

Terue Takatsuki 1), Toyofumi Fujiwara1), Yasunori Yamamoto1)

1) Database center for life science, Japan

● Abstract

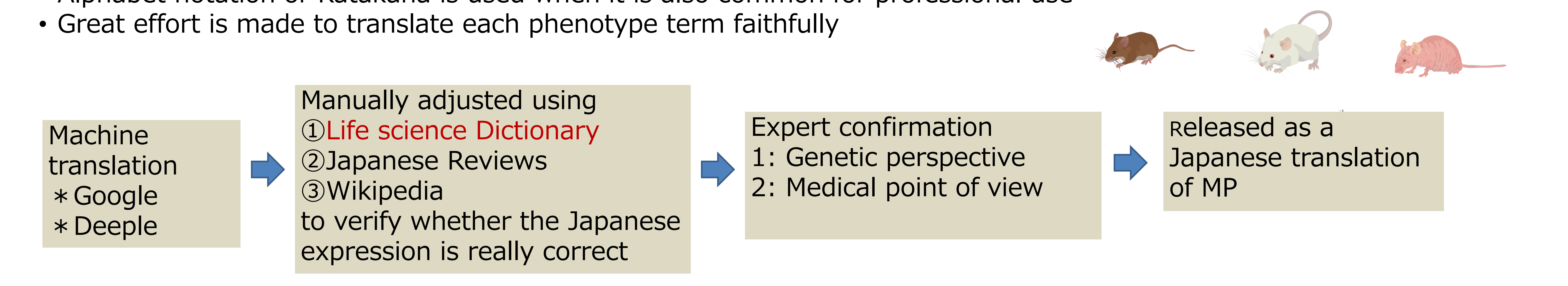
MGI (Mouse genome informatics) において作成されているMammalian phenotype ontology (MP) は、マウスの表現型やラットの表現型のアノテーションに広く利用されている。MPをアノテーションに用いる事で、表現型情報の網羅的検索、相互比較、異なるデータベース間での表現型情報共有などを改善する。MPの日本語化は、日本における表現型アノテーションを加速することを目的として、理研バイオリソース研究センター (RIKEN BRC) で着手され、現在はライフサイエンス統合データベースセンター (DBCLS) で、提供、更新等の作業が行われている。日本語化にあたっては、機械的翻訳に加え、日本語における適切な表現にするため、マニュアルでの確認作業を実施し、機械翻訳では難しい、日本語における哺乳類と人との表現の違い等を加味し、正しい表現を用いる様に修正を行っている。また、機械学習や検索精度の向上のため、訳に用いる日本語語彙の統一化も行っている。2023年度はMGIとの連携が実現し、作成した日本語訳はMGIにおいてもInternational_versionとして配布が開始された。このMPの日本語化において、マニュアルキュレーションで最も用いているライフサイエンス辞書について、開発者の退官により、DBCLSへの移管が行われることになり、現在その引継ぎ作業を実施している。

The Mammalian Phenotype Ontology (MP), developed by the Mouse Genome Informatics (MGI), serves as a widely utilized tool for annotating phenotypes in mice and rats. The application of MP in annotation enhances comprehensive searches for phenotypic information, facilitating comparisons across models. The initiative to make a Japanese translation of MP for Japanese users was spearheaded by the RIKEN BioResource Research Center (RIKEN BRC) with the goal of expediting phenotypic annotation activities in Japan. This ongoing effort receives support from the Life Science Integrated Database Center (DBCLS).

The procedure of Japanese translation of the Mammalian phenotype ontology.

訳の方法についてはGoogleやDeepleでの機械翻訳を最初に行い、その後ライフサイエンス辞書を用いた、マニュアルで確認修正作業を実施、更に遺伝学や医学的な専門用語については、専門家に監修を依頼している。日本語訳がある場合でも、専門的にアルファベット表記やカタカナ読みが一般的な場合には、そちらを採用し、検索語彙として優位になる様にしている。

- Machine translation was performed first, followed by manual curation
- Experts were involved to supervise genetics and medical terminology
- Alphabet notation or Katakana is used when it is also common for professional use
- Great effort is made to translate each phenotype term faithfully



The use case of RIKEN bioresource center

理研バイオリサーチセンターでは、既に公開されているMPの日本語訳をリソースの検索システムに取り込むことで、マウスリソースに対してアノテーションされているMPに対して、日本語で検索を行い、アノテーションしたマウスリソースが表示される仕組みを構築している。

日本語で検索を行えることで、国内のマウス利用者にとってより検索のしやすい仕組みを提供している。

At the RIKEN BioResearch Center, by incorporating the Japanese translation of MP into the resource search system, we can search in Japanese for MP annotation for mouse resources, and the corresponding mouse resources will be displayed. Being able to search in Japanese removes barriers for domestic users performing mouse-based research.

● ROIS-DS -JOINT2023-1 採択課題
哺乳類表現型語彙の日本語整備とオントロジーアライメントおよびそれを活用したバイオリソース検索システムの構築
Preparation of Japanese terms of mammalian phenotype ontology, the ontology alignment, and development of a bioresource retrieval system.

https://web.brc.riken.jp

MP:0005386(abnormal behavior)でアノテーションされたマウスの検索例

Repository and How to use MGI

現在MPの日本語訳については、日々増える新規タームに対して、定期的にフォローアップを行い、追加更新を継続して実施している。データはGithub (https://github.com/dbcls/MP_Japanese) で公開しており、誰でも自由に使う事ができる。日本語訳データについては、TSV形式及びRDF形式で公開している。またUseCaseとして、MGIとRGDが提供するMPがアノテーションされたファイルに対して、MPの英語ラベル、日本語ラベルを付与したファイルも定期的に提供を行っている。

- MP Japanese translation is being regularly updated
- Data are accessible to everyone and published in Github
- (https://github.com/dbcls/MP_Japanese)
- Formats are TSV and RDF (Resource Description Framework)

MPの日本語訳はMGIとの連携により、MGIのサイトでもInternational Versionとして配布されており、日本語訳の新規追加更新ファイルのアップデートに合わせて、MGI側でのデータ更新も行う様パイプラインを構築している。

In collaboration with MGI, the Japanese-translated MP data has been made accessible on MGI's download site (<https://www.informatics.jax.org/downloads/reports/index.html#pheno>).

Application of Life science dictionary

長年、広く使われてきたライフサイエンス辞書LSD (Life science dictionary) は、開発者の京都大学金子教授の定年により、DBCLSにサービスが移管されることになり、DBCLSにおける、LSDの管理体制の整備が必要となっている。そのため、従来の更新管理体制を一度引継ぎ、そこからより効率化した管理更新体制の構築を行う。MPの日本語訳化においても、LSDをマニュアルチェックの第一優先として利用している。

The Life Science Dictionary (LSD), which has been widely used for many years, will be transferred to DBCLS due to the retirement of its developer, Professor Kaneko of Kyoto University, and the LSD management system at DBCLS needs to be rebuilt. Therefore, the existing updating and management system will be taken over once, and a more efficient management and updating system will be established from there. In translating MP into Japanese, LSD is also used as the first priority academic words of Japanese for manual checking.

● Future plan

- MPの日本語訳の提供は、今後も定期的に見直しを行い、更新を行う予定である。日本語訳については、ユーザーからのより良いフィードバックを期待している。
Collect feedback from the Japanese translation users
- MPの日本語訳を整備することにより、今後日本語論文の自動テキストマイニング用の辞書としての活用することも可能である。
Expand use cases to include text-based knowledge extraction, auto-complete, dictionary-based workflows
- LLMを用いた機械翻訳部分の精度の向上を目指し、LLMの評価を行う
Evaluate LLM with the aim of improving the accuracy of machine translation
- ライフサイエンス辞書 (LSD) の活用
Utilization of life science dictionary (LSD)