

Facilitating Learning Analytics in Histology Courses with Knowledge Graphs

Jimmy Walraff^{1,†}, Andreas Coco^{1,†}, Guillaume Delporte^{1,†}, Merlin Michel^{1,†}, Allyson Fries², Valérie Defaweux² and Christophe Debruyne^{1,*}

¹Montefiore Institute of Electrical Engineering and Computer Science, University of Liège, Liège, Belgium

²Department of Biomedical and Preclinical Sciences, Faculty of Medicine, University of Liège, Liège, Belgium

Abstract

We report on an ongoing learning analytics project at the University of Liège, in which we want to analyze student interactions on Cytomine for a histology course. Cytomine provides tools for medical image annotation and an API that has been used for learning analytics. The problem, however, is that the data obtained from Cytomine has implicit semantics and requires many data preprocessing and integration steps. This poster presents the prototype KG we have built to address these problems. The KG adopts PROV-O to distinguish activities from their outcomes, addressing some of the issues faced in the past. We also demonstrate that the KG can be used in Jupyter notebooks, though learning analytics is left for future work. It did demonstrate that the data analysis process has become more declarative and transparent, as data is analyzed starting from SPARQL queries. We focused on one project in Cytomine, and future work consists of integrating additional projects. We also plan to investigate the development of more self-contained KG generation techniques as we have no direct access to the Cytomine application.

Keywords

KG Construction, Learning Analytics, Ontology Engineering

1. Introduction

Cytomine [1] is a Web-based image analysis software platform that facilitates collaborative exploration and analysis of large biological and medical image datasets. Cytomine provides tools for image annotation (see Figure 1). Its application facilitates collaboration and educational applications, as demonstrated by its use in histology courses at the University of Liège. Cytomine employs a MongoDB database for data storage and provides a fairly restricted API to engage with the various objects, such as the image annotations and tags created by its users.


While advantageous for object persistence, MongoDB's document-oriented storage model presents challenges for the *interconnected analysis* required in learning analytics research. Additionally, the various document types contain implicit relationships, so one must manually determine a user's subsequent annotations, for example. As such, prior learning analytics studies [2] relied on preprocessing pipelines to create CSV files for machine learning models, which led to various provenance issues (e.g., why were certain points omitted, amended, etc.).

SEMANTiCS 2024: 20th International Conference on Semantic Systems, September 17–19, 2024, Amsterdam, The Netherlands

*Corresponding author.

†These authors contributed equally.

 0000-0002-2780-7264 (A. Fries); 0000-0002-8928-1309 (V. Defaweux); 0000-0003-4734-3847 (C. Debruyne)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

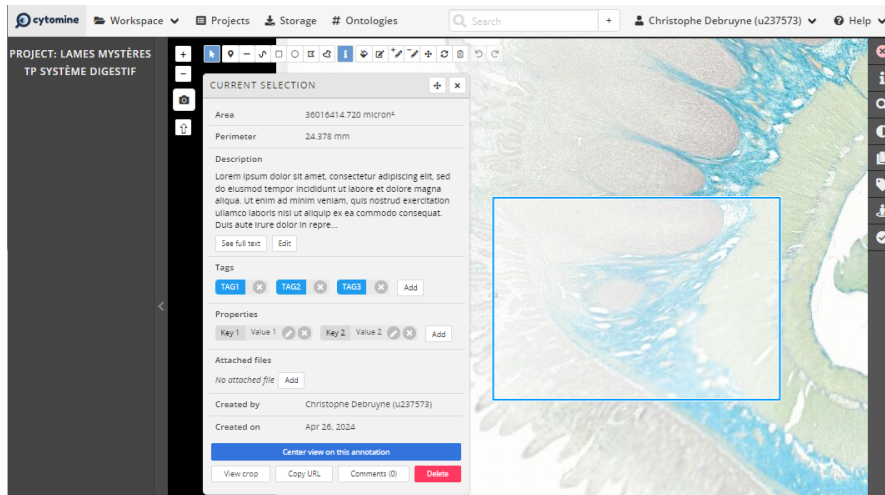


Figure 1: An example of an annotation in Cytomine created for this poster. In this example, one has selected an area on an image, entered a description, defined some tags, and provided some properties, a series of key-value pairs. Each annotation has a URL that can be shared with others.

This study aims to investigate the suitability of knowledge graphs (KGs) as a foundation for learning analytics research. It is hoped that KGs can render those implicit relationships explicit and that graph query languages are better suited to retrieve data for learning analytics. Another motivation for using KGs is that the tools used in learning activities are just that—tools. The data they store pertains to the tool. With KG technologies, we can integrate these data with (different) learning models, e.g., to analyze whether the *triple consistency*[3] between learning objectives, activities, and evaluations is met. In other words, KGs allow us to integrate these tools in a flexible manner to support learning analytics.

This paper briefly discusses our approach to integrating Cytomine’s data into a KG, demonstrates our KG in a Jupyter Notebook, and elaborates on future work. The potential of this study is substantial, as the feedback provided to students will guide their studies and enhance their performance. Moreover, the data will assist educators in effectively integrating digital microscopy into their pedagogical plan, thereby optimizing educational outcomes.

1.1. Related Work

There is little related work on the use of KGs for learning analytics. The learning analytics community seems to focus on using Linked Data to facilitate research, as can be observed in the LAK Data Challenge [4] and a Web-portal reported in [5]. [6] report on the potentials and challenges of KGs in learning analytics, but only mention anecdotal uses such as [7], who analyzed student enrollments in a university using a dataset enriched with Linked Datasets.

2. Approach: Building CytoGRAPH

The current iteration of the KG, dubbed CytoGRAPH, was built as follows:

Ontology Development The KG’s ontology was engineered with a middle-out approach where entities in the data (described below) were identified and aligned with the UoD of domain experts and existing ontologies. We adopted OWL 2 QL as we anticipate the KG to contain many assertions. The ontology we developed builds upon PROV-O [8] to model the interactions between users and images and a sequence of annotations on an image in one use session, GeoSPARQL [9] for representing the annotation’s geometries, and Web Annotation Vocabulary [10].¹ PROV-O was adopted as many of the core concepts aligned well with this ontology; entities are the resources used (e.g., the images) and produced (e.g., annotations) in the learning activities. The interactions of students are represented as activities. Both students and instructors are represented as agents.

Data Transformation We had no access to Cytomine’s MongoDB instance, though we could download the data via its API.² The data of one project consisting of 11 images, 588 users (pseudonymized), and 27185 annotations, 1571 properties, and 31507 descriptions. We used RML [11] with BURP [12] to generate RDF from the data. The University of Liège’s Cytomine instance has over 175 projects, which indicates the KG’s potential size.

Data Annotation While we have yet to create links to other datasets and even other institutional repositories (e.g., the e-learning platform), we have decided to represent geometries using `geo:wktLiterals` so that we can retrieve activities from certain areas on the images. As such, we enriched the data with a geometric dimension.

We recognize that our approach’s major limitation is its inability to transform the data stored in MongoDB. Moreover, Cytomine’s API is fairly restricted, allowing us to retrieve data when sufficient restrictions are placed (e.g., retrieving the annotations on a project-per-project basis). This limitation is beyond our control.

3. Results

The result of this study yielded a proof-of-concept KG for learning analytics. The KG can be explored with tools such as Ontodia [13], as shown in Figure 2. The KG currently contains information on over 27K annotations made by 587 users over one decade, which is for the sole project to which we have access.

To demonstrate that one could engage with the KG for learning analytics, we created a Jupyter Notebook that retrieved the number of annotations per contributor and used this to determine the optimal number of clusters using the Elbow Method, as shown in Figure 3.

4. Conclusions

We reported on the feasibility of creating a KG out of Cytomine, which required integrating CSV into RDF. The data we obtained from Cytomine was rather flat. Information about a user’s

¹The ontology, available at <https://chrdebru.github.io/papers/2024-09-semantic/ontology.owl>, is not yet made available using a persistent identifier. The ontology will be published in a future iteration of the KG construction.

²<https://doc.uliege.cytomine.org/dev-guide/api/reference>

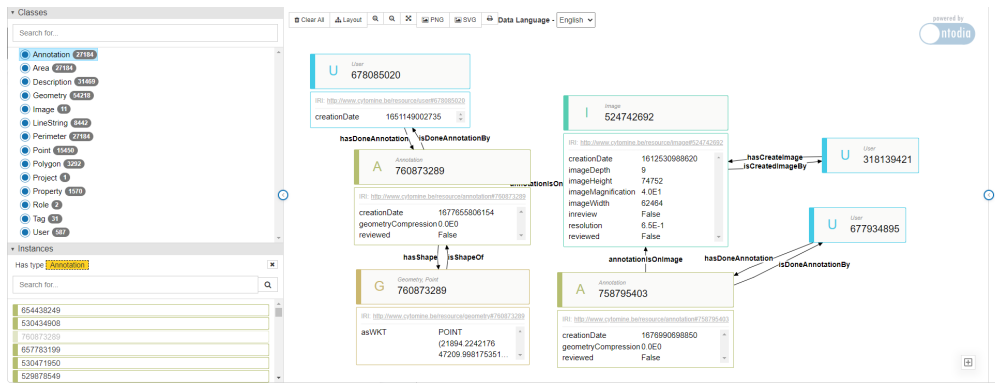


Figure 2: Ontodia is used to visualize concepts and their relationships in CytoGRAPH. This image illustrates relationships between users and their annotations of an image.

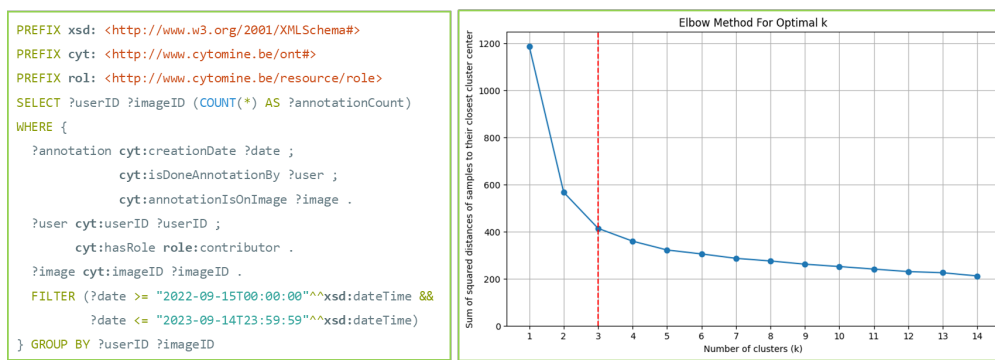


Figure 3: As a proof of concept, we showed domain experts how to interact with the KG using a Jupyter Notebook. Using the number of annotations per contributor (a type of user), we applied the elbow method to determine the optimal number of clusters (k). One can see that the optimal number of clusters seems to be three, as the elbow is the most pronounced at this specific number of clusters.

activity was implicitly stored but rendered explicit using PROV-O in the KG generation process. As users annotated slides and stored them with geometric coordinates, we adopted GeoSPARQL to use geospatial predicates. This allows us to analyze interactions on specific regions on slides, for example. The number of annotations within one project indicates our project's scale, knowing there are over 150 projects in Cytomine. Challenges that we will investigate include the evolution of this KG over time. As we currently have no access to the MongoDB instance, which is normal, we should investigate more elegant ways to generate the KG. One venue is to retrieve the data via rest calls in the mapping, which requires the development of bespoke RML iterators.

Acknowledgments

The authors wish to thank Ulysse Rubens from Cytomine Corporation.

References

- [1] U. Rubens, R. Hoyoux, L. Vanosmael, M. Ouras, M. Tasset, C. Hamilton, R. Longuespée, R. Marée, Cytomine: Toward an open and collaborative software platform for digital pathology bridged to molecular investigations, *PROTEOMICS – Clinical Applications* 13 (2019) 1800057.
- [2] A. Fries, M. Pirote, L. Vanhee, P. Bonnet, P. Quatresooz, C. Debruyne, R. Marée, V. Defaweux, Validating instructional design and predicting student performance in histology education: Using machine learning via virtual microscopy, *Anatomical Sciences Education* 17 (2024) 984–997.
- [3] V. R. Kovertaite, D. Leclercq, The triple consistency illustrated by e-tivities to help understand national and international policies in e-learning, *International Journal of Technologies in Higher Education* 3 (2006) 1–7.
- [4] M. d’Aquin, S. Dietze, E. Herder, H. Drachsler, D. Taibi, Using linked data in learning analytics, *eLearning Papers* 36 (2014) 1–9.
- [5] Y. Hu, G. McKenzie, J. Yang, S. Gao, A. Abdalla, K. Janowicz, A linked-data-driven web portal for learning analytics: Data enrichment, interactive visualization, and knowledge discovery, in: *Workshops at the 4th International Conference on Learning Analytics and Knowledge (LAK 2014)*, Indianapolis, Indiana, USA, March 24–28, 2014, volume 1137 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2014.
- [6] A. Zouaq, J. Jovanovic, S. Joksimović, D. Gašević, Linked data for learning analytics: Potentials and challenges, *Handbook of Learning Analytics* (2017) 347–355.
- [7] M. d’Aquin, N. Jay, Interpreting data mining results with linked data for learning analytics: motivation, case study and directions, in: *Third Conference on Learning Analytics and Knowledge, LAK ’13*, Leuven, Belgium, April 8–12, 2013, ACM, 2013, pp. 155–164.
- [8] S. Sahoo, T. Lebo, D. McGuinness, PROV-O: The PROV Ontology, *W3C Recommendation, W3C*, 2013. <https://www.w3.org/TR/2013/REC-prov-o-20130430/>.
- [9] R. Battle, D. Kolas, Geosparql: enabling a geospatial semantic web, *Semantic Web Journal* 3 (2011) 355–370.
- [10] R. Sanderson, P. Ciccarese, B. Young, Web Annotation Vocabulary, *W3C Recommendation, W3C*, 2017. <https://www.w3.org/TR/2017/REC-annotation-vocab-20170223/>.
- [11] A. Iglesias-Molina, D. Van Assche, J. Arenas-Guerrero, B. De Meester, C. Debruyne, S. Joza-shoori, P. Maria, F. Michel, D. Chaves-Fraga, A. Dimou, The RML ontology: A community-driven modular redesign after a decade of experience in mapping heterogeneous data to RDF, in: *22nd International Semantic Web Conference - ISWC 2023*, Athens, Greece, November 6–10, 2023, *Proceedings, Part II*, volume 14266 of *LNCS*, Springer, 2023, pp. 152–175.
- [12] D. Van Assche, C. Debruyne, Burping through RML test cases, in: *5th International Workshop on Knowledge Graph Construction co-located with ESWC 2024*, Hersonissos, Greece, May 27, 2024, volume 3718 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024.
- [13] D. Mouromtsev, D. S. Pavlov, Y. Emelyanov, A. V. Morozov, D. S. Razdyakonov, M. Galkin, The simple web-based tool for visualization and sharing of semantic data and ontologies, in: *ISWC 2015 Posters & Demonstrations co-located with ISWC-2015*, Bethlehem, PA, USA, October 11, 2015, volume 1486 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2015.