Performance Model's development: A Novel Approach encompassing Ontology-Based Data Access and Visual Analytics

Marco Angelini[1], Cinzia Daraio[1], Maurizio Lenzerini[1], Francesco Leotta[1], Giuseppe Santucci[1]

[1]{angelini, daraio, lenzerini, leotta, santucci}@diag.uniroma1.it
DIAG Department, Sapienza University of Rome, Via Ariosto, 25 00185, Rome (Italy)

Abstract
The quantitative evaluation of research is currently carried out by means of indicators calculated on data extracted and integrated by analysts who elaborate them by creating illustrative tables and plots of results.
In this paper we propose a new approach which is able to move forward, from indicators' development to performance model's development. It combines the advantages of the Ontology-based data Access (OBDA) integration with the flexibility and robustness of a Visual Analytics (VA) environment. A detailed description of such an approach is presented in the paper.

Introduction: An advanced models' development approach

In recent decades, the rapid changes taking place in the production, communication and evaluation of research have been signs of an ongoing transformation. It has been stated that "we are living a sort of Middle-Age guided by the information and communication technologies (ICT) revolution, or the so-called *forth revolution* as described by Floridi (2014) which emphasizes the importance of information" (Daraio, 2019, p. 636). Largely, the current Middle-Age of research evaluation might be understood as the transition from a traditional evaluation model, based on bibliometric indicators of publications and citations to a modern evaluation, characterized by a multiplicity of distinct, complementary dimensions. This step is guided by the development and increasing availability of data and statistical and computerized techniques for their treatment, including among others the recent advancements in artificial intelligence and machine learning. Daraio and Glänzel (2016) show that that the complexity of research systems requires a continuous information exchange.
These changes produce different effects (see further details and references in Daraio, 2019, Table 24.2, p. 644) i) on the *demand side* (those that ask for research assessment) including an increase of institutional and internal assessments, ii) on the *supply side* (those that offer research assessment) including proliferation of rankings, development of Altmetrics, open access repositories, new assessment tools and desktop bibliometrics), iii) on *scholars* (the increase of "publish or perish" pressure, impact on the incentives, behaviour and misconduct, and increasing critics against traditional bibliometric indicators), iv) on the assessment process (increasing the complexity of the research assessment) and on the indicators' development.
Daraio (2017a) showed that the formulation of models of metrics (in this paper we will use metrics and indicators as synonyms) is necessary to assess the meaning, validity and robustness of metrics. It was observed that developing models is important for *learning* about the explicit consequences of assumptions, test the assumptions, highlight relevant relations; and for *improving*, document/verify the assumptions, systematize the problem and the evaluation/choice done, explicit the dependence of the choice to the scenario. Moreover, there are several *drawbacks* in modelling, which have to be taken into account. The main pitfalls relate to the targets that are not quantifiable; the complexity, uncertainty and changeability of the environment in which the system works, to the limits in the decision context, and, last but not least, to the intrinsic complexity of calculation of the objective of the analysis.
In this paper we depart from the traditional approach to indicators' development, based on the selection of a specific set of indicators, collection of the relevant data, cleaning of the gathered

data, computation of the indicators and illustration of them in a plot or table. According to this traditional approach if you want to add a new data source or you want a different indicator you have to restart the process from the scratch.
We support an alternative approach based on an OBDA system for R&I data integration and access. An Ontology-Based Data Access (OBDA) system is an information management system constituted by three components: an ontology, a set of data sources, and the mapping between the two. An *ontology* in Description Logic (DL) is a *knowledge base*. It is a couple (pair) O=<TBox,ABox>, where TBox is the Terminological Box that represents the *intensional* level of the knowledge or the *conceptual* model of the portion of the reality of interest expressed in a formal way; and ABox is the Assertion Box that represents the *extensional* level of the knowledge or the *concrete* model of the portion of the reality expressed by means of assertions (instances). An ontology populated by instances and completed by rules of inference is defined as *knowledge base* (see e.g. Calvanese et al. 1998). The *data sources* are the repositories accessible by the organization where data concerning the domain are stored. In the general case, such repositories are numerous, heterogeneous, each one managed and maintained independently from the others. The *mappings* are precise specifications of the correspondence between the data contained in the data sources and the elements of the ontology. The main purpose of an OBDA system is to allow information users to query the data using the elements in the ontology as predicates.
The OBDA system, implemented with *Sapientia*, represents the ontology of multidimensional research assessment (Daraio, Lenzerini et al. 2015) and permits the extraction of relevant data coming from heterogeneous sources - maintained independently, and reasoning about the Performance Indicators (PI) of interest.
Daraio, Lenzerini et al. (2016a) showed the advantages of an OBDA system for R&I integration and Daraio, Lenzerini et al. (2016b) showed that an OBDA approach allows for an unambiguous specification of indicators according to its four main dimensions: ontological, logical, functional and qualitative. See also Lenzerini and Daraio (2019) where a detailed illustration of the usefulness of an OBDA approach for reasoning over the ontology about indicators of performance is reported. Even the simplest indicator of performance, such as number of publications has different conceptual aspects that the ontological commitment of the domain offers to the analyst (for additional details the reader is referred to Fig. 15.9 and 15.10 of Lenzerini and Daraio, 2019, pag. 368 and pag. 369).
The main contribution of this paper is making a step further, on our previous researches and to propose a new approach for the multidimensional assessment of research and its impact based on the combination of OBDA and Visual Analytics. This novel approach allows for the development and evaluation of performance models instead of the traditional indicators' building system.

Combining OBDA and Visual Analytics

The traditional way to define indicators relies on an *informal definition* of the indicator as the relationship between variables selected among a set of data collected and integrated "ad hoc", specific for the user needs (*silos based* data integration approach). This means that when a new indicator has to be calculated, the process of data integration has to restart since the beginning because the dataset created "ad hoc" for an indicator is not reusable for another one.

The contribution of an OBDM approach to overcome this traditional indicator development approach is twofold. First of all, it permits the *free* exploration of the *knowledge base* (or information platform) created to identify and specify new indicators, not planned or defined in advance by the users. This feature would be particularly useful to face two recent trends in user

requirements, namely *granularity* and *cross-referencing* (see Daraio and Bonaccorsi, 2017 for a discussion on university-based indicators). Secondly, it allows us to specify a given indicator in a more precise way as described in Lenzerini and Daraio (2019).
In this paper we develop further this approach combining it with the main strengths of Visual Analytics. Visual Analytics (Cook & Thomas, 2005, Keim et al., 2008) is "the science of analytic reasoning facilitated by interactive visual interfaces"; through the connection of the analytical calculation with visualization and interaction by the human user, this interdisciplinary approach enhances the exploratory analysis of data, allowing to represent multidimensional data in a simple way through innovative visual metaphors. Further it allows navigation in the data space, in order to obtain an overview of the eventually tunable to the required level of detail, the ability to apply complex analysis workflows that aim at explainability, the ability to obtain summary reports of the findings discovered during the analysis phase. See Figure 1 for an overview.

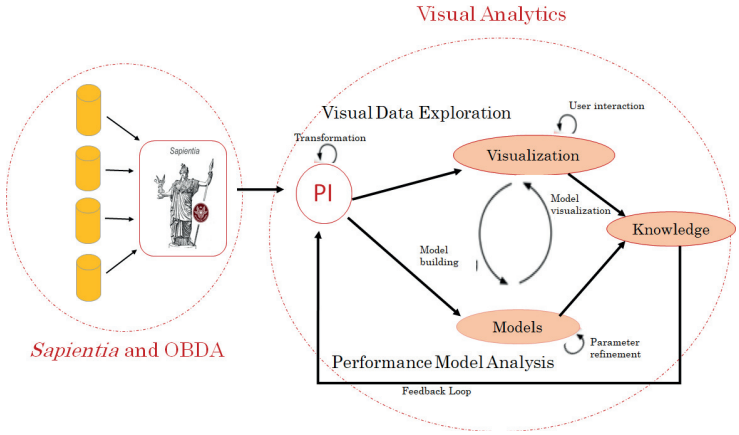

Figure 1. An illustration of our approach that combines *Sapientia*, OBDA and Visual Analytics. PI states for Performance Indicator.

The Visual Analytics approach developed in this paper allows us to move from Performance indicators development to Performance model development, by exploring and exploiting the modelling and the data features within the flexibility of a Visual Analytics environment.
This allows a multi-stakeholder viewpoint on the model of PI, the assessment of the sensitivity and robustness of the PI model in a multidimensional framework.
In the next section we outline the main features of *Sapientia* (the Ontology of Multidimensional Research Assessment). After that we present our Visual Analytics environment for the performance model's development together with an illustration of its potentialities. The final section concludes the paper.

OBDA at work through *Sapientia: The Ontology of multidimensional research assessment*
Sapientia, the Ontology of Multidimensional Research Assessment (Daraio et al. 2015, 2016a, 2016b), models all the activities relevant for the evaluation of research and for assessing its impact (see Figure 2 for an outline of its modules). For impact, in a broad sense, we mean any effect, change or benefit, to the economy, society, culture, public policy or services, health, the environment or quality of life, beyond academia.
The *Sapientia* ontology has been developed using the Graphol visual language (http://www.dis.uniroma1.it/~graphol/, Lembo et al. 2016), that can be easily translated into standard ontology languages like Owl.

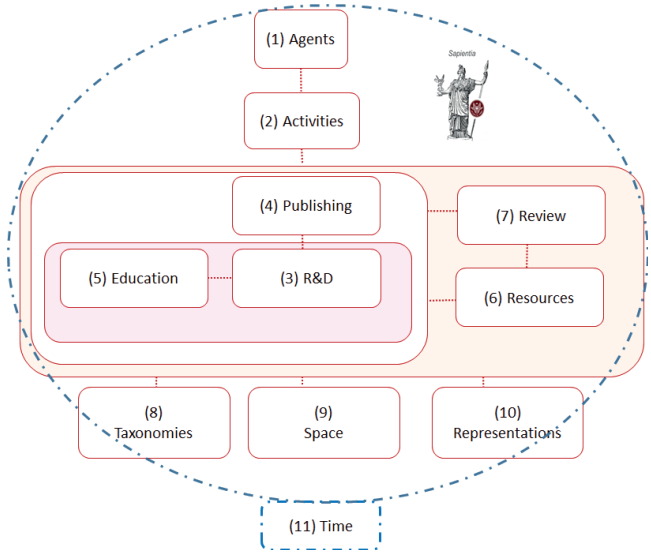

Figure 2. Modules of *Sapientia* 3.0. 1. Agents: describes all human actors and institutions involved in the education, research and innovation process. 2. Activities: describes the activities and projects the agents of the previous module are involved in. 3. R&D: describes the different products (e.g., publications, patents) that are produced in the knowledge production process. 4. Publishing: describes how knowledge products are published and made available to the public. 5. Education: introduces concept related to universities and courses. 6. Resources: describes all the ways and institution can be funded. 7. Review: describes the process entities related to the publishing activity. 8. Taxonomy: describes the elements that allows to define taxonomies applied to the different modules. 9 and 11: Space and Time: allow to describe respectively geographical entities and time instants and ranges. 10. Representation: allows to describe the fact that single instances of other modules can be represented in different ways by the different sources used in *Sapientia.*

Sapientia acquires information from multiple sources, whose content can be overlapping. The same entity modelled in the *Sapientia* ontology can be represented in more than one data source, and even one data source could present (due to internal inconsistencies or design choices) the same entity multiple times in different forms.

Hence, we have the need to identify duplicated items and integrate the information obtained for each entity from any of the available sources.

In particular, at the ontology level we have created the concept of Representation. Entities modelled in the ontology of which we have different views from different data sources may have their own representation, which specializes the general Representation concept. This makes it possible to keep track in the ontology, through the mappings, not only of the modelled entities, but also of the way in which the information relative to the entities has been gathered from the data sources.

Data acquisition from the external sources makes use of the web service standards (REST, SOAP) when available.

For less frequently updated sources and sources that do not implement an API, data acquisition leverages in some cases the open source edition of Pentaho Data Integration (http://community.pentaho.com/projects/data-integration/).

Imported data are saved in a relational database (MySql). Each source is modeled independently so that its peculiar structure can be fully exploited.

Sapientia extract information, among others, from the following datasets:

- *Scopus*. A very large abstract and citation database of peer-reviewed literature, containing information about scientific journals, books and conference proceedings. Scopus provides information about authors' affiliations as well. The available REST interface allows to retrieve: document information, document citations data, percentiles data end journal percentiles data.
- *ETER*. The ETER (European Tertiary Education Register) consortium acquired extensive information pertinent to tertiary educational institution of many European countries. Data have been acquired by the consortium for the years 2011-2016 and are publicly available (https://eter-project.com).
- *DBLP*. A service that provides open bibliographic information on major computer science journals and proceedings. Data is available through massive XML files.
- The *InCites* (https://incites.thomsonreuters.com) dataset contains research indicators organized on a geographical base. Data can be downloaded in the form of CSV files that are then imported using an ad-hoc procedure.
- *Geonames* (http://www.geonames.org/) is a dataset that contains information about geographical areas at any level. The dataset can be freely download, and has been employed to match geographical entities from the different data sources.
- *Web of Science* database is going to be included as well.

The data manipulation layer of the *Sapientia*, which allows to populate the ontology from the data sources, is composed of an indexing module, an entity-resolution module and a normalization module.

In general terms, the *indexing module* creates and maintains up to date the indices that are used by the *entity resolution module* to implement the blocking functionalities that allow to keep the time complexity of the entity-resolution algorithms under control. This module has the dual purpose of easing the definition of the mappings toward the *Sapientia* Ontology, and creating the basis for a common interface of the entity-resolution algorithms. Indices inside the *Sapientia* application are implemented using the Hibernate search (http://hibernate.org/search/) library and the Lucene indexer and searcher (http://lucene.apache.org/).

Entity resolution is the task of connecting matching entities between different data sources. As this kind of process is exponential in complexity with the number of data sources and entities per data sources, it is split in two phases:

- *Blocking*, which allows very quickly, by employing indices to create groups of potential matching entities
- *Entity matching*, which finds matching entities inside clusters identified by blocking.

After matching entities have been recognized by entity resolution, the *normalization step* is employed in order to provide a uniform representation for the information contained in different and heterogeneous data sources. These uniform representations are called mappable entities. These mappable entities are mapped to ontology entities through an operation called mapping. *Sapientia* uses the Mastro Ontology-Based Data Access (OBDA) management system (http://www.dis.uniroma1.it/~mastro/?q=node/1). The Sapientia ontology, however, is defined over a richer language than the one supported by Mastro. Hence, we used the OWL2DL tool in order to obtain a simplified version of the Sapientia ontology that conforms to the DL-light language supported by Mastro.

The definition of the mappings in Mastro is XML based. There are three types of ontology predicate mappings: concept, role and attribute.

As suggested by the names, the concept predicate mapping refers to entities, the role predicate mapping puts entities in relation, populating a role, while the attribute mapping relates an entity with a constant, which is the value of its attribute.

Some examples of extraction and mapping of relevant data

In order to show the potential of the proposed approach, we will show how indicators can be extracted from the ontology and grouped according to a specific level of analysis. In the illustration identified as European denoted Nomenclature of Territorial Units for Statistics (NUTS) code. The modules of the ontology interested in this query are:

- The *Agents module*, which contains the concept of University as a specialization of the concept of Organization. An Organization has an Organization State, which represents the evolution of the Organization in time, and that refers to the Residence.
- The *Space module* that contains the concept of Residence as a specialization of a Position. A Position has an Entrance, which is localized in an Address inside a City. The City is a Territory, and European Cities are European Territories that can be aggregated by NUTS codes.
- The *Taxonomy module* where an Organization is contained in a Taxonomic Unit. Each Taxonomic Unit has a State that has indicators as attributes.

For a specific university denoted by its Eter ID, we can for example compute the cardinality of academic staff with the following SPARQL query :

```
select ?academic_staff {
     ?org sapientia:has_place_in ?taxon_unit .
     ?org a sapientia:University .
     ?taxon_unit sapientia:has_state_of_taxonomic_unit ?state_tax .
     ?state_tax a sapientia:Present_state .
     ?state_tax sapientia:teacher_population ?academic_staff .
}
```

In order to group by a specific NUTS codes, it is possible to extend the previous query as follows:

```
select SUM(?academic_staff), ?nuts2 {
     ?org sapientia:has_place_in ?taxon_unit .
     ?org a sapientia:University .
     ?taxon_unit sapientia:has_state_of_taxonomic_unit ?state_tax .
     ?state_tax sapientia:teacher_population ?academic_staff .
     ?state_tax a sapientia:Present_state .
     ?org sapientia:has_state_of_organization ?org_state .
     ?org_state a sapientia:Present_state .
     ?org_state sapientia:has_residence ?resid .
     ?resid a sapientia:Legal_residence .
     ?resid sapientia:has_entrance ?entr .
     ?entr a sapientia:Address .
     ?entr sapientia:is_in_the_city ?city .
     ?city a sapientia:European_territory .
     ?city sapientia:is_territory_part_of ?region .
     ?region a sapientia:Small_europen_region .
     ?region sapientia: NUTS2ref ?nuts2 .
     ?region sapientia: NUTS2ref ?nuts1 .
     ?region sapientia: NUTS2ref ?nuts3
}
GROUP BY ?nuts2
```

Where the results have been grouped by NUTS2. It is possible to easily modify the query in order to group by other levels of NUTS. In a similar way, *mutatis mutandis*, it is possible to extract the data and indicators that will be used for the Performance Indicator and model development that is described in the next sections.

The Visual Analytics environment

This section describes the Visual Analytics environment and its main features. The developed solution uses Visual Analytics techniques to represent data from publications and education obtained from the OBDA approach described in the previous section, and complete the workflow. The system is implemented through Web technology. Clearly the large quantity of indicators and basic sizes for the different units of analysis, including the territorial ones, and

in the different years of analysis increases exponentially the cardinality of data to be analyzed; in this respect, the display part allows to obtain a visual overview of the data in a very simple form, and the interaction capabilities allow the user to navigate in this overview and conduct detailed analysis up to the desired level. The user is also supported in the discovery of any elements of analysis of interest through a process of *data exploration* that does not require a prior analysis goal.

In addition to the data exploration capacity there is a second area explicitly aimed at analyzing the model development and performance computed on these indicators, based on the definition and exploration of performance models. The environment is instantiated on European research and education institutions as a case study. The user can, on one hand, analyze the performance of the various institutions with respect to a performance model, in order to analyze the positioning of the institutions of interest; additionally, it allows to explore different performance models and to evaluate their goodness and fitness. Further, it is also possible to evaluate the goodness of the proposed models, analyzing their variability and conducting sensitivity analysis in order to evaluate which parameters of the model (whether inputs or resources, contextual factors or outputs) contribute more to the performance of the institution with respect to the chosen model. The following subsections will provide a description of the features of Visual Analytics environment.

Data Exploration Environment

The first panel that composes the Visual Analytics environment is the data explorer environment. This environment consists of three main views depicted in Figure 3.

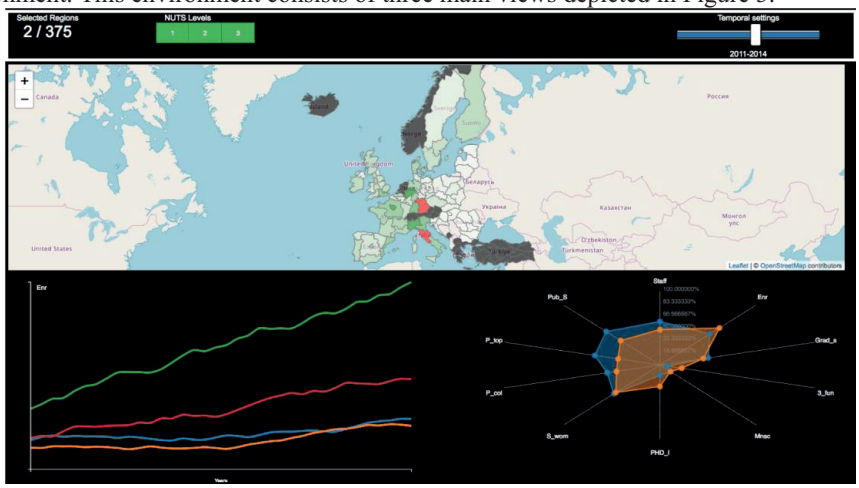

Figure 3. Data Exploration Environment

These three views are:

-*Geographic view*: which allows for geolocating of the different institutions with respect to territorial units on a geographic layer (using Leaflet.js framework, based on OpenStreetmap) is used. The map is navigable on 5 different levels of detail, where the first four follow the NUTS categorization from 0 (Nations) to 3 (Provinces) and the last one relates to single institutions. The user can at any time change the level of aggregation through a tab that shows the different available levels.

The color of each element of the map reflects an indicator (basic or derived) of, on a green scale that identifies the values (white: low value, dark green: high value). The gray color codifies the absence of data for the particular territorial unit. A slider allows the analyst to scroll through

the various years and conduct a temporal analysis on the available data, looking for institutions showing a high variability through a “time-lapse”.
-*Radar view*: this view follows the visual paradigm of the radar diagrams (Von Mayr, 1877), which represent the dimensions of a dataset one per axis, with the axes arranged in radial form starting from the center. The indicators are arranged one per axis and the graph presents several lines that join the points on each axis in the number of one per institution or territorial unit. When the user selects one or more territorial units, the corresponding splines are highlighted, in order to allow an easy visual comparison between the different territorial units selected on their different dimensions. It is also possible to highlight a dynamic average trend, consisting of a line that connects the different averages on the respective axes, in order to compare the performance of a territorial unit, or generally of a given unit of analysis, not only to other units but also to the aggregate behavior between the territorial units.
-*Linechart view*: This visualization allows analyzing the time course of the evaluation measures used for the units. It is possible to analyze both multiple territorial units to compare the trend of the same measure on them, and to analyze multiple measures on the same territorial unit, in order to have an overview of the progress of the unit itself, and a combination based on multiple territorial units and multiple measures. In this case the color-coding outlines all the measures belonging to each single territorial unit.
The combined use of these views, possibly guided by the definition of specific PIs, allows more powerful dynamic exploration of the model data of the territorial units compared to the classical approaches, making the user able to obtain an overview of the general trend and specific details on the individual units, subsequently allowing to refine the analysis through the visual selection of appropriate subsets of information. The approach therefore allows the exploration of s*pecific scenarios* chosen by the user in *real-time*, without precomputation, which better support the formation and validation of hypotheses and the identification of areas of interest on which to conduct further analysis or to be used for reporting activity.

Performance Model Analysis Environment
This environment is dedicated to the analysis of performances of the model used for analyzing the units. The visual environment is therefore more complex than the Data Exploration one, as shown in Figure 4.

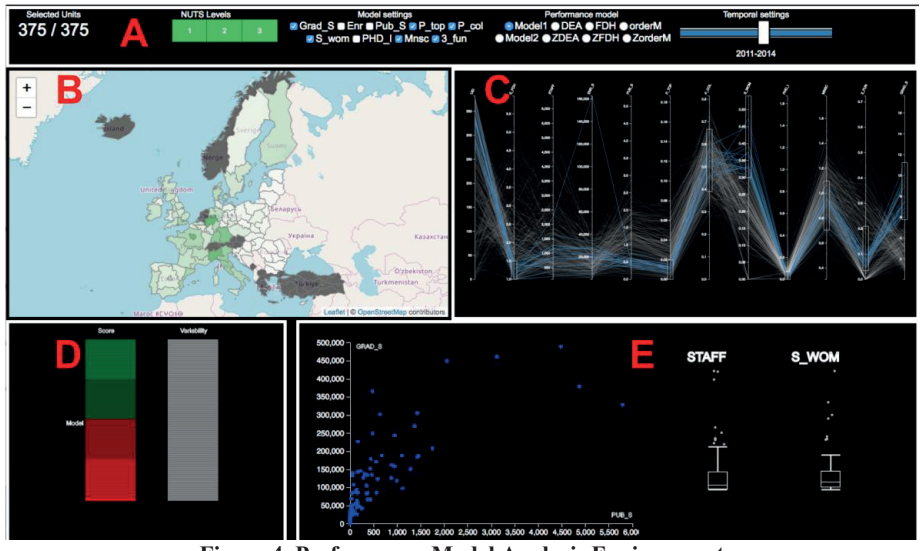

Figure 4. Performance Model Analysis Environment

The environment consists of a command bar (A), a geographical view borrowed from the Data Exploration environment (B), a view based on parallel coordinates (C), a view of the rankings due to the selected performance model (D) and finally a view based on scatter-plot and box-plot that allows to conduct sensitivity analysis on the parameters of the selected model (E). The features of the individual views are described below.

-*Command bar*: this area identifies the main analysis commands that will affect the selections in all remaining views. From left to right we have:

-the counter of the territorial units active with respect to the total (the territorial units contained in the current selection)

-a tab that allows to select the aggregation level on which to conduct the analysis

- the parameters and measures of the performance model, which can be activated using the appropriate checkboxes. This command allows to re-parameterize a model (among those available) in order to conduct a different type of analysis of performance.

-The model selector, which allows you to choose between 8 families of performance models, ranging from custom model defined by the Analyst (Model 1 and Model 2) to efficiency models, Data Envelopment Analysis (DEA), Free Disposal Hull (FDH), orderM, and their conditioned variants ZDEA, ZFDH, ZorderM. An overview on these performance models can be found in Daraio (2019).

-The time selector, which allows to evaluate the result of the chosen model with respect to a temporal interval that can be controlled by means of a slider.

-*Geographical view:* this visualization follows the same operating principle illustrated for the Data Analysis Environment. In this instance, however, the color linked to each individual territorial unit is proportional to the unit's performance score with respect to selected model. In this way the user can immediately get an overview of the different performance levels given the chosen hierarchical level, model and time interval. The user can zoom in on the map in order to get more details on individual portions of the map. It is also possible to use the map as a highlighting mechanism: by clicking on one or more units, these are highlighted in red on the map and in all other coordinated views, allowing to identify a subset of data of interests starting from geographical coordinates of the unit.

-*Parallel coordinates*: this view shows all the dimensions that are part of the model (inputs, possible conditioning factors, outputs) plus the year of analysis and the ID of the units. The purpose of this visualization is to explore the relationships that exist between these quantities, in order to decide whether or not to keep them in the selected model. From the visual point of view, each of the dimensions is represented as a vertical axis, and each unit as a line that joins the values it has on each axis. Through brushing operations on individual axes, it is possible to perform multi-filter operations on several dimensions, making possible to select very complex filtering expressions while maintaining the ease of creating these filters: by dynamically define new intervals on the various dimensions, and immediately verify the cardinality and the characteristics of the resulting subset, the analyst can explore several combinations and discover relations among dimensions (see Figure 5).

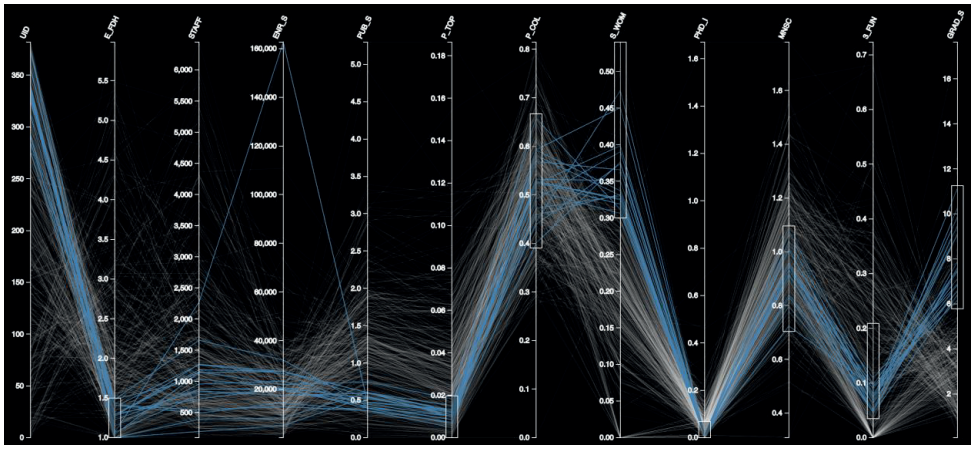

Figure 5. Example of parallel coordinates filter. Axes, from the left: UID is the institution id number, E_FDH is the FDH (in)efficiency score (equal to 1 means efficient; the higher it is, more outputs the unit could proportionally produce to become efficient) STAFF is number of academic staff in FTE (Full Time Equivalent), ENR_S is number of total enrolled students per academic staff, PUB_S is number of publications in WoS (fractional count) per academic staff, P_TOP is number of publications in top 10% of highly cited journals per academic staff, P_COL is percentage of papers done with international collaborations, S_WOM is share of women professors on total academic staff, PHD_I is PhD intensity, MNCS is Mean Normalized Citation Score (1 corresponds to the world average, >1 above (<1 below) world average), 3_FUN is share of third party funds in %, GRAD_S is total number of graduates per academic staff. The filter shows that among the most efficient units in teaching and research (i.e. E_FDH = [1 1.5]) there are those teaching oriented institutions (with the highest values of GRAD_S) in which the S_WOM is the highest ([0.30-0.50]): these are universities with almost zero PhD intensity that are able nevertheless to produce a small fraction of P_TOP publications with MNCS around the world average.

In addition, by drag and drop interaction, it is possible to exchange all the axes with each other, in order to better highlight any correlations, anti-correlations or similarity characteristics on specific subsets of data among dimensions. Any findings, as mentioned above, serve to better understands the results coming from the performance model used.

-*Rank analysis*: This view supports the tasks of exploring the performance scores of the individual units, and the sensitivity analysis on the model, in terms of estimating the contribution of each individual parameter of the model to the performance scores. The visualization is composed of two bars representing rankings, where the units are ordered according to the performance score from top (high performance score) to bottom (low performance score). Each unit is represented as a rectangle, whose color derives from the calculation of the distribution of the performance scores and from the assignment of a color to each of the 4 quartiles (the 3rd and 4th quartiles with deeper shades of green, the 1st and 2nd with deeper shades of red). An informative tooltip, activated by mouse-hover on each rectangle, allows to obtain accurate information on the performance of the unit. The second bar is initially completely gray, and is activated when individual elements (inputs, conditioning factors) of the model are selected / deselected from the command bar: in this way it is possible to evaluate the displacement in the rank of each single unit with respect to addition/deletion of a parameter of the model, and therefore be able to evaluate the stability of the model compared to the performance scores produced, and the sensitivity of the performance model in terms of contribution that any parameter produces in the ranking (see Figure 6).

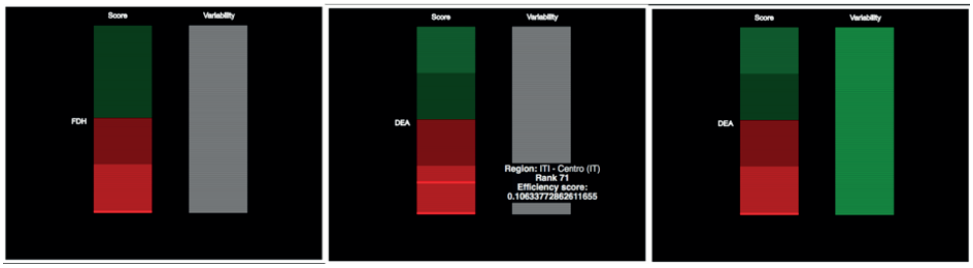

Figure 6. rank analysis obtained using a complete FDH model (left); the same chart is instantiated through a DEA model, and the tooltip reports the score for the "Italia Centro" territorial unit (center); finally, the result on the variability obtained by removing the output factor PUB_S and including P_TOP (right). As you can see, the whole bar is green, which means that the units rank remains stable with respect to this input, which could be replaced by another more significant input.

Sensitivity analysis: This view expands the sensitivity analysis capabilities, already introduced in the Rank Analysis view. The visualization uses two different visual paradigms to relate the different parameters that constitute the performance model: in the first one, a scatter plot, the relation between the conditioning factors (if present) and the outputs is reported. Input factors are instead reported as a distribution in the form of a box-plot for each input factor. The interactivity of this chart allows to select disjoint sets of values from each box-plot and inspect the propagated filter on the entire visual environment. It will be possible to analyze the relationship between the various elements of the performance model in a more precise and granular form, identifying from the distribution subsets of interest which will eventually correspond to the selection of a subset of units that respect the imposed constraints. The effect will therefore support the sensitivity analysis of the model but also support the explorative analysis of the data through filter operations based on factors of the model (see Figure 7).

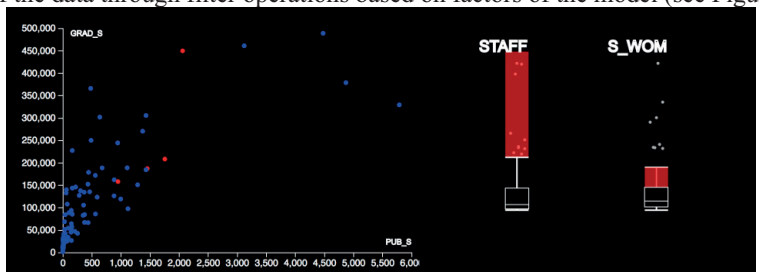

Figure 7. Example of data filtering: with respect to all the units, the selection is composed by high outliers for academic staff (STAFF) and the 4th quartile for percentage of women staff (S_WOM); the resulting points are highlighted in red in the scatter plot, and the unit can be identified by mouse-hover.

Conclusions

In this paper we leveraged on the research based on *Sapientia* and OBDA combining it with a Visual Analytics approach. The new approach proposed allows us to move from Performance indicators development to Performance model's development, by exploring and exploiting the modelling and the data features within the flexibility of a Visual Analytics environment. This allows a multi-stakeholder viewpoint on the model of PI, the assessment of the sensitivity and robustness of the PI model in a multidimensional framework as illustrated in the previous section.

Acknowledgments
The financial support of the Italian Ministry of Education and Research (through the PRIN Project N. 2015RJARX7), of Sapienza University of Rome (through the Sapienza Awards no. PH11715C8239C105 and KIMAR Project), and of Clarivate Analytics through the KOL Project, is gratefully acknowledged.

Selected References

Calvanese D., G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati (2007). Tractable reasoning and efficient query answering in description logics: The DL-Lite family. JAR, 39(3): 385–429.

Calvanese D., G. De Giacomo, M. Lenzerini, D. Nardi, and R. Rosati (1998). Description logic framework for information integration. In Proc. of KR, pages 2–13.

Daraio C. (2017), A framework for the assessment of Research and its Impacts, *Journal of Data and Information Science*, Vol. 2 No. 4, 2017 pp 7–42.

Daraio C. (2019), Econometric approaches to the measurement of research productivity, in *Springer Handbook of Science and Technology Indicators* edited by Glänzel W., Moed H.F., Schmoch H. and Thelwall M., forthcoming.

Daraio C., Bonaccorsi A., (2017), Beyond university rankings? Generating new indicators on universities by linking data in open platforms, *Journal of the Association for Information Science and Technology*, DOI: 10.1002/asi.23679

Daraio C., Glänzel W. (2016), Grand Challenges in Data Integration. State of the Art and Future Perspectives: An Introduction, *Scientometrics*, 108 (1), 391-400.

Daraio C., Lenzerini M., Leporelli C., Moed H.F., Naggar P., Bonaccorsi A., Bartolucci A. (2015). Sapientia: The Ontology of Multi-Dimensional Research Assessment, in Salah, A.A., Y. Tonta, A.A. Akdag Salah, C. Sugimoto, U. Al (Eds.), Proceedings of ISSI 2015 Istanbul: 15th International Society of Scientometrics and Informetrics Conference, Istanbul, Turkey, 29 June to 3 July, 2015, Bogaziçi University Printhouse, pp. 965-977.

Daraio C., Lenzerini M., Leporelli C., Moed H.F., Naggar P., Bonaccorsi A., Bartolucci A. (2016a), Data Integration for Research and Innovation Policy: An Ontology-based Data Management Approach, *Scientometrics*, 106 (2), 857-871.

Daraio C., Lenzerini M., Leporelli C., Naggar P., Bonaccorsi A. Bartolucci A. (2016b), The advantages of an Ontology-based Data Management approach: openness, interoperability and data quality, *Scientometrics*, 108 (1), 441-455.

Lenzerini M. & Daraio C. (2019), Challenges, Approaches and Solutions in Data Integration for Research and Innovation, in *Springer Handbook of Science and Technology Indicators* edited by Glänzel W., Moed H.F., Schmoch H. and Thelwall M., forthcoming.

Lembo, D., Pantaleone, D., Santarelli, V., & Savo, D. F. (2016). Easy OWL Drawing with the Graphol Visual Ontology Language. In KR (pp. 573-576).

Cook, K.A., Thomas, J.J. (2005). Illuminating the path: The research and development agenda for visual analytics. *Tech. rep*., Pacific Northwest National Lab. (PNNL), Richland, WA (United States)

Keim, D., Andrienko, G., Fekete, J.D., Gorg, C., Kohlhammer, J., Melancon, G. (2008). *Visual Analytics: Definition, Process, and Challenges*. Springer Berlin Heidelberg, Berlin, Heidelberg, 154–175

Von Mayr, G. (1877) *Die gesetzmässigkeit im gesellschaftsleben*. Didemburg.