Proceedings of the 11th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'25)
Paris, France – August, 2025
Paper No. ICBES 120
DOI: 10.11159/icbes25.120

ACUITEE: A Comprehensive Tool for Visualization, Editing and Curating textual Annotations in Clinical Data

Moussa Baddour[1], Olivier Dameron[2], Marie de Tayrac[3], Stéphane Paquelet[1], Paul Rollier[3], Thomas Labbe[4], Majd Saleh[1]
[1]Institute of Research and Technology b<>com
1219 Av. des Champs Blancs, 35510 Cesson-Sévigné, France
Moussa.Baddour@b-com.com; Stephane.Paquelet@b-com.com; majdsaleh84@gmail.com
[2] University of Rennes
Rue du Clos Courtel, 35000 Rennes, France
Olivier.Dameron@univ-rennes.fr
[3]University Hospital of Rennes
Rue Henri le Guilloux, 35000 Rennes, France
Marie.De.Tayrac@chu-rennes.fr ; Paul.Rollier@chu-rennes.fr
[4]Orange Labs
Rue du Clos Courtel, 35510 Cesson-Sévigné, France
Thomas.Labbe@orange.com

Abstract **-** Annotation and management of clinical data remains a critical but challenging task due to the complexity and diversity of medical records. Providing a tool to simplify, shorten, and improve the annotation work of clinicians is essential as it leads both to process optimization and better patients' characterization. We present ACUITEE (Annotation and Curation User Interface for Terms Extraction Engines), a web application that addresses these challenges. It offers a simple way to improve clinical data annotation workflows by integrating automatic analysis, manual processing, and real-time visualization of medical notes. Using advanced natural language processing (NLP) techniques for phenotypes extraction such as PhenoBERT and efficient string-matching algorithms, ACUITEE maps free-text medical notes to ontology terms and enables clinicians to validate or refine these annotations through a user-friendly interface. The system supports fully automated, semi-automated and manual annotation modes, providing flexibility for different use cases. A key feature of ACUITEE is its interactive annotation interface, which enables clinicians to validate, edit, and curate ontology terms with precision, thereby speeding up the annotation process while maintaining high accuracy. This paper outlines ACUITEE's architecture, features, and applications and demonstrates its potential to facilitate clinical data annotation through increased efficiency, adaptability, and user engagement. ACUITEE is freely available at https://acuitee.labs.b-com.com/. The source code is available at https://github.com/b-com/ACUITEE for local installation for protected health data.

Keywords**:** Annotation Curation, Visualization Tools, HPO, Phenotypes, Data Parsing, Information Extraction.

1. Introduction

Accurate annotations are essential for advancing medical research, training machine learning models for clinical applications, and improving patient care. However, the management and annotation of free-text clinical data pose significant challenges due to the various complexities as well as heterogeneity of medical records. Existing solutions often lack the balance between automation and manual curation, limiting their utility in real-world clinical environments. In addition, most of these tools are only available in English. That is where ACUITEE stands out, offering all the features of its integrated parsers, while supporting multiple languages. ACUITEE aims to address this gap by providing a seamless platform for visualization, editing, and manual correction of automatically generated annotations. This was underscored by our experiments and feedback from clinicians where the usability and annotations relevance of clinical reports management were highlighted, particularly in multilingual and resource-diverse settings.

2. Existing Solutions for Clinical Data Annotation

Existing tools for clinical data annotation can be categorized into three types: fully automated systems, semi-automated systems, and manual annotation tools. Each offers unique strengths and limitations:

2.1. Fully Automated Solutions leverage NLP [1] and machine learning to annotate data without human intervention. Examples include:

- MetaMap [2]: Maps text to UMLS (Unified Medical Language System) concepts but struggles with nuanced and ambiguous terms.
- cTAKES (Clinical Text Analysis and Knowledge Extraction System) [3]: Extracts information using terminologies like SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) [4] but requires domain-specific customization and struggles with complex syntax.
- Doc2HPO [5]: Automatically extracts Human Phenotype Ontology (HPO) terms from clinical text, but its precision can vary based on the quality and specificity of input data, and support just English language.
- ClinPhen [6]: Focuses on extracting phenotype terms directly from clinical notes, optimizing for rapid processing, but may require integration with additional tools for comprehensive annotation.

2.2. Semi-Automated Solutions combine automatic annotation with manual curation for a balance of efficiency and accuracy:

- Brat [7]: Supports manual and semi-automated workflows but requires technical expertise for advanced features.
- BioPortal Annotator [8]: Automatically annotates text using ontology concepts but depends on ontology quality and lacks robust customization.
- Tagtog [9]: Enables collaborative annotation but has a steep learning curve for customization and advanced NLP integration.
- RAG-HPO [10]: A hybrid tool designed for generating and refining HPO annotations. While it supports integration with clinical data workflows, it may require significant configuration for non-standard datasets.

2.3. Manual Annotation Tools rely entirely on human effort, ensuring precision but being time-consuming:

- EHOST [11]: Supports schema-based annotation but is impractical for large datasets.
- Knowtator [12]: Facilitates ontology-based annotation but lacks built-in NLP capabilities.

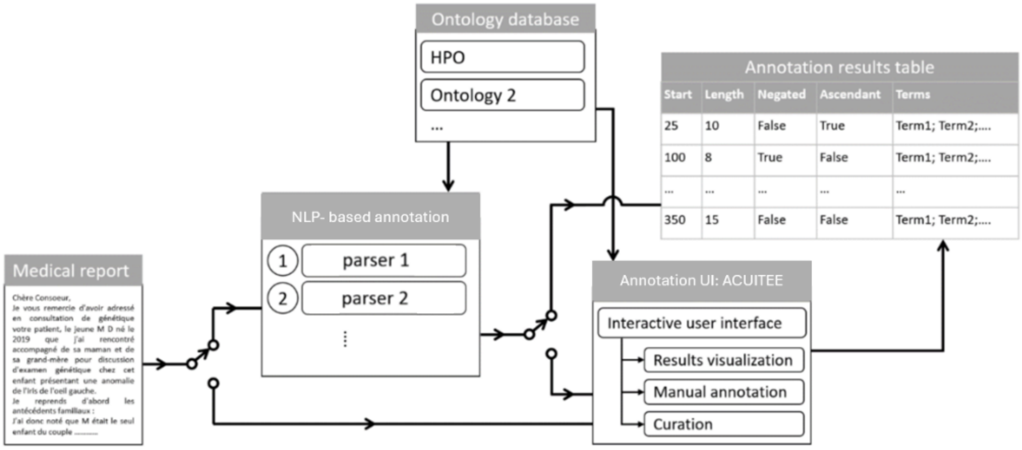
Key challenges with these solutions include scalability, accuracy in handling complex data, limited flexibility, complex interfaces, and potential privacy issues with cloud-based systems. ACUITEE addresses these challenges by combining robust NLP automation with interactive manual curation. Its features prioritize accuracy, scalability, and adaptability while supporting data privacy and multilingual environments. This makes ACUITEE a versatile tool for annotation of clinical data. Notably, ACUITEE ensures privacy by offering the capability to be installed locally, making it a versatile tool for the annotation of clinical data. The primary goal is to provide clinicians with simple, user-friendly access to state-of-the-art methods. Additionally, ACUITEE empowers users to refine and clarify the data corpus through manual annotation, ensuring continuous improvement. By focusing on clinician needs, ACUITEE bridges the gap between advanced technology and practical usability.

3. General Architecture

The architecture of ACUITEE is designed to facilitate extraction of ontology terms from medical notes. Figure 1 illustrates the system's architecture, where free-text medical notes serve as input, and annotated ontology terms are the output. The system includes the following components:

- Input Medical Note: A free-text record from genetic consultations
    - NLP-Based Parsers: Automatic parsing engines, such as a string-matching parser and the PhenoBERT-based parser, that extract reference ontology terms from the input text.

- Reference Ontologies: The system uses the Human Phenotype Ontology (HPO) [13] as the primary ontology source. Other ontologies can be integrated as needed.



Chère Consoeur,
Je vous remercie d'avoir adressé en consultation de génétique votre patient, le jeune M D né le 2019 que j'ai rencontré accompagné de sa maman et de sa grand-mère pour discussion d'examen génétique chez cet enfant présentant une anomalie de l'iris de l'oeil gauche.
Je reprends d'abord les antécédents familiaux :
J'ai donc noté que M était le seul enfant du couple ....

Start	Length	Negated	Ascendant	Terms
25	10	False	True	Term1; Term2;....
100	8	True	False	Term1; Term2;....
...	...	...	...	...
350	15	False	False	Term1; Term2;....

Fig. 1: General architecture of the proposed ontology extraction software.

The system supports three operational scenarios:

1. Fully Automated Extraction: Only automatic parsers are enabled, producing results without manual intervention.
2. Semi-Automated Extraction: both automatic and manual annotation/curation systems are enabled. Firstly, the selected automatic parser extracts the target ontologies and proposes the results to the clinician. Secondly, the clinician can use the manual curation/annotation system to:
    a. Explore the detailed results.
    b. Rate the terms proposed by the automatic parser (curation).
    c. Add new annotation units if the automatic parser failed to detect some ontologies.
    d. Edit the proposed automatic annotation results.
3. Manual annotation: Enables clinicians to annotate medical notes from scratch without automated assistance.

In a production environment, scenarios 1 and 2 can be used. On the other hand, in the construction phase, scenarios 2 and 3 can be used for building a database of labelled medical reports. We have adopted the semi-automated scenario (scenario number 2) in order to create a labelled database for two reasons:

a) It speeds up the annotation process and thus saves the clinicians' time.
b) It gives valuable human feedback to the machine learning-based parsers by reporting the clinicians' ratings of each proposed term.

4. The Annotation User Interface

ACUITEE is a web application equipped with an interactive user interface that implements three main functionalities:

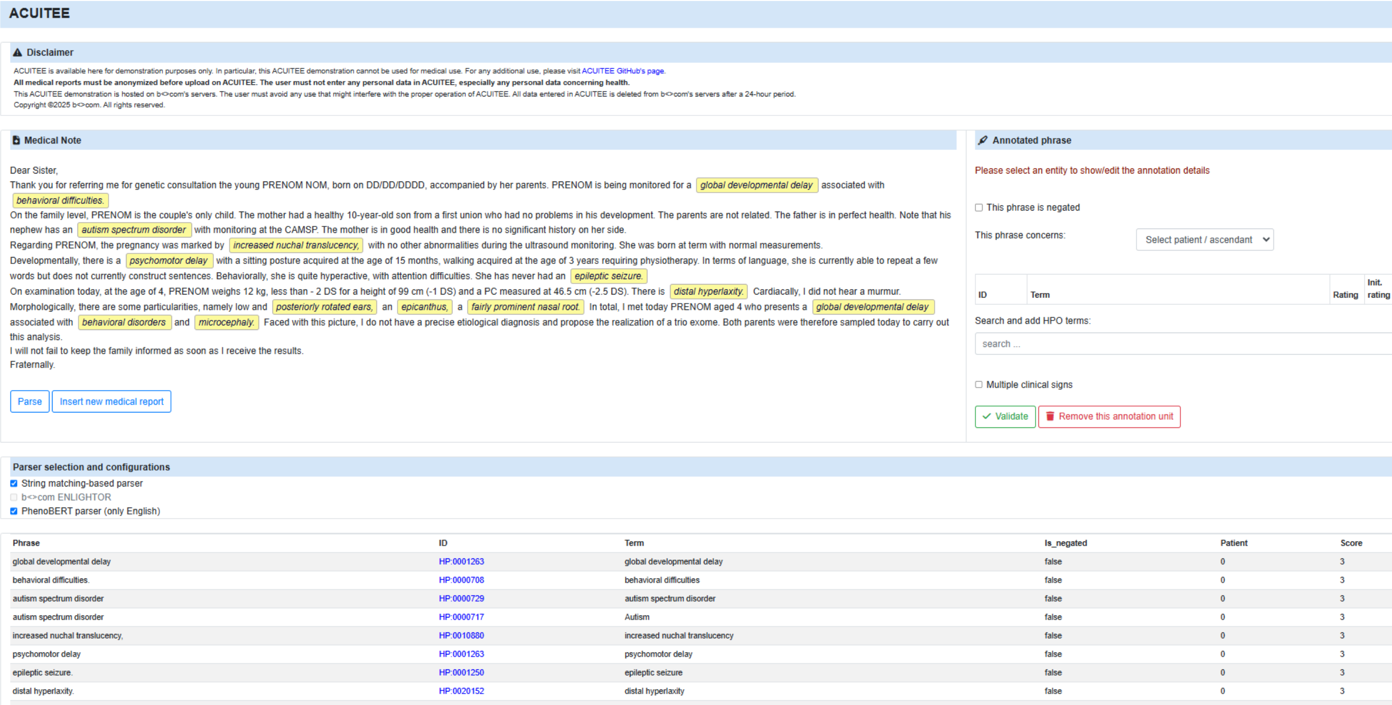
- Annotations visualization: highlights the detected spans with yellow color.
- Editing: adding new annotation units, modifying the properties of existing annotation units, searching for ontology terms, removing irrelevant annotation units, expanding/reducing the annotation span and merging several annotation units together.
- Manual curation: a stars rating system is used, where clinicians can rate any proposed term from 0 (0 ☆ ☆ ☆) to 3 (3 ★ ★ ★).

Figure 2 shows a snapshot of the user interface of ACUITEE. Before explaining the functionalities mentioned above, we introduce how the annotation units are represented.

4.1. Representation of an annotation unit

1. The annotated phrase e.g. “iris coloboma”.
2. Negation Status: Indicates whether the phrase confirms or negates the phenomenon.
3. The person concerned by the annotated phrase. This property can take one of the following five values:

a) The patient himself, b) An ascendant from the paternal line, c) An ascendant from the maternal line, d) Another patient discussed in the same medical note. For instance, one medical note could describe the phenotypes of twins studied in the same consultation, e) Other cases.



ACUITEE is available here for demonstration purposes only. In particular, this ACUITEE demonstration cannot be used for medical use. For any additional use, please visit ACUITEE GitHub's page.

All medical reports must be anonymized before upload on ACUITEE. The user must not enter any personal data in ACUITEE, especially any personal data concerning health.

This ACUITEE demonstration is hosted on b<>com's servers. The user must avoid any use that might interfere with the proper operation of ACUITEE. All data entered in ACUITEE is deleted from b<>com's servers after a 24-hour period.

Copyright ©2025 b<>com. All rights reserved.

Dear Sister,

Thank you for referring me for genetic consultation the young PRENOM NOM, born on DD/DD/DDDD, accompanied by her parents. PRENOM is being monitored for a *global developmental delay* associated with *behavioral difficulties.*

On the family level, PRENOM is the couple's only child. The mother had a healthy 10-year-old son from a first union who had no problems in his development. The parents are not related. The father is in perfect health. Note that his nephew has an *autism spectrum disorder* with monitoring at the CAMSP. The mother is in good health and there is no significant history on her side.

Regarding PRENOM, the pregnancy was marked by *increased nuchal translucency,* with no other abnormalities during the ultrasound monitoring. She was born at term with normal measurements.

Developmentally, there is a *psychomotor delay* with a sitting posture acquired at the age of 15 months, walking acquired at the age of 3 years requiring physiotherapy. In terms of language, she is currently able to repeat a few words but does not currently construct sentences. Behaviorally, she is quite hyperactive, with attention difficulties. She has never had an *epileptic seizure.*

On examination today, at the age of 4, PRENOM weighs 12 kg, less than - 2 DS for a height of 99 cm (-1 DS) and a PC measured at 46.5 cm (-2.5 DS). There is *distal hyperlaxity.* Cardiacally, I did not hear a murmur.

Morphologically, there are some particularities, namely low and *posteriorly rotated ears,* an *epicanthus,* a *fairly prominent nasal root.* In total, I met today PRENOM aged 4 who presents a *global developmental delay* associated with *behavioral disorders* and *microcephaly.* Faced with this picture, I do not have a precise etiological diagnosis and propose the realization of a trio exome. Both parents were therefore sampled today to carry out this analysis.

I will not fail to keep the family informed as soon as I receive the results.

Fraternally.

Please select an entity to show/edit the annotation details

☐ This phrase is negated

This phrase concerns: Select patient / ascendant

ID	Term	Rating	Init. rating

Search and add HPO terms:

☐ Multiple clinical signs

Parser selection and configurations

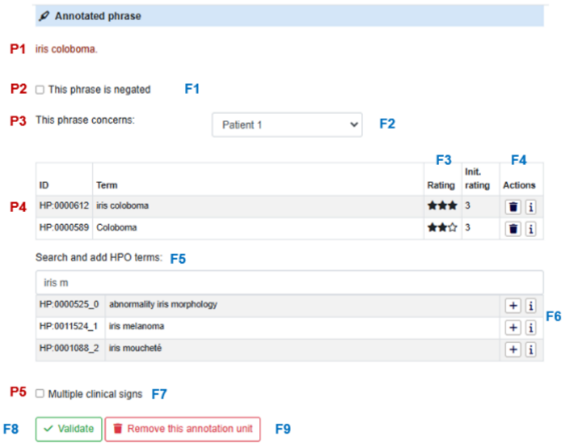
☑ String matching-based parser
☐ b<>com ENLIGHTOR
☑ PhenoBERT parser (only English)

Phrase	ID	Term	Is_negated	Patient	Score
global developmental delay	HP:0001263	global developmental delay	false	0	3
behavioral difficulties.	HP:0000708	behavioral difficulties	false	0	3
autism spectrum disorder	HP:0000729	autism spectrum disorder	false	0	3
autism spectrum disorder	HP:0000717	Autism	false	0	3
increased nuchal translucency,	HP:0010880	increased nuchal translucency	false	0	3
psychomotor delay	HP:0001263	psychomotor delay	false	0	3
epileptic seizure.	HP:0001250	epileptic seizure	false	0	3
distal hyperlaxity.	HP:0020152	distal hyperlaxity	false	0	3

Fig. 2: A snapshot of the user interface of ACUITEE.

4. Ontology Terms: A list of standard ontology terms associated with the annotated phrase. For example, the following standard HPO terms could be selected to describe the phrase “iris coloboma”: 1) coloboma and 2) cat eye and 3) abnormal iris morphology.
5. A flag indicating whether the selected standard terms represent multiple clinical signs or not.

These five properties of annotation units are illustrated in the right panel “Annotated phrase” in Figure 3. The following subsections present the three main functionalities of ACUITEE.



Annotated phrase

iris coloboma.

☐ This phrase is negated

This phrase concerns: Patient 1

ID	Term	Rating	Init. rating	Actions
HP:0000612	iris coloboma	★★★	3	
HP:0000589	Coloboma	★★☆	3	

Search and add HPO terms:

iris m

HP:0000525_0	abnormality iris morphology
HP:0011524_1	iris melanoma
HP:0001088_2	iris moucheté

☐ Multiple clinical signs

Fig. 3: The annotated phrase panel in ACUITEE.

4.2 Visualization of the annotation results

When the medical note is analyzed using some automatic parsing engine, the results are shown on the medical report where all the automatically annotated phrases are highlighted in yellow as shown in the left panel “Medical note” of Figure 2. The clinician can select any annotated phrase to see its detailed properties in the right-side panel “Annotated phrase”. This panel is shown in Figure 3 with associated labels to simplify the explanation. The visualized details of an annotation unit include (as denoted in Figure 3) :

P1: The highlighted phrase.

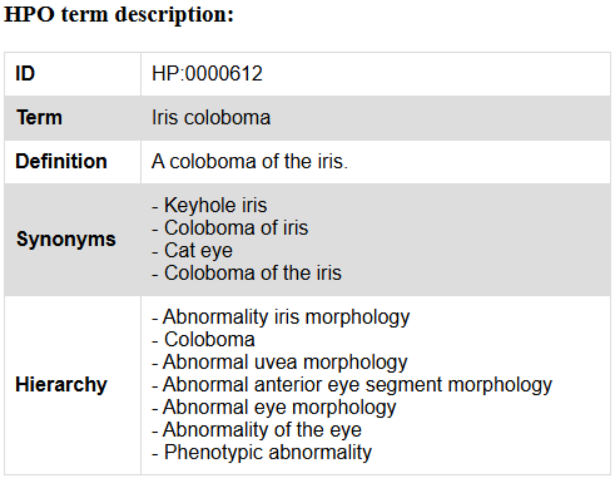
P2: The negation status i.e. the phrase is negated/not negated.

P3: The concerned person: the patient or one of his ascendants.

P4: List of proposed HPO terms (or terms from any other standard ontology).

P5: A binary value indicating whether the selected terms represent multiple clinical signs or not.

Moreover, detailed information about the HPO terms can be retrieved by pressing on the info button in the corresponding row. For instance, when the user clicks the info button associated with the HPO term “Cat eye,” a window, as depicted in Figure 4, will appear. This window displays information such as the term’s ID, definition, synonyms, and its hierarchical relationships with other HPO terms



HPO term description:

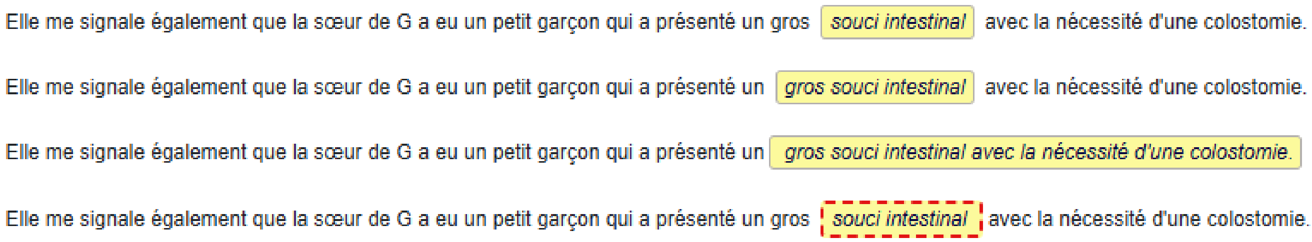
ID	HP:0000612
Term	Iris coloboma
Definition	A coloboma of the iris.
Synonyms	- Keyhole iris - Coloboma of iris - Cat eye - Coloboma of the iris
Hierarchy	- Abnormality iris morphology - Coloboma - Abnormal uvea morphology - Abnormal anterior eye segment morphology - Abnormal eye morphology - Abnormality of the eye - Phenotypic abnormality

Fig. 4: The information window showing detailed information about the HPO term “Iris coloboma”.

4.3 Editing the annotation results manually

The properties of any annotation unit can be manually edited by the clinician. The negation status and the “Multiple clinical signs” flag can be edited using the check boxes indicated respectively by F1 and F7 in Figure 3. The concerned person, i.e. patient or ascendant, can be changed using the combo box indicated by F2. The list of associated HPO terms can be modified using the function groups F4, F5 and F6. Particularly, the “trash box” button in the function group F4 can be used to delete the corresponding HPO term. The search box F5 can be used to look for HPO terms by typing some keywords. Then, the “+” button can be used to add the desired HPO terms. The validate button F8 can be used to indicate that the clinician has revised and validated the annotation unit. Consequently, the term will appear in blue in the “Medical note” panel. Finally, the entire annotation unit can be removed using the button “Remove this annotation unit”. Besides the aforementioned functionalities, the clinician can add new annotation units if needed, i.e. if the automatic parser has not succeeded to annotate an informative phrase, the clinician can highlight the desired phrase by mouse and then he hits “Enter” to add a new annotation unit. In addition, the annotation span can be expanded to the left or right or a sub-span can be selected as illustrated in Figure 5. The clinician can also merge multiple annotation units together. An example is given in Figure 6 where we have two annotated phrases:

- “retinal coloboma” a negated phrase which concerns an ascendant from the paternal line. It has two associated HPO terms: 1) Hole in the back of the eye and 2) Coloboma.
- “microphthalmia” a negated phrase which concerns an ascendant from the paternal line. It has one associated HPO term: abnormally small eyeball.



Elle me signale également que la sœur de G a eu un petit garçon qui a présenté un gros *souci intestinal* avec la nécessité d'une colostomie.

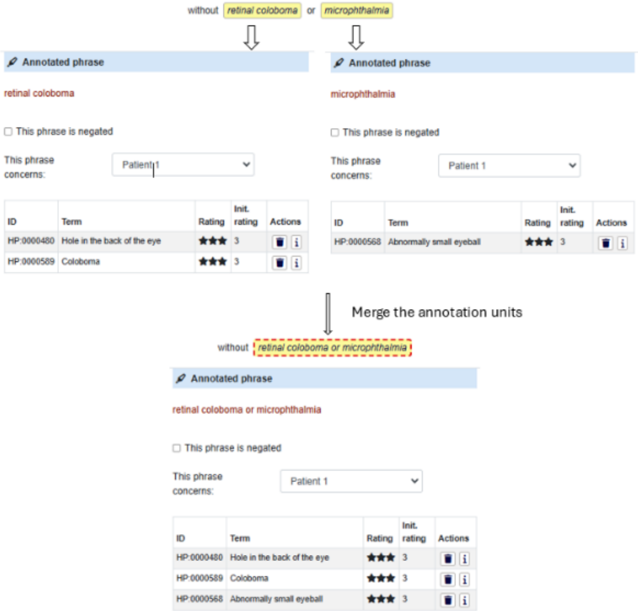
Elle me signale également que la sœur de G a eu un petit garçon qui a présenté un *gros souci intestinal* avec la nécessité d'une colostomie.

Elle me signale également que la sœur de G a eu un petit garçon qui a présenté un *gros souci intestinal avec la nécessité d'une colostomie.*

Elle me signale également que la sœur de G a eu un petit garçon qui a présenté un gros *souci intestinal* avec la nécessité d'une colostomie.

Fig. 5: Modifying the annotation span.

The clinician could prefer to fuse the two annotation units proposed by the parser in one unit. He can simply highlight both annotation units by mouse and hit Enter to merge them. As we observe from Figure 6, the new annotation unit contains all the HPO terms associated with the merged annotation units. In addition, since both merged units are negated, the resultant one is. Similarly, the concerned person is kept as “Ascendant: Paternal line” since both merged units share this value. However, when there is some conflict between the merged units, the results will be set to the default values.



without retinal coloboma or microphthalmia

Annotated phrase

retinal coloboma

☐ This phrase is negated

This phrase concerns: Patient 1

ID	Term	Rating	Init. rating	Actions
HP:0000480	Hole in the back of the eye	★★★	3	
HP:0000589	Coloboma	★★★	3	

Annotated phrase

microphthalmia

☐ This phrase is negated

This phrase concerns: Patient 1

ID	Term	Rating	Init. rating	Actions
HP:0000568	Abnormally small eyeball	★★★	3	

Merge the annotation units

without retinal coloboma or microphthalmia

Annotated phrase

retinal coloboma or microphthalmia

☐ This phrase is negated

This phrase concerns: Patient 1

ID	Term	Rating	Init. rating	Actions
HP:0000480	Hole in the back of the eye	★★★	3	
HP:0000589	Coloboma	★★★	3	
HP:0000568	Abnormally small eyeball	★★★	3	

Fig. 6: Merging multiple annotation units.

4.4 Manual curation of the terms proposed by the automatic parsers

Manual curation of the terms proposed by the automatic parsers is implemented by means of a 3-star rating system where the clinician can rate any proposed term from 0 (0 ☆ ☆ ☆) to 3 (3 ★ ★ ★). The curation system is designed to be used in the construction phase of the automatic parsers. Particularly, the intended parsers are based on machine-learning and thus they require a set of labelled data for training. Instead of building a training set of labelled medical notes from scratch, the proposed curation system enables to use the results proposed by multiple parsers even when those latter are only partially trained. In fact, this strategy has two advantages:

1. It speeds up building the training set of labelled medical notes in comparison with starting from scratch and thus saves the clinicians' time.
2. It provides valuable human feedback that can be used to adjust the behavior of the parsers.

In fact, the initial rating of the HPO terms proposed by an automatic parser is set to 3. Thus, the clinician doesn't need to change the rating unless the proposed term is irrelevant. However, if the clinician added a term manually, the initial rating is set to 3 assuming that the added term is completely relevant.

4.5 Ontology data sources

The full HPO ontology including the terms, synonyms and hierarchical relationships is stored locally on the server that runs the ACUITEE. In the current implementation, the HPO version of 08-02-2021 is used. It is important to use the same version of the HPO in order to avoid any conflicts that could take place if the terms used to manually annotate some medical notes have been changed in newer versions. Besides the aforementioned reason of using a local version of the ontology base, it also represents a good solution for the privacy of the medical notes since no data are required to be sent outside the local infrastructure. In fact, the HPO ontology is written in English (with only a subpart in other languages) while the target medical notes are in French. To enable the manual search using French keywords, a French version of the HPO dataset has been generated using Google neural translation models [17].

5 Parsing engines

Natural Language Processing (NLP) [1] has become a cornerstone of biomedical research, enabling the extraction and structuring of valuable information from unstructured clinical text. By leveraging techniques ranging from traditional string-matching algorithms to advanced transformer-based models, NLP facilitates tasks such as phenotype extraction, disease mapping, and automated annotation. These technologies empower clinicians and researchers to handle large-scale datasets efficiently, fostering advancements in precision medicine and healthcare delivery.

In the ACUITEE platform, we implemented two parsing engines to support diverse annotation workflows: a baseline string-matching parser and a transformer-based PhenoBERT [14] parser. This flexibility allows users to choose between parsers or combine their outputs for comprehensive analysis. New parsers will be integrated in future work.

5.1 The baseline string-matching parser

We implemented string-matching efficiently using the Aho-Corasick algorithm [15] through the Python library ahocorasick-rs (version 0.9.1) [16]. This setup supports searches in both English and French, leveraging the previously mentioned French-translated version of the HPO dataset. This parser is very efficient for explicit phenotypes, however more advanced technics should be used for implicit ones.

5.2 The phenoBERT-based parser

PhenoBERT is a hybrid method that uses advanced deep learning methods such as CNN [20] and BERT [18] to identify clinical disease phenotypes from free clinical text. In our previous work [19], we evaluated the performance of PhenoBERT in detail. Building on this, we developed an API that receives English medical notes as input and processes them through the transformer-based model PhenoBERT, functioning as the backend. The output from PhenoBERT is then reformulated into a structured format compatible with the ACUITEE platform.

6 Conclusion and perspectives

ACUITEE represents a comprehensive tool that combines visualization, editing, and manual curation to improve the efficiency and quality of annotation processes, particularly in free-text clinical data. Its scalable architecture and user-focused interface make it a promising solution for diverse clinical applications.

ACUITEE has been successfully deployed in the management of clinical data, demonstrating significant improvements in annotation accuracy and efficiency. Case studies have highlighted its impact in real-world environments, particularly in multilingual and resource-diverse settings.

Built using modern web technologies and APIs, the system offers high performance and scalability. Key technical features include backend integration for multiple parsers with minimal latency, local storage of ontology databases to ensure data privacy, and multilingual support, such as French translations of ontology terms for efficient use with French medical notes in string-matching-based parsers.

This tool provides a unified interface for production and experimental use cases, enabling users to test different parsers and generate curated datasets for machine learning projects. Future developments will focus on integrating new AI-driven extraction engines to further enhance annotation accuracy, expanding its application to other domains, and providing seamless integration of any ontology. Advanced visualization tools will also be introduced to simplify annotation workflows further.

Acknowledgements

We would like to thank our colleagues and contributors who assisted in the development and testing of ACUITEE.

References

[1] Manning, C. D., and Sch¨utze, H. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[2] Aronson, A. R., and Lang, F. M., “An overview of MetaMap: historical perspective and recent advances.” J Am Med Inform Assoc. 2010 May-Jun;17(3):229-36. doi:10.1136/jamia.2009.002733. PMID: 20442139; PMCID: PMC2995713.10.

[3] Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., and Chute, C. G. “Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications,” Journal of the American Medical Informatics Association, vol. 17, no. 5, pp. 507–513, 2010.

[4] SNOMED International, “SNOMED CT: The global language of healthcare,” [Online]. Available: https://www.snomed.org. [Accessed: Dec. 19, 2024].

[5] Liu, C., Peres Kury, F. S., Li, Z., Ta, C., Wang, K., and Weng, C. “Doc2Hpo: a web application for efficient and accurate HPO concept curation.” Nucleic Acids Research, vol. 47, no. W1, 2019, pp. W566-W570. doi:10.1093/nar/gkz386. . [Online]. Available: https: //doi.org/10.1093/nar/gkz386.

[6] Deisseroth, C. A., Birgmeier, J., Bodle, E. E., Kohler, J. N., Matalon, D. R., Nazarenko, Y., Genetti, C. A., Brownstein, C. A., Schmitz-Abe, K., Schoch, K., Cope, H., Signer, R., Undiagnosed Diseases Network, Martinez-Agosto, J. A., Shashi, V., Beggs, A. H., Wheeler, M. T., Bernstein, J. A., and Bejerano, G. “ClinPhen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis.” Genetics in Medicine, 2018.

[7] Stenetorp, P., Pyysalo, S., Topi˘c, G., Ohta, T., Ananiadou, S., and Tsujii, J. “BRAT: A Web-based Tool for NLP-assisted Text Annotation,” in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France, 2012, pp. 102–107.

[8] Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M.-A., Chute, C. G., and Musen, M. A. “BioPortal: Ontologies and Integrated Data Resources at the Click of a Mouse,” Nucleic Acids Research, vol.37, no. suppl 2, pp. W170–W173, 2009.

[9] Cejuela, D., McQuillan, D., Keane, G., Jardin, A., and Kruse, F. “tagtog: Interactive and Collaborative Text Annotation,” in *Proceedings of the 14th International Conference on Semantic Systems*, Vienna, Austria, 2018, pp. 310–314.

[10] Garcia, B., and Posey, J. “Improving Automated Deep Phenotyping Through Large Language Models Using Retrieval Augmented Generation.” medRxiv, 2024. doi:10.1101/2024.12.01.24318253.

[11] South, S., Shen, S., and Leng, S. “eHOST: Extensible Human Oracle Suite of Tools for Annotating Unstructured Clinical Text,” in *Proceedings of the 2012 AMIA Annual Symposium*, Chicago, USA, 2012, pp. 857–862.

[12] Ogren, T. “Knowtator: A Protégé Plug-in for Annotated Corpus Construction,” in *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, New York, USA, 2006.

[13] K¨ohler, S., Gargano, M., Paul, S., Schultz, J., Batzler, J., and Robinson, P. N. “The Human Phenotype Ontology in 2021”. Nucleic Acids Research, 49(D1), D1207–D1217. [Online]. Available: https://doi.org/10.1093/nar/gkaa1049.11

[14] Feng, Y., Qi, L., and Tian, W. “PhenoBERT: A Combined Deep Learning Method for Automated Recognition of Human Phenotype Ontology.” IEEE/ACM Trans Comput Biol Bioinform. 2023 Mar-Apr;20(2):1269-1277. doi: 10.1109/TCBB.2022.3170301.PMID: 35471885.

[15] Aho, A. V., and Corasick, M. J. “Efficient string matching: An aid to bibliographic search”. ACM Transactions on Information Systems, 2(3), 131–160. [Online]. Available: https://doi.org/10.1145/360825.360855.

[16] *ahocorasick-rs (version 0.9.1)*. (2021). [Online]. Retrieved from https://pypi.org/project/ahocorasick-rs/.

[17] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean, ”Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation,” arXiv preprint arXiv:1609.08144, 2016. [Online]. Available: https://arxiv.org/abs/1609.08144. [Accessed: Jan. 23, 2025].

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Minneapolis,MN, USA, June 2019, pp. 4171–4186. [Online]. Available: https://arxiv.org/abs/1810.04805.

[19] M. Baddour, S. Paquelet, P. Rollier, M. De Tayrac, O. Dameron and T. Labbe, “Phenotypes Extraction from Text: Analysis and Perspective in the LLM Era,” *2024 IEEE 12th International Conference on Intelligent Systems (IS)*, Varna, Bulgaria, 2024, pp. 1-8, doi: 10.1109/IS61756.2024.10705235.

[20] Keiron O’Shea and Ryan Nash. (2015). “An Introduction to Convolutional Neural Networks.” arXiv preprint arXiv:1511.08458.