MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220091

Hypergraph Based Data Model for Complex Health Data Exploration and Its Implementation in PREDIMED Clinical Data Warehouse

Christophe Cancé[f], Christian Lenne[b], Svetlana Artemova[a,b,c], PREDIMED group[i]
Pascal Mossuz[a,d,g], Alexandre Moreau-Gaudry[a,b,c,f]

[a] *CHU Grenoble Alpes (CHUGA), F-38000, Grenoble, France*
[b] *Clinical Investigation Center-Technological Innovation, Univ. Grenoble Alpes, INSERM CIC1406, CHUGA, Grenoble, France*
[c] *Public Health Department, CHU Grenoble Alpes, F-38000, Grenoble, France*
[d] *Department of Biological Hematology, Grenoble Alpes University Hospital, 38400 Grenoble, France*
[e]*Department of Pharmacy, CHU Grenoble Alpes, F-38000, Grenoble, France*
[f] *TIMC, Univ. Grenoble Alpes, CNRS, VetAgro'Sup, Grenoble INP, CHU Grenoble Alpes, F-38000, Grenoble, France*
[g] *IAB, Univ. Grenoble Alpes, CNRS, INSERM, Grenoble, France*

[i] *Pierrick Bedouch[aef], Jean-François Blatier[ac], Jean-Luc Bosson[abcf], Alban Caporossi[abfc], Sébastien Chanoine[aeg], Katia Charrière[ab], Brigitte Cohard[a], Jérôme Fauconnier[acf], Joris Giai[abcf], Pierre-Ephrem Madiot[a]*

Abstract

Within the PREDIMED Clinical Data Warehouse (CDW) of Grenoble Alpes University Hospital (CHUGA), we have developed a hypergraph based operational data model, aiming at empowering physicians to explore, visualize and qualitatively analyze interactively the complex and massive information of the patients treated in CHUGA.

This model constitutes a central target structure, expressed in a dual form, both graphical and formal, which gathers the concepts and their semantic relations into a hypergraph whose implementation can easily be manipulated by medical experts.

The implementation is based on a property graph database linked to an interactive graphical interface allowing to navigate through the data and to interact in real time with a search engine, visualization and analysis tools.

This model and its agile implementation allow for easy structural changes inherent to the evolution of techniques and practices in the health field. This flexibility provides adaptability to the evolution of interoperability standards.

Keywords:

Clinical Data Warehouse, data lake, hypergraphs, property graphs, flexibility, massive and complex data interactive exploration

Introduction

Structuring a health data warehouse requires the definition of a data model that allows both a precise and flexible description of the underlying concepts and the semantic links between them, due to the diversity of data sources and their high level of connectivity. The conclusive results of our approach in the PREDIMED proof of concept [2], allowed us to validate the use of an improved version of PREDIMED now in production.

Current modeling methodologies, regardless their application domain, are mainly based on the Entity-Association model, offering a good grasp on the objects to be modeled and the underlying constraints.

To explore data through a data model, an abstract model should be implemented and the most common way to do so are relational database management systems. For abstract models with concepts that are highly connected by various relations, this conversion can be quite challenging with such systems since the implemented structure and code are both dependent on the cardinality of the relations. While graphs allow to link a nodal element of a set to another element by an edge, hypergraphs generalize them by allowing to materialize sets (hyperclasses) and subsets (classes) of elements, corresponding to concepts using nodes, and edges representing the potential semantic relations of one to one or more classes, hyperclasses, elements they gather [9].

A rarely explored but suitable approach in health domain is to convert such an abstract model into a property graph database dealing with such concepts of vertices and edges [4, 5].

The PREDIMED project has chosen then to model all the data useful to the CDW as an Entity-Association hypergraph based data model and to implement it in a property graph database translating model concepts to graph vertices respectively without loss of semantics. This original approach allows the deployment of graphical tools to interactively query the data warehouse [2]. Its implementation in such NoSQL database offers a possibility of model's flexible modifications unlike rigid relational systems.

A formal definition language is associated with this graph model. The compilation of this language allows exports to commonly-used models such as I2B2 or OMOP, star models being poorer than the hypergraph model we have defined.

Hypergraph based models have been used for several decades by complex topological GIS systems [11] and have proven to be a powerful tool. A query represents a path within a hypergraph with possible constraints on the attributes of vertices or edges [10]. These queries are generated from a graphical interface to textual query language. In the following, we describe the architecture implementation, the model and the tools deployed or being deployed for various research projects using PREDIMED data. We show that our model is flexible and general enough to define health trajectories and instantiate them using the same modeling concepts.

Methods and implementation

Our approach since the beginning of the CDW project was to build the model in close collaboration with the physicians and researchers of the University Hospital. We then designed a software architecture and its implementation by offering them an evolving operational environment.

This system allows the user to graphically manipulate the many concepts and relationships in his or her field of expertise to explore the data lake while ensuring that he or she does not have to acquire knowledge of a computer language to do so.

Data design

Concerning the data modeling level, we defined a shared, readable and scalable target model (Fig 1), graphically describing the structure of the data lake we feed from thousands of source data tables, potentially isolated from each other. To do this, we reshape the basic information by transforming it into this flexibly structured target data warehouse that will be used at different levels in our software architecture (Fig 2).

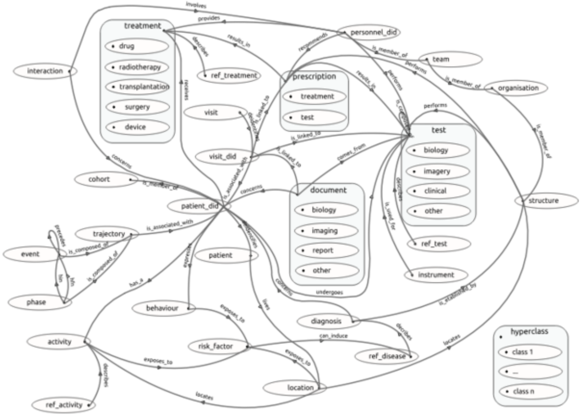


Fig1: Hypergraph based CHUGA's data lake model, composed of labeled nodal concepts linked together by semantic relations as directed edges (global view).

We have opted for a hypergraph based data model [9] that allows us to clearly represent the complex structure of concepts at different levels (nodes), their semantic relations (edges) and attributes, as close as possible to their intrinsic meaning in the health domain. In particular, assembling concepts into classes, possibly members of hyperclasses, simplifies the understanding of inheritance relations, common attributes, and intuitively generalizes the distributed relations between two hyper-sets to their sub-sets of elements. This notion simplifies the graph since a single link between two sets of classes replaces the NxM potential relations linking their respective N and M classes. The chosen model also allows for complex attributes (vectors, matrices, methods and execution parameters, etc.) at the various levels of classes and links, constituting an accurate reference catalog of the information collected in the data lake.

We have made this model directly operational by interfacing it with the functionalities of a NoSql property-graph database, natively optimized to express the information of complex, highly connected data structures. We have chosen a property graph database [14] suited to store information within named collections of nodes and directed edges, natively based on the JSON standard. Within these collections of nodes implementing concepts, we differentiate classes derived from the same set of concepts (or hyperclass) according to a specific high-level attribute from which all their elements inherit, allowing to simply implement classes and hyperclasses whatever their level in the data structure. These collections are indexed according to proven technologies in order to operate as close as possible to the "in memory" performances of the infrastructures that compute them, which opens the way to unprecedented interactive exploration possibilities compared to the usual performance of huge conventional databases. In addition to accessing the data via the user interfaces provided by the system, we have developed a generic high-performance RESTful API around this database in the form of microservices running on the server side. This API enables other applications to interact with the data lake using the standard REST/JSON protocol. Among these applications, a graphical interface, directly derived from the conceptual model, was developed in order to navigate in the data lake and explore it.

The design of the model is thus implemented in a spatial information management system (QGIS [13]) that we use to describe both graphically and formally the arcs and nodes of the model, as well as their attributes. This dual structure allows us to use GIS as a development framework to quasi-automatically generate the core environment for navigating through the data lake, as well as expressing its structure as a language [12]. In particular, the interactive webmapping functions are spontaneously adapted to select elements of the graph and access their attributes, without new developments being required to integrate the evolutions of the model.

Data navigation

We have linked this graphical navigation interface (Fig 3) to a specifically developed editor allowing to generate Cypher-like queries [10]. Thus, each graphical selection defining a path in the graph, from a class/hyperclass node to another one following an edge, composes a textual query in which the filter constraints on the attributes of the nodes or the link involved may be easily specified. The resulting query will be of the following form, generalizable to hyperclasses:

"source objects class[filters]—named relation[filters]—> target objects class[filters]", ...

This query language allows to generate, starting from any initial concept of the graph, a sequence of successive hops and filters, executable step by step, allowing to jump from selected sets of objects to others, according to the relations/arcs of the graph connecting them. The path corresponding to this sequence of elementary hops and successive filters on the attributes, potentially unlimited, is thus converted into a linear textual form with as many concatenations as successive elementary queries. The answer to a possibly complex question can thus be simply a matter of traversing a path in the graph with filters [2]. Each elementary query is transmitted and interpreted at the level of the API of the graph base, to be executed in the form of a graph traversal query starting from the objects of the current selection, resulting from the possible previous elementary query.

This approach allows us to query and explore step by step in an interactive way significant volumes of information and, thus, to make the results of these queries observable, traceable and thus easily reproducible.

Data exploration, visualization (Fig 3, 4, 5)

The set of objects selected at each stage of a navigation in the data lake can be easily projected on the fly in different application environments whose various complementary investigation functionalities are suited to explore their different facets. We thus offer possibilities of visualization and fine analysis of the information. For instance, we can consider only data related to each patient of a cohort, which allows to precisely control the inclusion of these patients in each clinical study. We can also produce interactive dashboards generated on the fly and linked to a Lucene [17] library based search engine, which can provide a statistical description of subsets of information. For this, we currently rely on the Elasticsearch/Kibana [15] package and the possibilities it offers to browse different information, visualize it and optionally export it to CSV text delimited format.

Data elaboration

The model and its API allow for elaborating new layers of information that benefit from the same navigation and analysis functionalities, in order to materialize the physicians' points of view in the form of health trajectories [6]. It is thus possible, for instance, to compute for a patient a sequence of events (such as test results, diagnosis, etc.) and related phases that we generate algorithmically from the facts gathered in the lake. In particular, the "CreateTrajectory(...)" method generates a trajectory for each patient of a cohort, made up of events and phases of interest whose algorithmic methods in the model describe the rules used by physicians to qualify and instantiate them, as well as the relations that materialize the sequence of their relative positions in time (precedences, coincidences, etc.) [7]. These elaborated data are used as a basis for trajectory representations in the form of Sankey diagrams in particular, as well as for the search for significant similarities between patient trajectories [5].

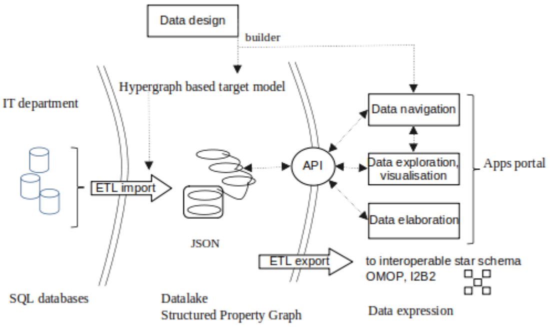
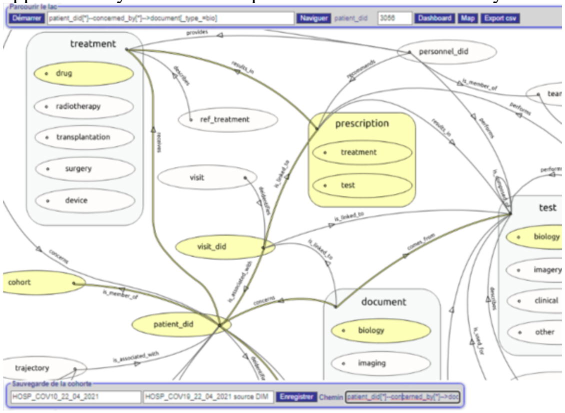


Fig 2: Data design driven layered architecture for clinical information processing in CHUGA's CDW

Results

Our work has resulted in a production application platform in continuity with the model designed with the medical experts, making it possible to interactively access (in a few seconds) to the underlying data. This data comprises several hundreds of million objects and semantic relations concerning approximately 1.7 million patients over the last 15 years.


Parcourir le lac

Démarrer | patient_did[*]--concerned_by[*]-->document[_type_=bio] | Naviguer | patient_did | 3056 | Dashboard | Map | Export csv

Sauvegarde de la cohorte

HOSP_COV10_22_04_2021 | HOSP_COV19_22_04_2021 source DIM | Enregistrer | Chemin | patient_did[*]--concerned_by[*]-->doc

Fig 3: Interactive graphic interface for navigation and exploration of CHUGA's CDW using Cypher-like auto-generated language to query information or retrieve datasets.

The multilingual navigation tool uses the graph model to accurately select information step by step, using an automatically generated Cypher-like query language (Fig 3).

Multimodal information on each patient can thus be collected in real time by medical experts, for instance to accurately refine the cohort of patients to be included within a study and possibly cross-reference the information to validate it.

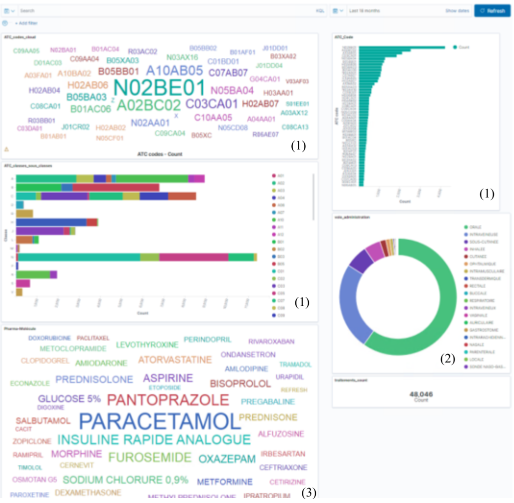


Fig 4: Interactive dashboard describing drug prescriptions concerning a cohort of patients, using Anatomical Therapeutic Chemical classification system (1), route administration (2) and chemical name information (3).

A cohort of patients resulting from a complex request can thus be established graphically to gather quickly the requested information. It is thus possible, for instance, to gather patients over 45 years old concerned by intracerebral hemorrhage while receiving anticoagulant treatment, then to observe their biological analyses in LDL and cholesterol, within a few minutes. The time required to retrieve all the related potentially massive data to such cohort within a dataset is now a few hours at most, whereas in practice, such an approach was often extremely difficult requiring weeks, if not impossible, given the huge amount of fragmented data within the hospital information system. The result is nonetheless accurate and reproducible.

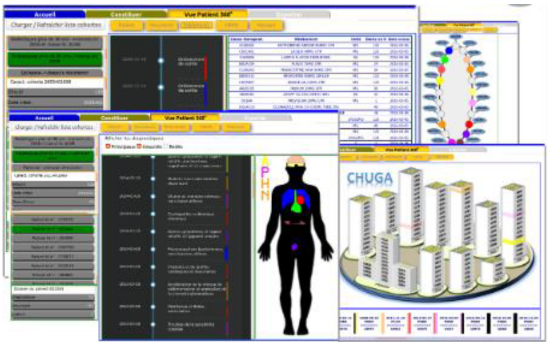


Fig 5: Dashboards gathering for each patient, clinical, biological and pharmacological related information as well as their pathway within the CHUGA's care units.

The extraction of cohort related datasets observed in depth by health specialists, allows, via a random analysis of patient data, to validate the projects datasets, to select and extract only relevant features to datasets of reasonable dimensions to be used by machine learning algorithms.

The general, highly connected and extensive structure of the model aims at progressively describing all the elements of interest to materialize the patients' long-term health trajectories.

The ability to gather and explore multimodal and multisource information within PREDIMED CDW allows it to be effectively involved in various clinical studies. For instance, first uses of PREDIMED enables the efficient provision of relevant data for a wide variety of projects such as:

- making predictions on emergency interventions around chronic hematology patients,
- daily regional cartographic tracking of the COVID-19 pandemic and of the emergence and evolution of clusters,
- identification of risk factors and their consequences in farmers, with analysis of their exposome, in particular their occupational activities and environment,
- study on the life quality at work for the network of care providers working in the emergency department during the Covid-19 pandemic.

All of the information in the data lake is described using various standard terminologies and high-level referentials, such as ICD10 for diseases coding or the Geonames [18] repository for geographic locations for instance. Our approach makes it possible to parameterize the interoperability and data exchange process of our data lake with the outside world.

Significant data transformation procedures are required, first to populate the data lake with information corresponding to the target model, and then to export parts of it as an interoperable standardized data set. These procedures involve an Extract Transform Load (ETL) software brick (Fig 3). The currently chosen ETL software [16] allows us to define different jobs to reshape and align the parametrized structures at the input and output of each of these processes, and to periodically run the validated and compiled version in an industrialized manner.

Our ability to produce an application environment in phase with the structure of a dataset in a semi-automatic way allows us to simply generate an interoperable datamart including both the information necessary for a study and an adapted application environment that allows its exploration (Fig 2, 3).

Discussion

Impact

The interactivity in the ability to analyze intermediate results while navigating the graph aims at empowering clinicians to observe and select variables and events of interest that can potentially be used then as features to create datasets, which can be then efficiently analyzed by complementary machine learning based algorithms.

As the basic structure of the model remains invariant even when the model evolves, the implementation variables only depend on the content of the class of classes (and hyperclasses), that of the class of links and their respective attributes, which constitute the primitive deployment parameters of our application architecture. The model is thus potentially extensible without any impact in terms of development on the implementation, independently of the potentially increasing number of classes of a hyperclass (to integrate new types of processing for instance) or the possibilities of links to new ontologies or other graphs.

This flexibility, made possible by a highly generic code, makes it virtually automatic to implement the coherence of the means of expressing information within the various software bricks of the platform, which constitute an application continuum driven by the model. This approach favors the implementation of complementary automatic and human control routines for the quality of data and services produced by the application platform. Thus, traversals of different paths supposed to lead to the same result as a response to a complex request can be used to evaluate the quality of the data.

The code developed to implement the original functionalities provided by PREDIMED platform is built around different interchangeable software bricks oriented towards "in memory, JSON Restful" technologies involved in the different layers, whether it is the property graph database, the search and visualization engine or the ETL, for which other Open Source alternatives exist.

Since the data lake navigation language we use has no adherence to the internal proprietary language of the graph database, the functional core hosting the data could be replaced without any other impact than the changing of a minor part of the API code that makes the graph database usable by other applications. Interchangeability provides robustness to the application architecture.

Plans and perspectives

Part of the work in progress aims at enriching the structure and data of the datalake with new information of interest computed from the unstructured data such as the textual data (examination reports, discharge letters, others) which represents 80% of health data.

In parallel, we are developing new means to describe from the physicians' point of view the dynamics of the phenomena in the form of trajectories composed of events and phases specific to their specialty, and also their temporal relations, from the information gathered in the lake [6].

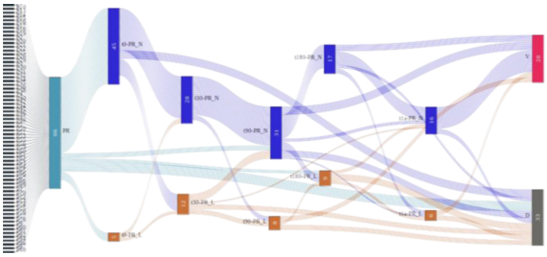


Fig 6: Survival diagram as a function of protein (PR) qualitative levels

As a proof of concept, we applied a method of qualitative calibration of biological values (PR, WBC, CRP, ...) on a cohort of patients with Acute Myeloblastic Leukemia. We proceeded to the temporal alignment of these biological events with respect to the date T0 of diagnosis (blasts > 20%). For these patients, we enriched the property graph dataset with these events to produce the above Sankey diagram (Fig 6), that illustrates as an example, the evolution of the PR rate at T0, T+10 days, T+1 month, T+3 months, T+6 months and T+1 year, and highlights the survival.

Beyond the existing layers gathering factual information, computed health trajectories of patients with elaborated data, an additional layer hosts information forged according to recommended interoperability standards, which is essential for our contribution to exchanges and to the constitution of high-level cohorts. Even if the graphical approach is largely privileged, the conceptual language approach of type Entities-Association which is natively associated to it, allows us, by the implementation of compilers of this language to open up to interoperable systems like OMOP, FHIR or I2B2 using code generation. These generators, currently under development, implement translation/correspondence rules towards the less complex concepts of these systems, and seem to us to be particularly well suited to implement the conversion of a source structure of increasing complexity to target standards, themselves evolving [12]. This work of structure alignment is largely supported by the shared work of other teams, such as the evaluation and selection of terminologies and ontologies carried out by D2IM in Rouen, France (https://www.hetop.eu/hetop/), the implementation work of the InterHop group (https://framagit.org/interhop/omop/snds-structural-mapping) or the international EHDEN initiative (https://www.ehden.eu/business-directory/) .

Limitations

An essential counterpart of the fluidity and flexibility of property graph operation is that no native control is performed on the referential integrity rules of the information. As this integrity is ensured upstream of the CDW feed via the ETL brick, we have opted for an "add and read only" mode for the information that we regenerate periodically (twice a week) according to the industrialized process already described.

Conclusions

The graph approach allows us to move forward in a rapid iteration cycle to provide tangible answers to clinicians' objectives in terms of visibility on complex and massive patients' data. This was achieved in the context of a pandemic that catalyzed the transition from a proof of concept to a production platform. We have worked particularly on the maturation of its architecture, its implementation and its human organization while being vigilant to remain in conformity with the various legal aspects. This dynamic acceleration of the PREDIMED platform's possibilities is sparking new initiatives in terms of research projects, which are also fertile for new methods of data exploration, offering clinicians new in-depth, configurable and interactive views of patients' medical data. PREDIMED allows non-IT specialists to use complex and massive data mining tools to effectively and rapidly explore and design patient information features for research projects. The ability to simply select and explore small amounts of information chosen from a large data set, in a differentiated way according to each study context, allows to successively express information of interest. It is then possible to reshuffle them, splice them, to elaborate a mature information that can be expressed at a higher level (Fig 2). PREDIMED thus provides agility and auto-improvement ability in expressing and understanding information while making the work traceable, reproducible, and transferable to other data-sets.

Acknowledgements

This work has been partially supported by MIAI@Grenoble Alpes (ANR-19-P3IA-0003). It has been performed as an operational platform since the independent supervisory authority in terms of the GDPR (CNIL in France) has authorized the CHUGA to proceed to the operational phase of PREDIMED, on October 10th 2019.

References

[1] S. Artemova and al., «PREDIMED: Clinical Data Warehouse of Grenoble Alpes University Hospital» *Studies in health technology and informatics,* vol. 264, p. 1421–1422, 2019.

[2] C. Cancé et al., "Cohort Creation and Visualization Using Graph Model in the PREDIMED Health Data Warehouse". MIE 2020, *Studies in Health Technology and Informatics*. https://www.doi.org/10.3233/SHTI200132

[3] Achard, P., Maugard, C., Cancé, C. *et al.* Medico-administrative data combined with agricultural practices data to retrospectively estimate pesticide use by agricultural workers. *J Expo Sci Environ Epidemiol* **30,** 743–755 (2020). https://doi.org/10.1038/s41370-019-0166-x

[4] D. Birtwell, H. Williams, R. Pyeritz, S. Damrauer, D. L. Mowery, Carnival: A Graph-Based Data Integration and Query Tool to Support Patient Cohort, MEDINFO, Lyon, 2019.

[5] M. Ganzinger et al., «A Concept for Graph-Based Temporal Similarity of Patient Data» doi:10.3233/SHTI190199

[6] D. Noël, M.Villanova-Oliver, J.Gensel, P.Le Quéau. Design patterns for modelling life trajectories in the semantic web. *15th International Symposium on Web and Wireless Geographical Information Systems (W2GIS 2017)* , May 2017, Shanghai, China. ⟨hal-01481127⟩

[7] Michel Dubois, Sylviane Schwer. Classification topologique des ensembles convexes de Allen. R.F.I.A. 2000, Reconnaissance des Formes et Intelligence Artificielle, Feb 2000, Paris, France. pp.59 - 68. ffhal-00575250f

[8] Stevens R, Goble CA, Bechhofer S. Ontology-based knowledge representation for bioinformatics. Brief Bioinform. 2000 Nov;1(4):398-414. doi: 10.1093/bib/1.4.398. PMID: 11465057.

[9] Bouillé F. (1977) The Hypergraph-Based Data Structure: A New Approach to Data Base Modelling and Application. In: Schneider H.J. (eds) GI — 7. Jahrestagun. Informatik — Fachberichte, vol 10. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-48908-2_3

[10] N. Francis, A. Green, P. Guagliardo, L. Libkin, T. Lindaaker, V. Marsault, S. Plantikow, M. Rydberg, P.Selmer, and A. Taylor, Cypher: An Evolving Query Language for Property Graphs, in: Proceedings of the 2018 International Conference on Management of Data, ACM,New York, NY, USA, 2018.

[11] Egenhofer, Max & Herring, John. (1991). High-level spatial data structures for GIS. 227-237.

[12] Boyd M., McBrien P. (2005) Comparing and Transforming Between Data Models Via an Intermediate Hypergraph Data Model. In: Spaccapietra S. (eds) Journal on Data Semantics IV. Lecture Notes in Computer Science, vol 3730. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11603412_3

[13] QGIS, https://www.qgis.org

[14] ArangoDB, https://www.arangodb.com/

[15] Elastic, https://www.elastic.co

[16] Talend, https://www.talend.com

[17] "*Introduction to Apache Lucene: Construction of Java Open Source Full Text Retrieval Systems*" by Koshi Sekiguti ; Gijutsu-Hyohron Co, Ltd; (ISBN 4774127809)

[18] Geonames, https://www.geonames.org/

Address for correspondence :

Christophe Cancé,
christophe.cance@univ-grenoble-alpes.fr.