Applied Ontologies for Managing Graphic Resources in Spectroscopy

Nikolai Lavrentev, Alexey Privezentsev, and
Alexander Fazliev[0000-0003-2625-3156]
Institute of Atmospheric Optics SB, RAS, Tomsk 634055, Russia
lnick@iao.ru, remake@iao.ru, faz@iao.ru

Abstract The report presents the tasks on graphical resources management thoroughly describing applied ontologies of GrafOnto research graphics collection used for solving problems of spectroscopy. The problems of ontology modularity and automatic classes` generation are being discussed. Examples of solving reduction problem as well as applied ontologies metrics are presented.

Keywords Research Graphical Resources Classification, Spectroscopic Graphical Resources Ontology

1 Introduction

In the middle of 2000s the emergence of digital scientific libraries with publications as well as Semantic Web approach oriented on semantic description of information resources induced the work on decomposition of resources into smaller parts that require the creation of semantic annotations oriented on the description of domains and various data representations used in them. Various forms of data representation are always used in scientific publications (text, tables, graphics, symbols (for example, formulas), etc …). On the other side researcher got the facilities for storing and presenting large amounts of information, although published data and information was needed for the control of this information quality. Virtual data centers in various domains appeared in the second half of the 2000-th. These data centers usually contained the published data represented in publications in tabular form. In the end of 2000s publications on scientific graphical resources' systematization started to appear Ref. [1–4]. An example of an approach to creating a collection of graphical resources in High Energy Physics is presented in Ref. [5].

The report presents the results of the final stage of scientific plots' systematization in three disciplines of spectroscopy. At the first stage we formed GrafOnto collection of graphical resources [6–10] describing the results of studies on the problems of a water molecule spectral lines' continuum and on spectral properties of weakly bounded complexes and absorption cross-sections used for the photochemical reactions rates' calculation. At the second stage the typification of plots and figures as well as the first version of GrafOnto resources ontology was done (see Ref. [11–13]).

In order to upload new datasets into GrafOnto system and support them one has to solve the tasks on managing graphical resources. These are such tasks as specification of informational resources' structure for spectroscopy problems and analysis of re-

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-2523/paper12.pdf

sources' validity, control of data completeness and trust estimation. The decision support system which used in management of the collection GrafOnto is based on ontologies describing the primitive and composite plots and figures. Description of these ontologies is the aim of this report.

2 GrafOnto Collection of Scientific Graphics

The collection is based on a digital library, containing more than a thousand articles. These articles are dedicated to spectroscopy research such as spectral lines' continuum, weakly bound complexes' properties and spectral functions in near and far ultra-violet range. A distinctive feature of the above problems of spectroscopy is that the major part of published data is represented in a form of plots, figures and images.

In order to create a collection, graphical objects should be manually extracted and converted into a digital form. Software used to upload, storage, view, search and integrate graphical resources into collection is original. At present, the collection contains about 3000 primitive plots included into 625 composite plots and 104 composite figures as well as about 4000 primitive plots ready for the upload. The uploaded plots describe properties of 19 molecules, 25 complexes and 50 mixtures. Almost a half of primitive plots characterize properties of a water molecule. Collections' plots are related to dozens of physical quantities (functions) and a dozen of physical quantities (arguments). Table 1 illustrates spectral lines' collections and a number of primitive plots related to these functions for substance groups. It is worth noting that, at present, only a part of the plots from the publication chosen by experts is uploaded into the collection. Other plots will be processed automatically after the software for machine processing of graphical resources is developed. The collection of plots that has already been created will be used as a data set for training a neural network aimed at automatic recognition of scientific graphics.

Table 1. Number of plots with the functions that are most frequently used in the collection

Functions	The number of the primitive plots			
	Mixture	Molecule	Complex	Total
Absorption Coefficient (cm^{-1})	20	103	17	140
Absorption Coefficient ($cm^{-1}atm^{-1}$)	14	38		52
Absorption Coefficient ($cm^2mol^{-1}atm^{-1}$)	95	567	27	689
Absorption Coefficient (dB/km)	14	128	11	153

3 Managing Graphical Resources in GrafOnto Collection

The principal tasks of graphical resources management are to control resources structure and data quality. An ontology knowledge base accumulating all computer-generated information on collection components is used for making decisions during the management.

Resources structure contains plots of various types, their description, substances, functions and their arguments, physical quantities' units, units table as well as coordinate systems and level of detail of their description, etc. Control of plots and figures validity is based on the analysis of calculated values of paired relations between cited plots and original plots related to them. Such a relation is characterized by a reference to publication, figure number and an identifier of a curve. Note that, at present, the collection of cited plots contains 693 primitive and 248 composite plots. The ontology describing the present state of the collection resources is presented below.

4 Applied Ontologies of Scientific Plots and Figures in Spectroscopy

Taxonomy of some of the most important artifacts of research publishing [5] includes concepts: figure (composite figure, plot (exclusion area plot, GenericFunctionPlot, histogram), diagram, picture. In our work we defined additional concepts characterized by the methods of acquiring physical quantities (FTP, Cell, etc …) as well as their types (Theoretical, Experimental, Fitting, Asymptotic), slang names of physical quantities, etc. and declared them as subclasses o GenericFunctionPlot class. These definitions are oriented on physical quantities used as plots' axes.

We defined the following hierarchy for forming ontologies in spectroscopy domain. Basic ontology of spectroscopy graphical resources contains three parts and each part is related to one of the three problems of spectroscopy. These problems are the following: problems of continuum absorption, weakly bound complexes as well as the specific task of spectral functions related to photochemical reactions in the atmosphere.

4.1 Basic Ontology and Applied Ontologies of Domain Problems

Basic ontology contains some classes and properties, which are used in applied ontologies of domain problems. In our case, these problems are weakly related to each other and are represented by the following independent modules: graphical resources of continuum absorption, weakly bound complexes and absorption profiles, defining rate of photochemical reaction. Each of these modules is split into three parts: the first part characterizes coordinate systems used in GrafOnto collection, the second one characterizes physical quantities, while the third one characterizes the substances, the properties of which are presented in the collection.

4.2 Main Classes

In ontologies classes define many resources presented in our work in a form of plots and figures from GrafOnto collection as well as in a form of description of their properties. All the classes are explicitly defined in OWL 2 syntax with the use of Manchester syntax for their definition.

Basic ontology classes

In the framework of the chosen model the main entities in spectroscopy are substances (**Substance** class) – molecules as well as complexes and mixtures, and methods of acquiring physical quantities' values (**Method**). Graphical representation is related to graphical system entity (**GraphicalSystem**). The components of a graphical system are, for example, the coordinate axes of plots representing physical entities. In GrafOnto each published plot or figure is related to the description of its properties (**Description**, **ResearchPlotDescription** classes). One of such properties is a bibliographic reference to a publication (**Reference**). The **Problem** class contains three individuals (Continuum, Complex and CrossSection) each identifying a problem related to a graphical resource.

Classes related to domain problems' ontologies

Domain problems are closely related to the tasks for their solution. GrafOnto collection contains graphical resources related to the problems mentioned in the introductory abstract of this paragraph. Classes of spectroscopy problems' ontologies contain numerous resources and their description. **PhysicalQuantity** class consists of two non-adjacent subclasses named **SystemPhysicalQuantityDepended** and **SystemPhysicalQuantityIndepended**. The first class contains physical quantities the dependency of which on other physical quantities is presented in plots and figures, while the second one contains physical quantities the dependency of which is presented in plots and figures.

In order to understand the names of ontology classes we have to describe the etymology first. A name may consist of several words. These words correspond to the names of individuals in the corresponding classes **MethodType**, **SystemPhysicalQuantityDepended** and **Substance**. Fig. 1 presents examples of schemes for creating subclasses names in A classes (**Physical quantity and related substances**) and B classes (**Substance and related physical quantities**) presented in simplified syntax.

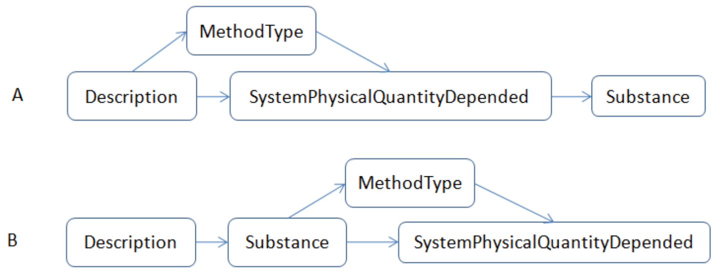

Fig. 1. Word order in the names of A and B groups' classes

For example, a class named **Description_Experimental_Absorption_Coefficient__cm2mol_1atm_1_** contains all the descriptions of measured absorption coefficients with $cm^2 molecule^{-1} atm^{-1}$ dimension for a series of substances being a subclass of **Physical quantity and related substances**. The third group of classes related to subclasses of **GraphicalSystem** class is not presented in this work.

4.3 Main Properties

Comments for all the properties used in natural language are presented in OWL 2 ontologies code. Here we present a simplified classification of some properties related to physical quantities and descriptions of plots and figures. Description of properties related to Description and CoordinateSystem classes as well as to Temperature and Pressure quantitative characteristics is omitted.

Table 2 lists ontology properties defining their domains and ranges. The last column of the table shows abbreviations of properties used in the scheme of individual presented in Fig. 2.

Qualitative properties characterized physical quantities are *hasOriginType*, *hasSourceType* and *hasMethodType*. The values of *hasOriginType* property indicate the origin of dataset related to the plot: it should be original and should be obtained by digitizing the curve of a primitive plot. The values of *hasSourceType* property can describe primary data, i.e. the data obtained by the authors of the publication as well as the previously published curves (i.e. cited) and commonly known curves (i.e. expert). The values of *hasMethodType* property characterize qualitative acquisition of datasets of primitive plot: Theoretical is a calculation using physical or mathematical model, Experimental is measurement, Fitting is a continuous curve creation using the method of fitting to experimental values.
The relations between plots and figures are defined by 6 properties (*has*{*OPPD, CPPD, OCPD, CCPD, MCPD*}, *hasPrototype*) . First five mereological properties describe composition of composite plots (*OPPD, CPPD*) and figures (*OCPD, CCPD, MCPD*). The value of *hasPrototype* property used in the description of cited primitive plot is the corresponding original plot. This property defines the descriptions that contain datasets with closely related values.

4.4 Main Types of Individuals

Being equivalents of figures and plots from published graphical resources on the above problems images generated in GrafOnto system are related to the description of their metadata making the most significant part of ontology individuals included in A-box. Typification of figures and plots given in Ref. [14] is defined by the property values. Abbreviation of corresponding values is used in the names of such individuals (for example, OCP – Original Composite Plot). Fig. 2 illustrates the structure of one of such plot types, i.e. original primitive plot. Ovals stand for ontology individuals, rectangles stand for literals and directed arcs stand for objective (OP) and determined (datatype – DTP) properties. Cited primitive plot have a similar structure with an addition of observations with hasPrototype, hasChild and hasParent properties. Special cases of individuals characterizing properties of coordinate system and its axes are shown in the lower part of Fig. 2. A series of individuals are related to the classes defined by enumeration of its individuals.

Table 2. Main properties of base ontology and spectroscopy problems ontology

Primitive Plot Description (PPD)			
Domain	Property	Range	Abbr
Description	hasReference	Reference	OP1
PrimitivePlotDescription	hasSubstance	Substance	OP2
PrimitivePlotDescription	hasSourceType	{Primary, Expert, Cited}	OP3
PrimitivePlotDescription	hasOriginType	{Digitized, Original}	OP4
PrimitivePlotDescription	hasCurveType	{Line, Point}	OP5
PrimitivePlotDescription	hasCS	CoordinateSystem	OP6
CitedPrimitivePlotDescription	hasCitedReference	Reference	OP7
CoordinateSystem	hasCSType	{2D-Decartes}	OP8
CoordinateSystem	hasX-axis	X-axis	OP9
CoordinateSystem	hasY-axis	Y-axis	OP10
Y-axis	hasMethod	Method	OP11
Y-axis	hasMethodType	{Theory, Experiment, Fitting}	OP12
Y-axis	hasPY-axis	PubPhysQuanDepended	OP13
Y-axis	hasSY-axis	SysPhysQuanDepended	OP14
X-axis or Y-axis	hasAxisScale	{Linear, Logarithmic}	OP15
X-axis	hasPX-axis	PubPhysQuanIndep	OP16
X-axis	hasSX-axis	SysPhysQuanIndep	OP17
CitedPrimitivePlotDescription	hasPrototype	OriginalPrimitive-PlotDescription	OP18
PrimitivePlotDescription	hasTemperature	float	DT1
PrimitivePlotDescription	hasPressure	float	DT2
PrimitivePlotDescription	hasSystemFigureNumber	integer	DT3
ResearchFigureDescription	hasOriginalImageOfPlot	URI	DT4
ResearchFigureDescription	hasOriginalPlotInformation	URI	DT5
FigureDescription	hasFigureCaption	string	DT6
FigureDescription	isPartOfFigureNumber	integer	DT7
ResearchFigureDescription	hasNumberOf Points	integer	DT9
ResearchFigureDescription	hasPlotCaption	string	DT12
Original Composite Plot Description (OCPD)			
OriginalCompositePlotDescrip-tion	hasOPPD	OriginalPrimitivePlotDescription	Op19
Cited Composite Plot Description (CCPD)			
CitedCompositePlotDescription	hasCPPD	PrimitivePlotDescription	Op20
Composite Figure Description (CFD)			
CompositeFigureDescription	hasOCPD	OriginalCompositePlotDescription	Op21
CompositeFigureDescription	hasCCPD	CitedCompositePlotDescription	Op22
CompositeFigureDescription	hasMCPD	MultipaperCompositePlotDescription	Op23

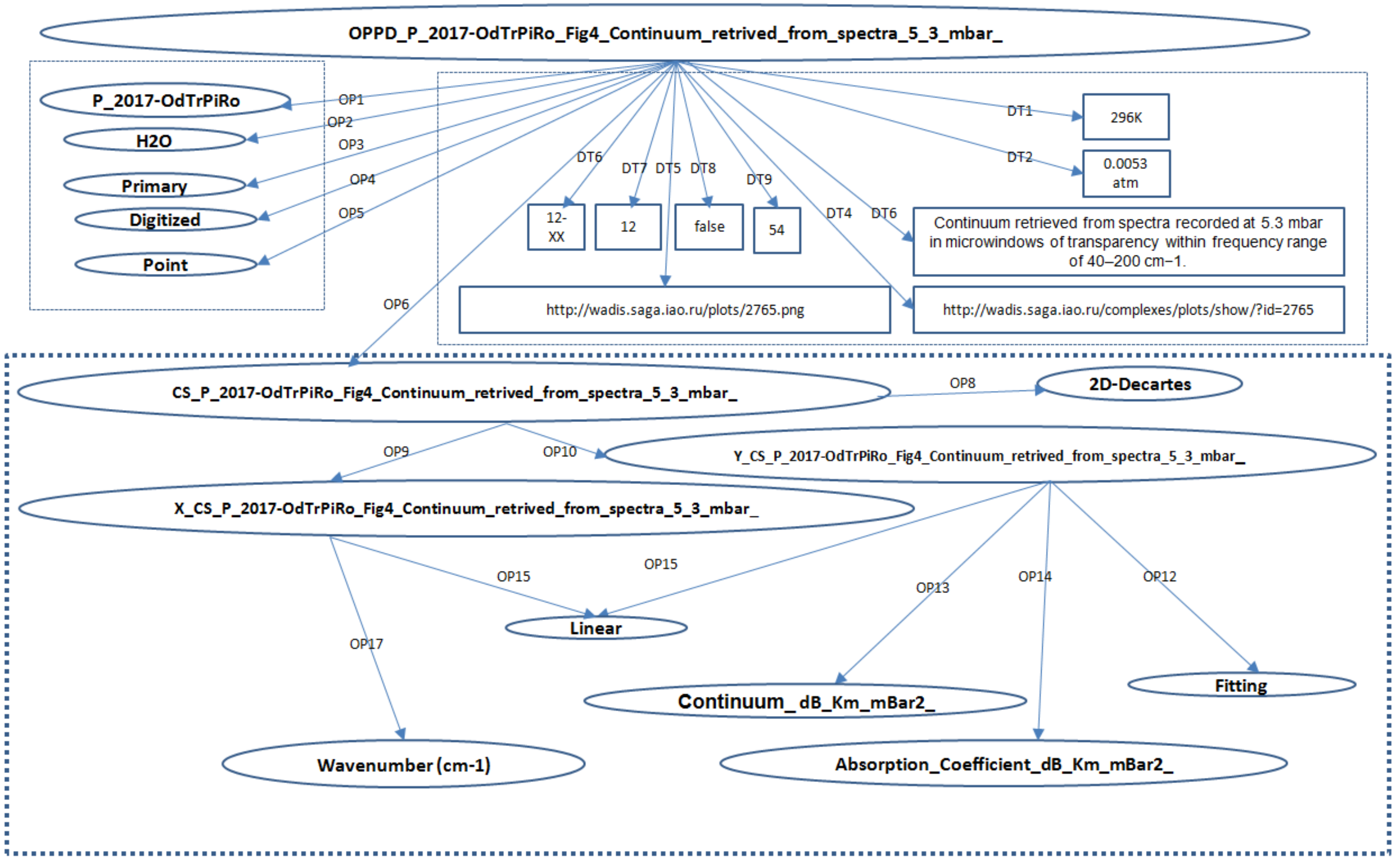

Fig. 2. Structure of individual characterizing original primitive plot from Fig. 4 in Ref [14]

4.5 Ontologies Metrics

Ontologies metrics are used for comparing ontologies of different parts of a domain or of different domains, characterizing quantitative and qualitative peculiarities of ontological description. In OWL ontologies the number of object properties characterizes the number of paired relations between individuals. Some of these individuals may have quantitative estimation. The estimated relations are described by certain (datatype) properties.

Table 3 contains metrics for applied ontologies of three spectroscopy problems as well as the unification of these ontologies (Σ Ontology). As for GrafOnto resources collection the equality of numbers characterizing the number of properties, their domains and ranges means that they are characterized by identical properties. However, the difference in classes` numbers indicates the use of a greater number of spectral functions in Continuum problem in comparison with Complex and Cross Section problems. As individuals of one and the same group of types are used applied ontologies we may conclude that ontology on Continuum problem describe the highest number of primitive plots.

Table 3. Ontology metrics, characterizing the spectroscopy graphical resources collection

	Continuum	Complex	Cross Section	Σ Ontology
Metrics				
Axiom	53279	12737	10704	84662
Logical axiom count	43180	10171	8634	68587
Declaration axiom count	8187	2134	1715	13135
Class count	295	205	55	552
Object properties count	24	24	24	24
Datatype properties count	14	14	14	14
Individual count	7910	1984	1641	12578
DL expressivity	ALCO(D)	ALCO(D)	ALCO(D)	ALCO(D)
Object property axioms				
Object properties domains	24	24	24	24
Object properties ranges	24	24	24	24
Datatype property axioms				
Datatype properties domains	14	14	14	14
Datatype properties ranges	14	14	14	14
Individual axioms				
Class assertion	426	204	206	854
Object properties assertion	27971	6231	5590	44152
Datatype properties assertion	14204	3314	2683	22516
Annotation axioms				
Annotation assertions	1912	432	355	2940

Metrics comparison of Σ Ontology with ontologies of tabular information resources Ref. [14] reveals that in our work on graphical resources ontology we managed to significantly increase the number of classes in one year. It clearly indicates that Σ Ontology contains the highest number of obvious answers on typical user requests.

5 Conclusion

The report presents applied ontologies of scientific plots and figures used for managing graphical resources in three problems of spectroscopy. Ontologies describe a collection of plots and figures published in the period from 1918 till 2018. Ontologies are created for managing structure of collection resources as well as for making decisions on such tasks as development, storage and systematization of plots and figures for solving such problems as continuum absorption and research of properties of weakly related complexes and cross sections absorption. Ontology as well as its individuals and classes are automatically generated with the enlargement of the collection.

The future of GrafOnto collection is related to automatic recognition of plots and figures used in spectroscopy as well as to the generation of applied ontologies characterizing validity analysis and confidence estimation of its resources.

References

1. Halpin, H. and Presutti, V.: An ontology of resources for linked data. Linked Data on the Web 2009, Madrid, Spain. ACM 978-1-60558-487-4/09/04
2. Thorsen, H. and Pattuelli, C.M.: Ontologies in the time of linked data. In Smiraglia, Richard P., ed. Proceedings from North American Symposium on Knowledge Organization **5**, 1–15 (2015).
3. Niknam, M. and Kemke, C., Modeling shapes and graphics concepts in an ontology. https://pdfs.semanticscholar.org/c20b/3b819ce253715bbfa9c2151a10ea87f718e4.pdf
4. Kalogerakis, E., Christodoulakis, S., and Moumoutzis, N.: coupling ontologies with graphics content for knowledge driven visualization. https://people.cs.umass.edu/~kalo/papers/graphicsOntologies/graphicsOntologies.pdf
5. Praczyk, P.A.: Management of scientific images: an approach to the extraction, annotation and retrieval of figures in the field of High Energy Physics. Thesis Doctoral, Universidad de Zaragoza (2013). ISSN 2254-7606
6. Voronina, Yu.V., Lavrentiev, N.A., Privezentzev, A.I., and Fazliev, A.Z: Collection of published plots on water vapor absorption cross sections. Proc. SPIE **10833** (2018). doi: 10.1117/12.2504586s
7. Lavrentiev, N.A., Rodimova, O.B., Fazliev, A.Z., and Vigasin, A.A.: Systematization of published research plots in spectroscopy of weakly bounded complexes of molecular oxygen and nitrogen. Proc. SPIE **10833** (2018). doi: 10.1117/12.2504327
8. Lavrentiev, N.A., Rodimova, O.B., and Fazliev, A.Z.: Systematization of published scientific graphics characterizing the water vapor continuum absorption: I. Publications of 1898–1980. Proc. SPIE **10833** (2018). doi: 10.1117/12.2504325
9. Lavrentiev, N.A., Rodimova, O.B., Fazliev, A.Z., and Vigasin A.A.: Systematization of published research graphics characterizing weakly bound molecular complexes with carbon dioxide. Proc. SPIE 104660E (2017). doi: 10.1117/12.2289932

10. Lavrentiev, N.A., Rodimova, O.B., and Fazliev, A.Z.: Systematization of graphically plotted published spectral functions of weakly bound water complexes. Proc. SPIE **10035** (2016). doi: 10.1117/12.2249159
11. Lavrentiev, N.A., Privezentsev, A.I., and Fazliev, A.Z.: Tabular and Graphic Resources in Quantitative Spectroscopy. In: L. Kalinichenko et al. (eds.) DAMDID/RCDL 2018, CCIS
12. Lavrentiev, N.A., Privezentsev, A.I., and Fazliev, A.Z.: Systematization of Tabular and Graphical Resources in Quantitative Spectroscopy. CEUR Workshop Proceedings, Selected Papers of the XX International Conference on Data Analytics and Management in Data Intensive Domains. Edited by Leonid Kalinichenko, Yannis Manolopoulos, Sergey Stupnikov, Nikolay Skvortsov, Vladimir Sukhomlin **2277**, 25–32 (2018).
13. Lavrentiev, N.A., Privezentsev, A.I., and Fazliev, A.Z.: Applied Ontology of Molecule Spectroscopy Scientific Plots. Proc. of Conference "Knowledge, Ontologies, Theories", DigitPro **2**, 36–40 (2017).
14. Odintsova, T.A., Tretyakov, M.Yu., Pirali, O., and Roy, P.: Water vapor continuum in the range of rotational spectrum of H_2O molecule: New experimental data and their comparative analysis. Journal of Quantitative Spectroscopy and Radiative Transfer **187**, 116–123 (2017). doi: 10.1016/j.jqsrt.2016.09.00