Méthodes du Web sémantique pour l'intégration de données en sciences de la vie

Olivier Dameron

PARTIE 1

Titre de la partie

1

Méthodes du web sémantique pour l'intégration de données en sciences de la vie

Olivier DAMERON[1]
[1] *Univ Rennes, CNRS, Inria, IRISA - UMR 6074, 35000 Rennes, France*

Les sciences de la vie sont à la fois intrinsèquement *compliquées* à cause du grand nombre d'éléments différents qui entrent en jeu, et *complexes* à cause de l'interdépendance forte de ces éléments (Bult 2006 ; Bodenreider and Stevens 2006). Elles cherchent typiquement à mettre en évidence aussi bien des règles générales que des signaux faibles, dans un domaine où s'entremêlent la variabilité intra et inter-individuelle ainsi qu'un gradient de situations allant du sain au pathologique. Cette situation se trouve répliquée à différents niveaux de granularité allant des molécules jusqu'aux organismes et aux écosystèmes. Pour ne rien arranger, les données sont généralement bruitées et incomplètes.

Jusqu'à récemment, la faible quantité de données disponibles sous forme numérique, et les capacités de traitement limitées imposaient la double contrainte de travailler sur des domaines fragmentés (précis et étroits, ou larges mais peu profonds), et d'avoir recours à des hypothèses simplificatrices (Blake and Bult 2006).

L'évolution des capacités d'acquisition des données ainsi que des méthodes et infrastructures permettant leur analyse (l'Internet, les grilles de calcul,...) a permis l'explosion de la production de données disponibles dans des domaines complémentaires comme les données omiques, les phénotypes, les pathologies, le micro et le macro-environnement (Aldhous 1993 ; Blake and Bult 2006 ; Cannata *et al*. 2005 ;

Méthodes du Web Sémantique pour l'intégration de données en sciences de la vie,
coordonné par Olivier DAMERON.

Bellazzi *et al*. 2011). Cette nouvelle situation laisse entrevoir la possibilité de dépasser l'ancienne approche fragmentée pour aborder la complexité des sciences de la vie de manière intégrée et systématique.

La section 1.1 présente les caractéristiques des bases de données en sciences de la vie, identifie les besoins liés à leur utilisation dans des analyses systématiques, et compare ces besoins aux principales approches courantes. La section 1.2 présente les principes généraux des technologies du Web Sémantique et montre comment ils pourraient répondre aux besoins précédents. Enfin, la section 1.3 présente les domaines de recherche actuels pour l'intégration de données omiques.

1.1. Besoins liés aux données en sciences de la vie

Aujourd'hui, nous sommes entrés dans une ère de production à grande échelle de données en sciences de la vie, et ces données sont disponibles sous forme électronique. Cependant, on s'aperçoit que cela ne suffit pas pour faire face à la complexité du domaine du vivant.

La manière de traiter ces données est ainsi devenue un domaine d'étude à part entière. Cela fait dire à Lenoir dès 1998 de façon un peu provocatrice que la biologie est devenue une science de l'information (et d'une certaine manière, cela s'applique à toutes les sciences expérimentales). Cela englobe deux aspects :

– le premier concerne évidemment le développement de méthodes d'analyse capables de traiter des données ayant toutes les caractéristiques détaillées au début de ce chapitre. On parle de *data science*.

– le second concerne la représentation des données et de leurs relations, ainsi que la façon de les interroger pour permettre l'analyse. On parle de *data engineering*.

Ce chapitre traite du second aspect.

1.1.1. *Bases de données en sciences de la vie*

Le portail BioMart[1] offre une interface permettant un accès unifié à plus de 800 ensembles de données biologiques hébergés dans une quarantaine de bases de données à travers le monde et couvrant notamment les aspects génomiques, protéomiques ou de cancer (Smedley *et al*. 2015). *Nucleic Acids Research* recense environ 1600 bases biologiques de référence[2] (Rigden and Fernández 2019). Ce « *déluge des données* » (Aldhous 1993) est l'appellation pour les sciences de la vie du phénomène

1. http://www.biomart.org/
2. http://www.oxfordjournals.org/our_journals/nar/database/cap/

plus général du « *Big data* » avec comme spécificités que les quantités de données y sont les plus importantes, et que ces données sont fortement interconnectées (Stephens *et al*. 2015). De plus, cette tendance va vraisemblablement s'amplifier (Stephens *et al*. 2015). À ces bases de référence, s'ajoute la multitude des bases spécifiques à chaque projet et des entrepôts de données cliniques (voir chapitre **??**).

Pour analyser les besoins liés aux données en sciences de la vie, il est ainsi nécessaire de prendre en compte la grande quantité de bases de référence en résolvant la contradiction entre d'une part leur nature complémentaire nécessitant des références croisées, et d'autre part le manque d'interopérabilité de leurs schémas et de leurs formats.

1.1.2. *Besoins*

L'enjeu de l'*intégration de données* consiste à établir puis à rendre systématique l'utilisation des liens entre éléments de domaines différents à divers degrés de granularité (par exemple des données omiques aux pathologies, ou inversement pour une seule espèce ou entre plusieurs espèces) (Bellazzi *et al*. 2011). La méta-analyse d'annotations hétérogènes en s'appuyant sur des connaissances pré-existantes a ainsi permis des découvertes qui auraient été hors de portée d'analyses individuelles (Zhang *et al*. 2010 ; Rho *et al*. 2011). La biologie des systèmes, la médecine de précision et la bioinformatique translationnelle reposent toutes sur l'utilisation systématique de ces liens (Al Kawam *et al*. 2018).

L'exploitation systématique des données résultant de la phase d'intégration rend l'automatisation de leur traitement encore plus nécessaire. En raison de la complexité intrinsèque des sciences de la vie, de grandes quantités d'éléments et des relations représentant leur interdépendance doivent ainsi être prises en compte (Baumgartner *et al*. 2007 ; Goble and Stevens 2008).

En plus d'être massive, cette exploitation systématique doit aussi aborder la complexité (Cannata *et al*. 2005 ; Bechhofer *et al*. 2013). L'analyse systématique de données intégrées nécessite une part d'interprétation, qui repose sur les connaissances du domaine (Blake and Bult 2006). Ces connaissances du domaine, aussi appelées expertise, peuvent être vues comme l'ensemble des règles modélisant dans quelles conditions des données peuvent être utilisées ou combinées pour inférer de nouvelles données ou de nouveaux liens entre données (Levesque 2013).

L'automatisation de l'analyse systématique de données en s'appuyant sur les connaissances du domaine repose sur les besoins suivants.

1.1.2.1. *Besoin 1 : identifier les ressources de manière interopérable*

La complémentarité de la multitude de bases des sciences de la vie repose sur le fait que plusieurs bases font référence aux mêmes entités.

Sur ce sujet, les sciences de la vie ont une longue tradition de normalisation qui leur donne un avantage sur d'autres domaines (par exemple pour désigner les personnes ou les entités géographiques). Curieusement, cela repose cependant sur une approche non-coordonnée où différentes bases utilisent chacune leurs identifiants propres. Ces différentes bases coexistent et font alors couramment référence aux identifiants des bases concurrentes, puis des bases supplémentaires se spécialisent dans la gestion des correspondances (Côté *et al*. 2007). Les bases les plus utilisées finissent par s'imposer comme des standards *de facto* pour leur communauté (typiquement, la désignation des gènes se fait entre spécialistes de chaque espèce) en étant reprises par des organisations indépendantes fédérant la communauté et chargées de maintenir une liste d'identifiants non ambigus et de gérer leur obsolescence (par exemple, le *Gene Ontology Consortium* pour la représentation des processus biologiques, des localisations cellulaires ou des fonctions moléculaires).

On retrouve une auto-organisation similaire pour la représentation des références entre les communautés.

Plus récemment, le besoin de mettre en place des solutions techniques pour assurer l'interopérabilité entre les bases de données et permettre l'automatisation de l'analyse des données a conduit à des initiatives plus génériques comme les *Life science identifiers* (LSID) (Clark *et al*. 2004), identifiers.org (Juty *et al*. 2012 ; Wimalaratne *et al*. 2015), ou les travaux de McMurry (McMurry *et al*. 2017) et de Pierce (Pierce *et al*. 2019).

1.1.2.2. *Besoin 2 : décrire les ressources*

Une fois les entités identifiées, les bases de données servent à représenter leur description de façon structurée (par opposition à des descriptions en texte libre).

La description d'une entité recouvre trois axes :

– ses **caractéristiques** qui sont ses propriétés : par exemple pour un gène son nom usuel, ses synonymes, ses coordonnées de début et de fin ;

– ses **relations avec d'autres entités** elles-mêmes désignées par leurs identifiants : par exemple pour un gène le chromosome et le brin sur lequel on le trouve, le taxon auquel il se rapporte, les transcrits qu'il peut produire. On voit que la description d'une entité fait typiquement intervenir plusieurs relations différentes, pointant vers d'autres entités de la même base de données ou de bases de données extérieures (dans ce dernier cas, on parle d'*entity matching*).

– les **classes** (aussi appelées « catégories ») auxquelles elle appartient. Ces classes dépendent en général des valeurs de certaines des relations vues précédemment. Par exemple, un gène peut appartenir à la classe des gènes codant pour des protéines (donc en fonction de la nature de ses transcrits, par opposition à la classe des gènes qui codent pour des ARN de transfert par exemple) et/ou à la classe des gènes homéotiques (qui régulent le développement des structures anatomiques, donc ici selon la fonction

des transcrits du gène). À la différence du point précédent, on ne fait intervenir ici que la relation d'instanciation, qui pointe non plus vers d'autres entités mais vers des classes d'entités issues des bases de connaissances.

La description d'une entité forme ainsi un graphe. Les entités pouvant être en relation les unes avec les autres, ou appartenir aux mêmes classes, les descriptions d'un ensemble de données forment souvent un *graphe connexe*.

La description d'une entité à travers ses relations avec d'autres entités et des classes fait donc référence à des éléments de bases de données et de bases de connaissances extérieures (voir besoin 1).

1.1.2.3. *Besoin 3 : combiner des descriptions fractionnées*

À cause de la complexité des sciences de la vie, la description d'une entité fait typiquement intervenir de nombreux paramètres. Il est donc peu courant que la description exhaustive d'une entité soit disponible dans une seule base. Le besoin 2 indiquait que la description d'une entité peut faire référence à des éléments d'autres bases. Le besoin 3 indique que cette description est potentiellement elle-même fragmentée dans différentes bases.

La nécessité de combiner les descriptions partielles d'une entité venant de plusieurs origines découle donc des besoins 1 et 2.

1.1.2.4. *Besoin 4 : interroger ces descriptions*

Les trois premiers besoins portaient respectivement sur l'identification des ressources et la représentation de leurs descriptions. Cependant, accéder à ces descriptions ne suffit pas, et il est nécessaire de pouvoir les interroger de façon automatique afin de réaliser des analyses exhaustives.

Il est ainsi nécessaire de disposer d'un langage de requêtes s'appliquant aussi bien aux descriptions localisées sur une base de données (besoin 2) qu'à celles qui résultent de l'intégration de plusieurs bases (besoin 3).

1.1.2.5. *Besoin 5 : raisonner sur ces descriptions grâce à des connaissances*

Enfin, les *connaissances* du domaine interviennent aussi bien dans le mécanisme d'interrogation lui-même (besoin 4) *via* les classes des entités et les relations entre ces classes (besoin 2) que dans l'analyse du résultat des requêtes (Bechhofer *et al*. 2013). Sur ces deux points, Robert Stevens note respectivement : « *The complex biological data stored in bioinformatics databases often require the addition of knowledge to specify and constrain the values held in that database* » et « *Much of biology works by applying prior knowledge [...] to an unknown entity* » (Stevens *et al*. 2000). Là encore, il faut être capable d'en tenir compte dans les traitements automatiques permettant de réaliser des analyses exhaustives.

Ces connaissances symboliques du domaine sont formalisées dans des bases de connaissances sous forme d'*ontologies*, qui sont des représentations formelles des connaissances dans lesquelles les classes essentielles sont combinées par des règles qui décrivent leur structure et les relations entre elles (Bard and Rhee 2004). Les sciences de la vie ont une longue histoire de modélisation des connaissances sous forme d'ontologies (Cimino 1998, 2005 ; Smith 2005) et de leur utilisation pour annoter des données (Stevens *et al*. 2000 ; Bard and Rhee 2004 ; Blake and Bult 2006 ; Bodenreider and Stevens 2006). Ces ontologies sont donc déjà disponibles (Stevens *et al*. 2000 ; Cimino and Zhu 2006), et leur (ré-)utilisation est facilitée par des portails comme BioPortal (Noy *et al*. 2009 ; Whetzel *et al*. 2011).

À partir des connaissances en complément des descriptions des données, on peut définir le *raisonnement* comme un ensemble de méthodes guidant le parcours automatique du graphe de données ou son enrichissement. Ce mécanisme de raisonnement permet de parcourir des relations entre deux entités, entre une entité et une classe, et entre deux classes, le tout réparti entre plusieurs bases. Les règles de parcours automatique sont conditionnées par le degré de formalisation des ontologies, qui va de simples hiérarchies à des organisations sémantiquement riches (Cimino and Zhu 2006).

Comme pour le besoin 4, il est nécessaire de disposer d'un mécanisme de raisonnement s'appliquant aussi bien aux descriptions et aux ontologies localisées sur une base de données (besoin 2) qu'à celles qui résultent de l'intégration de plusieurs bases (besoin 3).

1.1.3. *Approches courantes : InterMine et BioMart*

Aujourd'hui pour les sciences de la vie, les principales bases de données et de connaissances sont là, mais réparties dans des silos peu interopérables.

Ce constat avait conduit au cours des dernières décennies à des solutions spécifiques aux sciences de la vie pour permettre un accès unifié aux différentes bases. InterMine repose sur une approche centralisée où les données de différentes bases sont intégrées dans un seul entrepôt au format prédéfini (Kalderimis *et al*. 2014). Il existe une trentaine de tels entrepôts, organisés par espèces[3]. BioMart repose sur une approche décentralisée à travers une fédération d'une quarantaine de bases de données[4] (Smedley *et al*. 2015).

Il faut noter le succès limité rencontré par ces initiatives, dû en partie aux difficultés de mise en œuvre face à des bases hétérogènes et qui évoluent rapidement et indépendamment les unes des autres. Ces solutions sont spécifiques aux sciences de la

3. `http://registry.intermine.org/`

4. `http://www.biomart.org/community.html`

vie. Elles regroupent quelques dizaines de bases et il est manifeste qu'aucune d'elles ne permet le passage à l'échelle au regard des 1600 bases de référence. De plus, si ces approches visent à répondre aux besoin d'intégration (besoins 1 à 4), elles offrent peu de capacités de raisonnement basé sur des connaissances (besoin 5).

Face à ces limitations, rendre les données biologiques réutilisables et interopérables constitue un enjeu majeur (Goodman *et al*. 2014). Partant du constat que de nombreux autres domaines ont eux aussi connu une explosion des données, il faut se demander si le problème de l'intégration de données biologiques ne dépasse pas les sciences de la vie et devrait être abordé d'une manière globale et générique.

1.2. Le Web Sémantique

Les initiatives d'intégration et d'analyse de données en sciences de la vie que nous avons vues à la section précédente se sont développées en parallèle à l'émergence du *Web Sémantique* dans la communauté informatique au début des années 2000 (Berners-Lee *et al*. 2001 ; Ruttenberg *et al*. 2007 ; Schulz *et al*. 2013).

Le Web Sémantique est une extension du Web classique qui met l'accent sur la représentation des données dans un format pour en rendre le sens explicite et permettre leur traitement automatique en les associant à des ontologies (Shadbolt *et al*. 2006 ; Berners-Lee *et al*. 2007). Ce format doit permettre une représentation fine des données, leur intégration et leur interprétation. Les principes sont :

– **affiner la granularité de l'information** : alors que le Web classique est organisé autour des documents, le Web Sémantique est centré sur les éléments de données contenus dans ces documents en les identifiant chacun par un *URI*[5] spécifique (maintenant un IRI[6]) ;

– **représenter explicitement les relations** en les identifiant également par des URI : alors qu'une des clés du succès du Web classique est la représentation de liens non-typés entre documents (*via* les `href` en HTML), le Web Sémantique repose sur des relations typées (et elles-mêmes identifiées par leur URI) entre données ;

– **permettre la généralisation et la prise en compte des connaissances** du domaine grâce à des relations spéciales : l'*instanciation* entre une donnée atomique (qui est anecdotique) et une généralité (qui est universelle) et la *subsomption* entre deux généralités. Par exemple, la relation d'instanciation permet d'indiquer que la pathologie d'un patient avec toutes ses spécificités est un élément de l'ensemble « maladie d'Alzheimer » (qui est identifié par `DOID:10652` dans la *Disease Ontology*). La relation de subsomption permet d'indiquer que les cas de maladie

5. URI : *Uniform Resource Identifier*

6. IRI : *Internationalized Resource Identifier*, extension des URI aux caractères non ASCII permettant de prendre en compte les signes diacritiques ou les alphabets non-occidentaux.

d'Alzheimer (`DOID:10652`) forment un sous-ensemble des maladies neurodégénératives (`DOID:1289`).

1.2.1. *Techniques*

Dans le cadre du Web Sémantique, le W3C a proposé plusieurs recommandations (qui sont des standards *de facto*) relatives à la représentation (RDF, RDFS et OWL), l'intégration et l'analyse (SPARQL et OWL) de données et des connaissances.

1.2.1.1. *URI (et IRI) pour identifier les données*

Pour répondre au premier besoin d'identifier les données, le Web Sémantique repose sur les URI, qui sont une extension des URL[7], dont ils conservent la syntaxe. Par exemple, `http://purl.uniprot.org/uniprot/P35558` est l'URI de la protéine humaine PCKGC dans la base UniProt.

Les URL sont donc des cas particuliers d'URI et on conçoit bien que donner l'adresse d'un document est une bonne façon de l'identifier (c'est LE document qui se trouve à cette adresse). Néanmoins, ce mécanisme est mal adapté pour désigner les données à l'intérieur d'un document (sauf à utiliser la partie locale d'un URL, après le `#`), représenter les relations entre ces données (ici la partie locale de l'URL ne suffit plus), ou gérer des descriptions de données fragmentées dans plusieurs documents. Inversement, tous les URI ne sont pas nécessairement des URL et ne désignent pas forcément un document accessible sur le Web. Si ce n'est pas obligatoire, il peut néanmoins être pratique que l'URL correspondant à un URI désigne l'adresse de la description de l'entité en question. Ce mécanisme s'appelle la *déréférenciation d'URI*. Cette description peut même prendre différentes formes en utilisant le mécanisme de négociation de contenu du protocole HTTP : en entrant `http://purl.uniprot.org/uniprot/P35558` dans un navigateur, on obtient une page HTML avec une description de PCKGC destinée aux utilisateurs humains. La même adresse avec le *mime type* `application/rdf+xml` renvoie une description de PCKGC au format RDF (la même que l'on peut obtenir à l'URL `http://purl.uniprot.org/uniprot/P35558.rdf`)

Les utilisateurs peuvent créer des URI comme bon leur semble, du moment qu'ils sont syntaxiquement valides, sans avoir besoin de posséder le nom de domaine ni de déployer un site. Cela confère une grande facilité d'utilisation aux URI. Ce n'est cependant pas parce qu'un utilisateur a la possibilité de créer l'URI `http://purl.uniprot.org/uniprot/B12345` que c'est automatiquement judicieux : 1) cela ne suffit pas à ajouter une entité à la base UniProt, 2) les administrateurs d'UniProt (ou n'importe qui d'autre, d'ailleurs) pourraient très bien

7. URL: *Uniform Resource Locator*, permet de représenter l'adresse d'un document sur le Web.

décider de créer le même URI pour désigner une autre entité, et on aurait alors une ambiguïté, et 3) il lui aurait été tout aussi simple de créer un URI comme `http://www.univ-rennes1.fr/odameron/B12345` qui aurait évité les deux problèmes précédents. Si la capacité à créer facilement des URI permet d'éviter les ambiguïtés, il convient de favoriser l'interopérabilité en réutilisant des URI existants pour éviter d'avoir plusieurs URI qui désignent la même entité. Dans « *Cool URIs don't change*[8] » en 1998, Tim Berners-Lee avait commencé à aborder les bonnes pratiques liées aux URI (heureusement, l'URL n'a pas changé depuis), et ce document avait ensuite été étendu dans une note du W3C « *Cool URIs for the Semantic Web*[9] » dix ans plus tard.

Dans la suite de ce chapitre, et conformément à l'usage, une *ressource* désigne n'importe quelle entité qui peut être identifiée par un URI :

– des entités réelles (`http://www.wikidata.org/wiki/Q42` désigne Douglas Adams),

– des groupes d'entités (`http://purl.uniprot.org/uniprot/P35558` désigne l'ensemble des protéines humaines PCKGC), y compris imaginaires (`http://www.wikidata.org/wiki/Q7246` désigne l'ensemble des licornes),

– des constructions intellectuelles (`http://purl.obolibrary.org/obo/DOID_1289` pour les maladies neurodégénératives).

Enfin, les URI sont bien adaptés au traitement automatique mais sont difficiles à manipuler pour les humains. Les *CURIE* (*Compact URI*) sont des versions abrégées composées d'un préfixe et d'un identifiant local. Par exemple, `uniprot:P35558` est la version abrégée de `http://purl.uniprot.org/uniprot/P35558` après avoir associé la valeur `http://purl.uniprot.org/uniprot/` au préfixe `uniprot`. C'est ce mécanisme qui intervenait lorsqu'on a utilisé `DOID:10652` pour identifier la maladie d'Alzheimer. L'utilisateur est libre de choisir le nom des préfixes, puisque ceux-ci sont ensuite systématiquement remplacés par leur valeur pour transformer les CURIE en URI complets. Pour plus de lisibilité, on utilise cependant souvent les mêmes préfixes, et le site `http://prefix.cc` permet de retrouver le fragment d'URI qui leurs est associé.

1.2.1.2. *RDF pour décrire les données*

RDF[10] (Resource Description Framework) est un formalisme permettant de représenter les éléments décrivant une entité sous forme d'un ensemble de *triplets* :

– le *sujet* est le premier élément du triplet. C'est l'URI de la ressource que l'on décrit ;

---

8. `https://www.w3.org/Provider/Style/URI.html`
9. `https://www.w3.org/TR/cooluris/`
10. `http://www.w3.org/RDF/`

– le *prédicat* est le second élément du triplet. C'est l'URI de la relation que l'on utilise pour décrire le sujet ;

– l'*objet* est le troisième élément du triplet. C'est une des valeurs possibles du prédicat pour le sujet. Cette valeur peut être un URI si le prédicat décrit une relation entre deux ressources, ou une chaîne de caractères représentant une simple chaîne de caractères (par exemple pour le nom d'une protéine), un nombre (par exemple pour la position de début d'un gène), une date, une valeur booléenne, etc. Si un prédicat a plusieurs valeurs (par exemple un facteur de transcription qui régule plusieurs gènes), il faut utiliser autant de triplets qu'il y a de valeurs. Cela permet ainsi à plusieurs bases de se compléter facilement, pourvu qu'elles utilisent bien les mêmes URI pour désigner les mêmes ressources.

RDF fournit un prédicat spécial (`rdf:type`) pour représenter la relation d'instanciation entre une ressource et une classe. On trouve aussi souvent le prédicat `rdfs:label` pour décrire une ressource par une chaîne lisible.

On peut visualiser un triplet comme un arc (le prédicat) allant du sujet à l'objet. Les triplets d'un *dataset* forment ainsi un graphe orienté, constitué de composantes connexes lorsque deux triplets partagent le même sujet, le même objet ou encore lorsque l'objet de l'un est le sujet de l'autre. La figure 1.1 montre une vue simplifiée de la description de la réaction chimique `BR5566` de Reactome qui transforme l'oxaloacetate en phosphoenolpyruvate. On y voit notamment que c'est une instance de la classe `BiochemicalReaction` qui vient de l'ontologie BioPAX, qu'elle est associée à une chaîne par la relation `displayName`, qu'elle fait partie de la voie métabolique de néoglucogénèse et que certaines relations peuvent avoir plusieurs valeurs. Ainsi, elle consomme (*via* la relation `left`) deux molécules dont l'une est l'oxaloacetate (`SM1312`) et en produit (*via* la relation `right`) trois, dont du phosphoenolpyruvate (`SM1316`). Enfin, il faut remarquer que les éléments intervenant dans la description de cette réaction chimique font référence à des ressources d'autres *datasets* figurées ici par des couleurs différentes, ce qui illustre la capacité d'intégration de RDF. Par exemple `taxon:9606` désigne *Homo sapiens* dans la *NCBI taxonomy of species*, l'enzyme PCKGC qui catalyse la réaction est la protéine `P35558` dans UniProt, et les molécules consommées et produites par la réaction sont associées à des ressources dans l'ontologie ChEBI.

Il existe plusieurs formats de fichiers pour représenter des données RDF : N-triples (un triplet par ligne, les URI sont donnés en entier), Turtle (extension de N-triples permettant d'utiliser des CURIE et des commentaires), XML-RDF pour une sérialisation en XML, et JSON-LD pour une sérialisation en JSON, ou encore RDFa pour inclure du RDF dans des pages HTML. La figure 1.2 montre la représentation en Turtle d'une partie des triplets de la figure 1.1. Enfin, puisque traiter des données massives est une des raisons d'être du Web Sémantique, il est également possible de dépasser les limitations des fichiers en stockant des données RDF dans des *triplestores*, qui sont

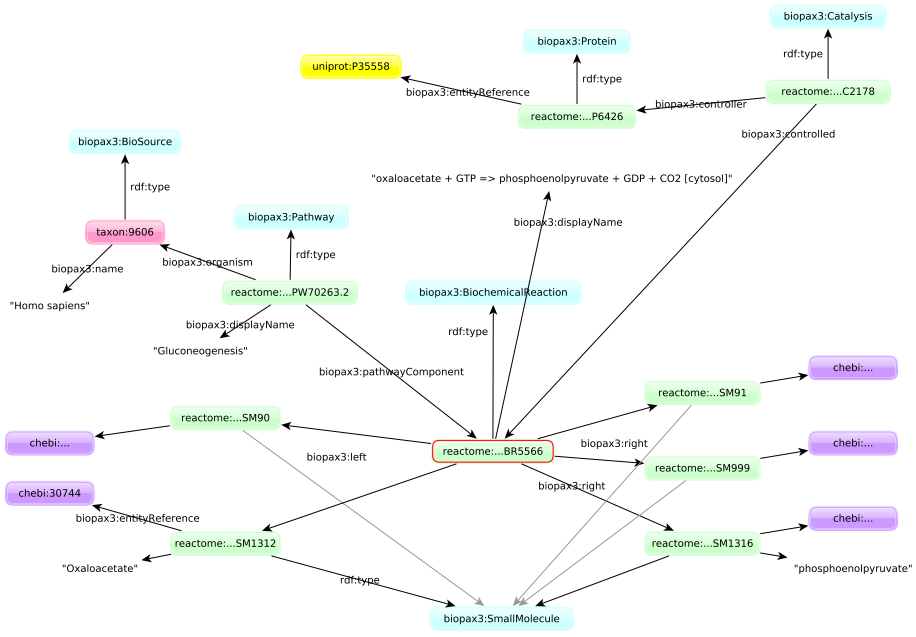

Figure 1.1. *Vue simplifiée des triplets RDF décrivant la ressource `BR5566` (entourée en rouge) de la base Reactome. Les ressources sont représentées par des rectangles, et les littéraux par des chaînes de caractères. Les prédicats sont les arcs, reliant le sujet à l'objet de chaque triplet. Certaines relations `rdf:type` sont grisées pour faciliter la lecture. Les ressources de la base Reactome font référence à des ressources d'autres bases de données comme la hiérarchie des espèces du NCBI (en rose), des protéines dans UniProt (en jaune), des molécules de ChEBI (en violet) ou encore pour l'instanciation des classes de l'ontologie BioPAX (en bleu).*

l'équivalent pour RDF des bases de données. Outre leur fonction de stockage des données RDF, les triplestores permettent de les interroger en utilisant le protocole et le langage SPARQL *via* des *endpoints* (voir Section 1.2.1.4).

1.2.1.3. *RDFS et OWL pour décrire les connaissances*

Alors que RDF est adapté pour décrire des entités, leur(s) type(s) et des relations entre entités, RDFS[11] (RDF schema) et OWL[12] (Ontology Web Language) permettent

11. `http://www.w3.org/TR/rdf-schema/`
12. `http://www.w3.org/2001/sw/wiki/OWL`

```
# prefix declarations for CURIEs
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix reactome: <http://www.reactome.org/biopax/61/48887#> .
@prefix biopax3: <http://www.biopax.org/release/biopax-level3.owl#> .
#...

# RDF triples
reactome:BR5566 rdf:type biopax3:BiochemicalReaction .
reactome:BR5566 biopax3:left reactome:SM90 .
reactome:BR5566 biopax3:left reactome:SM1312 .
reactome:BR5566 biopax3:right reactome:SM91 .
# ...

reactome:SM1312 rdf:type biopax3:SmallMolecule .
reactome:SM1312 biopax3:name "Oxaloacetate" .
# ...
```

Figure 1.2. *Représentation en syntaxe RDF Turtle d'une partie des triplets de la figure 1.1. Chaque triplet est représenté sur une ligne dans l'ordre sujet, prédicat, objet, et correspond à un des arcs de la figure 1.1. Il n'y a pas d'ordre particulier entre les triplets.*

de décrire les classes ou les prédicats qui constituent les ontologies. RDFS et OWL sont basés sur une sémantique ensembliste.

RDFS fournit deux ressources particulières : `rdfs:Class` et `rdfs:Property` et deux prédicats : `rdfs:subClassOf` et `rdfs:subPropertyOf`. `rdfs:Class` permet d'indiquer qu'une ressource (par exemple `taxon:9606`) est un ensemble d'instances (c'est l'ensemble de tous les Homo sapiens). `rdfs:subClassOf` permet d'indiquer l'inclusion de la sous-classe dans sa super-classe (par exemple ici que tous les Homo sapiens sont aussi des éléments du genre Homo). Elle correspond à la relation de subsomption que nous avons déjà vue. Le principe est le même pour les prédicats avec `rdfs:Property` et `rdfs:subPropertyOf`. Les ontologies en RDFS sont donc des *taxonomies*, c'est à dire des ensembles de classes reliées par la relation de subsomption. La figure 1.1 présentait la description d'une réaction chimique en faisant référence notamment à des entités des ontologies *NCBI Taxonomy of species* (en rose) et ChEBI (en violet). La figure 1.3 reprend cette description et montre comment ces deux ontologies rendent explicites les relations de subsomption entre ces entités et leurs super-classes. La figure montre également que RDF et RDFS se combinent naturellement pour permettre l'intégration des données et des connaissances.

OWL fournit deux ressources : `owl:Class` et `owl:Property`, qui sont des sous-classes (*via* `rdfs:subClassOf`) de `rdfs:Class` et de `rdfs:Property`. OWL

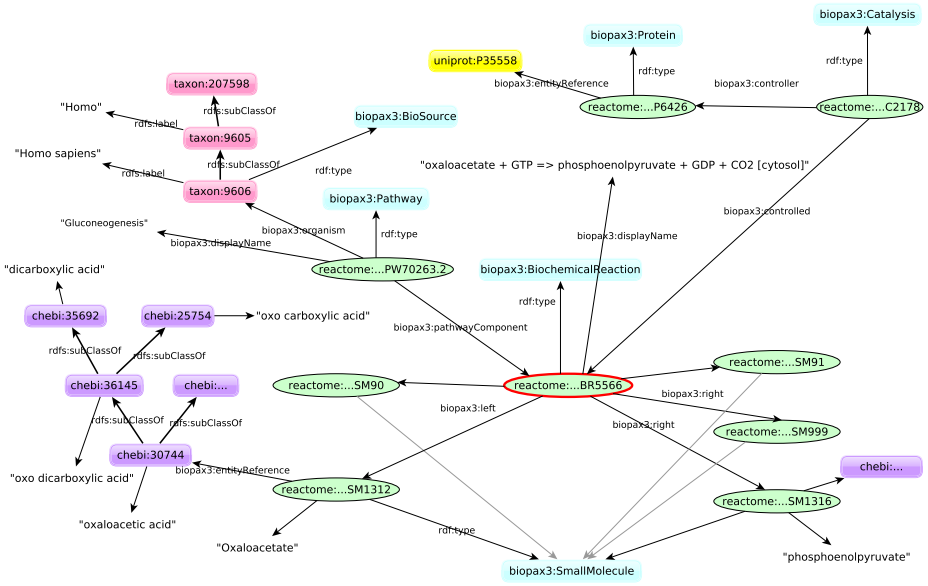

Figure 1.3. *Connaissances associées à la description de la réaction chimique* `BR5566` *entourée en rouge. Les instances sont représentées par des ellipses, et les classes par des rectangles. Les relations* `rdf:type` *entre les rectangles et* `owl:Class` *ne sont pas représentées pour ne pas surcharger la figure. Les relations de subsomption* `rdfs:subClassOf` *sont mises en valeur en gras. Les connaissances issues de ChEBI sont en violet, et celles de la NCBI Taxonomy of species en rose. Remarquez que la transitivité de* `rdfs:subClassOf` *permet de généraliser dans les deux ontologies, et que l'héritage multiple est permis (ici dans ChEBI).*

permet ensuite d'indiquer si deux classes sont disjointes ou équivalentes, ou encore de définir de nouvelles classes par intension (c'est-à-dire en énumérant ses membres de façon exhaustive) : la classe des éléments liés à au moins une instance d'une classe par un prédicat (par exemple, la classe des pizzas au fromage est l'ensemble des pizzas dont au moins un ingrédient est une instance de fromage), ou la classe des éléments dont toutes les valeurs d'un prédicat sont des instances d'une classe (par exemple, la classe des pizzas végétariennes). Les ontologies en OWL sont donc bien plus riches que les taxonomies en RDFS. La figure 1.4 montre une partie de la représentation en OWL de la classe `BiochemicalReaction` de BioPAX. On y voit d'une part que pour les instances de cette classe, toutes les valeurs des relations `participant` (notamment les relations `left` et `right` vues aux figures 1.1 et 1.3) doivent être des instances directes ou indirectes de la classe `PhysicalEntity`, et

d'autre part que `BiochemicalReaction` est disjointe des classes `Degradation` et `ComplexAssembly` (mais pas de `Transport` par exemple, ce qui explique que certaines réactions de conversion puissent être des réactions biochimiques de transport). Un raisonneur OWL déduira de cette ontologie que si une réaction est une instance de `BiochemicalReaction`, elle ne peut pas être une instance de `Degradation` ou de `ComplexAssembly`, et toutes les valeurs des relations `participant` doivent être des instances de `PhysicalEntity`. Inversement, si une instance de `BiochemicalReaction` viole une de ces contraintes, le raisonneur détectera une incohérence.

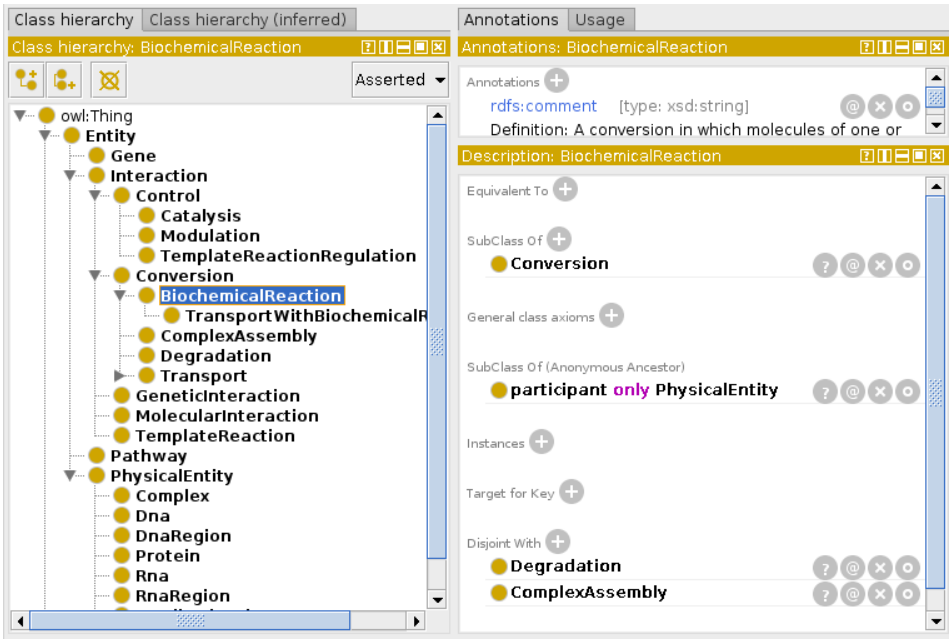

Figure 1.4. *Représentation en OWL de la classe `BiochemicalReaction` de BioPAX, visualisée avec l'éditeur d'ontologies Protégé. La partie gauche montre la hiérarchie des classes. La partie droite montre les contraintes associées à la classe sélectionnée. Ici on voit que `BiochemicalReaction` est une sous-classe de `Conversion`, qu'elle est disjointe de `Degradation` et de `ComplexAssembly`, et que toutes la valeurs de la relation `participant` (ou de ses sous-relations) doivent être des instances de la classe `PhysicalEntity`.*

RDFS permet de représenter des taxonomies à partir d'inclusions des sous-classes dans leurs super-classes. OWL permet des descriptions plus fines et donc des raisonnements plus riches à partir d'union, d'intersection et de compléments ensemblistes,

de quantifieurs existentiels ou universels, et de définitions nécessaires et suffisantes. RDFS est bien adapté à des raisonnements simples sur de grosses bases de connaissances, tandis qu'OWL est adapté aux raisonnements plus complexes, moyennant des temps de calcul possiblement plus longs. Le formalisme d'OWL est constitué d'entités et de prédicats spéciaux représentés en RDF (tout comme on utilise `rdf:type` en RDF pour désigner l'instanciation, ou `rdfs:subClassOf` en RDFS pour désigner la subsomption). Ainsi, toute assertion RDFS ou OWL est aussi une assertion RDF valide. OWL est lui-même une extension basée sur RDFS, ce qui fait que toute assertion en OWL est également valide en RDFS. On peut donc utiliser des outils RDFS sur des ontologies en OWL, même s'ils ne seront pas capables d'exploiter les éléments spécifiques à OWL.

1.2.1.4. *SPARQL pour interroger*

L'interrogation des données en RDF (et donc également des connaissances en RDFS et en OWL puisque ces formats reposent sur RDF) est possible grâce au langage SPARQL[13] (Pérez *et al*. 2009). Il faut souligner que SPARQL1.1 permet de prendre en compte la plus grande partie de la sémantique de RDFS.

La requête SPARQL de la figure 1.5 permet de retrouver dans Reactome toutes les réactions chimiques consommant de l'oxaloacétate chez l'humain (voir Figure 1.1). Il est de plus possible d'effectuer des raisonnements basés sur la relation de subsomption en exploitant les connaissances contenues dans les ontologies. La requête 1.6 généralise ainsi celle de la figure 1.5 en permettant de retrouver toutes les réactions chimiques consommant un acide oxo-carboxylique ou une de ses sous-classes, directe ou indirecte (voir Figure 1.3).

Il est possible d'interroger en SPARQL aussi bien des fichiers RDF (grâce au projet Jena[14]), que des données contenues dans des triplestores *via* leur *endpoints* (triplestore fait référence aux aspects de stockage et d'indexation, tandis qu'*endpoint* fait référence à l'interrogation des données stockées dans un triplestore). Ces derniers utilisent des index pour garantir de bonnes performances même sur de gros volumes de données. Les principaux logiciels permettant de déployer un *endpoint* sont Virtuoso[15], Fuseki[16], RDF4J[17] et Corese (Corby *et al*. 2004).

1.2.1.5. *Raisonneurs OWL pour utiliser les connaissances*

OWL ne dispose pas d'un langage de requêtes, mais n'en a pas vraiment besoin non plus car le raisonnement qu'il permet consiste principalement à déterminer si une

13. `http://www.w3.org/TR/sparql11-overview/`
14. `https://jena.apache.org/`
15. `https://virtuoso.openlinksw.com/`
16. `https://jena.apache.org/documentation/fuseki2/index.html`
17. `https://rdf4j.eclipse.org/`

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX reactome: <http://www.reactome.org/biopax/61/48887#>
PREFIX biopax3: <http://www.biopax.org/release/biopax-level3.owl#>
PREFIX taxon: <http://identifiers.org/taxonomy/>

SELECT ?reaction ?reactionName
FROM <http://rdf.ebi.ac.uk/dataset/reactome>
WHERE {

  ?reaction rdf:type biopax3:BiochemicalReaction .
  OPTIONAL { ?reaction biopax3:displayName ?reactionName . }
  ?reaction biopax3:left reactome:SmallMolecule1312 . # oxaloacetate

  ?pathway biopax3:pathwayComponent ?reaction .
  ?pathway biopax3:organism taxon:9606 .
}
```

Figure 1.5. *Requête SPARQL permettant de retrouver dans Reactome toutes les réactions chimiques consommant de l'oxaloacétate chez l'humain (voir Figure 1.1). En SPARQL, les noms des variables commencent par un point d'interrogation. Le moteur SPARQL détermine les combinaisons de valeurs des variables qui satisfont les contraintes de la requête à partir du dataset. Il est possible d'utiliser l'endpoint de l'EBI à* `https://www.ebi.ac.uk/rdf/services/sparql`

entité est une instance d'une classe, ou si une classe est une sous-classe d'une autre classe. Les principaux raisonneurs sont HermiT[18], Pellet[19] et Racer[20].

Il faut noter que même si la plupart des bio-ontologies sont représentées en OWL, peu d'entre elles en utilisent l'expressivité. La plupart sont en fait des taxonomies RDFS déguisées en OWL (ce qui est possible puisque toutes les classes OWL sont aussi des classes RDFS), même si des travaux ont montré qu'elles bénéficieraient des primitives supplémentaires de OWL (Stevens *et al*. 2007 ; Aranguren *et al*. 2007 ; Hill *et al*. 2013).

De plus, SWRL[21] (Semantic Web Rule Language) permet de représenter des règles d'inférence contenant des variables (OWL ne dispose pas de variables).

18. `http://www.hermit-reasoner.com/`
19. `https://github.com/stardog-union/pellet`
20. `https://github.com/ha-mo-we/Racer`
21. `http://www.w3.org/Submission/SWRL/`

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX reactome: <http://www.reactome.org/biopax/61/48887#>
PREFIX biopax3: <http://www.biopax.org/release/biopax-level3.owl#>
PREFIX taxon: <http://identifiers.org/taxonomy/>
PREFIX chebi: <http://purl.obolibrary.org/obo/CHEBI_>

SELECT ?reaction ?reactionName
FROM <http://rdf.ebi.ac.uk/dataset/reactome>
WHERE {

  ?reaction rdf:type biopax3:BiochemicalReaction .
  OPTIONAL { ?reaction biopax3:displayName ?reactionName . }
  ?reaction biopax3:left ?reactant .

  ?reactant biopax3:entityReference ?reactantChEBI .
  ?reactantChEBI rdfs:subClassOf* chebi:25754 . # oxo carboxylic acid

  ?pathway biopax3:pathwayComponent ?reaction .
  ?pathway biopax3:organism taxon:9606 .
}
```

Figure 1.6. *Requête SPARQL permettant de retrouver dans Reactome toutes les réactions chimiques consommant un acide oxo-carboxylique ou une de ses sous-classes, directe ou indirecte chez l'humain. Le raisonnement consiste à suivre la relation `rdfs:subClassOf` zéro, une ou plusieurs fois à partir de `?reactantChEBI`. Il est représenté en appliquant l'étoile de Kleene à la relation `rdfs:subClassOf` ligne 17. La hiérarchie de classes autour de l'acide oxo-carboxylique est basé sur l'ontologie ChEBI (voir Figure 1.3).*

1.2.2. *Mise en œuvre*

1.2.2.1. *bio2rdf, BioPortal et ressources disponibles*

Les sciences de la vie constituent un domaine d'application privilégié pour les technologies du Web Sémantique (Cannata *et al*. 2008 ; Post *et al*. 2007 ; Bellazzi 2014), et plusieurs équipes travaillent sur l'analyse de données de sciences de la vie grâce au Web Sémantique, notamment au sein du groupe d'intérêt du W3C *Semantic Web Health Care and Life Sciences* (HCLSIG[22]).

22. `http://www.w3.org/wiki/HCLSIG`

Les principales bases de données et de connaissances comme UniProt[23] ou les bases de l'EBI[24] et du NCBI (Anguita *et al*. 2013) sont en train d'adopter les technologies du Web Sémantique et sont accessibles *via* des *endpoints*. De même, BioPortal (Noy *et al*. 2009 ; Whetzel *et al*. 2011) qui recense les principales ontologies en sciences de la vie les rend accessibles *via* SPARQL (Salvadores *et al*. 2012). Enfin, le projet Bio2RDF [25] promeut des conventions simples pour convertir et intégrer les bases qui n'ont pas encore entamé leur adoption de RDF (Belleau *et al*. 2008 ; Cheung *et al*. 2009 ; Callahan *et al*. 2013).

Les technologies du Web Sémantique font maintenant partie intégrante de la médecine personnalisée et de la bioinformatique translationnelle (Bellazzi *et al*. 2011 ; Chen *et al*. 2012). Plusieurs travaux ont montré que ces technologies peuvent être utilisées pour intégrer et interroger des informations sur le génotype et le phénotype (Sahoo *et al*. 2007 ; Taboada *et al*. 2012). De plus, Holford *et al.* ont proposé une architecture basée sur le Web Sémantique pour intégrer des données omiques sur le cancer et des connaissances biologiques (Holford *et al*. 2012).

1.2.2.2. *LOD*

L'initiative des *données liées* (Bizer *et al*. 2009) et particulièrement le projet *Linked Open Data* portent sur l'intégration de sources de données dans des formats du Web Sémantique. Le célèbre nuage des données liées[26] montre bien le rôle majeur des sciences de la vie à travers à la fois le nombre de bases disponibles et la densités de leurs références croisées. Ces dernières années, les technologies du Web Sémantique ont fait la preuve de leur pertinence pour répondre au besoin d'intégration de données (Hill *et al*. 2013 ; Livingston *et al*. 2015).

De plus, la mise en correspondance entre identifiants de bases différentes est également facilitée par des initiatives comme identifiers.org[27] (Wimalaratne *et al*. 2015) qui fournit un registre associant un couple (base de données, identifiant) dans la base à un URI. Enfin, une fois ces bases intégrées, il devient possible de les interroger (relativement) facilement grâce à des *requêtes fédérées*, qui traitent ces différentes bases comme s'il s'agissait d'un graphe unique virtuel (Cheung *et al*. 2009). Ce point est détaillé à la section 1.3.3.

---

23. `https://sparql.uniprot.org/`
24. `https://www.ebi.ac.uk/rdf/`
25. `https://github.com/bio2rdf`
26. `https://lod-cloud.net/`
27. `http://identifiers.org/`

1.3. Perspectives

Les technologies du Web Sémantique constituent une solution pertinente aux besoins identifiés à la section 1.1. Elles sont actuellement utilisées avec succès pour aborder la complexité et l'hétérogénéité des données en sciences de la vie en permettant d'intégrer les bases de données et de connaissances de référence.

Néanmoins, les nouvelles perspectives ainsi ouvertes s'accompagnent de nouveaux enjeux.

1.3.1. *Faciliter l'appropriation par les utilisateurs*

Si les technologies du Web Sémantique sont couramment utilisées par les data scientists, leur adoption par les biologistes ou médecins demeure limitée. Les facteurs principaux sont techniques et psychologiques. L'aspect technique est lié à la difficulté d'apprendre des langages comme RDF et SPARQL, et à celle de les mettre en œuvre en écrivant des scripts pour convertir les données brutes, ou en déployant des *endpoints*. L'aspect psychologique est lié à la facilité d'utilisation des tableurs, et aux techniques de traitement et d'analyse qui sont limitées mais intuitives. Ces deux facteurs expliquent pourquoi les utilisateurs finaux se sentent dépossédés de leur données une fois qu'elles sont converties en RDF : ils perdent l'accès direct à *leurs* données et se retrouvent obligés de passer par un data scientist. Il est regrettable que nous soyons actuellement coincés entre continuer à utiliser des tableurs qui ne répondent pas aux besoins d'intégration, ou s'appuyer sur les technologies du Web Sémantique qui rendent plus compliquée et rigide la phase exploratoire de l'analyse des données par des gens dont l'expertise repose sur ce que représentent ces données.

Faciliter l'appropriation des technologies du Web Sémantique par les utilisateurs au-delà des data scientists passe par deux leviers :

– faciliter l'intégration des données de leurs projets au nuage des bases de données et de connaissances de référence. Cette intégration doit être bi-directionnelle : d'une part pour permettre d'exploiter ces bases de référence lors de l'analyse des données d'un projet, et inversement, incorporer les données du projet au nuage des données liées pour en favoriser la réutilisation (même si ce n'est pas nécessairement sous forme de données ouvertes). Des initiatives comme datalift (Scharffe *et al*. 2012) vont dans ce sens. Datalift est une plateforme qui prend en entrée des données dans différents formats comme CSV ou XML, et les met à disposition en les intégrant aux données liées.

– faciliter la création de requêtes portant (éventuellement) sur les données de l'utilisateur et les données des bases de référence. Plusieurs travaux ont déjà été menés dans ce sens pour synthétiser la structure d'un *dataset* (van Dam *et al*. 2015) ou autour

de la création de requêtes comme SPARQLassist (McCarthy *et al*. 2012), SPARKLIS (Ferré 2014), AskOmics[28] ou SPARQLbuilder (Yamaguchi *et al*. 2014) mais aucun ne s'est encore imposé comme une solution idéale.

1.3.2. *Faciliter l'appropriation par les programmes : données FAIR*

Un des deux enjeux de l'appropriation des technologies du Web Sémantique par les utilisateurs (Section 1.3.1) est de leur permettre de combiner leurs propres jeux de données aux bases de référence. On perçoit bien que cela s'accompagne d'un problème de passage à l'échelle. L'initiative FAIR vise à aborder cet écueil en rendant les données *Findable*, *Accessible*, *Interoperable* et *Reusable* (Wilkinson *et al*. 2016). Cette démarche englobe également les workflows d'analyse de données (voir section **??**).

L'approche FAIR a montré sa pertinence dans le domaine des sciences de la vie (Rodríguez-Iglesias *et al*. 2016 ; Brandizi *et al*. 2018 ; Sima *et al*. 2019). Il faut souligner que cela s'accompagne d'efforts pour valoriser la production de jeux de données réutilisables (Pierce *et al*. 2019).

1.3.3. *Requêtes fédérées*

RDF permet de tirer parti des références croisées entre bases (*via* le mécanisme d'*entity matching* évoqué dans le besoin 2) sans avoir à fusionner ces bases dans un seul *dataset*. On dispose ainsi de plusieurs *datasets* exposés sous forme de *endpoints* et faisant référence les uns aux autres, selon le principe des données liées. Il est bien sûr possible d'interroger chaque *endpoint* séparément, mais une question peut également nécessiter de combiner des informations venant de plusieurs *endpoints* (voir besoin 4).

SPARQL permet de créer des *requêtes fédérées* englobant plusieurs *endpoints* et traitant leurs données comme s'il s'agissait d'un seul ensemble virtuel de triplets (Cheung *et al*. 2009 ; Kratochvíl *et al*. 2019). C'est le moteur SPARQL qui a alors la charge d'envoyer les fragments de requêtes aux *endpoints* pertinents, et surtout d'effectuer l'intégration des résultats. Cette dernière étape est potentiellement compliquée car elle peut nécessiter beaucoup d'unions et de jointures. De telles requêtes ont plusieurs avantages : elles permettent à l'utilisateur de tirer pleinement parti de la complémentarité des bases (y compris pour combiner les données d'un de ses projets aux bases de connaissances), et le déchargent de la phase complexe d'intégration. Plusieurs travaux récents ont porté sur l'application des requêtes fédérées à l'analyse des données en sciences de la vie (Hasnain *et al*. 2017 ; Djokic-Petrovic *et al*. 2017 ; Zaki and Tennakoon 2017 ; Lombardot *et al*. 2018). La figure 1.7 montre une version fédérée de la

28. `https://askomics.org`

requête vue à la figure 1.6 en combinant un *endpoint* pour Reactome et un *endpoint* pour ChEBI.

```
1  PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2  PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
3  PREFIX biopax3: <http://www.biopax.org/release/biopax-level3.owl#>
4
5  PREFIX chebi: <http://purl.obolibrary.org/obo/CHEBI_>
6  PREFIX chebilocal: <http://purl.obolibrary.org/obo/CHEBI_>
7  PREFIX chebiremote: <http://purl.obolibrary.org/obo/CHEBI:>
8  PREFIX taxon: <http://identifiers.org/taxonomy/>
9
10
11 SELECT DISTINCT ?acid ?acidLabel ?reaction ?reactionName
12 WHERE {
13   ?acid rdfs:subClassOf* chebilocal:25754 .
14   OPTIONAL { ?acid rdfs:label ?acidLabel . }
15
16   BIND (URI(REPLACE(str(?acid), "_", ":")) AS ?acidRemote) .
17
18   SERVICE <https://www.ebi.ac.uk/rdf/services/sparql> {
19     ?reactant biopax3:entityReference ?acidRemote .
20
21     ?reaction biopax3:left ?reactant .
22     ?reaction rdf:type biopax3:BiochemicalReaction .
23     OPTIONAL { ?reaction biopax3:displayName ?reactionName . }
24
25     ?pathway biopax3:pathwayComponent ?reaction .
26     ?pathway biopax3:organism taxon:9606 .
27   }
28 }
```

Figure 1.7. *Requête SPARQL combinant un endpoint local avec ChEBI et un endpoint distant avec Reactome et permettant de retrouver toutes les réactions chimiques consommant un acide oxo-carboxylique ou une de ses sous-classes, directe ou indirecte chez l'humain. Il s'agit d'une version fédérée de la requête de la figure 1.6. La clause* `SERVICE` *indique l'URL du endpoint distant (ligne 18) et la partie de la requête à y envoyer (lignes 19 à 27). Remarquez que la variable* `acidRemote` *est commune entre la partie locale et la partie distante de la requête. L'ontologie ChEBI et la base Reactome n'utilisent pas exactement les mêmes préfixes pour les URI des molécules, obligeant à une ré-écriture peu élégante (ligne 16).*

Parmi les freins à l'adoption des requêtes fédérées, on retrouve principalement des problèmes de performance (Abdelaziz *et al*. 2018). C'est d'autant plus le cas

que les requêtes fédérées dont on aurait besoin en sciences de la vie semblent être plus complexes que celles typiquement envisagées dans les *benchmarks* (Saleem *et al*. 2019). Cela indique en contrepoint que les technologies du Web Sémantique permettent d'aborder des requêtes que l'on n'envisageait pas auparavant. Il y a donc là une piste de recherche avec de forts enjeux pour les prochaines années, et les avancées en terme d'architecture de données et de moteurs de requêtes fédérées auront un impact bien plus large que la sous-communauté « sciences de la vie » du Web Sémantique.

1.4. Synthèse

Ce chapitre a permis de faire ressortir le rôle majeur que jouent l'intégration et l'analyse de données en sciences de la vie afin de rendre les analyses exhaustives et de dépasser les analyses ponctuelles et partielles dans chaque sous-domaine. Cela a conduit à dégager des besoins en termes de représentation des données, de structuration de données, de requêtes et de raisonnement. Ces besoins relèvent à la fois du *data engineering* et de la *data science* et sont résolument informatiques. Il semble donc pertinent d'envisager un cadre général plutôt que des solutions *ad hoc*. Les technologies du Web Sémantique ont démontré qu'elles constituent une solution viable à ces besoins, et leur adoption à large échelle par la communauté des sciences de la vie est en cours. Le succès de cette approche réside notamment dans le fait que l'on effectue maintenant des scénarios d'analyse de données qui n'étaient pas envisageables auparavant. Les prochains défis concernent à la fois les aspects informatiques à travers l'amélioration des performances des requêtes notamment fédérées, et les aspects d'utilisabilité pour permettre l'adoption à large échelle par des utilisateurs qui sont des spécialistes de leur domaine mais pas nécessairement du Web Sémantique.

Il est ainsi apparu d'une part que les sciences de la vie constituent un domaine privilégié pour développer des solutions, et d'autre part que ces solutions ont un impact potentiel affectant également l'ensemble de la communauté du Web Sémantique.

1.5. Bibliographie

Abdelaziz, I., Mansour, E., Ouzzani, M., Aboulnaga, A., Kalnis, P. (2018), Lusail: A system for querying linked data at scale, *in* International Conference on Very Large Data Bases (VLDB 2018), vol. 11, pp. 485–498.

Al Kawam, A., Sen, A., Datta, A., Dickey, N. (2018), Understanding the bioinformatics challenges of integrating genomics into healthcare, *IEEE journal of biomedical and health informatics*, 22(5), 1672–1683.

Aldhous, P. (1993), Managing the genome data deluge, *Science (New York, N.Y.)*, 262(5133), 502–503.

Anguita, A., García-Remesal, M., de la Iglesia, D., Maojo, V. (2013), NCBI2RDF: Enabling full RDF-based access to NCBI databases, *BioMed research international*, 2013, 983805.

Aranguren, M. E. n., Bechhofer, S., Lord, P., Sattler, U., Stevens, R. (2007), Understanding and using the meaning of statements in a bio-ontology: recasting the gene ontology in OWL, *BMC bioinformatics*, 8, 57.

Bard, J. B. L., Rhee, S. Y. (2004), Ontologies in biology: design, applications and future challenges, *Nature reviews. Genetics*, 5(3), 213–222.

Baumgartner, W. A., Cohen, K. B., Fox, L. M., Acquaah-Mensah, G., Hunter, L. (2007), Manual curation is not sufficient for annotation of genomic databases, *Bioinformatics (Oxford, England)*, 23(13), i41–i48.

Bechhofer, S., Buchan, I., De Roure, D., Missier, P., Ainsworth, J., Bhagat, J., Couch, P., Cruickshank, D., Delderfield, M., Dunlop, I., Gamble, M., Michaelides, D., Owen, S., Newman, D., Sufi, S., Goble, C. (2013), Why linked data is not enough for scientists, *Future Generation Computer Systems*, 29(2), 599–611.

Bellazzi, R. (2014), Big data and biomedical informatics: A challenging opportunity, *Yearbook of medical informatics*, 9(1). In press.

Bellazzi, R., Diomidous, M., Sarkar, I. N., Takabayashi, K., Ziegler, A., McCray, A. T. (2011), Data analysis and data mining: current issues in biomedical informatics, *Methods of information in medicine*, 50(6), 536–544.

Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P., Morissette, J. (2008), Bio2RDF: towards a mashup to build bioinformatics knowledge systems, *Journal of biomedical informatics*, 41(5), 706–716.

Berners-Lee, T., Hall, W., Hendler, J. A., O'Hara, K., Shadbolt, N., Weitzner, D. J. (2007), A framework for web science, *Foundations and Trends in Web Science*, 1(1), 1–130.

Berners-Lee, T., Hendler, J., Lassila, O. (2001), The semantic web, *Scientific American*, 284(5), 34–43. `http://www.sciam.com/2001/0501issue/0501berners-lee.html`.

Bizer, C., Heath, T., Berners Lee, T. (2009), Linked data–the story so far, *International Journal on Semantic Web and Information Systems*, 5(3), 1–22.

Blake, J. A., Bult, C. J. (2006), Beyond the data deluge: Data integration and bio-ontologies, *Journal of Biomedical Informatics*, 39(3), 314–320.

Bodenreider, O., Stevens, R. (2006), Bio-ontologies: current trends and future directions, *Briefings in Bioinformatics*, 7(3), 256–274.

Brandizi, M., Singh, A., Rawlings, C., Hassani-Pak, K. (2018), Towards FAIRer biological knowledge networks using a hybrid linked data and graph database approach, *Journal of integrative bioinformatics*, 15(3). In press.

Bult, C. J. (2006), From information to understanding: the role of model organism databases in comparative and functional genomics, *Animal Genetics*, 37(suppl.

1), 28–40.

Callahan, A., Cruz-Toledo, J., Dumontier, M. (2013), Ontology-based querying with Bio2RDF's linked open data, *Journal of biomedical semantics*, 4 Suppl 1, S1.

Cannata, N., Merelli, E., Altman, R. B. (2005), Time to organize the bioinformatics resourceome, *PLoS Computational Biology*, 1(7), 0531–0533.

Cannata, N., Schröder, M., Marangoni, R., Romano, P. (2008), A semantic web for bioinformatics: goals, tools, systems, applications, *BMC bioinformatics*, 9 Suppl 4, S1.

Chen, H., Yu, T., Chen, J. Y. (2012), Semantic web meets integrative biology: a survey, *Briefings in bioinformatics*, 14(1), 109–125.

Cheung, K.-H., Frost, H. R., Marshall, M. S., Prud'hommeaux, E., Samwald, M., Zhao, J., Paschke, A. (2009), A journey to semantic web query federation in the life sciences, *BMC bioinformatics*, 10 Suppl 10, S10.

Cimino, J. J. (1998), Desiderata for controlled medical vocabularies in the twenty-first century, *Methods of information in medicine*, 37(4–5), 394–403.

Cimino, J. J. (2005), In defense of the desiderata, *Journal of biomedical informatics*, 39(3), 299–306.

Cimino, J. J., Zhu, X. (2006), The practical impact of ontologies on biomedical informatics, *Methods of information in medicine*, .

Clark, T., Martin, S., Liefeld, T. (2004), Globally distributed object identification for biological knowledgebases, *Briefings in bioinformatics*, 5(1), 59–70.

Corby, O., Dieng-Kuntz, R., Faron-Zucker, C. (2004), Querying the semantic web with corese search engine, *in* R. L. de Mantaras, e. L. Saitta, (eds), Proc. of the 16th European Conference on Artificial Intelligence (ECAI 2004), pp. 705–709.

Côté, R. G., Jones, P., Martens, L., Kerrien, S., Reisinger, F., Lin, Q., Leinonen, R., Apweiler, R., Hermjakob, H. (2007), The protein identifier cross-referencing (picr) service: reconciling protein identifiers across multiple source databases, *BMC bioinformatics*, 8, 401.

Djokic-Petrovic, M., Cvjetkovic, V., Yang, J., Zivanovic, M., Wild, D. J. (2017), PIBAS FedSPARQL: a web-based platform for integration and exploration of bioinformatics datasets, *Journal of biomedical semantics*, 8(1), 42.

Ferré, S. (2014), Expressive and scalable query-based faceted search over SPARQL endpoints, *in* International Semantic Web Conference (ISWC), pp. 438–453.

Goble, C., Stevens, R. (2008), State of the nation in data integration for bioinformatics, *Journal of biomedical informatics*, 41(5), 687–693.

Goodman, A., Pepe, A., Blocker, A. W., Borgman, C. L., Cranmer, K., Crosas, M., Di Stefano, R., Gil, Y., Groth, P., Hedstrom, M., Hogg, D. W., Kashyap, V., Mahabal, A., Siemiginowska, A., Slavkovic, A. (2014), Ten simple rules for the care and feeding of scientific data, *PLoS computational biology*, 10(4), e1003542.

Hasnain, A., Mehmood, Q., Sana E Zainab, S., Saleem, M., Warren, C., Zehra, D., Decker, S., Rebholz-Schuhmann, D. (2017), BioFed: federated query processing over life sciences linked open data, *Journal of biomedical semantics*, 8(1), 13.

Hill, D. P., Adams, N., Bada, M., Batchelor, C., Berardini, T. Z., Dietze, H., Drabkin, H. J., Ennis, M., Foulger, R. E., Harris, M. A., Hastings, J., Kale, N. S., de Matos, P., Mungall, C. J., Owen, G., Roncaglia, P., Steinbeck, C., Turner, S., Lomax, J. (2013), Dovetailing biology and chemistry: integrating the Gene Ontology with the ChEBI chemical ontology, *BMC genomics*, 14, 513.

Holford, M. E., McCusker, J. P., Cheung, K.-H., Krauthammer, M. (2012), A semantic web framework to integrate cancer omics data with biological knowledge, *BMC bioinformatics*, 13 Suppl 1, S10.

Juty, N., Le Novère, N., Laibe, C. (2012), Identifiers.org and miriam registry: community resources to provide persistent identification, *Nucleic acids research*, 40(Database issue), D580–D586.

Kalderimis, A., Lyne, R., Butano, D., Contrino, S., Lyne, M., Heimbach, J., Hu, F., Smith, R., Stepán, R., Sullivan, J., Micklem, G. (2014), Intermine: extensive web services for modern biology, *Nucleic acids research*, 42(Web Server issue), W468–W472.

Kratochvíl, M., Vondrášek, J., Galgonek, J. (2019), Interoperable chemical structure search service, *Journal of cheminformatics*, 11(1), 45.

Levesque, H. J. (2013), On our best behaviour, *in* Proceedings of IJCAI2013 conference.

Livingston, K. M., Bada, M., Baumgartner, W. A., Hunter, L. E. (2015), Kabob: ontology-based semantic integration of biomedical databases, *BMC bioinformatics*, 16, 126.

Lombardot, T., Morgat, A., Axelsen, K. B., Aimo, L., Hyka-Nouspikel, N., Niknejad, A., Ignatchenko, A., Xenarios, I., Coudert, E., Redaschi, N., Bridge, A. (2018), Updates in rhea: Sparqling biochemical reaction data, *Nucleic acids research*, 47(D1), Interoperable chemical structure search service.

McCarthy, L., Vandervalk, B., Wilkinson, M. (2012), Sparql assist language-neutral query composer, *BMC bioinformatics*, 13 Suppl 1, S2.

McMurry, J. A., Juty, N., Blomberg, N., Burdett, T., Conlin, T., Conte, N., Courtot, M., Deck, J., Dumontier, M., Fellows, D. K., Gonzalez-Beltran, A., Gormanns, P., Grethe, J., Hastings, J., Hériché, J.-K., Hermjakob, H., Ison, J. C., Jimenez, R. C., Jupp, S., Kunze, J., Laibe, C., Le Novère, N., Malone, J., Martin, M. J., McEntyre, J. R., Morris, C., Muilu, J., Müller, W., Rocca-Serra, P., Sansone, S.-A., Sariyar, M., Snoep, J. L., Soiland-Reyes, S., Stanford, N. J., Swainston, N., Washington, N., Williams, A. R., Wimalaratne, S. M., Winfree, L. M., Wolstencroft, K., Goble, C., Mungall, C. J., Haendel, M. A., Parkinson, H. (2017), Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data, *PLoS biology*, 15(6), e2001414.

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., Musen, M. A. (2009), Bioportal: ontologies and integrated data resources at the click of a mouse, *Nucleic acids research*, 37(Web Server issue), W170–W173.

Pérez, J., Arenas, M., Gutierrez, C. (2009), Semantics and complexity of sparql, *ACM Trans. Database Syst.*, 34(3), 16:1–16:45.
URL: *http://doi.acm.org/10.1145/1567274.1567278*

Pierce, H. H., Dev, A., Statham, E., Bierer, B. E. (2019), Credit data generators for data reuse, *Nature*, 570(7759), 30–32.

Post, L. J. G., Roos, M., Marshall, M. S., van Driel, R., Breit, T. M. (2007), A semantic web approach applied to integrative bioinformatics experimentation: a biological use case with genomics data, *Bioinformatics (Oxford, England)*, 23(22), 3080–3087.

Rho, K., Kim, B., Jang, Y., Lee, S., Bae, T., Seo, J., Seo, C., Lee, J., Kang, H., Yu, U., Kim, S., Lee, S., Kim, W. K. (2011), GARNET - gene set analysis with exploration of annotation relations, *BMC bioinformatics*, 12 Suppl 1, S25.

Rigden, D. J., Fernández, X. M. (2019), The 26th annual nucleic acids research database issue and molecular biology database collection, *Nucleic acids research*, 47(D1), D1–D7.

Rodríguez-Iglesias, A., Rodríguez-González, A., Irvine, A. G., Sesma, A., Urban, M., Hammond-Kosack, K. E., Wilkinson, M. D. (2016), Publishing fair data: An exemplar methodology utilizing phi-base, *Frontiers in plant science*, 7, 641.

Ruttenberg, A., Clark, T., Bug, W., Samwald, M., Bodenreider, O., Chen, H., Doherty, D., Forsberg, K., Gao, Y., Kashyap, V., Kinoshita, J., Luciano, J., Scott Marshall, M., Ogbuji, C., Rees, J., Stephens, S., Wong, G. T., Wu, E., Zaccagnini, D., Hongsermeier, T., Neumann, E., Herman, I., Cheung, K.-H. (2007), Advancing translational research with the semantic web, *BMC Bioinformatics*, 8(3).

Sahoo, S. S., Bodenreider, O., Zeng, K., Sheth, A. (2007), An experiment in integrating large biomedical knowledge resources with RDF: Application to associating genotype and phenotype information, *in* Proceedings of the WWW2007 Workshop on Health Care and Life Sciences Data Integration for the Semantic Web.

Saleem, M., Szárnyas, G., Conrads, F., Bukhari, S. A. C., Mehmood, Q., Ngonga Ngomo, A.-C. (2019), How representative is a sparql benchmark? an analysis of rdf triplestore benchmarks, *in* Proceedings of ACM Conference (TheWebConf WWW19).

Salvadores, M., Horridge, M., Alexander, P. R., Fergerson, R. W., Musen, M. A., Noy, N. F. (2012), Using SPARQL to query Bioportal ontologies and metadata, *in* Proceedings of the International Semantic Web Conference ISWC 2012, vol. 7650 of *Lecture Notes in Computer Science*, pp. 180–195.

Scharffe, F., Atemezing, G., Troncy, R., Gandon, F., Villata, S., Bucher, B., Hamdi, F., Bihanic, L., Képéklian, G., Cotton, F., Euzenat, J., Fan, Z., Vandenbussche, P.-Y., Vatant, B. (2012), Enabling linked data publication with the datalift platform, *in* AAAI 2012, 26th Conference on Artificial Intelligence, W10:Semantic Cities, July 22-26, 2012, Toronto, Canada.

Schulz, S., Balkanyi, L., Cornet, R., Bodenreider, O. (2013), From concept representations to ontologies: A paradigm shift in health informatics?, *Healthcare informatics research*, 19(4), 235–242.

Shadbolt, N., Hall, W., Berners Lee, T. (2006), The semantic web revisited, *IEEE Intelligent Systems*, pp. 96–101.

Sima, A. C., Stockinger, K., de Farias, T. M., Gil, M. (2019), Semantic integration and enrichment of heterogeneous biological databases, *Methods in molecular biology (Clifton, N.J.)*, 1910, 655–690.

Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., Arnaiz, O., Awedh, M. H., Baldock, R., Barbiera, G., Bardou, P., Beck, T., Blake, A., Bonierbale, M., Brookes, A. J., Bucci, G., Buetti, I., Burge, S., Cabau, C., Carlson, J. W., Chelala, C., Chrysostomou, C., Cittaro, D., Collin, O., Cordova, R., Cutts, R. J., Dassi, E., Genova, A. D., Djari, A., Esposito, A., Estrella, H., Eyras, E., Fernandez-Banet, J., Forbes, S., Free, R. C., Fujisawa, T., Gadaleta, E., Garcia-Manteiga, J. M., Goodstein, D., Gray, K., Guerra-Assuncao, J. A., Haggarty, B., Han, D.-J., Han, B. W., Harris, T., Harshbarger, J., Hastings, R. K., Hayes, R. D., Hoede, C., Hu, S., Hu, Z.-L., Hutchins, L., Kan, Z., Kawaji, H., Keliet, A., Kerhornou, A., Kim, S., Kinsella, R., Klopp, C., Kong, L., Lawson, D., Lazarevic, D., Lee, J.-H., Letellier, T., Li, C.-Y., Lio, P., Liu, C.-J., Luo, J., Maass, A., Mariette, J., Maurel, T., Merella, S., Mohamed, A. M., Moreews, F., Nabihoudine, I., Ndegwa, N., Noirot, C., Perez-Llamas, C., Primig, M., Quattrone, A., Quesneville, H., Rambaldi, D., Reecy, J., Riba, M., Rosanoff, S., Saddiq, A. A., Salas, E., Sallou, O., Shepherd, R., Simon, R., Sperling, L., Spooner, W., Staines, D. M., Steinbach, D., Stone, K., Stupka, E., Teague, J. W., Dayem Ullah, A. Z., Wang, J., Ware, D., Wong-Erasmus, M., Youens-Clark, K., Zadissa, A., Zhang, S.-J., Kasprzyk, A. (2015), The BioMart community portal: an innovative alternative to large, centralized data repositories, *Nucleic acids research*, 43(W1), W589–W598.

Smith, B. (2005), New desiderata for biomedical terminologies, *in* Ontologies and Biomedical Informatics, Conference of the International Medical Informatics Association.

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., Robinson, G. E. (2015), Big data: Astronomical or genomical?, *PLoS biology*, 13(7), e1002195.

Stevens, R., Egaña Aranguren, M., Wolstencroft, K., Sattler, U., Drummond, N., Horridge, M., Rector, A. (2007), Using OWL to model biological knowledge, *International Journal of Human Computer Studies*, 65(7), 583–594.

Stevens, R., Goble, C. A., Bechhofer, S. (2000), Ontology-based knowledge representation for bioinformatics, *Briefings in bioinformatics*, 1(4), 398–416.

Taboada, M., Martínez, D., Pilo, B., Jiménez-Escrig, A., Robinson, P. N., Sobrido, M. J. (2012), Querying phenotype-genotype relationships on patient datasets using semantic web technology: the example of cerebrotendinous xanthomatosis, *BMC medical informatics and decision making*, 12, 78.

van Dam, J. C., Koehorst, J. J., Schaap, P. J., Martins Dos Santos, V. A., Suarez-Diez, M. (2015), Rdf2graph a tool to recover, understand and validate the ontology of an rdf resource, *Journal of biomedical semantics*, 6, 39.

Whetzel, P. L., Noy, N. F., Shah, N. H., Alexander, P. R., Nyulas, C., Tudorache, T., Musen, M. A. (2011), BioPortal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications, *Nucleic acids research*, 39(Web Server issue), W541–W545.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B. (2016), The fair guiding principles for scientific data management and stewardship, *Scientific data*, 3, 160018.

Wimalaratne, S. M., Bolleman, J., Juty, N., Katayama, T., Dumontier, M., Redaschi, N., Le Novère, N., Hermjakob, H., Laibe, C. (2015), SPARQL-enabled identifier conversion with identifiers.org, *Bioinformatics (Oxford, England)*, 31(11), 1875–1877.

Yamaguchi, A., Kozaki, K., Lenz, K., Wu, H., Kobayashi, N. (2014), An intelligent sparql query builder for exploration of various life-science databases, *in* Proc. of the 3rd International Conference on Intelligent Exploration of Semantic Data, vol. 1279, pp. 83–94.

Zaki, N., Tennakoon, C. (2017), BioCarian: search engine for exploratory searches in heterogeneous biological databases, *BMC bioinformatics*, 18(1), 435.

Zhang, Y., De, S., Garner, J. R., Smith, K., Wang, S. A., Becker, K. G. (2010), Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information, *BMC medical genomics*, 3, 1.