey of related work is presented in the following section. We introduce our proposed data model in Section III and then elaborate its flexibility and extensions in Section IV. Finally, we conclude the paper and outline future work in Section V.

II. Related work

Recently, the World Wide Web Consortium (W3C) attempted to standardize the description of tabular data [15], such that the tabular data is structured into rows, each of which contains information about some thing. Each row contains the same number of cells providing values of properties of the thing described by the row. The W3C initiative broadly classifies tabular data into three main models: a *simple table*; consisting of columns, rows and cells with no form of annotation, an *annotation table*, i.e., a table annotated with additional metadata, and a *group of tables* comprising of a set of tables and a set of annotations that relate to the dataset.

Similarly, the INSPIRE Observation and Measurements standards (O&M) aims to normalize the representation of records of scientific measurement [16]. It introduces the notion of observation as an event whose result is an estimation of the value of some property(ies) of a feature-of-interest, obtained using a specified procedure. O&M defines a core set of properties for an observation, and these include the feature of interest, observed property, result value, procedure (the instrument, algorithm or process used), event specific parameters (e.g. instrument setting), phenomenon time, etc. One physical realization of this model is the tabular data. Users of this standard are not only able to describe features and properties but also to organize and store data. While the O&M standard was developed in the context of geographic information systems, the model is not limited to spatial information.

The Statistical Data and Metadata Exchange (SDMX) initiative [17] also sets standards that can describe and facilitate the exchange of statistical data and metadata. Based on the standard, every dataset will have a data structure definition, which specifies the organization of a data set. In addition, each column in the table can either be a function as a dimension, a measure, or an attribute. They may also play a role based on a set of roles defined in the standard e.g. identity, time format, frequency. Every column in the table is also based on a concept which has to be defined before the creation of the column. Different organizations can implement the standard and use it to exchange datasets. For instance, many of the datasets in Eurostat are implementations of this standard. Typically, a group of data providers defines an implementation of the standard which is used within the group, e.g., Balance of Payments data exchange, National Account data exchange.

Pangaea [2] is a repository for managing tabular data. It is an information system aimed at archiving, publishing, and distributing data related to earth science fields. The challenge of managing these heterogeneous data was met through a flexible data model. In this model, a dataset is modelled as a collection of data series and a data series consists of one or several data points for one parameter (table column). Information about the parameters, e.g., parameter unit and collection method is documented. This information can be used to parse, store and read the actual tabular data, which is stored independently of its description.

It is clear that tabular data has become widely used not only in generic domains but also in scientific data. One of these domains is biodiversity data. As a consequence, a number of repositories have been developed. In the following, we present some of these repositories, focusing on how they model and organize data. BEFdata [18] is a software platform providing support for interdisciplinary data sharing and harmonization for collaborative research projects [19]. It provides functionalities for the upload, validation, and storage of data from a formatted Excel workbook. A collection of columns (variables) in the main Excel sheet then establishes a dataset. During data upload, the Excel sheet containing the main tabular data is decomposed into its sheet-cells at the database level so that each and every single primary data value is stored independently in a database table row. Each value is thus uniquely identified in this integrating table by its source table identifier, its source table variable identifier, and its source table row identifier.

In addition, other repositories exists e.g., the Biodiversity Exploratories BExIS (BE BExIS) [20], BCO-DMO [21] that archive tabular data either as dump of the original files or in some relational forms, and provide some functionality for describing the structure of the tabular data [11][12]. In BE BExIS, tabular data (referred to as primary data) is a collection of *"observation"* entities so that each observation record is a set of values related to a specific observation. The data structure introduces the list of variables, so that each variable at least has a name, data type and a description. These information are stored as part of the metadata of the dataset. BExIS keeps track of all editing and deletions of the observations of datasets by means of a versioning mechanism.

One further direction with reference to modeling tabular data is to semantically enhance the data by using different methods, such as taxonomies [22], metadata, and ontologies [12]. For example, a wide range of metadata standards have been established over the last decade, such as EML [23] for ecological data and ABCD [24] for collection data. Ontologies can be viewed as extensions of metadata standards and are the most fundamental approach to address the problem of semantic heterogeneity. The goal of an ontology is to describe not only data, but the knowledge behind the data. One quarter of the existing repositories for biodiversity data uses ontologies, such as OBOE (Extensible Observation Ontology), as a flexible solution for standardizing attributes and their relationships [10][13][25][26].

III. Core Dataset Model

The current work is a continuation of [14], which presents a general purpose conceptual model for scientific data management. A dataset, in the model, plays the role of a data container for observations, measurements, simulations, and other supported forms of data. The meaning of data is determined by its bound data structure, which in turn determines the columns of the dataset by introducing the variables. The variables define among others the name, data type, unit of measurement, methodology and procedure of obtaining data, and measurement scale. The reusable elements of variables such as units of measurements, unit conversion information, data types, and data validation rules are factored out into *Data Container* concepts, to make data sharing, integration, and cross querying easier.

Copyright (c) IARIA, 2015. ISBN: 978-1-61208-415-2

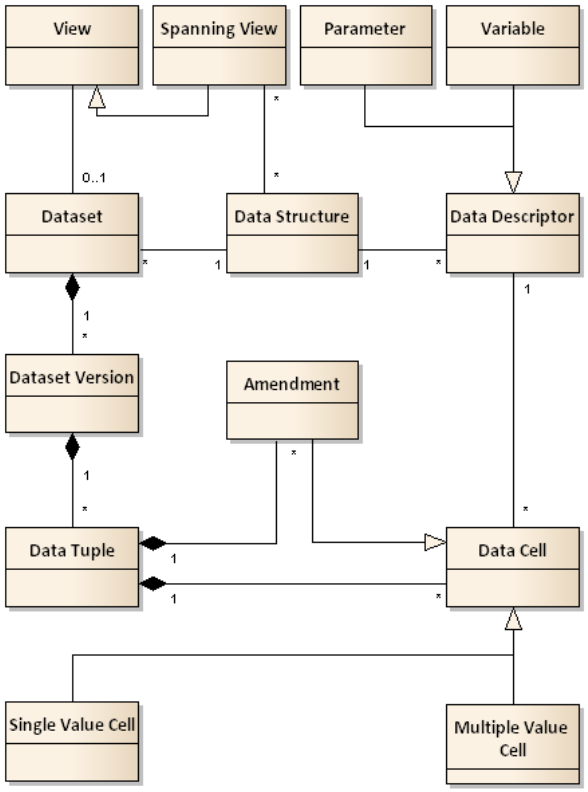

Figure 1. Conceptual model

In this paper, we look at the internals of the dataset and explain its elements in more detail. A *Dataset*, in our design, is a set of, possibly duplicate, *Tuples*. Each tuple is a collection of *Data Cells* containing the *Data Items* as shown in Fig. 1. Each cell is a compound data structure able to hold single or multiple values resulting from observations, measurements, computations, simulations, or any other means of data acquisition. In addition to the data values, the cells contain sampling, result times, and descriptions about the values, and most importantly the link to their formal description, which is captured by the concept *Data Descriptor*. The reason why sampling and result time are captured separately is that in physical object samplings, the sample may have been taken in a time different than the measurement or observation. This time difference is a considerable factor for some analyses, e.g., in soil sample water or gas containment.

As the model in Fig. 1 shows, each data cell should be associated with its corresponding variable or parameter. The variables and parameters are generalized under the the *Data Descriptor* concept, as they share almost all of their attributes. The only difference between them is that the parameters are considered to be auxiliary data to a variable. An example of such an auxiliary data would be the GPS location of a tree, whose diameter at breast height is measured. Data descriptors act as table headers to determine the name, data type, unit of measurement, methodology, and other important attributes of the columns of datasets. Factoring out the variables, units, and data types not only encourages reuse, but also establishes a foundation for data harmonization, integration, and discovery. For example, an analysis process that needs data from multiple datasets may merge the relevant columns by converting their units of measurement to a single consistent one, or searching for datasets containing temperature variables having values above 20 degree Celsius may also return datasets containing temperature values greater than 68 degree Fahrenheit. More sophisticated dataset integrations can be powered by annotating the variables with ontologies and applying semantic matching algorithms to find equivalent columns among datasets.

IV. Dataset Model Extensions

The base model is capable of materializing a table, but it may not be enough for some special requirements. In addition to the basic tabular form of the datasets, the following extensions are available to all datasets.

Amendments are special kinds of data cells scientists can attach to specific tuples, as shown in Fig. 2 as *Amendment Class* inherited from *Data Cell* and associated with *Data Tuple*. Like a usual data cell, they have their own data descriptor linked to them, hence all other attributes like unit of measurement, methodology, measurement scale, and so on. Different tuples may have different numbers of amendments each linking to their designated data descriptor. Capturing exceptional observations would be an example of using amendments. There is no need for all the tuples to have the same set of amendments. Also there is no need for the amendments of various tuples to be associated to the same variables.

Although we have tried to enrich the data descriptor class with as many attributes as possible, there are cases where scientists need more data about the variables. For example, if the values of a column are obtained using a special model of a sensor, which has a known exceptional error margin, the scientist may be interested in capturing the sensor model or the error margin as a property of the column, to use it in the analyses to be done on the column. Also the measurement system calibration, configuration, and environmental parameters are proper candidates to be modeled using extended properties. These kinds of information are column level in the scope of the dataset that contains data. Fig. 2 shows an example of this extension by attaching error, rounding indicator, and resolution properties respectively to the Soil_N, Tmp (temperature), and Time variables. To summarize it, an *Extended Property* is a user defined, dataset specific attribute whose value applies to a single column.

Sometimes, the scientists need to reduce the size of a dataset by means of removing some of the columns or filtering out the data tuples in order to perform a fast experimental analysis on the data. *Views* are proper tools to extract a subset of datasets namely for processing, sharing, or sampling purposes. Also the views can be used for security or digital right management, so that a small insensitive portion of data is exposed to the public and the original dataset is kept secure. The views can filter both the visible columns and data tuples. A view applies to a single dataset, but a *Spanning View* applies to multiple datasets that use the same structure. View 1 shown in Fig. 3, has filtered all the variables of Fig. 2's sample dataset, except the Soil_Moi., Depth, and Hu variables, as well as the data tuples matching the $Depth < -10$ predicate. View 2 in the same figure has only hidden the Depth, Pos., Hu., and Temp variables.

Copyright (c) IARIA, 2015. ISBN: 978-1-61208-415-2

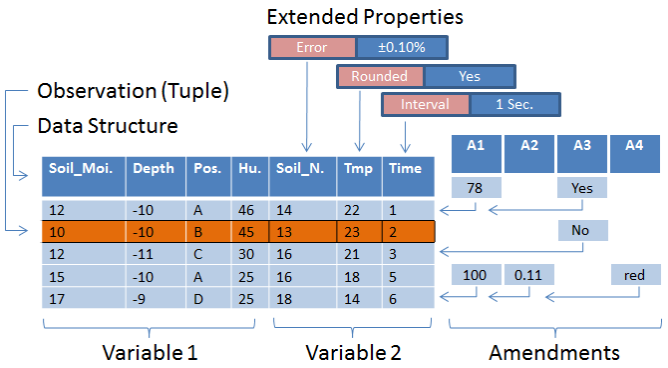

Figure 2. A sample dataset with its elements described. Soil_Moi. is Soil Moisture, Pos. is Plot code, Hu. is Humidity, Soil_N is Soil Nitrogen, and Tmp is the Temperature..

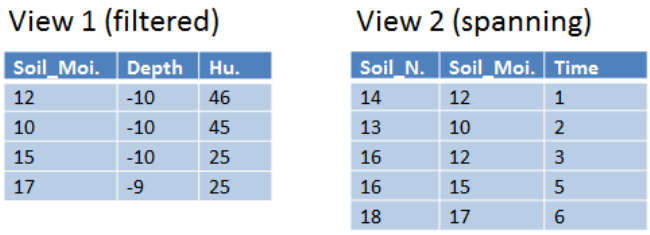

Figure 3. Two exemplary views. View 1 filters some of the variables and tuples. View 2 filters some of the datasets variables.

A small but useful customization feature of the model is its multi-lingual support for variable names. It helps multi-language teams working on the same dataset have their native names on the columns. This feature potentially reduces the effect of information loss and naming inaccuracy caused by translating the domain terminologies.

As shown in Fig. 1, each dataset can have multiple versions. The data tuples belong to the versions. This versioning scheme helps to freeze the versions so that they are accessible for later processing and citations, independent of the following changes. Technical details of the versioning are not in the scope of this paper, but as a short description, it provides a check-in, check-out mechanism, computes and stores the difference between the versions. The information collected in the core and extended mode entities can serve an additional role of being treated as metadata. For example, a dataset export tool can, in addition to the actual data, extract some parts of the variables as metadata and serialize them alongside with. The datasets have metadata at three levels. Cell level metadata captures how the value was obtained, when it was sampled if so, when the result was ready, and a free-text description. Structural level metadata are handled by defining the variables, parameters and extended properties under a data structure. The dataset version level metadata are captured by user-defined or standard metadata schemas e.g., EML or ABCD. The version level metadata is describing the whole dataset version as a unit of data and may consist of various aspects among them authorship, geographical extent, copyrights, sensors or measurement tools, or software configuration. Each version of a dataset may have its own metadata, so that changes in the metadata are aligned with changes in the data.

In addition to the mentioned capabilities, the variables are able to be linked to semantic elements such as terminologies, taxonomies, or ontologies. This features make the model a proper candidate for automatic schema matching, data integration, multi-project joint analyses and so on.

V. Conclusion and Future Work

Biodiversity data has become more and more important, therefore, there is a growing need to develop new platforms and infrastructures that facilitate creating, storing, reusing, and sharing scientific data. To this end, in this paper, we introduced a data structure for tabular scientific data. In particular, we presented the core elements in the model, including dataset, dataset versions, tuples, and data cells as well as the possible extensions to these core elements. The model can be used to enforce the structure and type of information to be collected as well as a base for data validation. The attributes assigned to the variables, e.g., unit of measurement, semantic annotations, and the unit conversion information can be used in data integration efforts. Datasets published using this model allow the following researchers to obtain the data with its structure and the meaning of the elements, so that they can run similar analyses to validate or reproduce the original work, or use it in their own work. In addition, the dataset versions provide a strong framework for dataset citation.

The model lacks some features like user-defined data types for the cells and versioning the views. Currently, there is a predefined set of data types introduced to the model, so that all the cells, whether single or multiple value, accept data of those types only. It would be an improvement to allow the model users to define their own scalar or complex data types and use them in their dataset modeling needs. As described, the views can reduce the amount of visible data of target datasets. A useful feature of the model would be to apply the versioning concept to the views too, so that they can be attributed or cited independently guaranteeing access to the same subset of datasets over time. In our future work, we are going to extend the model to address these shortcomings.

Acknowledgments

This work was partly funded by the German Science Foundation through the projects BExIS++, AquaDiva (subproject INFRA1), and Biodiversity Exploratories (subproject Data Management).

References

[1] “Dryad Digital Repository,” accessed 07/05/2015. [Online]. Available: http://www.datadryad.org/

[2] M. Diepenbroek, H. Grobe, M. Reinke, U. Schindler, R. Schlitzer, R. Sieger, and G. Wefer, “PANGAEA – an information system for environmental sciences,” Computers & Geosciences, vol. 28, 2002, pp. 1201–1210.

[3] “Scientific Data,” accessed 07/05/2015. [Online]. Available: http://www.nature.com/sdata/

[4] “Earth System Science Data, The Data Publishing Journal,” accessed 07/05/2015. [Online]. Available: http://www.earth-system-science-data.net/

[5] “Biodiversity Data Journal,” accessed 07/05/2015. [Online]. Available: http://biodiversitydatajournal.com/

[6] J. Kratz and C. Strasser, “Data Publication Consensus and Controversies,” F1000Research 2014, vol. 3:94, 2014. [Online]. Available: http://f1000research.com/articles/3-94/v3

Copyright (c) IARIA, 2015. ISBN: 978-1-61208-415-2

[7] H. Allen, P. Gearan, and K. Rexer, "6th Annual Data Miner Survey," Rexer Analytics, Tech. Rep., 2011.

[8] J. King and R. Magoulas, 2013 Data Science Salary Survey Tools, Trends, what pays (and what doesn't) for Data Professionals, 1st ed. O'Reilly Media, 2014.

[9] J. Kattge, K. Ogle, G. Bönisch, S. Dìaz, S. Lavorel, J. Madin, K. Nadrowski, S. Nöllert, K. Sartor, and C. Wirth, "A generic structure for plant trait databases," Methods in Ecology and Evolution, vol. 2, no. 2, 2011, pp. 202–213. [Online]. Available: http://dx.doi.org/10.1111/j.2041-210X.2010.00067.x

[10] S. Bowers, "Scientific Workflow, Provenance, and Data Modeling Challenges and Approaches," Journal on Data Semantics, vol. 1, no. 1, 2012, pp. 19–30. [Online]. Available: http://dx.doi.org/10.1007/s13740-012-0004-y

[11] T. Lotz, J. Nieschulze, J. Bendix, M. Dobbermann, and B. König-Ries, "Diverse or uniform? Intercomparison of two major German project databases for interdisciplinary collaborative functional biodiversity research," Ecological Informatics, vol. 8, 2012, pp. 10–19. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S157495411100094X

[12] K. Bach, D. Schäfer, N. Enke, B. Seeger, B. Gemeinholzer, and J. Bendix, "A comparative evaluation of technical solutions for long-term data repositories in integrative biodiversity research," Ecological Informatics, vol. 11, no. 0, 2012, pp. 16 – 24, data platforms in integrative biodiversity research. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1574954111000987

[13] R. Walls, J. Deck, R. Guralnick, S. Baskauf, R. Beaman, and et al., "Semantics in support of biodiversity knowledge discovery: An introduction to the biological collections ontology and related ontologies," PLoS ONE, vol. 9, 2014.

[14] J. Chamanara and B. König-Ries, "A conceptual model for data management in the field of ecology," Ecological Informatics, vol. 24, 2014, pp. 261–272.

[15] J. Tennison, G. Kellogg, and I. Herman, "Model for Tabular Data and Metadata on the Web," April 2015, accessed 07/05/2015. [Online]. Available: http://www.w3.org/TR/tabular-data-model/

[16] "Guidelines for the use of Observations & Measurements and Sensor Web Enablement-related standards in INSPIRE Annex II and III data specification development," 2014. [Online]. Available: http://inspire.ec.europa.eu/documents/Data_Specifications/D2.9_O&M_Guidelines_v2.0.pdf

[17] SDMX Statistical Working Group, and SDMX Technical Standards Working Group, "Guidelines on Modelling a Statistical Domain for Data Exchange in SDMX," 2015, accessed 07/05/2015. [Online]. Available: http://sdmx.org/

[18] "Biodiversity and Ecosystem Functioning Data," accessed 07/05/2015. [Online]. Available: https://github.com/befdata/befdata

[19] K. Nadrowski, K. Pietsch, M. Baruffol, S. Both, J. Gutknecht, H. Bruelheide, H. Heklau, A. Kahl, T. Kahl, P. Niklaus, W. Kröber, X. Liu, X. Mi, S. Michalski, G. von Oheimb, O. Purschke, B. Schmid, T. Fang, E. Welk, and C. Wirth, "Tree Species Traits but Not Diversity Mitigate Stem Breakage in a Subtropical Forest following a Rare and Extreme Ice Storm," PLoS ONE, vol. 9, no. 5, 05 2014, p. e96022.

[20] "Exploratories for Large-scale and Long-term Functional Biodiversity Research," accessed 07/05/2015. [Online]. Available: http://www.biodiversity-exploratories.de/startseite/

[21] "The Biological and Chemical Oceanography Data Management Office," accessed 07/05/2015. [Online]. Available: http://www.bco-dmo.org/

[22] "ITIS - Integrated Taxonomic Information System," accessed 07/05/2015. [Online]. Available: http://www.itis.gov

[23] The Knowledge Network for Biocomplexity (KNB), "Ecological Metadata Language (EML)," accessed 07/05/2015. [Online]. Available: https://knb.ecoinformatics.org/#external//emlparser/docs/index.html

[24] "Access to Biological Collection Data (ABCD)," accessed 07/05/2015. [Online]. Available: http://wiki.tdwg.org/ABCD

[25] J. Madin, S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, and F. Villa, "An ontology for describing and synthesizing ecological observation data," Ecological Informatics, vol. 2, 2007, pp. 279–296.

[26] Y. Shu, D. Ratcliffe, M. Compton, G. Squire, and K. Taylor, "A semantic approach to data translation: A case study of environmental observations data," Knowledge-Based Systems, vol. 75, 2015, pp. 104–123.

Copyright (c) IARIA, 2015. ISBN: 978-1-61208-415-2