Scientific Visualization, 2023, volume 15, number 5, pages 50 - 63, DOI: 10.26583/sv.15.5.05

Using Data Fabric Architecture to Create Personalized Visual Analytics Systems in the Field of Digital Medicine

S.I. Chuprina[1]

Perm State University, Perm, Russia

[1] ORCID: 0000-0002-2103-3771, chuprinas@inbox.ru

Abstract

The paper describes an innovative approach of applying data factory technologies, which have found their successful practical application to create new generations of corporate data warehouses, are proposed to be used to develop personalized visual analytics environments. The approach provides access to digital data of a particular person from both heterogeneous distributed external and local sources not on the principles of their consolidation, but on the principles of virtual integration. The application of the described approach is demonstrated by the example of the development of visual analytics tools in the field of digital medicine. The emphasis is on the technological aspects of applying the concept of data factories for the development of personalized visual analytics services, means of semantic integration of heterogeneous data sources and their adaptation not only to heterogeneous structured and unstructured data sources, but also to the semantic content of various subject areas.

Keywords: Data Fabric, data federation, ontology engineering, Semantic Web, semantic integration, knowledge graph, 4P medicine, visual analytics.

1. Introduction

In recent years, modern approaches to creating a new generation of data warehouses (DW) use Data Fabric (DF) architecture and have been actively developing [1]. To integrate data from heterogeneous structured or unstructured sources, both Internet and corporate ones, as well as local resources of enterprises or clients DF technology leverages methods, which are performed not on the principles of consolidation, but on the principles of virtual data integration (federalization) without physically moving them to a single data storage. So, DF provides a unified view of all enterprise data without data centralization and instead of replacing or rebuilding the existing infrastructure DF adds a new, an integration and orchestration layer built on top of multiple heterogeneous data sources like text collections, web resources, relational databases, data warehouses, data lakes, IoT, legacy systems, etc. This abstract layer is a knowledge graph that may be implemented as machine learning (ML) and/or ontology-powered resource.

The knowledge graph describes concepts and relationships in terms of the subject area of data sources, which should be integrated, as well as metadata about them, including both information about the storage format and the native instrumental environment, and the specification of their semantic and pragmatic content, as well as information about synonyms (considering the context) and abbreviations.

In this paper, we propose the using of data factory technology to create personal, not just corporate, visual analytics systems in the field of digital medicine. Tuning to the specifics of subject areas and data sources is carried out thanks to an ontological layer of knowledge describing the semantic content of data and semantic relationships between them in terms of established concepts of the relevant subject area. Because volume of personal data sources and velocity of their updating are not comparative with Big Data ones we suggest using visual

ontology editors to create ontology layer and if its necessary use ML methods to design of ontology by means of ontology learning methods.

The presented approach makes it possible to perform semantic integration of data from heterogeneous sources, unify the means of their processing and analysis, significantly simplify access to end–user data, providing the user the opportunity to uniformly access medical and other personal data stored in different places and in different formats – in the form of familiar queries in natural language (NL) or languages by analogy with queries in search engines of Internet services. This greatly contributes to the successful implementation in practice of all the main components of the concept of 4P medicine [2], which are:

1. Personalization (individual approach to each patient).
2. Prediction (creating a probabilistic health forecast).
3. Prevention (prevention of the development of diseases).
4. Participativeness (motivated participation of the patient).

As shown below, the using of data factory technology to create modern visual analytics systems in the field of digital medicine and personal, not just corporate, data factories is well justified.

The paper's content is deliberately presented from the standpoint of "human-centricity", and the emphasis is not on the person of the doctor, but on the person of the patient ("patient-centricity" [2]) and the availability at his/her request of all accessible data associated with the person from different sources at the right time and in the right place location ("personal patient data factory"). From a technological point of view, there is no difference between the described approach "personal patient data factory" and the "personal data factory of a doctor", which is based on semantic integration of heterogeneous data about patients of a particular doctor for full-fledged analytics and medical decision-making.

It should be noted that serious consideration of information security issues in accordance with the concept of the proposed approach is the subject of a separate discussion and out of the scope of this paper.

2.The Rationale of Using DF Technology to Implement the Principles of 4P Medicine

Gartner identifies data fabric implementation as one of the top strategic technology trends for 2022 and expects that by 2024, data fabric deployments will increase the efficiency of data use while halving human-driven data management tasks (read more at: https://www.cxotoday.com/hardware-software-development/data-fabric-architecture-is-key-to-modernizing-data-management/).

A data factory can be defined as a virtual area with an integrated development environment (IDE), the architecture and services of which provide a semantic relationship between data stored in heterogeneous sources, regardless of the format of their storage and the specifics of native data source systems. Source data remains in its original place, and DF is able to deliver it to the right place as a virtual view and at the right time of "on-demand access", performing the necessary preprocessing, aggregation and analytics on demand.

To tackle the problems of integrating heterogeneous data and their visual analytics, modern intelligent data factories are designed and implemented mainly on the principles of model–driven architecture (MDA – Model Driven Architecture). In addition to models based on UML and domain-specific languages, ontology models have been increasingly used for these purposes lately. As a tool environment for creating data factories, you can use existing domestic solutions, in particular, the Russian DataFabric KGL platform (http://datafabric.cc/), created by a resident of Skolkovo LLC "DATAFABRIC".

As noted in the review of the analytical agency TAdviser (https://www.tadviser.ru/index.php/ Статья:AI:_ot_dannyh_–_known), DataFabric has created a universal corporate ontology platform DataFabric KGL, which allows you to unify access to all data in the enterprise using an open source data virtualization platform. The Da-

taFabric KGL (or Logical DW) of an enterprise is implemented on the basis of a graph (semantic) data model in terms of the corresponding domain. It allows federally accessing heterogeneous data from different sources without the need for their preliminary collection, aggregation and storage: data is accessed in the terminology of the subject area through an abstraction layer in the form of a business glossary and does not require performing any operations at the physical level of data storage.

Let's look at the problems of distributed data processing in the field of digital medicine. It is no secret that the human body intensively generates large amounts of data across the lifecourse, which have one of the most important characteristics of Big Data – "Variety" (diversity, heterogeneity). Another thing is that not all of this data is regularly digitized and stored, and the person usually has problems accessing his/her own personal digital data stored in different environments. For example, even the data available to the patient in the personal account of a concrete Medical Information System (MIS), firstly, is actually not fully available, which prevents the implementation of the principles of prediction and prevention, if operational analysis of the data outside the framework of this MIS is required. Secondly, personal data reuse within the environment of third-party MIS of other public or private medical institutions, as well as available analytical platforms for both individual researchers and large medical institutions are often impossible or very difficult.

The difficulties are explained by the fact that it requires not just data conversion due to differences in formats, but automated preprocessing and semantic normalization, aggregation and transformation of heterogeneous data. These problems are solved in the environment of traditional DW, built on the principles of consolidation, however, it requires the organization of an ongoing management process of DW in order to synchronize updates between sources and DW, otherwise there is a problem of "freshness" of data. Such ongoing maintenance is costly not only for individuals, but also for entire organizations, although recently there is hope that the expansion of domestic cloud storage services will lead to a reduction in the cost of their services. Nevertheless, it is important to note that if we are talking about personal, and not corporate, DW, then the need to provide data for analytics is "on demand" (and not ongoing) comes first. Along with retrospective personal medical data, the most up-to-date information should be available for monitoring and analysis, which can be stored both in third-party MIS databases and on the local personal computer (for example, the results of laboratory tests in pdf format, which have been obtained on the initiative of the patient himself without consulting a doctor).

We believe that any person should rightfully have access to most of his or her personal digital data, including the digital data from third-party storage systems with the possibility of their virtual integration in the process of analysis with data stored in different formats on the personal computer. This is necessary for health monitoring, operational analysis and decision-making, for example, to urgently consult a doctor or reasonably promptly adjust your diet. It seems that there can be no question of any real participativeness without the person being the rightful owner of the own digitized personal data with the right to promptly provide the data to third parties, for example, employees of sanatorium-resort institutions or ambulance doctors.

In terms of volume, the majority of data streams over 100 GB per day can be classified as Big Data, which is not always typical for most personal medical data at the current stage of IT technology development. However, over time, with the development of IoT and Ubiquitous Computing as well as improvement, growth in demand and availability of wearable personal devices, volume of the digital footprint of an increasing number of people will be able to grow to the volume of Big Data. In addition to purely medical data, so-called "health diaries", which are kept by the patient himself locally on the personal computer must be considered for the purposes of analytics. The patient's personal data can also include data from various questionnaires about environmental, psychological and social living conditions, lifestyle and human activity, including professional factors. To semantic integration and analysis all of mentioned heterogeneous data must be preprocessed automatically.

The second main characteristic of Big Data, Velocity, implies not only the speed of data generation/growth/change, which is also not typical for personal medical data yet, but also the need for their high-speed processing, which is essential for the implementation of all the basic principles of 4P medicine.

As for the third main characteristic, Variety, where diversity is understood as the heterogeneity of data and the need for simultaneous and unified processing of various types of both structured, semi-structured and unstructured data, it is fully inherent in the specifics of a personal data and corresponds to the goals, objectives and key features of building data factories. It is important to note that this third characteristic has been widely recognized as the main driver of the development of modern Big Data technologies.

To emphasize the validity of the use of DF technologies for the analysis of personal data to solve the problems of 4P medicine, we note the similarity of their goals and objectives, which are the automatic extraction of new useful information and knowledge from the source data, including previously unknown patterns in order to automatically solve analytical problems to support evidence-based decision making and make predictions to prevent some risks, in particular. Modern DF technologies are the best suited for monitoring and integrating the patient's fresh personal data with previously received results of predictions in order to assess health risks and prevent the development of diseases. The methods and tools used in DF, in particular, the methods and tools of the Semantic Web [1] and Visual Analytics do not require the construction and maintenance of a new data warehouse based on the principles of consolidation.

3.The Semantic Integration Concept Based on the Principles of Federalization

It is known that data integration systems are capable of providing integration at various levels: physical, logical and semantic. Integration of data at the physical level usually comes down to converting data from various sources into some single format of their physical representation. Data integration at the logical level provides the ability to access data from various sources in terms of a single global schema that describes their joint representation considering structural and behavioral (in the case of object models) properties without considering the semantic properties of the data. Data integration at the semantic level provides support for a unified representation of data based on their semantic properties in the context of a common knowledge graph about the related subject area. The end-user interface hides all the technical aspects of storing data sources such as location, storage format, structure (if the data source is structured) and others.

Figure 1 shows a simplified diagram of two approaches to data integration for comparison: the traditional one based on the principles of consolidation (ETL – Extract, Transform, Load) and the virtual one based on the principles of federalization. In BI (Business Intelligence) applications and OLAP systems, in the ETL process, all the necessary data available from different sources are loaded into a single datastore even before users start requesting them. Before loading, to bring heterogeneous data to a single representation, which is often called "normalization" of data, DW use a global reference data schema. As a result of ETL processes, physical and logical data integration is provided. After uploading data to a single repository, access to sources is not required, however, in corporate DW there is a problem of "freshness" of data.

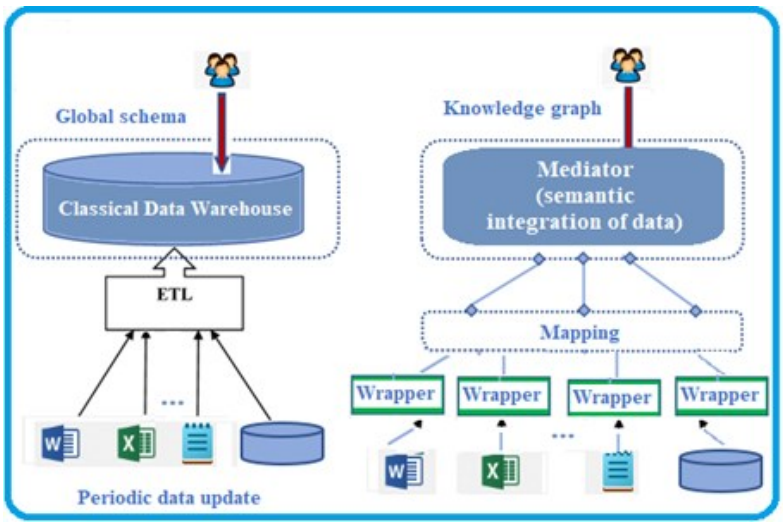

Figure 1: Data integration schemes based on the principles of ETL consolidation (left) and federalization (right)

Unlike consolidation, federalization allows you to extract data from various sources, combine them and perform analytics in real time without the need for physical movement: the data remains with their owners, and access to them is carried out on demand, i.e. in response to a request when there is a need in the analysis process. Thus, federalization supports a single virtual space for a variety of heterogeneous data sources, ensures their semantic integration and does not require the cost of creating and maintaining a single physical DW. The user formulates NL query in terms of the knowledge graph, which allows for semantic search and semantic integration of data, considering the description of semantic relationships between data elements presented in the knowledge graph. All necessary data transformations are carried out in the process of extracting them from sources. It is important to note that virtual integration helps to ensure the rules of the security policy and license restrictions, if direct copying of data from source systems is prohibited.

Logically, virtualization is carried out due to an additional intermediate layer that isolates the physical storage of data from applications. The central component of the integration system based on the principles of federalization is the so-called mediator, which integrates data received from adapters (wrappers). Adapters are components that ensure uniform interaction of the intermediary with data sources (in terms of a single model). The intermediary supports a unified user interface based on a global knowledge graph describing metadata and semantic content of data across various sources, and also supports mapping between global and local representations of data. A user query formulated by means of a unified interface is automatically decomposed into a set of subqueries addressed to the necessary data sources. Based on the results of their processing, a complete response to the request is synthesized as a virtual view on the personal computing device.

It is thanks to the architecture described above that data factories are able to implement semantic integration based on the principles of federalization. If we talk about the possible disadvantages of DF, they may be associated with additional costs for access to numerous data sources, especially if they are large, which affects the performance of analytical services. However, as noted above, two of the three main characteristics of Big Data (Volume and Velocity) are not the main features of the patients' personal data. In addition, the semantic coherence of data from different sources is guaranteed, at least, by their belonging to the same person that helps to eliminate the specified problem.

Approaches to the implementation of systems of semantic integration of heterogeneous data based on intermediaries (mediators) in the early 90s were based on the use of relational data models: the main metamodel was ODMG–93; then from the end of 90s to 2004-2005, ontologies were mainly used for these purposes: the main metamodel was ontology description language; since 2004, the focus had shifted to application of Semantic Web technologies

and RDF as the main meta–model. The Semantic Web has raised the role of ontologies to a new, one might say, "worldwide" level, because they play a key role in the automatic semantic markup and integration of data from different Internet resources, regardless of which natural language a particular web resource is presented in.

The development of Semantic Web technologies made virtual integration more promising. Although it should be noted that recently there has been a convergence of different approaches to data integration [3]. When implementing modern corporate data factories, integration based on the principles of federalization is used, where ontologies act as a knowledge graph. It is this approach that we propose to use, expanding its application to support the principles of 4P medicine and the creation of patient personal data factories, which enabled analytical services based on the semantic integration of heterogeneous personal data.

4.Technological Aspects of Building Personal DF for Visual Analytics Systems

As noted above, data factories are built on the principles of semantic virtual integration of heterogeneous data sources based on their consistent model, presented in the form of a knowledge graph (in our case, in the form of ontology). DF services allow the user to pose a natural language query (NL query) to the entire virtual data space in terms of this knowledge graph, without caring about the place of data real storage. And because this data can be stored in different formats (plain text, database, web resource, etc.) then for their unified processing by means of Natural Language Processing (NLP) methods it is necessary storing not only data semantics description, synonyms (considering the context) and abbreviations but also the metadata (including data about storage format and native data environment).

Since we are discussing the creation of not corporate, but personal DF with an emphasis on the field of medicine, an important aspect of our approach lies in the possibility of the semantic integration personal data and results of data analysis stored in a variety of information system environments, for example, public or private MIS, with the person-related data from private or public cloud (for example, with personal data generated by wearable devices) and from personal files stored locally (on user's personal computer, on flash media, etc.). In the last case, local personal files usually have text or spreadsheet format and contain the personal data about monitoring blood pressure, daily diet, physical activity, etc. On the one hand, the processing of medical data is complicated by the fact that in this area there are many terms and abbreviations that are not generally accepted, but well-established and constantly circulating in this or that community of professionals. Moreover, different concepts may correspond to the same designations. For example, "CP" can mean "Chest Pain" when administered by a cardiologist or ambulance physician, but it can also mean "Cerebral Palsy" when administered by a neurologist or pediatrician. And on the other hand, the processing of unstructured local personal files is complicated by the fact that a person can use household, not professional vocabulary.

We use traditional NLP services, which utilized ML algorithms to identify and correct data quality issues (such as missing values, duplicates, and inconsistencies). Because the complexity of medical data processing, we integrate the using of traditional ML tools, methods and tools of ontology engineering and open NL resources such as lexico-syntactic patterns, medical digital dictionaries and reference books as well as our own ones to automate mapping and intelligent transformation of data when integrating.

Based on our experience of prior developments (for example, [4]) and using SciVi visual analytics platform as IDE [5-6] as well as the collaborative projects with BIOGENOM LLC [7], we propose to use not one single large ontology as a knowledge graph, but a coherent set (family) of ontologies related to a particular domain, which represents the single global integrated schema (called mediated schema) of underlying sources. This set of ontology has a common model, which is described below. Ontologies are stored in the so-called smart repository managed by a meta-ontology containing metadata about the ontologies themselves

(date, owners, location, binding to the category of described data such as "Group of diseases", "Medical records", "Questionnaires", "Reference Books", "Health Diary", etc.). Meta-ontology is used for semantic indexing when searching and mapping ontologies. Various methods and means of ontology integration are well studied and actively used in practice [8].

Figure 2 shows a fragment of ontology of personal data that acts as a knowledge graph layer providing the semantic integration of heterogeneous data (some related data has a database format, but the other ones have text or xml format). As the system presents one global integrated schema, users enable pose their queries in terms of the corresponding application domain.

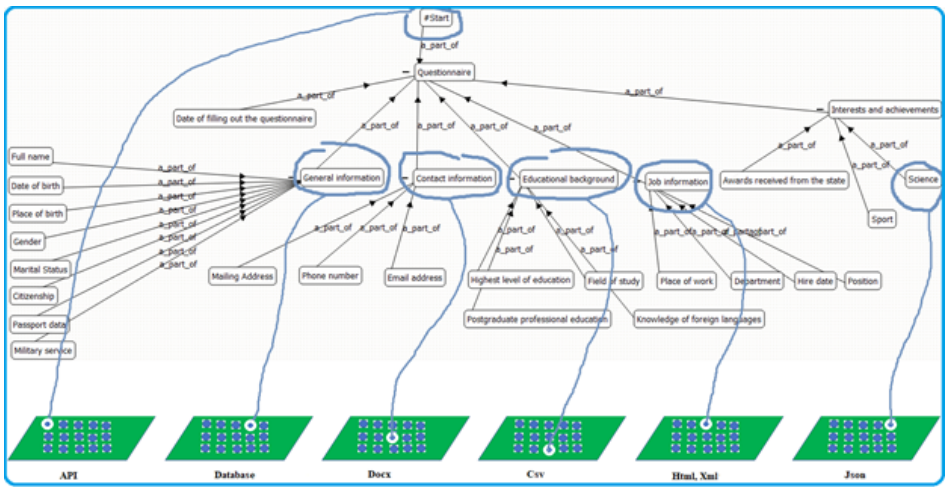

Figure 2: Fragment of the personal data ontology layer providing the semantic integration of heterogeneous data

According to DF virtual integration approach we use the solution that utilizes a wrapper layer known from mediated systems to provide data semantic integration (see Figure 1). Firstly, mediator's mechanisms find the corresponding ontologies in the repository and, if necessary, establish a semantic correspondence between them, which is stored in the form of a mapping specification based on the principles of OBDA mapping [9]. Such correspondence in the case of the ontologies remain unchanged is established one time and stored in the repository as a metadata for a further use. If some ontology has been subjected to change, the special mechanism monitors this fact and automatically starts re-mapping.

Then, the mediator uses adapters to organize access to the necessary data and their aggregation in response to a request. The user's query is decomposed into sub-queries to individual sources based on their descriptions. These sub-queries are sent to the wrappers of individual sources, which will execute them over local models and schemas of sources. The mediator receives answers from wrappers, combines them into one answer, and sends it to the user as a virtual view. You can find the typical architecture of a mediation system and more detailed description in [10].

To design the ontology, we utilize our own visual ontology editor named ONTOLIS [4, 11-12] that unlike most other ontology editors focuses not only on knowledge engineers but also on untrained users. This visual editor itself is also implemented as an ontology driven solution and makes it easy to adapt the visualization tools provided by it to the individual preferences of end-users. ONTOLIS has been developed to design so-called lightweight ontologies [13].

Lightweight ontology represents a thesaurus of domain concepts and a basic set of supported types of relations between them, but does not include axiomatics. That is, axiomatics is not explicitly specified as a separate component of the ontology model, and ONTOLIS has no tools to define it. But in ONTOLIS, any node or arc of an ontology graph can have attributes, which provide a way to store additional data associated with the nodes and the arcs in a graph. We use this opportunity to uniform interpreting the semantics of various types of rela-

tionships represented by arcs between concepts represented by nodes during logical inference by means of an ontology-driven mechanism implementation.

In accordance with our approach, the personalized visual analytics system under development includes a special uniform reasoner mechanism to interpret the specification of data stored in the attributes. For this, all data specifications, which can be stored in the attributes, are categorized and belong to some category. When the reasoner finds the keyword, which represents concrete category the interpretation mechanism is running and the corresponding functions are called automatically from so-called smart repository. The set of categories is expandable. To expand this set it is necessary to add new interpretation function with corresponding metadata to smart repository.

All specifications stored in the attributes have a json-like format, which allows a unified way to describe different actions across reasoning such as calling a library function to interpret the semantics of a certain type of relationship and others according metadata categories of attributes. Specification of the category begins with unique identifier, for example, the path to the function is specified after the keyword "path", and the “def” is used for denoting the definition of a concept represented by some node. When the reasoner bypasses the ontology graph and find a non-empty attribute of the current node or arc, it interprets the metadata specification in accordance with corresponding category. For example, if the "path" specification is presented in the attribute description, then the corresponding function from the system repository is automatically executed and specified arguments of the function are extracted from the ontology.

Figure 3 depicts fragment of the ontology for assessing the risk of atherosclerosis (arcs have no name because they represent the same type – causal relationship). This ontology was also created within the ONTOLIS environment. Because one of the important tasks of any ontology driven solution is to validate the ontology we demonstrate here how reasoner uses the attribute properties of ontology nodes and arcs storing the specification of the path to the needed functions and models to run the dialog form providing domain expert to evaluate the validity of relations established between different factors of risk and the correctness of models evaluating the common effect of the different factors to the same risk. Because intellectual property protection, the Figure 3 does not show formulas for calculating the assessment of the joint influence to the same risk of several factors at once.

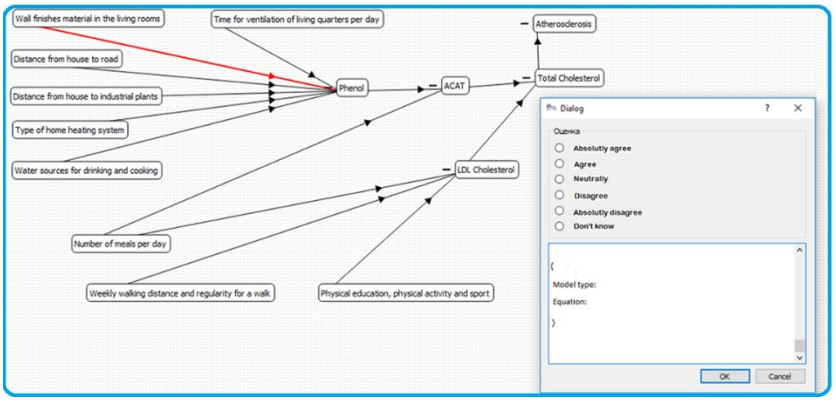

Figure 3: Fragment of the knowledge graph for assessing the risk of atherosclerosis using the attribute properties of nodes and arcs of ontology

Typically, a reasoner is designed as a program that infers logical consequences from a set of explicitly asserted facts or axioms to automated support the reasoning to tackle such tasks as classification, debugging and querying. In our case, the reasoner automatically interprets all different attribute specifications (including those that interpret the semantics of relationships and different types of restrictions) while all the elements of the ontology are accessible by navigating through the ontology graph during the inference engine performance. So, in some sense, our approach to implementing inference engine based on automatic procedural

interpretation of categorized attributes of ontology graph acts the same role as the interpreting of axioms, which are represented in declarative form and used across the different solutions based on heavyweight ontologies.

Accessing other services of the system is performed similarly. For example, when solving some analytical problem, the reasoner can find related specifications in the attributes of nodes or arcs of the current applied ontology for launching concrete machine learning methods and the corresponding library functions stored in the system repository are called. Despite the fact that our approach assumes, if necessary, using a whole family of interrelated ontologies instead of single one, all of them have a common model, which allows to perform the reasoning different ontologies in a uniform way. In particular, the same function is used for different ontologies to interpret "a_part_of" ("part-whole") type relationship. This type of relationships you can see in Figure 2. This figure illustrates also that the attributes of the node "start" store the specification an API call to launch the corresponding interpreter function.

The common ontology model is described below. We use the such model to implement ontology driven visual analysis tools that are easily adaptable to such personal preferences as:

— data source and its location;

— type of graph that displays the result of data analysis;

— data structure (description of the correspondence of data structure fields to graph elements indicating which data measurements are postponed along the axes of abscissa, ordinate and application, if the graph is three-dimensional, what is the displayed interval, etc.);

— color scheme used for rendering.

Fragment of the lightweight ontology describing the software entities utilized for visualization, their properties and using, as well as corresponding data sources is demonstrated in Figure 4 (you can see nodes representing csv files with statistical data on the growth of diseases in China and the USA indicated as data sources). The set of domain concepts in the model of this ontology includes such root concepts as "Data Source" (data source), "Data Field" (data field), "Chart" (chart), "Chart View" (chart plotting area), "Color" (color), etc., part of the taxonomy of chart types in NChart3D and the instances of data sources and charts. The represented ontology model has been successfully used to ontology driven visual data analysis tools development in different real-world projects for a long time.

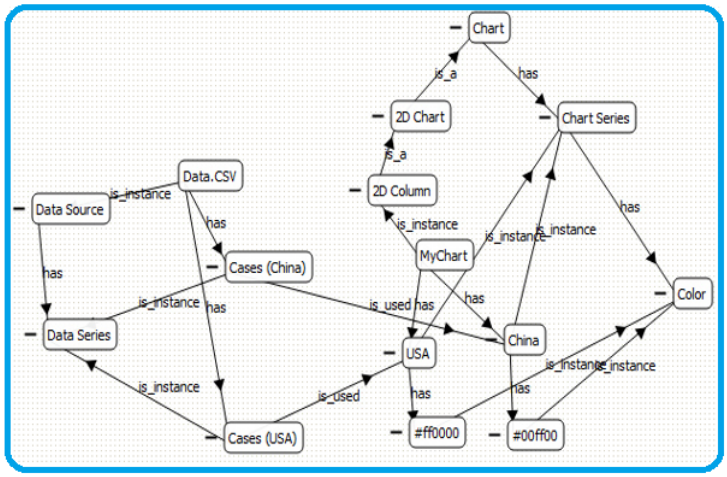

Figure 4: Fragment of the applied ontology for ontology-driven adaptable visual analytics tools development

The model supports the following set of types of relationships between concepts:

1. "is_a" ("subclass–class") describes the relationship between child and parent concepts and support an inheritance of properties of the parent class.
2. "has"– describes one concept as a property or attribute of another one.

3. "is_instance" describes the relationship of belonging of some instance (individual) to some class ("instance-class" relationship).

4. "is_used" describes the using one instance by other ones by means of calling during the program performing.

The "is_used" relationship is very important when developing ontology-driven software solutions to tackle visual analytics problems, since it determines the actual interactions between instances of software entities described by the ontology and affects the scenario of program execution so to be able managing the program behavior in dynamics. As noted above, the set of axioms described in the ontology model we use is empty, because, as our experience with using ontologically managed services of the SciVi visual analytics platform has shown the leveraging of applied lightweight ontologies is enough for the purpose of adapting visualization tools to the specifics of analyzed data and personal preferences of the users.

It is known that using interactive and cognitive computer graphics tools contributes to greater human involvement in the analysis process than just a simple look at the visual image. In the context of the topic under discussion, such tools work more to implement the principle of participativeness of 4P medicine. In our practice, traditionally, to automate visualization tools development and to customize them to the personal preferences of both end-users and developers, we use the methods and tools of SciVi visual analytics platform [5-6] enabling personalized visual analytics tools according to the ontological user profile. We have expanded SciVi repository with new medical-oriented visualization models.

As an example of visualization model that, in our opinion, is adequate for the implementation of interactive cognitive graphics tools in the personal data factory environment (as a part of visual analysis services based on the virtual integration of heterogeneous medical data), may be a model of the human body in the form of a subway diagram. For example, in Underskin Body Map created by Sam Loman (see Figure 5), instead of the usual metro lines, the main physiological lines of the human body are displayed, marked with different colors (human arteries are indicated in red, the venous system is indicated in blue, etc.). To have explanations the colors for marking the human muscular system, central nervous system, respiratory system, lymphatic circulation system, excretory and digestive systems see the in-figure legend.

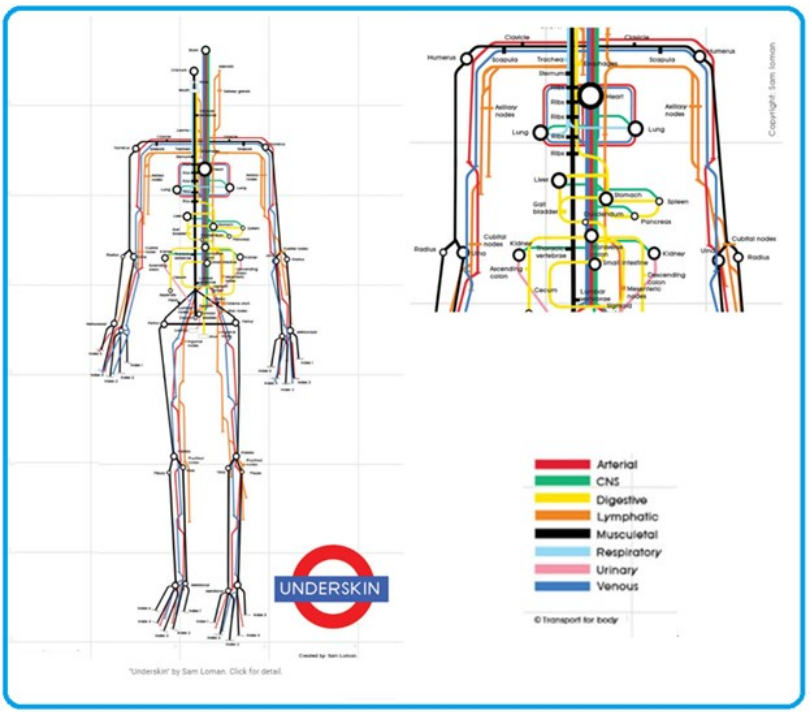

Figure 5: Underskin Body Map created by Sam Loman (https://vizworld.com/2010/03/sam-lomans-underskin-visualization/)

The main nodes of all the listed physiological systems of the human body are singled out separately in something similar to the stations on the metro line diagram. The digital healthcare industry could use visualization techniques such as "anatomical subway maps" to better convey medical concepts to patients, explain problems with their current health status, and assist doctors in analyzing heterogeneous patient data.

This method of visualizing the physiological systems of the human body scales well and is suitable for rendering on mobile device screens. Using the SciVi semantic filtering mechanism and the semantic integration services of data factories described above, it is possible to output all needed personal information available from the virtual space in relation to the certain organ systems and the main physiological systems of the human body in the simple and understandable view.

The size of the pictograms may correspond to the volume of available data about the particular organ or organ system in humans, the color intensity – about the level of health risks of the related organ or physiological system in humans according to the results of integrated analysis of the accessible medical data, genetic and environmental factors, data on human nutrition or lifestyle. The corresponding analysis is performed by machine learning methods and/or by knowledge engineering methods. Causal, temporal and other types of relations are automatically considered to assess the risks of developing diseases and healthy longevity. This kind of visual integration on one screen of a large volume of heterogeneous data based on the principles of cognitive compression of information is quite natural, visual and understandable not only to professionals. Means of interactive graphics (clicking on the icon of the relevant organ or the "metro line") allow, in addition to the initial and aggregated information, to see an explanation of the conclusions made about health risks.

The use of such visualization models greatly contributes to the implementation of such principles of 4P medicine as prediction and prevention, because the interaction with this kind of graphics, provided the completeness and consistency of the analyzed data, can help to identify in advance in a person a set of those key factors that affect to the occurrence some disease, and formulate recommendations to change a lifestyle and/or the need to contact specialists to solve problems that require the participation of a doctor. Recommendations for lifestyle correction and preservation of current results in a personal data factory can help accompany the patient on the way to changing his/her daily behavior and forming new habits in terms of sleep, nutrition, movement, stress management, etc.

As practice shows, the use of lightweight ontologies can reduce the threshold for entry of newcomers into the field of ontological engineering, labor intensity and the time for developing ontologies [13]. It is important to emphasize that one of the main advantages of developing ontology driven solutions is that such software systems easily adapt to changes in the subject area of the tasks being solved. For the adaptation, it is enough to change the ontologies without changing the source code of the software platform. This is fundamentally important when creating data factories both for adaptation to new data sources and their semantic integration on the principles of federalization, and for automating the development of DF and its services in general. In addition, through the use of ontological engineering methods, the adaptation of the functioning of analytical services of data factories to the personal preferences of users is simplified, which is also aimed at implementing one of the most important principles of 4P medicine – personalization. These principles are implemented especially effectively when DF has a microservice architecture.

So, from traditional enterprise-level data factories, "personal data factories" inherit mainly the following:

1. Data sources, using the capabilities of modern graphical interfaces (graphical APIs), receive end-to-end integration.
2. Microservice architectures are used instead of a single block of software platforms.

3. Along with local data sources (personal collections of documents, as well as person-related data from MIS of both public and private clinics), the largest number of possible cloud solutions are used in the "personal data factory" environment.
4. Information flows are orchestrated (based on special services).
5. The quality of information is improved through unification and virtualization.
6. The quality of analysis, monitoring, hypothesis testing, risk assessment and predicative analytics is improved through the use of advanced scientific visualization tools and visual cognitive data analysis by AI methods.
7. Regardless of the type and volume of the data source, it is provided with quick access (to sites, relational databases, data warehouses, data lakes, etc.).
8. Providing secure and delimited access to different categories of users.

But unlike traditional enterprise-level data factories, the "personal data factory" contains only those data that effectively work to perform global end-to-end (across the entire set of all data sources) analytical NL-queries about the concrete person. The time of doctors for a comprehensive analysis of information obtained about the patient from various third-party sources can significantly be reduced due to DF and modern AI methods.

5.Conclusion

The approach proposed in the paper, called "personal data factory", scales well, and the use of microservice architecture allows effectively expanding the functionality of the visual analytics systems through a unified way of integrating new services and data resources to the DF platform. At the same time, there is no need to make changes to the existing infrastructure of data sources: an additional semantic layer based on ontologies is added between DF services and data sources, which deals with metadata management and access to heterogeneous data based on the principles of federalization. DF services help optimize data storage, processing and analysis, improve the quality of service of personal information resources and hardware and allow building ontological profile of the health status of the owner of the "personal data factory".

This profile is automatically updated being available for further semantic analysis, health monitoring and contextual search, for example, in order to get advice from a psychologist or recommendations on nutrition and physical activity. The information presented in the ontological profile aggregated from various sources in relation to various systems of human organs and interactive visual analysis tools allow you to visually track the dynamics of changes in the state of health and promptly recommend to consult a doctor of the right specialty.

The described approach to ontology-driven data integration has a synergistic effect, because it allows through semantic integration to expand the capabilities of search services compared to the capabilities of modern Internet search engines. In particular, the search service becomes more expert due to the ability to give answers like "YES"/"NO", automatically determining the need for such an answer option. In addition, the semantic power of search engines increases in the case when no resource contains a complete answer to the query and automatic semantic aggregation of search results from different resources is required. So, instead of issuing a list of links to different resources each containing only some part of the required information, a single and more complete response to the query is generated on the principles of semantic integration using NLP tools and Semantic Web technology. This opens up new opportunities for the personalized IT industry and not only in relation to the field of 4P medicine.

6.Acknowledgments

This study received support from the Federal State Task Program by Scientific and Technological Center of Unique Instrumentation of the Russian Academy of Sciences (FFNS-2022-0010). This work was performed using the equipment of the Shared Research Facilities

of the Scientific and Technological Centre of Unique Instrumentation of the Russian Academy of Sciences.

References

1. Patel A., Debnath, N.C., Bhushan, B. (Eds.). Semantic Web Technologies: Research and Applications (1st ed.). CRC Press. 2022. 404 p. DOI:10.1201/9781003309420

2. Martyushev-Poklad A.V., Yankevich D.S., Panteleev S.N., Pryanikov I.V., Guliev Y.I. Healthcare information systems and organizational model of care: current situation and opportunities for progress // Vrach i informazionnie technologies. 2020. Vol. 5. P. 6-16. DOI:10.37690/1811-0193-2020-5-6-16 (in Russian)

3. Business Intelligence and Analytics: On-demand ETL over Document Stores / M. Souibgui, F. Atigui, S. B. Yahia, S. S.-S. Cherfi // In book: Research Challenges in Information Science. 2020. Vol. 385. P. 556–561. DOI: 10.1007/978-3-030-50316-1_38

4. Chuprina S.I., Postanogov I.S. Enhancing Legacy Information Systems with a Natural Language Query Interface Service // Vestnik of Perm State University. 2015. Vol. 2 (29). P.78-86. (in Russian)

5. Ryabinin K., Chuprina S. Development of ontology-based multiplatform adaptive scientific visualization system // Journal of Computational Science. Elsevier. 2015. Vol. 10. P. 370-381. DOI: 10.1016/j.jocs.2015.03.003

6. Ryabinin K., Belousov K., Chuprina S. Novel Circular Graph Capabilities for Comprehensive Visual Analytics of Interconnected Data in Digital Humanities // Scientific Visualization. 2020. Vol. 12(4). P. 56–70. DOI: 10.26583/sv.12.4.06

7. Chuprina S.I. Data Fabric Technologies as the Basis of Intelligent Analytical Platforms of Preventive and Resort Medicine Digital Environment // (Problems of Balneology, Physiotherapy, and Exercise Therapy. 2023. Vol. 100(3). P. 219. DOI:10.17116/kurort202310003222 (in Russian)

8. Noy, N.F. Ontology Mapping // Handbook on Ontologies. Springer, Berlin. 2009. P. 573–590. DOI:10.1007/978-3-540-92673-3_26

9. Chuprina S., Postanogov I., Nasraoui O. Ontology Based Data Access Methods to Teach Students to Transform Traditional Information Systems and Simplify Decision Making Process // Procedia Computer Science. 2016. Vol. 80. P. 1801–1811. DOI: 10.1016/j.procs.2016.05.458

10. Macura, M. Integration of Data from Heterogeneous Sources Using ETL Technology // Computer Science. 2014. Vol. 15 (2). P. 109-132. DOI: 10.7494/csci.2014.15.2.109

11. Chuprina S.I., Zinenko D.V. Adaptable Visual Ontology Editor ONTOLIS // Vestnik of Perm State University. 2013. Vol. 3 (22). P. 106-110. (in Russian)

12. Chuprina S., Nasraoui O. Using Ontology-Based Adaptable Scientific Visualization and Cognitive Graphics Tools to Transform Traditional Information Systems into Intelligent Systems // Scientific visualization. 2016. Vol. 8(1). P. 23-44.

13. Davies J. Lightweight Ontologies // Theory and Applications of Ontology: Computer Applications. 2010. P. 197–229. DOI:10.1007/978-90-481-8847-5_9