

「Wiki をもちいたデータベース設計」

有田 正規

東京大学大学院新領域創成科学研究科、理化学研究所植物科学研究センター

慶應義塾大学先端生命科学研究所

e-mail: arita@k.u-tokyo.ac.jp

Wikipedia のために開発された MediaWiki は多くの生物学関連プロジェクトで利用されているが、いずれも「各ページが独立に編集され、機能する」という Wiki 本来の仕様をそのまま踏襲している。これらをバイオデータベースと呼ぶのは間違いである。テキストファイルの編集を容易にただけに過ぎない。データの整合性をチェックする機能が欠けており高度な検索を施すこともできない Wiki の場合、むしろ、全データを一枚のファイルにまとめたテキスト形式よりも使いづらくなる可能性が大きい。

Wiki に最も足りない点は「各データは一回しか登録しない」(つまり修正するときは一箇所だけ修正すればよい)というデータベースの原則である。例えば、どの Wiki にも他ページの内容を集計して表示する機能が存在しない。ページ間での情報再利用(コピー)も許さない。だから、ページのインデックスや分類表、統計結果を作成することができない。こうした再利用機能の欠如は、Wikipedia のような百科事典の作成には不要であるが(むしろ剽窃を助長しないので好ましいかもしれない)、情報検索により必要部分を切り出す機能を本質とするデータベースとは相容れない発想になってしまう。

我々は、MediaWiki を機能拡張してデータベースを作成する方法を紹介したい。ページ内を文字列検索する機能を Wiki コードとしてページ内に記述できるようにすることで、データベースに類似の機能(たとえば統計の算出や、矛盾の検出)を持たせられる。具体例として 7000 分子のフラボノイドを収録したデータベースの構築を紹介したい。Wiki のページは化合物、植物種、文献の3情報から構成され、ページ間で相互に情報を自動収集してビューを構成する。また、Wiki は基本的にテキストデータしか扱えないが、型変換を適切に行なうことで質量数などの数字も扱える。特に提案したいのが、ページ内検索機能の導入によってなかば必然的に導入される機能である。例えば、インデックスや統計結果のようにページの中身が全て他ページに依存する情報はページタイトルも含めて全て自動生成、更新すべきであろう。またそうしたページが検索エンジンにかかるような工夫もしたい。こうした機能を洗い出していくと、インターネット上で情報を配信する RSS やユーザの入力に応じたビューを提供する CGI といった機能が、ページ内検索の帰結として示唆されることを紹介する。