LOD上の情報をクラス間関係で切り取る基盤技術開発

ライフサイエンス統合データベースセンター
山口 敦子

理化学研究所 情報基盤センター
小林 紀郎

理化学研究所 バイオリソースセンター
桝屋 啓志

ライフサイエンス統合データベースセンター
山本 泰智

大阪大学 産業科学研究所
古崎 晃司

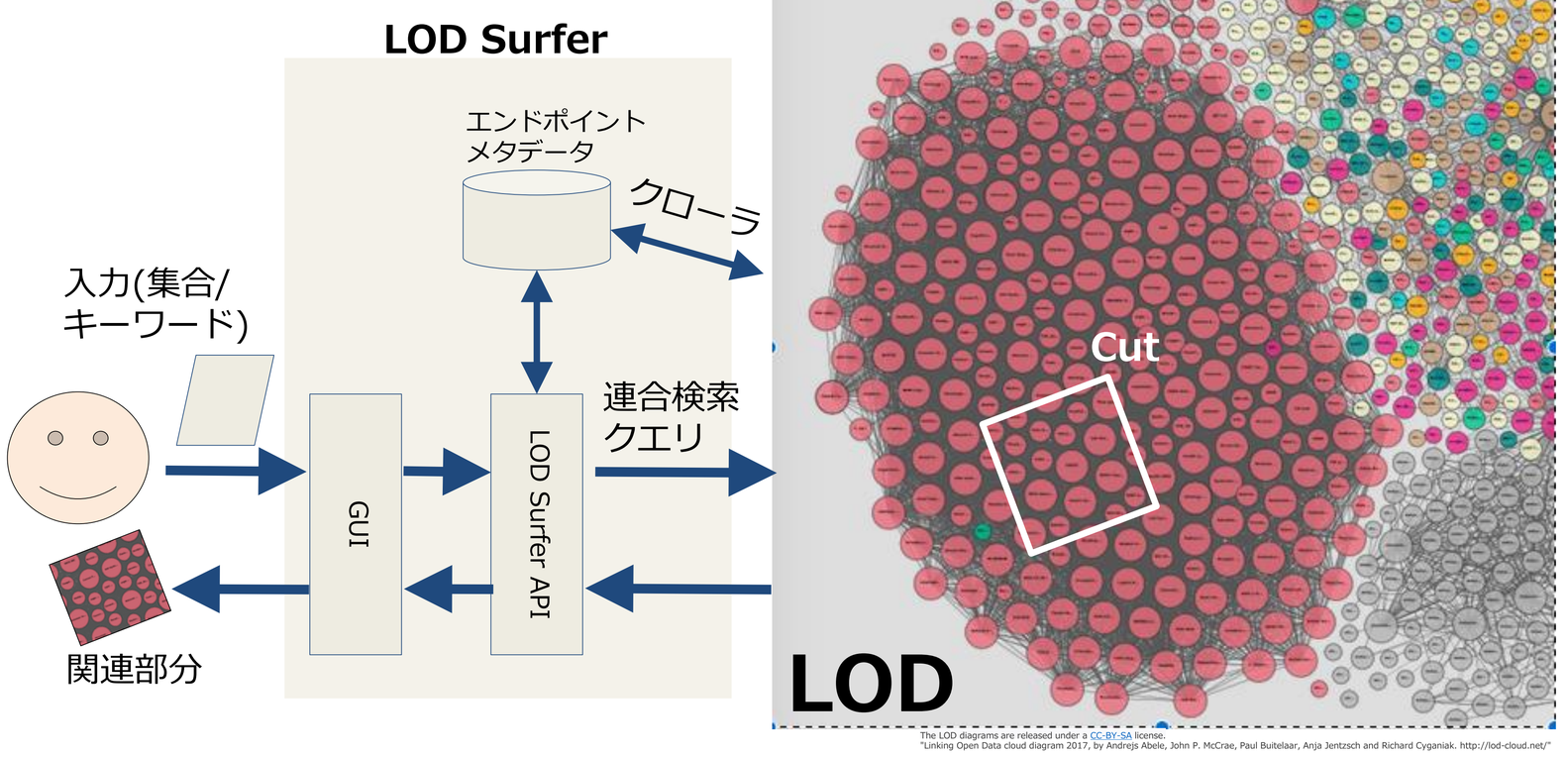
Linked Open Data (LOD)上の情報を自在に切り取るために

発表者らはクラス間関係提示を用いた対話的なGUI を介してLOD 上の情報をクラス間関係で切り取り、ユーザが着目したLOD データを取得するシステム LOD Surfer の開発をすすめている．このツールを実現するため，これまでSPARQL Builder のために開発してきた「クラス間関係探索」「メタデータ取得蓄積」の技術に加え，「連合検索」「クラスグラフ解析」などの必要な技術を開発している．本発表では LOD Surfer の全体像を展望しつつ，その実現に必要となるそれぞれの技術を紹介する．

LOD Surfer

LOD Surfer とは，対話的なGUI を介して，ユーザが興味をもつデータやキーワードに対して，関連するLOD上の情報をクラス間関係で切り取り，出力するシステムである．

入力されたデータやキーワードは，システム内で，LOD上のクラスに紐づけられ，さらに，そのクラスを始点としたクラス間関係探索により，入力と関連する部分をLODから切り取り提示することができる．



LODクラスグラフ

対話的にユーザへ関連クラス情報を提示したり，クラス間関係を効率的に探索できるように，LOD Surfer は事前にSPARQLエンドポイント(LOD の標準WebAPI)をクロールし，どんなデータ(クラス)がどういう構造(プロパティ)で記述されているかを取得し，メタデータとして蓄積している．このメタデータからLOD全体の構造を見渡し，効率的探索法を設計するため，LODクラスグラフを構築し，解析を試みた．

LODクラスグラフ

頂点＝クラス
辺＝プロパティとエンドポイント情報

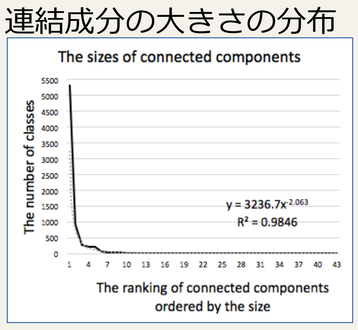
クラス間関係＝
クラスグラフ上の二つの頂点をつなぐパス

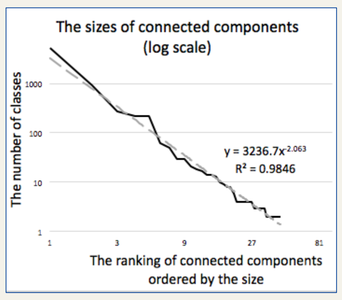
LODクラスグラフ解析

対象：43エンドポイント，76データセット
解析内容：連結成分を解析し，複数のエンドポイントにまたがったものがどのくらいあるか，また，複数のエンドポイントにまたがった連結成分に対し，どのクラスがつなぐ役割を果たしているかを明らかにする．

LODクラスグラフ解析結果

連結成分の大きさの分布





連結成分の大きさトップ3

	大きさ	EP数	エンドポイント情報
1	5327	32	31 eagle-i関連, EBI RDF platform
2	925	6	Bio2RDF, EBI RDF platform, WikiPathways, LinkedCT, DisGeNet and Organic Edunet
3	269	1	Bio2RDF

連結成分1におけるつなぐクラス

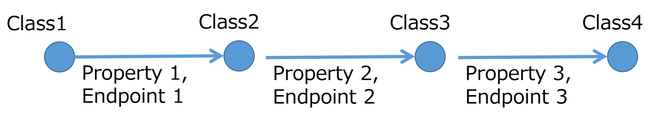
ラベル	オントロジー	EP数	分割数
Core Laboratory	VIVO	27	16
Software	ERO	23	11
Laboratory	VIVO	19	20
protocol	OBI	13	3
Mus musculus	NCBI Taxon	11	4
monoclonal antibody reagent	ERO	8	4
algorithmic software suite	ERO	5	4
human subject	ERO	3	4
Technology Transfer Office	ERO	2	3

連結成分2におけるつなぐクラス

ラベル	オントロジー	EP数	分割数
BiochemicalReaction	BioPax	2	3
SmallMoleculeReference	BioPax	2	3
ModificationFeature	BioPax	2	3

共通オントロジーの利用がデータセットをつなぐカギとなっている

クラスパスと連合検索



素朴な方法

それぞれの辺に対し，対応するエンドポイントに問い合わせ，返ってきた答えを結合する
→一つでも巨大な答えが出るクラスの組み合わせがあると遅くなる

提案手法

メタデータを利用し，小さい答えが返ってくると予想される順に問い合わせ，返ってきた答えを順次結合する

今後の課題

GUIの設計および実装
クローラの性能向上
APIの返答速度向上
連合検索システムの実装