Българска Академия на Науките

Институт по информационни и комуникационни технологии
Секция: "Лингвистично моделиране и обработка на знания"

Издателство "Пенсофт" ЕООД

Автореферат

за придобиване на образователната и научна степен

"Доктор"

по докторска програма "Информатика" (01.01.12)
професионално направление 4.6 "Информатика и компютърни науки"

OpenBiodiv: отворена система за управление на знанието за биологичното разнообразие

Автор:
маг. Виктор Сендеров

Научен ръководител:
проф. д-р Любомир Пенев
Научен консултант:
доц. д-р Кирил Симов

София, 2019 г.

Това изследване е осъществено благодарение на финансовата подкрепа на Е.С. в рамките на изследователски грант No. 642241 по проект BIG4 на програмата Мария-Склодовоска Кюри, част от “Хоризонт 2020”.

Увод

Актуалност и значимост на темата

Необходимостта от създаване на интегрирана информационна система, обхващаща знание за биологично разнообразие, датира от 1985 г. Тогава е учредена Работна група по таксономични бази данни (TDWG), която в последствие е преименувана на Стандарти за информационни технологии за биологичното разнообразие, но запазва съкращението TDWG[1]. През 1999 г. е създаден Глобален институт за информация за биологичното разнообразие (GBIF), след като Организацията за икономическо сътрудничество и развитие (OECD) стига до извода, че е необходим международен механизъм за достъп до данни и информация за биоразнообразието в световен мащаб (*What is GBIF?*). Декларацията Bouchout (*Bouchout Declaration* 2014) ознаменува резултатите от финансирания от Европейския Съюз проект *pro-iBiosphere*, който продължава от 2012 г. до 2014 г. и е посветен на задачата за създаване на интегрирана информационна система за биологичното разнообразие. Декларацията "Бухаут" призовава към свободното предоставяне на научни знания за биологичното разнообразие като Linked Open Data (LOD). Успореден процес в САЩ започва още по-рано със създаването на Архитектура за глобални имена, GNA (Patterson et al., 2010; Pyle, 2016).

През 2014 г., в рамките на европейската програма "Хоризонт 2020", между академични институции и частни фирми е създаден консорциумът BIG4. BIG4 цели да допринесе за напредъка на науката за биологичното разнообразие. Мисията на проекта е: "BIG4 - Биосистематика, информатика и генетика на големите 4 групи насекоми: обучение на учените и предприемачите от утрешния ден" (University of Copenhagen et al., 2014). Важен член на консорциума е академичното издателство и софтуерна компания "Пенсофт". Пенсофт публикува няколко десетки добре познати таксономични списания с отворен достъп[2]. Като подписваща страна на декларацията "Бухаут", Пенсофт е естествен кандидат за изграждането на Отворена система за управление на знанията за биоразнообразието (OBKMS).

Поради тези причини настоящата дисертация за придобиване на докторска степен е базирана в академичното издателство "Пенсофт" и в Институт по информационни и комуникационни технологии на Българска академия на науките (ИИКТ-БАН).

Преглед на литературата по проблема

Поради интердисциплинарния характер на дисертацията, в рамките на литературната справка ще се прегледат две области: (а) "бази от знания и свързани отворени данни", а също така и (б) "публикуване на знания и данни в сферата на биоразнообразието".

[1]Историята на TDWG стартира през 1985 г. и може да се види на `http://old.tdwg.org/past-meetings/`. Много от препратките на тази уеб страница за съжаление са невалидни и страницата се нуждае от известна поддръжка.

[2]Напр. ZooKeys, PhytoKeys, MycoKeys, Biodiversity Data Journal (BDJ).

Бази от знания и свързани отворени данни

Въвеждат се на термините "база от знания (knowledge base)" и "система, използваща знание (knowledge-based system)". В дисертацията двата термина се използват взаимозаменяемо. Понякога се изписва по-дългият вариант – система, използваща знания/ система от знания – когато се подчертават аспекти на базата от знания, които не са свързани с технологичните аспекти на базата данни – напр. потребителски интерфейс, източници на информация и т.н. Понятието се дефинира посредством изрични дефиниции, а също и чрез разглеждане на реализации на няколко системи от знания на практика. Терминът е широко обсъждан още през осемдесетте години на 20 век (Jarke et al., 1989) и началото на деветдесетте години (Harris et al., 1993). Значението, влагано му тогава, е "използване на идеи, както от системите за управление на бази данни (DBMS), така и от изкуствения интелект (ИИ) за създаване на тип компютърна система, наречена *knowledge management system* (KBMS)". Harris et al., 1993 пише, че характеристиките на KBMS са, че тя съдържа "предписани правила и факти, от които полезни изводи могат да бъдат извлечени от машина за правене на извод (inference engine)". Трябва да се отбележи, че това тълкуване идва от времето на първото поколение ИИ (rule-based systems). В по-ново време бе постигнат напредък във включването на статистически техники в базите данни (Mansinghka et al., 2015). Независимо от това, в този проект се работи с класическото разбиране за KBMS, основано на логически правила. С други думи под *база от знания* се разбира "подходяща база данни, тясно интегрирана с логически слой, който позволява правенето на логически изводи и придава семантика на данните".

Друго относително по-ново развитие е приложението на принципите на свързаните отворени данни (Heath and Bizer, 2011). Повечето съществуващи бази от знания наблягат на социални аспекти, които правят данните взаимосвързани и годни за повторно използване. Примери са Freebase (Bollacker et al., 2008), която неотдавна беше включена в WikiData (Vrandečić and Krötzsch, 2014; Pellissier Tanon et al., 2016), DBPedia (Auer et al., 2007), както и Wolfram|Alpha (*Wolfram|Alpha, Making the wolrd's knowledge computable*), а също така и Google Knowledge Graph (Singhal, 2012). Общото между тези системи е, че акцентът се поставя не само върху логическия слой, който позволява изводи, а и върху единното информационно пространство: тези системи действат като интегрират информация от множество източници и следват принципите на свързаните отворени данни, Linked Open Data (LOD). Семантичната мрежа (Semantic Web, Berners-Lee et al., 2001) предлага редица препоръки за LOD (Heath and Bizer, 2011), които, когато се прилагат правилно, гарантират, че публикуваните в Мрежата данни са годни за използване от всички потребители на Мрежата. Принципите на свързаните данни и тяхното приложение към OpenBiodiv се разглеждат в глава 3.

Водени от тези тенденции, модерните бази от знания наблягат повече върху свързването на данни, отколкото върху разработването на сложни механизми за логически изводи. Критика на идеята за интегриране на логически слой в базата данни се състои в наблюдението, че интегрирането води до повишена сложност (Barrasa, 2017). Критиката може да бъде обобщена с две точки. Първо, поставянето на логика близо до данните (особено, когато е прекомерно мощна за задачата) може да доведе до драстично намаляване на производителността[3]. Второ, нови техники (например машинно обучение) правят съществуващия логически слой излишен.

[3]Ефективността на логическия слой OWL се сравнява с по-слабата RDF схема (RDFS) в глава 3.

От наша гледна точка, данните са обектът, който е много по-ценен, докато стратегията за вадене на изводи (дали е логически слой, основан на правила, или техника за статистическо машинно обучение) може да бъде подменяна с напредването на компютърните науки. Тези идеи водят до интересна главоблъсканица в избора на технология за база данни, разгледана в следващите раздели.

И накрая, всяка система, базирана на знания, трябва да включва задължително и компоненти на потребителския интерфейс и/или интерфейс за програмиране на приложения (API), както и приложения (apps). Те служат като точка на контакт между човека и машината и са от решаващо значение за успеха на всяка такава система.

Публикуване на знания данни за биоразнообразието

В биомедицинската област отдавна се работи по извличането на информация и откриването на знания от първична литература (напр. Rebholz-Schuhmann et al., 2005; Momtchev et al., 2009; Williams et al., 2012). Областта на биоразнообразието, и по-специално биологичната систематика и таксономия (от тук нататък в дисертацията съкратено наричана *таксономия*), също се движи в посока към семантизация (напр. Agosti, 2006; Patterson et al., 2006; Kennedy et al., 2005; Penev et al., 2010a; Tzitzikas et al., 2013). Академичната издателска дейност е моделирана чрез Онтологичните издателски и референтни онтологии, SPAR Ontologies (Peroni, 2014). Онтологиите на SPAR са колекция от онтологии, включващи, наред с другото, библиографската онтология FaBiO (Peroni and Shotton, 2012) и DoCO, онтология за компонентите на даден документ (Constantin et al., 2016). Онтологиите на SPAR осигуряват класове и свойства за описанието на академични статии с общо предназначение. Таксономичните статии и техните компоненти, от друга страна, са моделирани чрез TaxPub XML Document Type Definition (DTD) и неформално се нарича "XML схема" (Catapano, 2010). TaxPub е XML схемата на няколко важни таксономични списания (напр. ZooKeys, PhytoKeys, Biodiversity Data Journal) и служи като концептуален шаблон за онтология OpenBiodiv-O (глава 2), разработена в хода на дисертацията.

Таксономичната номенклатура е дисциплина с много дълга традиция. Тя се трансформира в модерната си форма с публикуването на линеевата система (Linnaeus, 1758). Вече към началото на миналия век са били използвани стотици таксономични термини (Witteveen, 2015). Понастоящем именуването на групи от организми се регулира от Международния кодекс за зоологична номенклатура (ICZN, International Commission on Zoological Nomenclature, 2017) и Международния кодекс за номенклатура на водорасли, гъби и растения, известен като Кодекс на Мелбърн (*International code of nomenclature for algae, fungi and plants (Melbourne code)* 2012). Поради тяхната сложност (напр. ICZN има 18 глави и 3 приложения), се оказва непреодолимо до този момент предизвикателство да се създаде непротиворечива онтология базирана на кодексите за биологична номенклатура. Опити за преодоляването на това предизвикателство са сравнително пълната онтология на NOMEN (Dmitriev and Yoder, 2017) и не до там завършените Термини за статута на таксономичните номенклатури, TNSS.

Съществуват няколко проекта, целящи моделирането на по-широката област на биологичното разнообразие. Darwin-SW (Baskauf and Webb, 2016) адаптира предишните термини "DarwinCore" (Wieczorek et al., 2012) като Resource Description Framework (RDF). Двата модела се занимават основно с данни за наличие на организми. Моделирането и формализирането на строго таксономичния домейн е обсъдено от Berendsohn, 1995 и по-късно в напр. Franz and Peet, 2009; Sterner

and Franz, 2017. За отбелязване е и XML-базираната схема за прехвърляне на таксономични концепции (Taxonomic Names and Concepts Interest Group, 2006) и вече невалидната Taxon Concept Ontology. Наскоро в общността на TDWG се поднови интересът към таксономичните концепции, чрез създаването на Група за имена и таксономични концепции. Груповите дискусии могат да бъдат достъпени под `https://github.com/tdwg/tnc`. Интересното е, че в първата дискусия в GitHub се обсъди OpenBiodiv-O и възможността за приемането му като стандарт на TDWG.

През юни 2015 г. се стартира проектът OpenBiodiv. До старта на проекта са публикувани редица важни статии по темите за свързване на данни и споделяне на идентификатори в областта биологичното разнообразие (Page, 2008), относно обединяването на филогенетични знания (Parr et al., 2012), и по темата на таксономичните имена и връзката им със семантичната мрежа (Page, 2006; Patterson et al., 2010), както и относно обобщаването на изследванията на биологичното разнообразие (Mindell et al., 2011). Дискусия относно OBKMS се намира в научния блог iPhylo (Page, 2014, 2015). Правните аспекти на OBKMS са обсъдени от Egloff et al., 2014. Освен това няколко системи за интегриране на данни за биоразнообразието са разработени от различни групи. Някои от най-важните са UBio, Глобални имена, BioGuid, BioNames, Профил на таксони на Пенсофт и Плази[4].

Основни изводи от литературната справка

Основните изводи от разгледаните източници са обобщени както следва:

1. Биоразнообразието се занимава с различни видове данни: таксономични, биогеографски, филогенетични, визуални, описателни и други. Тези данни са трудно достъпни в не взаимно-свързани хранилища за данни.

2. Базите данни за биологичното разнообразие се нуждаят от универсална система за именуване на таксономични концепции поради недостатъците на линеевите имена за модерната таксономия. Етикетите за таксономични концепции са предложени като решение, разбираемо от човека. А глобално-стабилни уникални идентификатори (GUID) на таксономични концепции – като решение, удобно за обработка в машинен вид.

3. Налице е основа от цифровизирана полу-структурирана информация за биологичното разнообразие в Мрежата с подходящи лицензи, чакащи да бъдат интегрирани като база от знания.

Цел и задачи

Предвид огромния международен научен интерес към отворената система за управления на знания и данни за биоразнообразието, тази дисертация стартира проект OpenBiodiv, който да допринесе за създаването на системата, като се концентрира върху знание, извлечено от научната литература.

Целта на проекта е създаването на формален семантичен модел на областта на публикуването на знание за биологичното разнообразие и приложението на този модел за създаване на система за свързани отворени данни за биоразнообразието.

За да се завърши системата, са формулирани следните задачи:

[4]UBio: http://ubio.org/; Глобални имена: http://globalnames.org/; BioGuid: http://bioguid.org/; BioNames: http://bionames.org/; Профил на таксон на Пенсофт: http://ptp.pensoft.eu/; Плази: http://plazi.org/wiki/.

Задача 1: Онтология. Изучаване на информатиката на биоразнообразието и публикуването на данни и разработване на онтология, която позволява интегрирането на данни за биологичното разнообразие от различни източници.

Задача 2: Софтуерна архитектура. Формализация на OpenBiodiv като основана на знания система и създаването на интегрираната ѝ софтуерна архитектура.

Задача 3: Свързани отворени данни. Създаване на свързани отворени данни (LOD) въз основа на публикувани таксономични статии, използващи онтологията от Задача 1.

Задача 4: Софтуерна библиотека. Разработване на инструменти за преобразуване на таксономичните публикации в семантичния модел на онтологията с цел подпомагане на Задача 3.

Задача 5: Методи за работа. Разработване на практически методи за работа за непрекъснато преобразуване на таксономичните данни в таксономични публикации и по този начин актуализиране на набора от данни LOD.

Задача 6: Уеб портал. Създаване на уеб портал с примерни приложения в допълнение към базата от знания.

Методология

В този раздел се очертават изборите, направени преди започване на фазата на проектиране и изпълнение на системата. Те включват, но не се ограничават до напр. парадигмите за програмиране и за база данни.

Избор на парадигма на база данни за OpenBiodiv

OpenBiodiv е формулирана като система, основана на знание, с фокус върху структурирането и взаимовръзката на данни за биоразнообразието. Две от възможните технологии за бази данни, с които системата може да се реализира, са семантична бази данни (triple stores), като напр. GraphDB (Ontotext, 2018) и графична[5] бази данни (labeled property graphs), като напр. Neo4J (Neo4J Developers, 2012). И двата типа бази данни всъщност се основават на модела на математическия граф и могат да бъдат наричани "графични". Семантичните (графични) бази данни предлагат много прост модел за данни: всеки факт, съхраняван в такава база данни, се представя като тройка от подлог *subject*, сказуемо *predicate* и пряко допълнение *object*. Subjects са винаги идентификатори на ресурси, докато objects могат да бъдат други идентификатори или буквални стойности (литерали, напр. низове, числа и т.н.). Връзките се дават от predicates (посочени също като идентификатори). Тези връзки се наричат предикати или свойства (*properties*). По този начин може да се визуализира граф, чиито върхове са идентификатори или буквални стойности и, чиито ребра са свойства. Семантичните бази данни имат уникалната черта, че логическият слой също се изразява като тройки, съхранени в базата данни. Този логически слой, известен като онтология (*ontology*), не

[5]Тук трябва да се внимава: графична в случая означава базирана на граф (т.е. съвкупност от ребра и върхове), а не: на графика.

само отговаря за извличането на нови факти от данните (извод), но също така формализира и семантиката (смисловия модел) на знанието.

Не-семантичните графични бази данни, наричани просто графични, предлагат по-свободен модел за данни, като позволяват и ребрата на графа от знания да имат етикети. Например, в граф, чиито върхове са два града A и B, които са свързани посредством свойството *свързани с пътм*, е възможно допълнително да се приложи стойността "500 км" към това свойство. По този начин посочваме, че дължината на пътя, свързващ градовете, е 500 км. Този разширен модел не е по-изразителен от простия семантичен, разгледан в предишния параграф. Т.е. няма твърдение, което може да бъде изказано с разширения модел, а да не може да бъде изказано посредством семантичния модел. Сложни взаимоотношения в обикновен triple store могат да бъдат изразени чрез преобразуването на сложни свойства в нови ресурси, които имат собствени свойства. Този процес е известен като реификация (*reification*). Например, двата града A и B могат да се свържат към друг връх, R, посочващ пътя. R ще има три свойства: *start*, *end* и *length*. Стойността (object) на *start* ще бъде A, на *end* ще бъде B, а *length* ще бъде литералът "500 км" или числото 500.

Разликите между графичните и семантичните бази данни са обобщени в Таблица 1. След внимателни разсъждения, семантична база данни е избрана на технология за база данни. Това решение е информирано от широката наличност на висококачествени онтологии и RDF модели в областта на биологичното разнообразие (Baskauf and Webb, 2016; Peroni, 2014) и от популярността на семантичната мрежа (Berners-Lee et al., 2001) в научната общност.

Въпреки това смятаме, че несемантичните графи са по-свободен и по-естествен модел на данни и са напълно подходящи за информатиката за биоразнообразието. По-специално, те осигуряват много по-естествен формализъм за изразяване на взаимоотношенията между таксономичните концепции (обсъдени в глава 2). Също така напоследък семантични бази данни, различни от RDF, като WikiData, стават популярни. Ето защо считаме, че приложимостта на RDF за OpenBiodiv трябва постоянно да бъде преоценявана.

Избор на източници на информация

Биоразнообразието и свързаните с биоразнообразието данни имат два различни "цикъла на живот" (*pro-iBiosphere project final report* 2014). В класическия вариант, след като е направено наблюдение на жив организъм, то е записвано в тетрадка и след това бележка за наблюдение е публикувана в научна статия или монография. Не-правителствената организация "Плази", както и Biodiversity Heritage Library (Miller et al., 2012), полагат усилия за дигитализиране на публикации, публикувани по този начин на хартиен носител (Agosti et al., 2007). За тази цел са разработени няколко специални XML схеми (вж. Penev et al., 2011 за преглед), от които TaxPub (Catapano, 2010) и TaxonX са най-широко използвани (Penev et al., 2012). Дигитализирането на публикациите съдържа няколко стъпки. След сканиране и оптично разпознаване на символи (OCR), се извършва text-mining. Тази процедура маркира елементи (semantic markup), които могат да бъдат извлечени и предоставени за бъдещо използване и повторно използване (Miller et al., 2015).

В днешно време знание и данни за биоразнообразието и се публикуват в цифров формат като семантично подобрени публикации, enhanced publications (EP, Claerbout and Karrenbach, 1992; Godtsenhoven et al., 2009; Shotton, 2009). Според Claerbout and Karrenbach, 1992, "EP е публикация, която е разширена с данни

Таблица 1: Разлики между бази данни върху семантични графи (напр. GraphDB) и не-семантични графи (напр. Neo4j).

Критерий	Семантичен граф	Не-семантичен граф
Семантика	Съхранява се в самата база данни като OWL или RDFS-изрази. Осигурява единно пространство за данни. Изисква експерти-онтолози да извличат знания.	Формална семантика обикновено липсва. Бързо внедряване. Унифицирано пространство за данни по-трудно постижимо.
Извод	Осигурен от самата база данни от нейната онтология или формулиран посредством SPARQL заявки. С общо предназначение, по-бавен.	Външен за базата данни. Трябва да се напише за всяка конкретна задача. Със специално предназначение. По-бърз.
Общност	Има богата и зряла общност от онтолози и инженери по знания. Много онтологии за различните дисциплини. Стремяща се към стандарти.	Моделите се създават ad-hoc от програмисти за дадена задача. Постигането на интероперативност на данните изисква усилия и не е от първостепенно значение. Стремяща се към работещи приложения.

от изследвания, допълнителни материали и допълнителни данни. Тя има структура, базирана на обекти, с изрични връзки между обектите. Един обект може да бъде (част от) статия, набор от данни, изображение, филм, коментар, модул или връзка към информация в база данни". По този начин семантично подобрените публикации са "родени в Интернет и в семантичната мрежа" за разлика от техните хартиени предшественици.

Актът за публикуване в цифров, подобрен формат, се различава основно от публикуването в печатен източник. Основната разлика е, че цифрово публикуваният документ може да бъде структуриран в такъв формат, че да е подходящ както за машинна обработка, така и за човешкото око. В сферата на науката за биологичното разнообразие, списания на Пенсофт като ZooKeys, PhytoKeys и Biodiversity Data Journal (BDJ) от години предлагат публикуване на EP (Penev et al., 2010b).

Предвид факта, че публикациите на Пенсофт и Плаци покриват голяма част от таксономичната литература както по обем, така и по хронология, и че публикациите на тези две издателства са достъпни като семантични EP, периодичните издания на Пенсофт и Плаци бяха избрани като основни източници на информация за OpenBiodiv.

Освен това в системата е включен таксономичният гръбнак на GBIF GBIF Secretariat, 2017 като източник за интеграция на данни. Допълнителни разяснителни детайли са дадени в глава 3.

Избор на методология и среда за програмиране

През 2016 г., въз основа на резултатите от pro-iBiosphere и на съществуващи разработки в областта на информатиката за биологичното разнообразие, е публикуван план за докторантурата (Senderov and Penev, 2016). Тази публикация може да се счита за първата спецификация на проекта на OpenBiodiv. Въпреки това, в хода на разработването на системата, дизайнът на системата бе променен итеративно чрез обратна връзка от сътрудници от проекта BIG4 project[6] Тези промени са в духа на отворената наука (*open science*) и на *agile разработката на софтуер* (Beck et al., 2001). Този итеративен подход се различава от подхода waterfall, където след фазата на проектиране, спецификациите "са замразени" през дълга изпълнителна фаза.

През последните години програмният език R се използва широко в областта на науката за данни data science (R Core Team, 2016). R има богата библиотека от софтуерни пакети, включваща пакети за обработка на XML (Wickham et al., 2018), за достъп до API (Wickham, 2017) и се фокусира върху отворена наука (Boettiger et al., 2015). Възможностите на R като функционално-ориентиран и интерпретиран език облекчават подхода за итеративно разработване на софтуер, очертан в предходния параграф. Освен това R се използва широко в общността на информатиката за биологичното разнообразие. Поради тази причина средата на софтуера R е избрана за основна програмна среда.

Отворена наука и Семантична Мрежа

Следните методологии правят резултатите от изследването отворени и възпроизводими.

[6]Кандидатът Виктор Сендеров е част от Международната мрежа за обучение на "Мария Склодовска-Кюри" BIG4: Биосистематика, информатика и геномика на големите 4 групи насекоми: обучение на утрешните изследователи и предприемачи.

OpenBiodiv трябва да бъде разгледан от гледна точка на *Open Science*. Съгласно Kraker et al., 2011 и *Was ist Open Science?*, шестте принципа на Отворената наука са: отворена методология, отворен код, отворени данни, отворен достъп, отворени рецензии и отворени образователни ресурси. Целта на Отворената наука е да осигури достъп до всички изследователски продукти: данни, открития, хипотези, код и т.н. Това отваряне гарантира, че научният продукт може да бъде възпроизводим и проверим от други учени (Mietchen, 2014). Съществува голям интерес към разработването на процеси и инструменти, които дават възможност за възпроизводимост и проверка. Тези проблеми са разгледани напр. в специален брой в Nature, посветена на възпроизводими изследвания (*Challenges in irreproducible research* 2010). Поради това изходният код, данните и публикациите на OpenBiodiv ще бъдат публикувани открито.

Освен това OpenBiodiv се разглежда като неразделна част от Семантичната Мрежа (Berners-Lee et al., 2001). Семантичната мрежа е визия за бъдещето на Мрежата, където са свързани не само документи, но и данни.

Структура на дисертацията

Във Въведение е дадена мотивировка за съществуването на системата OpenBiodiv, както и обобщение на нейните цели и задачи.

В глава 1 е представена формалната спецификацията и дизайнът на системата, както и нейната архитектура; тази глава формира Задача 2, но в логическия ход на дисертацията е хубаво да бъде разгледана първо. Следващите глави обсъждат изпълнението на OpenBiodiv. Глава 2 предлага формална концептуализация на областта на публикуването на данни за биологичното разнообразие. Въвежда се централният научен резултат на дисертацията – онтологията OpenBiodiv (OpenBiodiv-O) и се покрива Задача 1. Глава 3 описва отворените данни, които са генерирани въз основа на OpenBiodiv-O и формира Задача 3. Глава 4 подробно описва софтуерния пакет RDF4R (R пакет за работа с RDF), който бе използван за създаване на Linked Open Data (OpenBiodiv-LOD) и формира Задача 4. В глава 5 се обсъждат два казуса за внасяне на данни в OpenBiodiv от важни международни хранилища, Задача 5. Глава 6 обсъжда уеб-сайта, който се подготвя да служи на на OpenBiodiv-LOD и приложенията му и покрива Задача 6. В Заключение резултатите на дисертацията са обобщени и научните ѝ и научно-приложни приноси са изтъкнати. Разисква се публикуването и популяризирането на резултатите.

Глава 1

Архитектура на OpenBiodiv

Въвеждат се компонентите на OpenBiodiv, които ще бъдат разгледани подробно в следващите глави. Описва се, как взаимодействат тези компоненти, за да се формира базираната на знанието система OpenBiodiv.

1.1 Какво е OpenBiodiv?

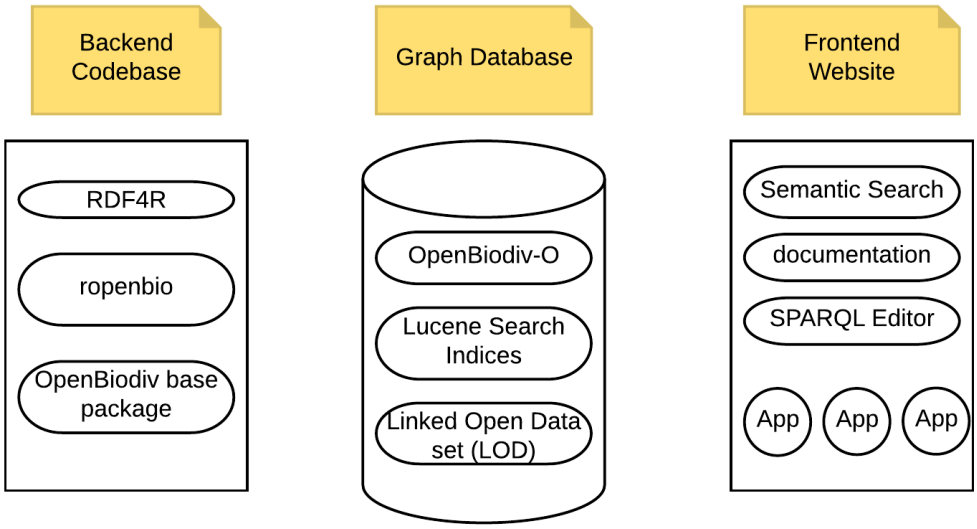
Разбирането за OpenBiodiv като система, базирана на знание, може да се обобщи по следния начин: OpenBiodiv е база данни от взаимосвързана информация за биологичното разнообразие, заедно с логика и приложения, позволяващи на потребителите не само да се допитват до данните, но и да откриват допълнителни факти, свързани с данните. Основните източници на информация в OpenBiodiv са списанията на академичния издател "Пенсофт", таксономичната информация от Плаци и таксономичният гръбнак на Global Information Biodiversity Facility (GBIF).

Изследователският проблем на архитектурата на OpenBiodiv може да се формулира като проектиране на семантична графична база данни на основата на RDF, която е с отворен достъп и включва информация, предоставяна от Пенсофт, Плаци и GBIF и позволява на потребителите на системата да задават сложни заявки.

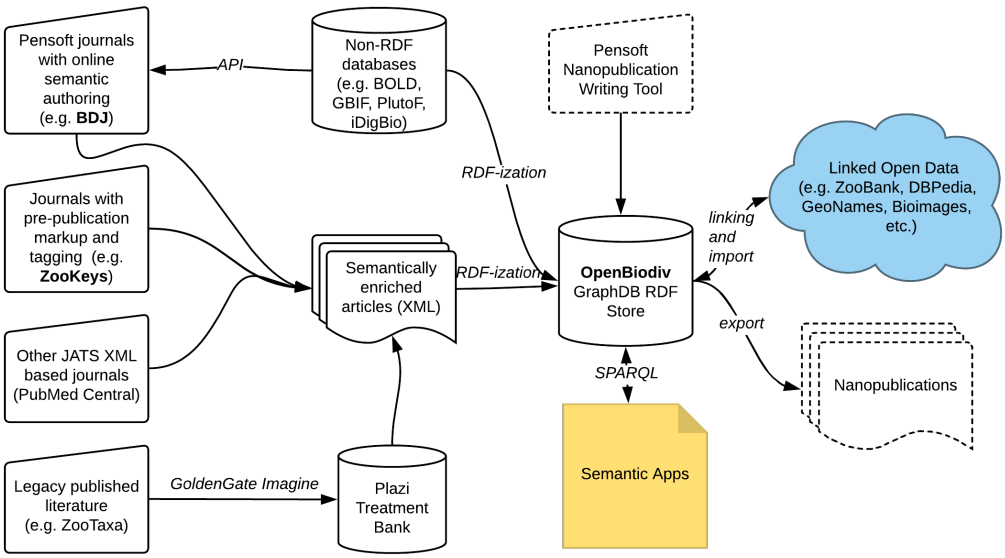
OpenBiodiv се състои от (1) семантична графична база данни, (2) програмен код, осигуряващ функционирането на базата и (3) динамична уеб страница (front-end), улесняваща достъпа до основната база от знания (Фиг. 1.1). OpenBiodiv позволява динамичното вмъкване на данни от хранилища за данни за биологичното разнообразие като BOLD или GBIF. OpenBiodiv извлича от таксономични списания знание (напр. ZooKeys, Biodiversity Data Journal (BDJ), PhytoKeys, MycoKeys и т.н.[1]). В същото време знание под формата на факти (triples) се извлича от Plazi TreatmentBank, архив на литература за биологичното разнообразие, съдържащ над 200 хиляди таксономични дискусии[2] и актуализиран всеки ден. Не на последно място, тези факти са взаимосвързани чрез таксономичния гръбнак на GBIF (GBIF Secretariat, 2017). След това извлеченото знание се съхранява в нашата семантична база данни (Фиг. 1.2).

[1] Списанията могат да бъдат достъпени под https://pensoft.net/browse_journals

[2] Таксономичната дискусия е специален раздел в биологична публикация, която описва и дискутира вид или по-висок таксон. TreatmentBank е достъпен под https:////plazi.org/resources/treatmentbank/



Фигура 1.1: Компоненти на OpenBiodiv.



Фигура 1.2: Поток на информация в пространството за данни за биологичното разнообразие. Пунктираните линии са компоненти, които все още не са създадени.

1.2 Семантична база данни

Основен резултат от усилията на OpenBiodiv е създаването на семантична база данни, базирана на знания, извлечени от архивите на Пенсофт и Плаци и таксономичния гръбнак на GBIF и достъпни под http://graph.openbiodiv.net/. Следва обсъждане на компонентите на базата данни.

Централният резултат от усилията по OpenBiodiv е създаването на формален модел на областта за публикуването на знание за биоразнообразието. Този формален модел е онтологията OpenBiodiv-O (Senderov et al., 2017). Изходният код на онтологията и придружаващата документация могат да бъдат достъпени под https://github.com/pensoft/openbiodiv-o. Детайлна дискусия е представена в глава 2.

Използвайки OpenBiodiv-O и инфраструктурата, описана по-нататък в тази глава, са създадени свързани отворени данни, включващи приблизително 200 хиляди записа от Плаци, пет хиляди статии от Пенсофт, както и таксономичния гръбнак на GBIF (над милион биологични имена). Данните са достъпни онлайн чрез работния инструмент на семантичната база данни http://graph.openbiodiv.net. Разясняват се подробно в глава 3.

1.3 Backend

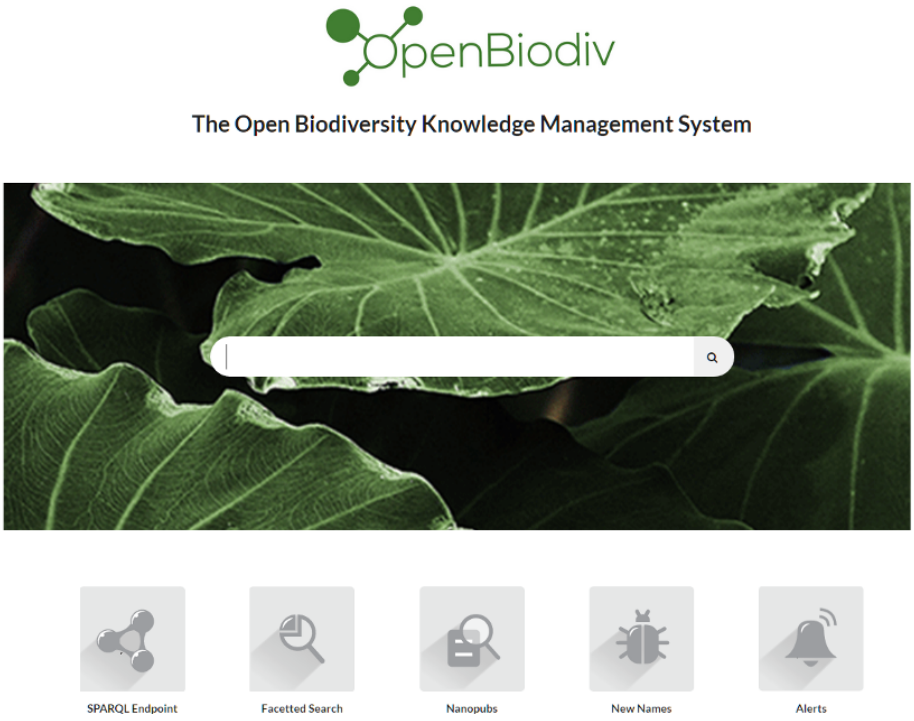
За да се попълва семантичната база данни, е необходимо да се създаде инфраструктура, която преобразува необработени данни (текст, изображения, таблици с данни и т.н.) в структуриран семантичен формат. OpenBiodiv предоставя инфраструктура за трансформиране на научни публикации за биоразнообразието в твърдения под формата на RDF с помощта на инструментите, описани в този раздел.

Едно от по-големите технически предизвикателства за OpenBiodiv е трансформирането на информация за биологичното разнообразие (напр. таксономични имена, метаданни, фигури и т.н.), съхранявани като полу-структуриран XML, в напълно структурирани семантични знания под формата на RDF. За да се реши това предизвикателство, е разработен R пакет, който позволява създаването, манипулирането и записа в семантична база данни на създадения RDF. Този пакет е достъпен под лиценз с отворен код на GitHub под https://github.com/vsenderov/rdf4r и е описан в глава 4.

В комбинация с пакета RDF4R, програмният код съдържа още един R пакет, `ropenbio` и базисен програмен код от скриптове и документация, необходими за стартиране на базата данни. `ropenbio` използва пакета RDF4R за преобразуване на полу-структуриран XML в RDF. Той съдържа преобразуванията, необходими за тази реализация. Той е достъпен под https://github.com/pensoft/ropenbio. Базисният софтуерен код координира извикването на `ropenbio`, съдържа скриптове за автоматично импортиране на нови ресурси и други подробности. Той е достъпен под https://github.com/pensoft/openbiodiv. Генерирането на OpenBiodiv-LOD с помощта на тези пакети е обсъдено в глава 3.

1.4 Работни процеси

Език за екологични метаданни (EML) е популярен формат за описване на екологични данни (Michener et al., 1997). Хранилища на данни за биоразнообразието, като GBIF и DataOne, използват този формат за метаданните, които съхраняват.



Фигура 1.3: Бета версия на потребителския интерфейс.

Автоматичното преобразуване на EML файл в data paper ръкопис от Biodiversity Data Journal[3] е възможно с помощта на системата OpenBiodiv (Senderov et al., 2016). Този работен процес е описан подробно в глава 5.

Един от важните видове данни за биологичното разнообразие са данни за наблюдения на организми, occurrence data. Това са данни, които документират наличието на правилно таксономично идентифициран организъм на дадено място и време. Такива данни се съхраняват в международни хранилища като BOLD, GBIF, PlutoF и iDigBio. За да се улесни публикуването на такъв тип данни, е разработен работен процес за импортиране на такива записи от тези бази данни в таксономична статия (taxonomic paper) в списанието Biodiversity Data Journal (Senderov et al., 2016). Този работен процес е описан подробно в глава 5.

1.5 Интерфейс

В допълнение към предоставения endpoint за база данни с възможност за търсене, се разработва уеб сайт, позволяващ семантично търсене и капсулиращ специфични задачи, пакетирани като приложения (`http://openbiodiv.net`). Бета версията вече е в действие. Фиг. 1.3. Дискусия е представена в глава 6.

[3]Data paper (Chavan and Penev, 2011) е научна статия, обсъждаща научни данни.

1.6 Дискусия

Проектирането на системата е стартирано през втората половина на 2015 г., след разгледане на различни алтернативи за база данни (Neo4J, GraphDB, WikiBase) и различни технологии за RDF-изация на знание. Освен това е направен избор на източници за информация и типове данни. След анализ на основните модели данни и онтологии, е публикувана спецификацията на системата в Senderov and Penev, 2016 като отворен проект за дисертация (PhD project plan). По време на имплементацията се оказва, че първоначалният план се нуждае от доразвиване, за да отговаря на изменящите се изисквания на системата и на новите предизвикателства. По тази причина през втората и третата година се преминава към модела agile, където спецификацията е разбита на по-малки user-stories, които биват реализирани ad-hoc в рамките на едно- или двумесечни спринтове. Визия за бъдещето на системата е работата по нея да се поеме от agile team, който да се възползва от пълния арсенал на методологията.

Глава 2

Онтология OpenBiodiv-O

OpenBiodiv трансформира информация за биоразнообразието от научни публикации и академични бази данни в семантичен вид. В тази глава е представена OpenBiodiv-O (Senderov et al., 2018) – онтологията, формираща модела на OpenBiodiv за знания и извод. OpenBiodiv-O предоставя концептуален модел на структурата на академична публикация в сферата на биоразнообразието и на съдържаните в нея таксономични концепции. За първи път формално е концептуализирана областта на публикуването на знание за биоразнообразието.

Чрез разработването на онтология, насочена към биологичната таксономия са запълнени празнините между онтологии като Darwin-SW и онтологиите за семантично публикуване като онтологиите "SPAR". Счита се, че е предимство да се моделира самият таксономичен процес, а не някакво конкретно състояние на знанието.

Изходният код и документацията са достъпни под лицензa CC BY[1] от GitHub[2]. Следва увод в областта на биологичната таксономия и биоразнообразието.

2.1 Концептуализация на областта

Представена е историята на модерната биологична таксономия, като се започне с Карл Линей (1707-1778), който предлага модерното групиране на организма на *царства, класове, разреди, родове* и използването на латинските биномиални имена в *Systema Naturae* (Linnaeus, 1758). Подчертава се, че работата на ученитетаксономи за описване и организиране на биоразнообразието далеч не е пълна. Това информира създаването на новата онтология не като статично формализиране на съществуващата биологична таксономия в компютърно-четима форма, а като формализиране на *научния процес на биологичната таксономия.*

След това се описва подробно, как протича научният процес в биологичната таксономия. Започва се с въвеждането на таксономични концепции и начина, по който се формират. Таксономичната концепция е научна хипотеза (Deans et al., 2012), че определена група от организми съществува в природата. Тя се формира чрез изследване на екземпляри и задължително включва научен критерий, по който те да се групират, често наричан видова концепция (Mallet, 2001). Исторически погледнато, организмите могат да бъдат групирани по външен вид и вътрешно устройство (концепция за морфологични видове) или репродуктивно поведение (биологична видова концепция), но напоследък фокусът се е насочил към групиране въз основа на генетична свързаност (филогенетични и геномни видови концепции).

[1]Creative Commons Attribution 4.0 International Public License
[2]https://github.com/vsenderov/openbiodiv-o/blob/master/LICENSE.md

След това се описват единиците на биологичната таксономия и начина, по който те са регламентирани от международните кодекси International Commission on Zoological Nomenclature, 1999; *International code of nomenclature for algae, fungi and plants (Melbourne code)* 2012). Кодексите уреждат по-ниските рангове: видове, род, семейство, разред; по-високите рангове (напр. отдел, царство, домейн и т.н.) могат да бъдат използвани от изследователите, както сметнат за подходящо. Това води до множество конкуриращи се гледни точки.

Публикуването на таксономични концепции е неразделна стъпка в научния работен поток на всеки таксоном. Описва се структурата и типовете таксономични публикации, като се акцентира специално върху секцията "Таксономична дискусия", раздела в таксономичната публикация, където таксономичната концепция е дефинирана.

Направена е литературна справка, като се разглеждат опитите областта на биологичната таксономия да бъде формално концептуализирана. Интересни са SPAR Ontologies, Peroni, 2014) и TaxPub XML Document Type Definition (Catapano, 2010).

Концептуализацията основно е повлияна освен от практиката и от кодексите (International Commission on Zoological Nomenclature, 1999; *International code of nomenclature for algae, fungi and plants (Melbourne code)* 2012), а така и от стандартите, създадени от TDWG (напр. Darwin Core, DwC, Wieczorek et al., 2012).

Направен е и обзор на областта на концептуалната таксономия (Berendsohn, 1995; Franz and Peet, 2009; Sterner and Franz, 2017), която представлява нова гледна точка за това, как трябва да протича процесът на описване на видове в биологичната таксономия, с оглед на напредъка на информационните технологии.

2.2 Методи

OpenBiodiv-O е изразена посредством RDF чрез използване на RDF Schema (RDFS) и Web Ontology Language (OWL).

За да се разработи онтологията, е използван следният процес: (1) анализ на областта и идентифициране на важните класове обекти и техните взаимоотношения (наричани свойства); (2) анализ на съществуващите информационни модели и онтологии и идентифициране на липсващите класове и свойства за успешно формализиране на областта.

2.3 Резултати

OpenBiodiv-O е *споделена формална спецификация на концептуализация* на областта на биоразнообразието по смисъла на Gruber, 1993; Obitko, 2007; Staab and Studer, 2009. Тяхното разбиране за онтология е въведено в Background.

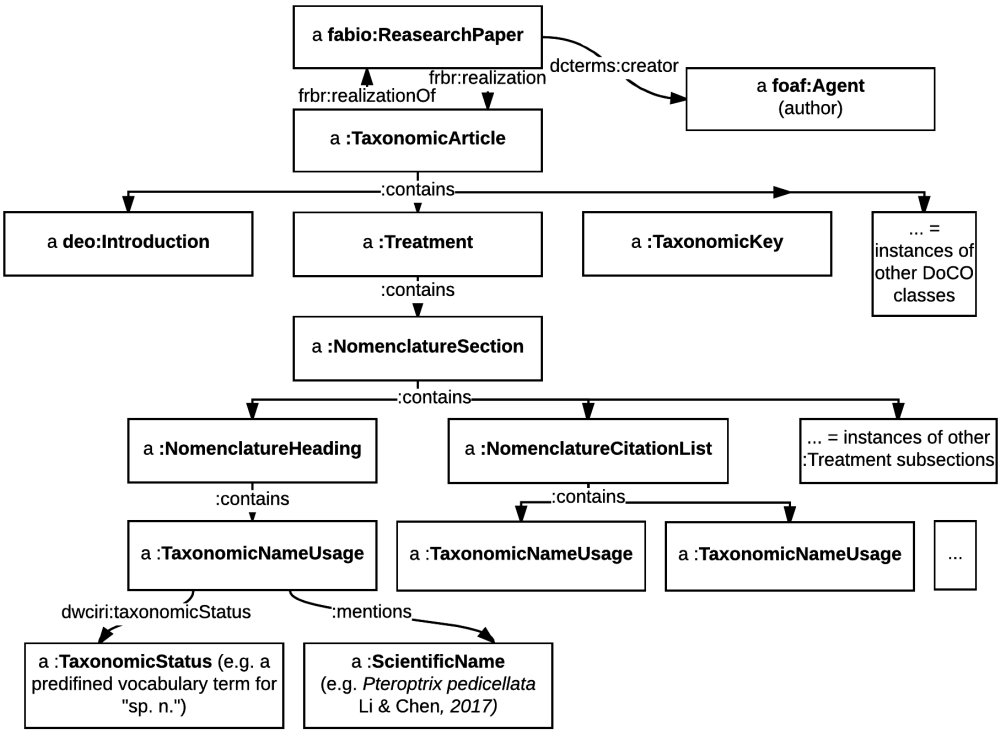
Има няколко домейна, в които моделираните ресурси падат. Първият е научната област за публикуване на данни за биоразнообразието. Втората област е тази на таксономичната номенклатура. Третата област е на по-широката таксономия (например таксономични понятия и техните взаимоотношения, видови събития, черти).

Публикуване на данни. Рамката на онтологиите "SPAR" е разширена посредством като въвеждането на нова категория за таксономичните статии, техните подраздели, както и нов клас за споменаване на таксономично име (вж. следващ подраздел). Тези нови класове са обобщени в Таблица 2.1.

Таблица 2.1: Нови класове в областта на публикуването на данни за биоразнообразието.

Class QName	Comment
`:Treatment`	section of a taxonomic article
`:NomenclatureSection`	subsection of Treatment
`:NomenclatureHeading`	contains a nomenclatural act
`:NomenclatureCitationList`	list of citations of related concepts
`:MaterialsExamined`	list of examined specimens
`:BiologySection`	subsection of Treatment
`:DescriptionSection`	subsection of Treatment
`:TaxonomicKey`	section with an identification key
`:TaxonomicChecklist`	section with a list of taxa for a region
`:TaxonomicNameUsage`	mention of a taxonomic name

Графичното представяне на взаимоотношенията между ресурси на класовете, свързани с публикуването, които OpenBiodiv въвежда, може да се намери в диаграмата на фиг. 2.1.



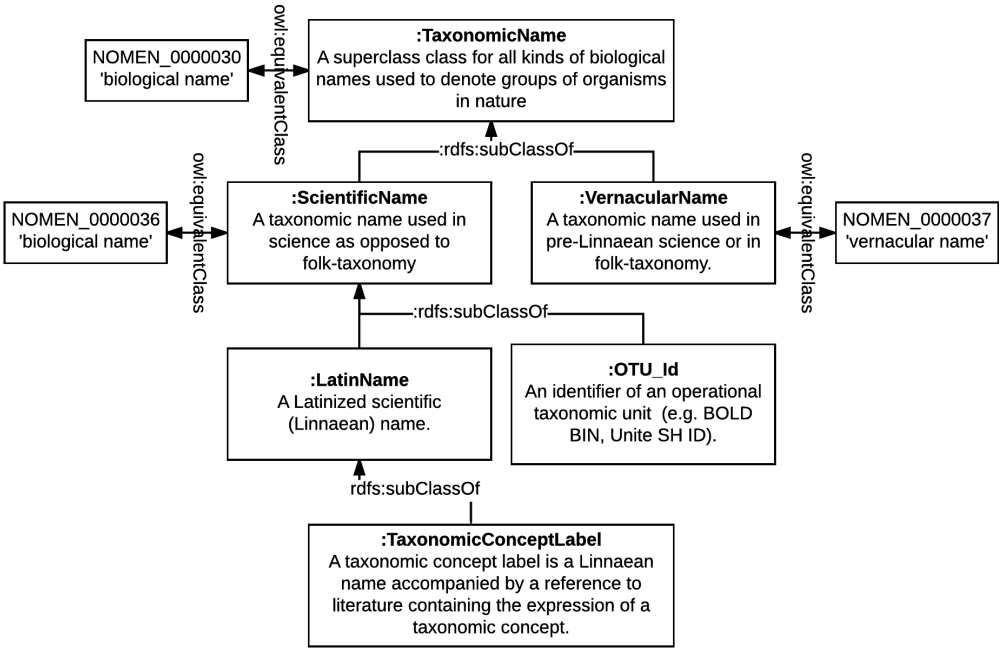
Фигура 2.1: Графична представяне на взаимоотношенията между ресурсите, които OpenBiodiv въвежда за публикуване на данни за биоразнообразието.

Класовете и свойствата, които се въвеждат, са в съответствие с модела на функционалните изисквания за библиографски записи (FRBR), използван от SPAR. Таксономичната статия се третира като конкретен израз/запис, FRBR Expression, на абстрактното понятие работа, FRBR Work, представляващо интелектуалното

съдържание на статията. Таксономичната дискусия се третира подобно на Въведение, Методи, Резултати и т.н., т.е. също е FRBR Expression и DEO discourse element. Таксономична концепция е съответният абстрактен ресурс от клас FRBR Work на дадена таксономична дискусия.

Биологичната номенклатура. Биологичната номенклатура е система с над 200 годишна традиция датираща до преди времето на информатиката и дори до преди времето на Теорията за еволюцията на Дарвин. Много е трудно да се моделира поради сложността си и само частично е обхванатата от онтологиите NOMEN и TNSS (въведени в подраздел “Previous Work”). С OpenBiodiv-O се използва подход "отдолу-нагоре"за моделиране на използването на таксономични имена в статиите. Където е възможно, понятията от OpenBiodiv-O са подравнени към NOMEN.

Йерархия от класовете на таксономичните имена е поместена вув фиг. 2.2. Освен това се въвежда taxonomic name usage (`TaxonomicNameUsage`).



Фигура 2.2: Йерархията, включваща както традиционните таксономични наименования, така и таксономични концептуални етикети и оперативни таксономични единици.

Въвежда се Taxonomic Concept Label (`TaxonomicConceptLabel`). Етикетът на таксономична концепция (TCL) е линеево име плюс позоваване на публикация, където дискутираният таксон е дефиниран. Връзката се осъществява чрез ключовата дума “sec.” (Латински за (*secundum* Berendsohn, 1995). Напр. *Andropogon virginicus* var. *tenuispatheus* Blomquist, 1948. Тук Blomquist, 1948 е валидна библиографска справка за публикацията, в която концепцията е дефинирана.

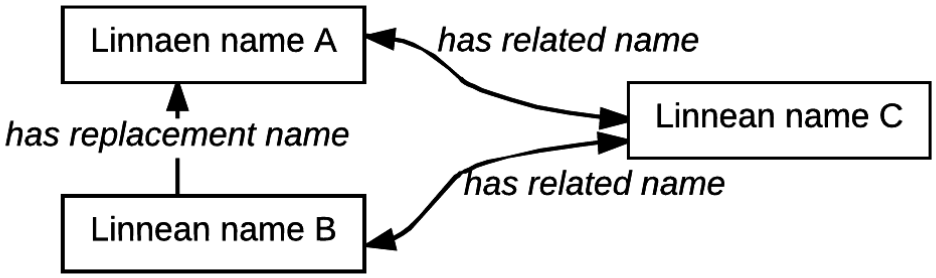
Извадени са съкращения от таксономични термини от около 4000 статии в четири таксономични списания (ZooKeys, Biodiversity Data Journal, PhytoKeys и MycoKeys) и е създаден таксономичен речник на термините, който обхваща осемте най-често срещани случая Таблица 2.2). Латинските съкращения, които

са класифицирани в тези класове, могат да бъдат намерени на страницата на OpenBiodiv-O GitHub. (вж. Методи за повече подробности).

Таблица 2.2: Речник с таксономични термини.

Vocabulary Instance QName	Example Abbrev	Comment
`:TaxonomicUncertainty`	*incertae sedis*	Taxonomic Uncertainty
`:TaxonDiscovery`	*sp. n.*	Taxonomic Discovery
`:ReplacementName`	*comb. n.*	Replacement Name
`:UnavailableName`	*nomen dubium*	Unavailable Name
`:AvailableName`	*stat. rev.*	Available Name
`:TypeSpecimenDesignation`	*lectotype designation*	Type Specimen Designation
`:TypeSpeciesDesignation`	*type species*	Type Species Designation
`:NewOccurrenceRecord`	*new country record*	New Occurrence Record (for region)

Въз основа на анализ на термините за таксономични статуси, са идентифицирани два модела за съответствие между кодирането на латинизираните научни имена (Фиг. 2.3). Моделът *заместващо име*, изпълняван чрез свойството `replacementName`, показва, че вместо едно латинизирано име, трябва да се използва друго. Този модел обхваща голямо разнообразие от случаи в кодексите, като например поставянето на един вид таксон в нов род (нова комбинация), поправката на наименование поради номенклатурни причини (ново име), или прилагането на Принципа на приоритет за откриването на синоними ("syn nov. International Commission on Zoological Nomenclature, 2017).

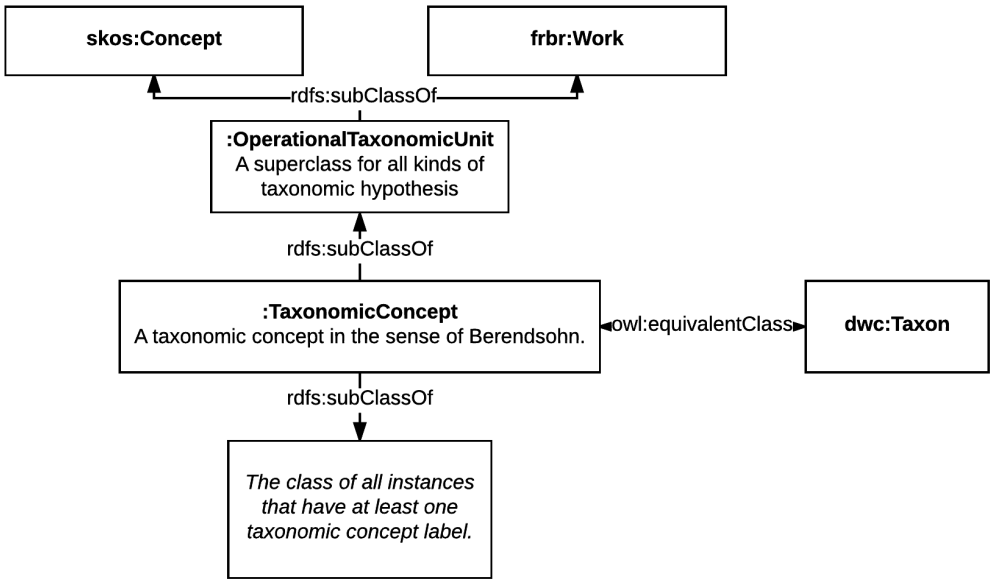


Фигура 2.3: Веригите на *заместващи имена* могат да бъдат проследени, за да се намери използваното понастоящем име. *Свързано име* показва, че две имена са свързани по някакъв начин, но не кой е предпочитан.

Другият модел е този на *свързани имена* (`relatedName`). Това е по-широк модел, който показва, че две имена са някак си свързани. Например, те могат да бъдат синоними, едното да замества другото, или да сочат към таксономично свързани таксономични понятия. Например, *Harmonia manillana* (Mulsant, 1866) е свързано с *Caria manillana* Mulsant, 1866 *Harmonia manillana* (според Poorani and Booth, 2016 лектотипус на *Harmonia manillana* (Mulsant, 1866) sec. Poorani and Booth, 2016 носи името *Caria manillana* Mulsant, 1866).

Както се вижда от фиг. 2.2, таксономичните имена на OpenBiodiv-O са приравнени към NOMEN имена.

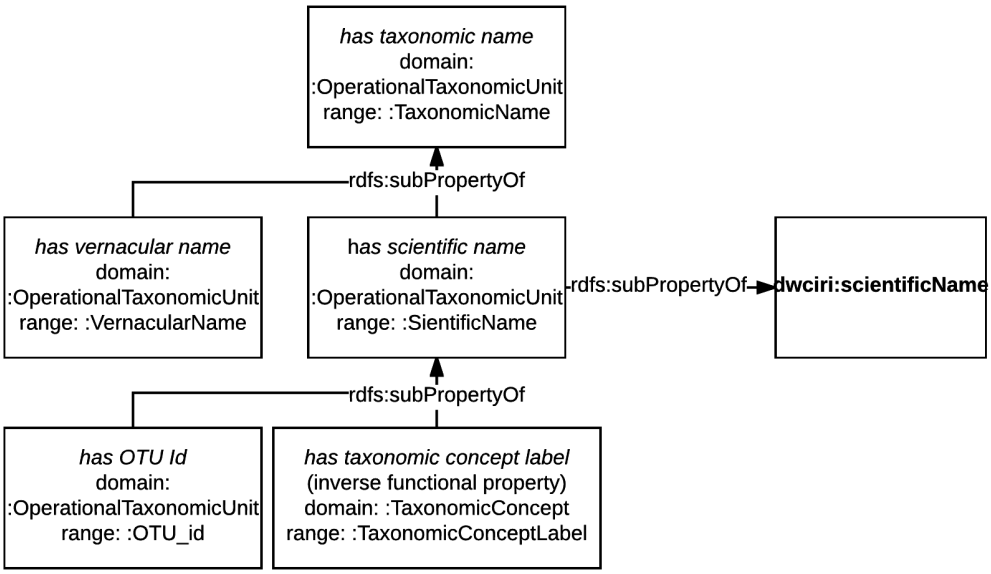
В OpenBiodiv, таксономичните имена не са носители на семантична информация за таксоните. Тази задача се изпълнява от нов клас, Taxonomic Concept (`TaxonomicConcept`). Таксономична концепция е теорията, която таксономът формира около таксон посредством научна таксономична публикация. Тя винаги има етикет, състоящ се от таксономичното име и библиографско позоваване на статията, в която името е описано. Въвежда се и по-общ клас, оперативно таксономично звено (`OperationalTaxonomicUnit`), което може да се използва за всички видове таксономични хипотези, включително такива, които нямат правилен таксономичен етикет. Класовата йерархия е илюстрирана на Фиг. 2.4. Свойствата на таксономичните имена са илюстрирани в Фиг. 2.5. Двата начина са изразяване на взаимоотношения между таксономични концепции са дадени във Фиг. 2.6).



Фигура 2.4: Таксономичната концепция е от клас `skos:Concept`, `frbr: Work`, `dwc:Taxon` и има поне един етикет.

Простите отношения между таксони не са подходящи за машинен извод. Ето защо Franz and Peet, 2009, позовавайки се на Koperski et al., 2000 предлагат да се използва езикът RCC-5, за изразяване на взаимоотношенията между таксономичните понятия. От техни колеги е разработена програмата Euler (Chen et al., 2014), която използва Answer Set Programming (ASP) за разсъждения и извод по таксономичните взаимоотношения, стъпвайки на RCC-5. Изводната машина за RCC-5 не е част от OpenBiodiv, тъй като тази задача може да бъде изпълнена от euler; въпреки това, ние OpenBiodiv-O предоставя RCC-5 терминологичен речник.

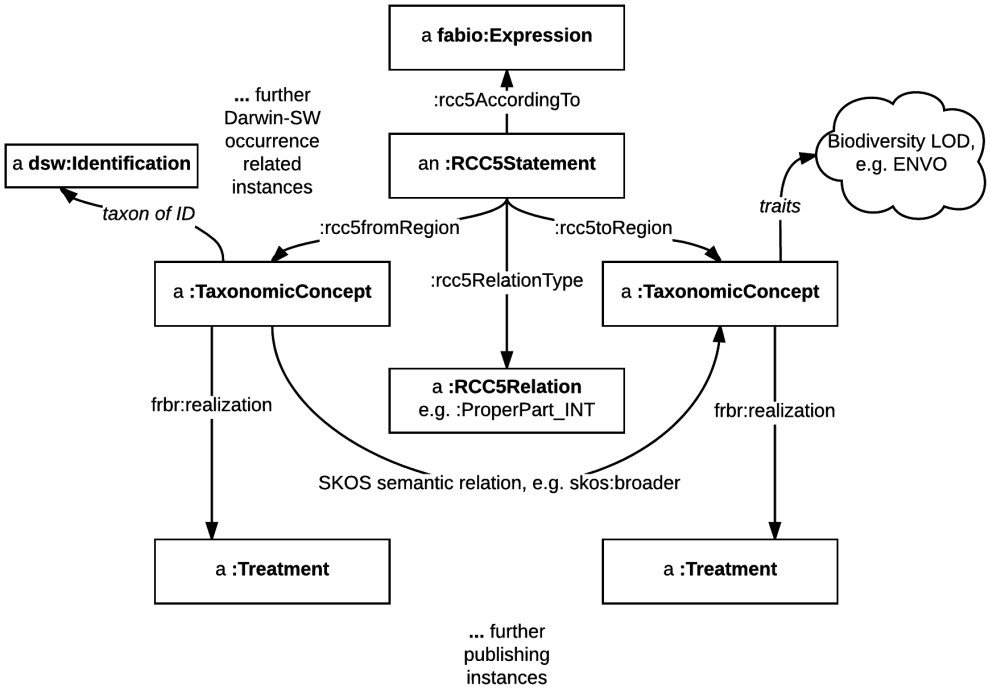
Таксономичните понятия са приведени в съответствие с DarwinCore (DwC) и е проведена дискусия за това как са представени таксономичните понятия, свързани както с прости отношения (SKOS), така и в подходящ за извод вид (RCC-5). Също така се обсъждат взаимоотношенията между биологичните имена и таксономичните концепции. Обсъжда се как OpenBiodiv-O е първият по рода онтологичен модел на таксономична публикация. Тя проправя пътя към създаване на граф от знания за биологичното разнообразие.



Фигура 2.5: Йерархията на свойствата съвпада с йерархията на таксономичните имена и е приравнена към DarwinCore.

2.4 Заключение

Главата предоставя концептуализация на таксономичния процес и формализация в OpenBiodiv-O. Въвеждат се класове и свойства в областта на публикуването на биоразнообразие и биологичната систематика и ги привежда в съответствие с важните онтологии, специфични за този домейн.



Фигура 2.6: За да изразите RCC-5 връзка между концепциите, създайте `RCC5Statement` и използвайте съответните свойства, за да свържете две таксономични концепции през него. Освен това, таксономичните понятия са свързани със качества (например екология в ENVO), събития (например Darwin-SW) и са абстрактният клас на Таксономичните дискусии.

Глава 3

Свързани отворени данни

Създадени са свързани отворени данни OpenBiodiv-LOD, съдържащи информация за биологичното разнообразие, извлечена от списанията на Пенсофт и базата данни на Плаци. Данните са интегрирани с помощта на таксономичния гръбнак на GBIF. Като онтология се използва новата OpenBiodiv-O, разработен в хода на дисертацията. Предлага се на общността на информатиката за биоразнообразието да използва OpenBiodiv-LOD като централна точка на семантичния граф за знания за биоразнообразието. OpenBiodiv-LOD са достъпни под http://graph.openbiodiv.net.

OpenBiodiv-LOD е синтетичен набор от данни. Той не съдържа предварително непубликувани данни. Вместо това той интегрира в една база от знания информация, която преди това е била оповестена в академични списания и бази данни. Интеграцията позволява материализация на скрити взаимоотношения. В следващите няколко параграфа ще обсъдим източниците на информация, които са комбинирани в OpenBiodiv-LOD и видовете ресурси, които са извлечени, както и общия модел на данните. Също така се обсъждат принципите на Linked Open Data, които свързват всичко заедно. Главата завършва с много примери за заявки в набора от данни и с техническа дискусия за начина, по който тя е генерирана.

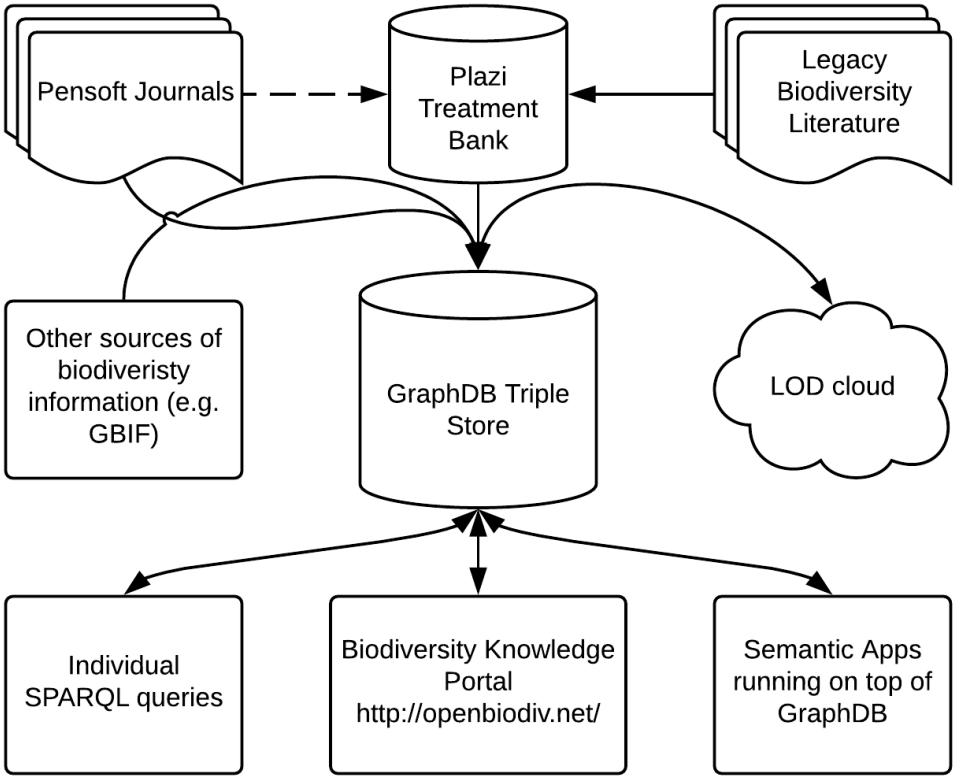
3.1 Източници на данни

Данните в OpenBiodiv към времето на писането на дисертацията идват от три основни източника: таксономичния гръбнак на GBIF (GBIF Secretariat, 2017), и научни статии публикувани от Пенсофт и Плаци (Fig. 3.1).

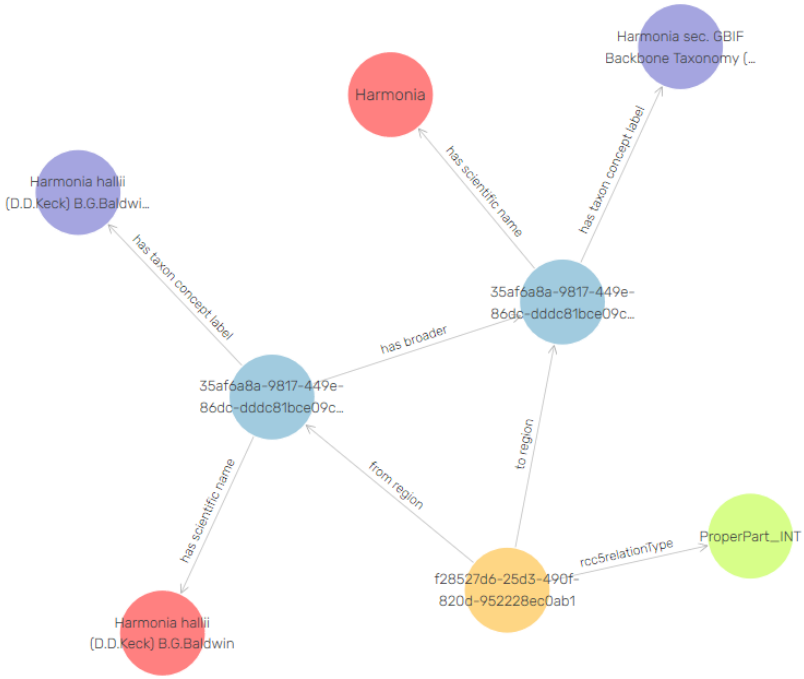
GBIF е най-голямото международно хранилище на данни за наблюдения на организми (occurrence data). GBIF позволява на потребителите да търсят в тяхната система, като използват таксономичната йерархия, Nub (GBIF Secretariat, 2017). Nub е база данни, която организира таксономични концепции в йерархия, обхващаща всички биологични имена, събрани от GBIF. Тя е синтетична, т.е. алгоритмично генерирана класификация. По този начин гръбнакът GBIF *не представлява* експертен консенсус за това как таксоните са йерархично подредени според еволюционните критерии в природата, но въпреки това е много полезен на практика.

За да псе предоставят същите възможности на OpenBiodiv, концепциите от Nub са импортирани като `openbiodiv:TaxonomicConcept` по OpenBiodiv-O (фиг. 3.2).

Всички валидни статии от таксономичните списания, публикувани от Пенсофт и упоменати в Таблица 3.1 бяха конвертирани в RDF и съхранени в графа от знания на биологичното разнообразие. В допълнение, всички валидни таксономични дискусии (treatments) на Плаци бяха конвертирани в RDF и също съхранени в графа. Процедурата по RDF-изиране се повтаря всяка седмица и



Фигура 3.1: Опростен модел на архитектурата на OpenBiodiv от глава 1 фокусираща върху източниците на информация.



Фигура 3.2: Илюстрация на представянето на йерархична информация, импортирана от GBIF като таксономични концепции в OpenBiodiv.

по този начин семантичната база данни винаги съдържа най-новите статии и таксономични дискусии. RDF-изацията е възможна благодарение на факта, че списанията на Пенсофт публикуват статии в TaxPub XML (Catapano, 2010), докато Плаци публикува дискусиите си в TaxonX (Penev et al., 2011) (Fig. 3.3). И двете схеми са стандартни и общо-достъпни.

Таблица 3.1: Списания ма Пенсофт, които са превърнати в RDF.

Journal Name	Submission Style	Number of Articles
ZooKeys	Word document	3829
PhytoKeys	Word document	537
MycoKeys	Word document	127
Biodiversity Data Journal	Web based (ARPHA)	490
Journal of Orthoptera Research	Word document	32

Таблица 3.2: Типове данни, маркирани в TaxPub and TaxonX и тяхната кореспонденция към RDF типовете, които се използват в OpenBiodiv.

Datatype	TaxPub	TaxonX	RDF Type
Article metadata	T	T	`fabio:JournalArticle` and related
Keyword group	T	F	`openbiodiv:KeywordGroup`
Abstract	T	T	`sro:Abstract`
Title	T	F	`doco:Title`
Author	T	T	`foaf:Person`
Introduction section	T	F	`deo:Introduction`
Discussion section	T	T	`orb:Discussion`
Treatment section	T	T	`openbiodiv:Treatment`
Nomenclature section	T	T	`openbiodiv:NomenclatureSection`
Materials examined	T	T	`openbiodiv:MaterialsExamined`
Diagnosis section	T	T	`openbiodiv:DiagnosisSection`
Distribution section	T	T	`openbiodiv:DistributionSection`
Taxonomic key	T	T	`openbiodiv:TaxonomicKey`
Figure	T	T	`doco:Figure`
Taxonomic name usage	T	T	`openbiodiv:TaxonomicNameUsage`

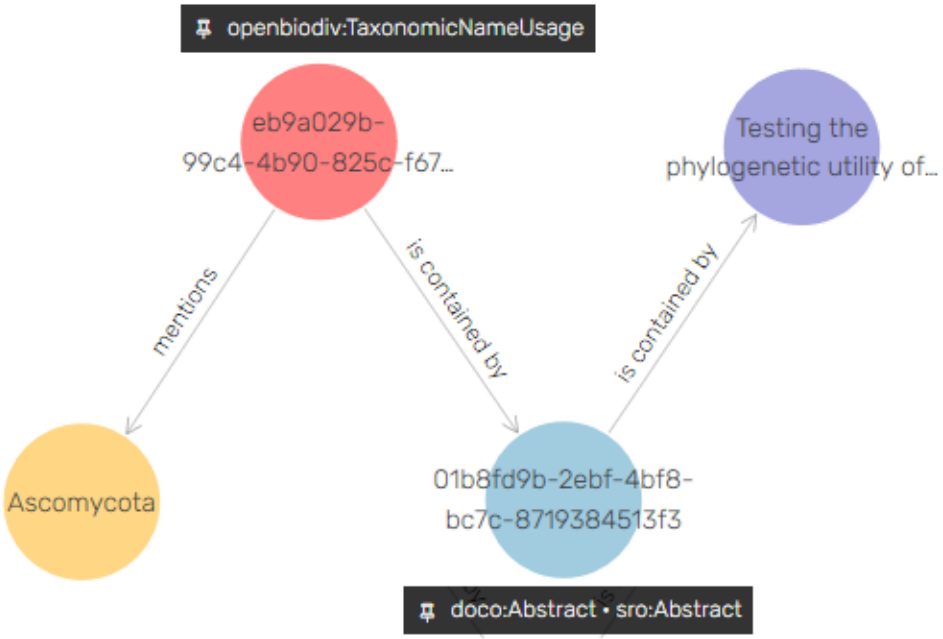
3.2 Свързани отворени данни

Linked Open Data (LOD, Heath and Bizer, 2011) е идея на Семантичната Мрежа (Berners-Lee et al., 2001), целяща да гарантира полезността на данни, публикувани в Мрежата, като улеснява тяхното повторно намиране и употреба от трети лица. В тази секция са разяснени принципите на Свързаните Отворени Данни и тяхното приложение в OpenBiodiv (Heath and Bizer, 2011).

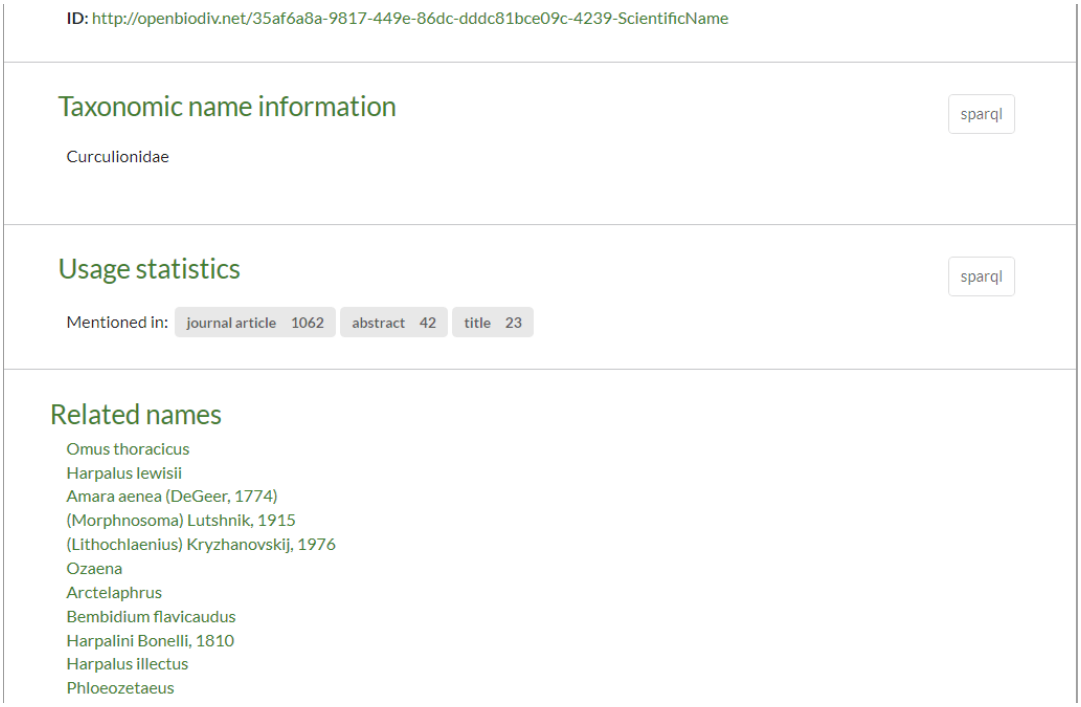
Връзка между модела OpenBiodiv-O, подробно разяснен в глава 2, и допълнителните онтологии е дадена във Fig. 3.5.

3.3 Примерен SPARQL

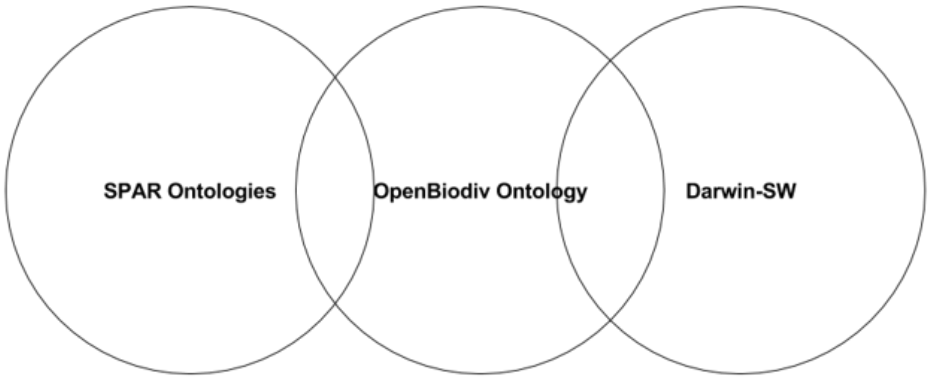
В тази секция е илюстрирано, как представеният модел за данни, заедно с инкорпорираната в базата данни информация е достатъчен за да се изпълнят някои



Фигура 3.3: Употреба на таксономични имена в OpenBiodiv (taxonomic name usage).



Фигура 3.4: Визуализация.



Work	Treatment	Identificiation
Expression	Nomenclature	Organism
Journal	Nomenclature Heading	Token
Journal Article	Taxonomic Name	Occurrence
Research Paper	Latin Name	Event
FrontMatter	Taxonomic Concept Label	Agent
Body Matter	Taxonomic Concept	Location
Back Matter	Taxonomic Status	
Section	Vocabulary	
Table	RCC5 Vocabulary	
Figure		

Фигура 3.5: Връзка между OpenBiodiv-O и други публично-достъпни онтологии.

много интересни заявки върху графа за знания.

Ето кратък списък с типовете заявки, които са илюстрирани:

1. Заявка върху структурата на статията: Тъй като в OpenBiodiv LOD статиите са разбити по техните компоненти (вж напр. Таблица 3.2) и споменаванията на таксони (taxonomic name usages) винаги са свързани със специфична част на статията, могат да се правят запитвания, използващи тази структура.

2. Запитване за таксономични концепции

3. Размито търсене с Lucene: използван е Lucene connector (Ontotext, 2018) на GraphDB, за размити търсения, където се допускат правописни грешки.

4. Отговаряне на експертни въпроси: поради машината си за извод, OpenBiodiv може да функционира като експертна система. Напр. за оценка на научната загуба след пожара в Museu Nacional в Бразилия, системата може да бъде запитана относно списък с видовете, чиито екземпляри бяха загубени в пожара в Рио де Жанейро.

5. Проверка на валидността на таксономично име: проверк, дали дадено таксономично име е валидно, т.е. дали е уместно да се употребява в научен контекст, или е било премахнато или заменено от друго име.

3.4 Генериране на данните

В тази секция се разкриват техническите подробности за това, как данните бяха създадени. Използват се следните инструменти

1. Пакет RDF4R

2. Пакет ROpenBio

3. Пакет TSV4RDF разработен от Пенсофт

4. Базисен код на OpenBiodiv

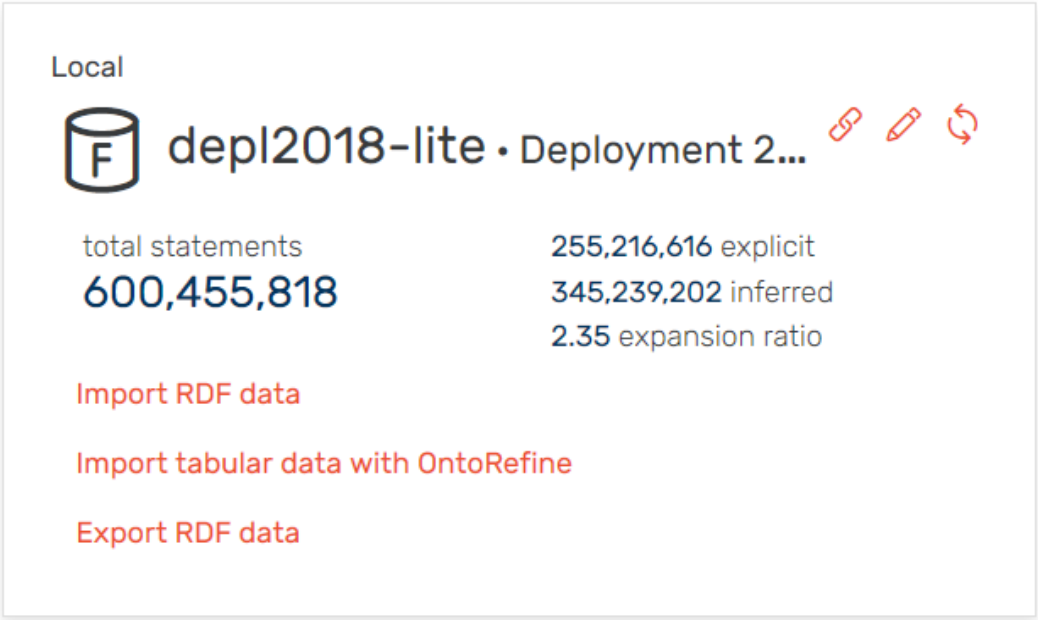
3.4.1 Трансформация от XML в RDF

За трансформация от XML в RDF е използвана йерархичната структура на XML, за да се реши проблемът рекурсивно. Решението е реализирано посредством Алгоритъм 1.

Algorithm 1 Екстрактор

```
1: procedure EXTRACTOR(XML Node X)
2:     a ← extract atoms of X                    ▷ Atoms extraction
3:     r ← construct RDF from a                  ▷ RDF construction
4:     C ← find relevant sub-nodes of X          ▷ Recursively applies itself
5:     R ← apply Extractor on each C_i ∈ C
6:     return r ⋃ R
7: end procedure
```

Текстовите полета в XML са наричени атоми. Тяхното извличане се осъществява с помощта на езика XPATH. След извличане на атомите, те могат да бъдат



Фигура 3.6: Брой твърдения.

събрани обратно във формата на RDF. Проблемът за RDF-изирането на цял XML файл е реализиран посредством разбиването му на малки проблеми, за които решението е тривиално.

За да работи Extractor, трябва да се дефинира схемата, в която XML е кодиран. Спецификацията включва, какви XML възли се търсят и къде се намират. След това рекурсивно се указва за всеки възел, кои под-възли се търсят и тяхното местоположение по XPATH спрямо техния родителски възел. И накрая, за всеки възел трябва да се укажат адресите на атомите и да се напише конструктор, който ги сглобява.

Спецификацията се извършва в R6 клас на R. Дефинираме и схемите на Пенсофт и Плаци: TaxPub[1] и TaxonX[2].

Извлечените RDF твърдения се изпращат в GraphDB, който се намира на адрес `http://graph.openbiodiv.net/`. OpenBiodiv-O автоматично материализира допълнителни твърдения посредством машината за извод. Допълнително се изпълняват следното допълнително правило:

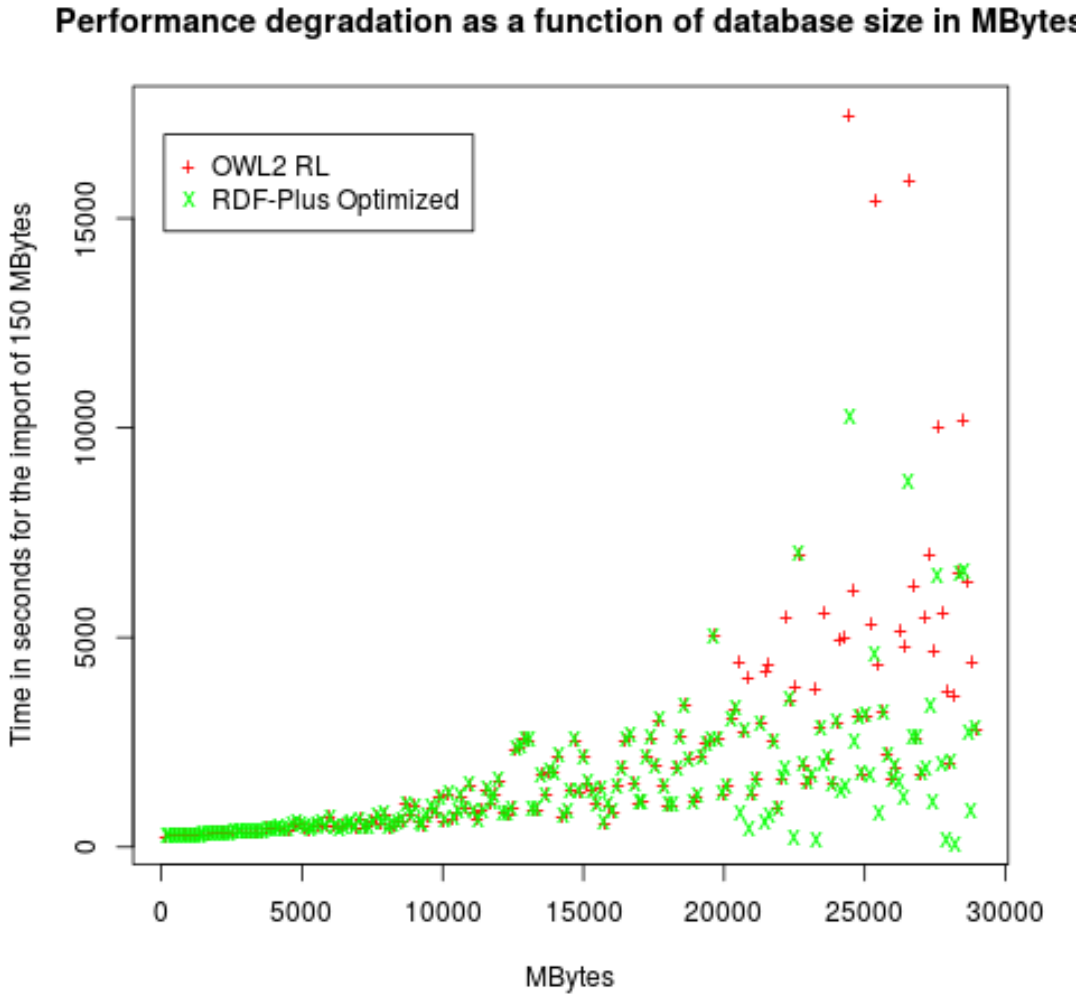
Ъпдейт на заместващите имена: изгражда връзките от тип `replacementName` между вмъкнатите имена.

3.5 Дискусия

Установено е, че обемът данни, който се обработва, в комбинация с пълна OWL логика води до неприемлива производителност. В тази секция се разглежда този проблем (Фиг. 3.6).

Направени са някои основни изводи и са очертаваме насоки за бъдещо развитие.

[1] `https://github.com/pensoft/ropenbio/blob/redesign/R/taxpub.R`
[2] `https://github.com/pensoft/ropenbio/blob/redesign/R/taxonx.R`



Фигура 3.7: Визуализация на деградацията на производителността.

Глава 4

Библиотека за R за работа с RDF

RDF4R (`rdf4r`) е R пакет за работа с RDF. Тя е разработена като част от проекта OpenBiodiv, но е напълно свободна от специфичен за OpenBiodiv код и може да се използва за общи цели, изискващи инструменти за работа с RDF данни в средата за програмиране R (R Core Team, 2016).

В тази глава се разяснява функционалността на пакета.

Основните функции на пакета са:

1. Връзка със семантична база данни
2. Функции за преобразуване на SPARQL заявки в R функции
3. Работа с литерали и идентификатори
4. Работа с префикси
5. Създаване и сериализация на RDF
6. Основна терминология на семантични понятия

Дискутирани са приликите и разликите на RDF4R с подобни пакети като `rdflib`. Обсъждат се, как са използвани модели от функционалното програмиране за създаване на RDF4R. Освен това се обсъждат, как се използват модели от обектно-ориентирано програмиране за създаване на RDF4R.

Глава 5

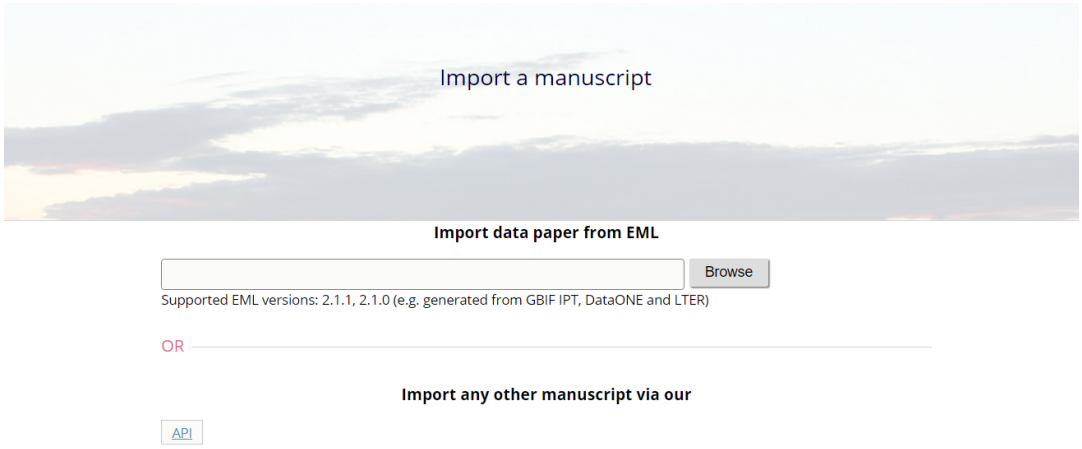
Работни процеси

В тази глава се обсъждат два автоматизирани работни процеса за обмен на данни за биологичното разнообразие, разработени като част от OpenBiodiv: (1) автоматично внасяне на записи за наблюдения на видове (occurrence data) в ръкописи и (2) автоматично генериране на ръкописни материали от екологичните метаданни (EML). Работните потоци са представени на webinar за организацията iDigBio и публикувани като научна статия (Senderov et al., 2016). Слайдовете от презентацията под github.com/vsenderov/idigbio-webinar.

Работен процес 1: Автоматизиран импорт на occurrence data в ръкописи, разработени в ARPHA Writing Tool Разработен е автоматичен импорт на occurrence data в таксономична статия в системата ARPHA от международните бази данни GBIF, BOLD, iDigBio и PlutoF (Фиг. 5.1).

Работен процес 2: Автоматизирано генериране на ръкопис от EML метаданни в ARPHA Writing Tool Създаден е работен процес, който позволява на авторите да създават автоматично ръкописни материали от метаданни, съхранявани в EML (фиг. 5.2).

Фигура 5.1: Работният поток от информационните портали GBIF, BOLD Systems, iDigBio и PlutoF чрез елементи на потребителски интерфейс в AWT.



Фигура 5.2: Полето за потребителски интерфейс за качване на EML файлове в ARPHA.

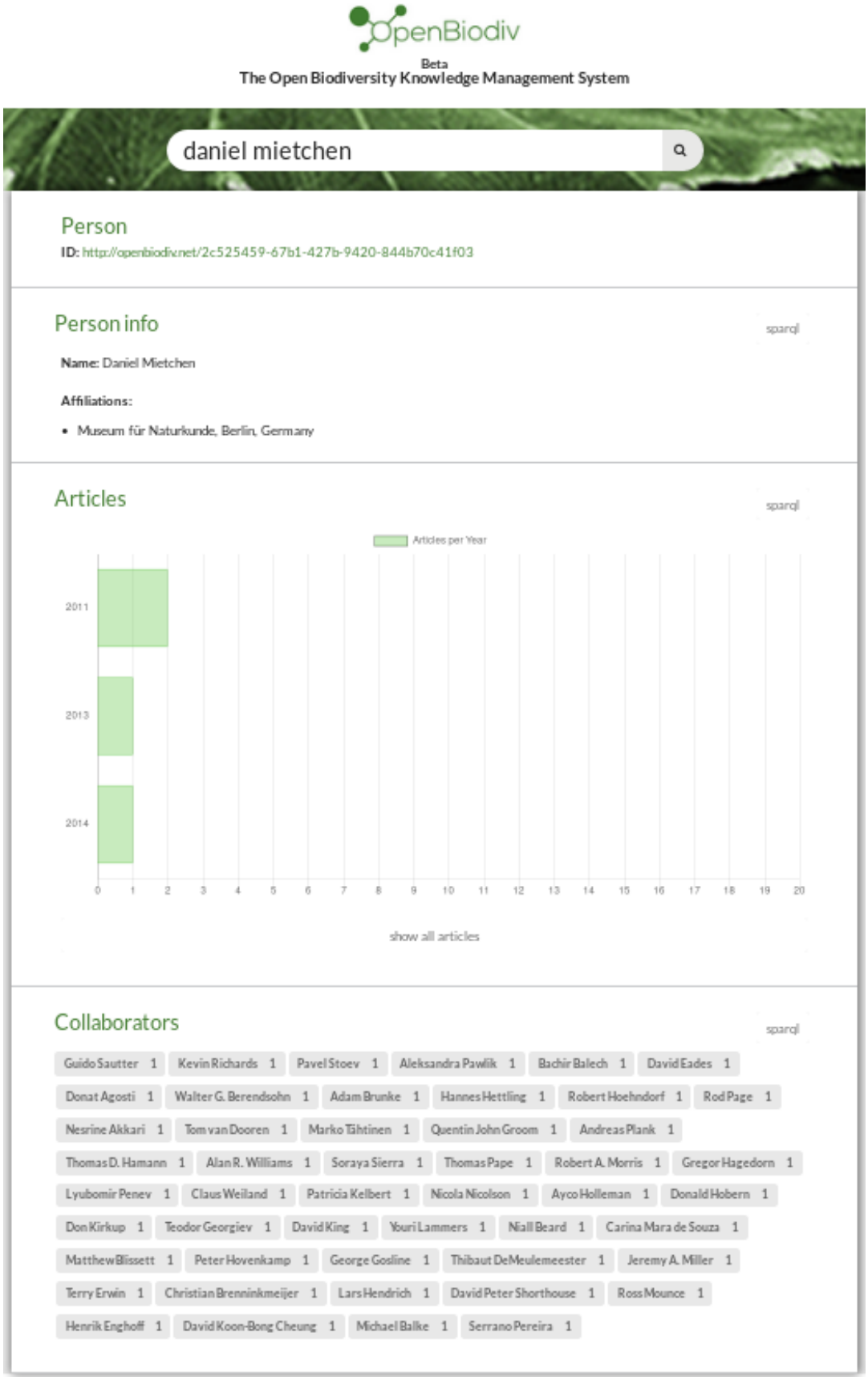
Глава 6

Уеб портал

Под openbiodiv.net може да се стигне до основния портал, който дава достъп до ресурсите на OpenBiodiv. Този портал е разработен от Пенсофт в подкрепа на OpenBiodiv и представя два визуални елемента на потребителя: лентата за търсене и списък с икони на приложения в долната част. Освен това, под graph.openbiodiv.net (също достъпен от иконката SPARQL) може да се достигне работния плот OpenBiodiv за на SPARQL.

Тези функции на потребителския интерфейс (UI) са предназначени да улеснят трите типа потребители на системата:

1. Основно ниво: използва лентата за търсене.
2. Ниво на специалист: използва приложения.
3. Power user: използва работната плот за SPARQL или R.



Фигура 6.1: Илюстрация на основната употреба на OpenBiodiv за търсене на информация за човек.

Заключение

Резултати

Считаме, че представеният научен труд изпълнява поставената цел и задачи.

Резултат 1. Основният резултат на дисертацията е създаването на концептуален модел на областта на публикуването на знание за биологичното разнообразие и логическото му формализиране му под формата на онтологията OpenBiodiv-O. Този резултат е подробно дискутиран в глава 2 и в Senderov et al., 2018 и изпълнява Задача 1. Изходният код на онтологията е достъпен под github.com/pensoft/openbiodiv-o.

Резултат 2. Създаването на софтуерната архитектура на OpenBiodiv, очертана в глава 1 и Senderov and Penev, 2016. Този резултат отговаря на Задача 2.

Резултат 3. Създаването на свързани отворени данни, OpenBiodiv-LOD, състоящи се от трансформирани в RDF и интегрирани в база данни твърдения за биоразнообразието. Твърденията са извлечени от XML на научни статии, публикувани от Пенсофт, XML таксономични дискусии, публикувани от Плаци, и CSV-дъмп на таксономичния гръбнак на GBIF. OpenBiodiv-LOD са достъпни под `graph.openbiodiv.net` и са описани в глава 3. Този резултат отговаря на Задача 3.

Резултат 4. Софтуерен пакет за средата за програмиране R, RDF4R. RDF4R дава възможност за манипулиране на RDF в R. С негова помощ е осъществено трансформирането на научни публикации от полу-структуриран XML формат в структуриран семантичен RDF. Този резултат е обсъден в глава 4 и Senderov et al., 2016 и изпълнява Задача 4. Пакетът е достъпен онлайн като свободен софтуер под github.com/pensoft/rdf4r. Освен това допълнителен изходен код (неоптимизиран), описващ XML схемите на Пенсофт и Плаци и работещ в тандем с RDF4R за конвертиране на XML в RDF може да бъде намерен под github.com/pensoft/ropenbio.

Резултат 5. Процеси за работа, които позволяват обогатяването на XML статии, публикувани от Пенсофт, чрез автоматично импортиране на данни от основните международни хранилища за данни за биологичното разнообразие: BOLD, GBIF, iDigBio , както и PlutoF. Освен това, благодарение на резултатите на тази дисертация е възможно автоматично да се създадат ръкописи от метаданни, кодирани в Ecological Metadata Language (EML). Обсъждането на тези автоматизирани работни процеси – автоматично генериране на data papers и автоматичен импорт на occurrence records - се извършва в глава 5. Този резултат изпълнява Задача 5.

Резултат 6. Уеб-сайт openbiodiv.net, съдържащ семантична търсачка и приложения. Уеб-сайтът е обсъден в глава 6 и изпълнява Задача 6.

Дискусия, изводи и бъдещи насоки

OpenBiodiv-LOD са свързани отворени данни за биоразнообразието на основата на научни публикации. OpenBiodiv-O е базата на свързаните отворени данни OpenBiodiv-LOD. Чрез разработването на онтология, насочена към биологичната таксономия, са запълнени празнините между онтологиите за ресурси от биологичното разнообразие като Darwin-SW и онтологиите за семантично публикуване като SPAR. Счита се, че е правилно да се моделира самият таксономичен процес, а не някакво конкретно състояние на знанието, тъй като процесът по описването на биоразнообразието на земята далеч не е завършил. На дадения етап, покритието на онтологията е достатъчно, за да може тя да бъде основата за създаването на LOD. В този смисъл онтологията е завършена. От друга страна добавяне на класове и свойства за нови типове данни за биоразнообразието е възможно и желателно.

Свързаните отворени данни, подобно на онтологията, вече представляват солиден ресурс за биолозите, тъй като включват информация от повечето статии, публикувани от Пенсофт и Плаци и наброяват над 600 милиона твърдения. Подобно на онтологията, те следва да бъдат разширявани. Статия, относно генерирането на данни по онтология е подготвена и представа пред научното списание Cybernetics and Information Technologies.

Софтуерният пакет RDF4R беше използван успешно за създаването на LOD, и като такъв, той може да се счита за завършен. Както всеки един софтуерен пакет, обаче, той би следвало да бъде поддържан и развиван. Очаква се публикуването му под формата на Software Description Paper.

Уеб-сайтът е все още в бета-версия. Функционалността, която работи отлично е семантичната търсачка. За някои основни типове данни има създадени макети (templates) за визуализация. Въпреки всичко, сайтът търпи разширение и повечето потребители на системата използват езика за SPARQL за търсене.

Важен извод, който може да се направи като обобщение на постигнатите резултати е, че е възможно използването на семантичен граф за интегрирането на голям обем от данни за биологичното разнообразие. Неочаквано ни се отдаде възможност да илюстрираме мощта на графа при анализа на щетите от трагичния пожар в Museu Nacional в Рио де Жанейро. Освен това илюстрирахме, че е възможно да се напишат сравнително прости логически изводи, позволяващи проверка на валидността на дадено таксономично име.

Поради значителния размер на данните е установено, че макар и използването на семантичен граф да е възможно, някои от първоначално избраните технологии са неприложими или трудно приложими. Извършено е наблюдението (вж. глава 3), че практическото приложение на пълния логически модел OWL е трудно, поради проблеми с бързодействието. Вместо него се спряхме на по-маломощната, но по-бърза RDFS. Друго наблюдение, което направихме е, че макар и програмната среда R да дава известни предимства при бързото създаване на прототипа на системата, при увеличаването на сложността на програмния код, необходимо при реално-съществуващата система за покриване на всички частни случаи, език с динамични типове като R създава главоболия при дебъгинг. Същевременно останахме впечатлени от мощния инструментариум на функционалното програмиране, който R предоставя.

Основна трудност се оказа дизамбигуацията на ресурси като имена на автори или таксономични имена. Във функционалния дизайн на пакета RDF4R заложихме модули, които позволяват при търсенето на идентификатор за даден ресурс да вмъкнем списък от функции/ правила за неговата дизамбигуация. Въпреки

това имахме само ограничен успех с дизамбигуацията на базата на правила и по тази причина в производствената система тя е изключена към момента.

Имайки предвид тези и други "пауки" бъдещето развитие на проекта OpenBiodiv може да се очертае по следния не непременно изчерпателен начин:

1. Като непосредствени цели да разширят LOD и онтологията с нови типове данни и нови източници на данни, използвайки съществуващия фреймуърк. Такива данни са напр. геномните данни, данните за място-находища на видове, (био-)географски данни, визуални данни, описателни данни и пр.

2. Да се търси още по-тясна интеграция с други съществуващи хранилища за данни за биоразнообразието освен GBIF. Напр. BioImages, iNaturalist, BOLD, и т.н.

3. Като по-дългосрочна задача да се изследва преминаването от семантичен граф към технология, където машината за извод е разграничена от семантичния граф като WikiData или Neo4j. Освен увеличеното бързодействие, това ще даде допълнителна гъвкавост на проекта като например позволи използването на машини за извод различни от RDF-базирането като напр. Euler.

4. Да бъде продължена разработката на системния софтуер с още по-широко приложение на функционалното програмиране и портирането му към функционален език като напр. Haskell или O'Caml.

5. Да бъде изследван проблемът за дизамбигуацията и сродните проблеми за named entity recognition на интересни ресурси от биоразнообразието, както и различни задачи от разпознаването на образи, от гледната точка на машинното обучение.

6. Разширяване на уеб-сайта на системата с повече темплейти и нови приложения.

Основни научни и научно-приложни приноси

Резултатите, дискутирани в предишните два раздела, обуславят следните научни и научно-приложни приноси:

1. Научен принос: създаване на онтология и формален модел на областта на публикуване на знание за биологичното разнообразие.

2. Научно-приложен принос: анализ на източниците на информация и създаване на OpenBiodiv-LOD.

3. Научно-приложен принос: имплементация на софтуерните модули на OpenBiodiv.

OpenBiodiv-O запълва уникалната ниша между библиографски онтологии като SPAR и онтологии за биоразнообразието като Darwin-SW и като такава без съмнение е от голям научен интерес за общността на информатиката на биоразнообразието. Работата има и сериозен научно-приложен характер, като в допълнение на онтологията предоставя и свързани отворени данни, използващи онтологията, както и програмно обезпечение за потребителите и разработчиците на системата.

Списък с публикации

Публикации в международни научни списания

1. Viktor Senderov and Lyubomir Penev. 2016. “The Open Biodiversity Knowledge Management System in Scholarly Publishing”. *Research Ideas and Outcomes* 2, no. e7757 (). ISSN: 2367-7163. doi:10.3897/rio.2.e7757. http://rio.pensoft.net/articles.php?id=7757. Намерени 3 уникални цитата от Franz and Sterner, 2018, Ordynets et al., 2017 и Burt and Mengual, 2017.

2. Sarah Faulwetter, Evangelos Pafilis, Lucia Fanini, Nicolas Bailly, Donat Agosti, Christos Arvanitidis, Laura Boicenco, Terry Catapano, Simon Claus, Stefanie Dekeyzer, Teodor Georgiev, Aglaia Legaki, Dimitra Mavraki, Anastasis Oulas, Gabriella Papastefanou, Lyubomir Penev, Guido Sautter, Dmitry Schigel, Viktor Senderov, Adrian Teaca, and Marilena Tsompanou. 2016. “EMODnet Workshop on mechanisms and guidelines to mobilise historical data into biogeographic databases”. *Research Ideas and Outcomes* 2, no. e10445 (). ISSN: 2367-7163. doi:10.3897/rio.2.e10445. http://rio.pensoft.net/articles.php?id=10445. Намерен уникален цитат от Pyron, 2018.

3. Pedro Cardoso, Pavel Stoev, Teodor Georgiev, Viktor Senderov, and Lyubomir Penev. 2016. “Species Conservation Profiles compliant with the IUCN Red List of Threatened Species”. *Biodiversity Data Journal* 4 (): e10356. ISSN: 1314-2828, 1314-2836. doi:10.3897/BDJ.4.e10356. http://bdj.pensoft.net/articles.php?id=10356. Индексация в WoS SCOPUS, както и в SJR 0.465. 4 уникални цитата от Bachman et al., 2018, Lin et al., 2017, Li et al., 2017 и Milano et al., 2017.

4. Viktor Senderov, Teodor Georgiev, and Lyubomir Penev. 2016. “Online direct import of specimen records into manuscripts and automatic creation of data papers from biological databases”. *Research Ideas and Outcomes* 2 (): e10617. ISSN: 2367-7163. doi:10.3897/rio.2.e10617. http://rio.pensoft.net/articles.php?id=10617. Намерен уникален цитат от Ordynets et al., 2017.

5. Lyubomir Penev, Daniel Mietchen, Vishwas Chavan, Gregor Hagedorn, Vincent Smith, David Shotton, Éamonn Ó Tuama, Viktor Senderov, Teodor Georgiev, Pavel Stoev, Quentin Groom, David Remsen, and Scott Edmunds. 2017b. “Strategies and guidelines for scholarly publishing of biodiversity data”. *Research Ideas and Outcomes* 3, no. e12431 (). ISSN: 2367-7163. doi:10.3897/rio.3.e12431. http://riojournal.com/articles.php?id=12431 Намерени 8 уникални цитата от Tennant et al., 2017, Marwick and Birch, 2017, Kissling et al., 2018, Mathieu, 2018, Шашков and Иванова, 2018, Шашков et al., 2017, Filippova et al., 2017 и Филиппова et al., 2017.

6. Lyubomir Penev, Teodor Georgiev, Peter Geshev, Seyhan Demirov, Viktor Senderov, Iliyana Kuzmova, Iva Kostadinova, Slavena Peneva, and Pavel Stoev. 2017a. “ARPHA-BioDiv: A toolbox for scholarly publication and dissemination of biodiversity data based on the ARPHA Publishing Platform”. *Research Ideas and Outcomes* 3, no. e13088 (). ISSN: 2367-7163. doi:10.3897/rio.3.e13088. http://riojournal.com/articles.php?id=13088

7. Emmanuel Arriaga-Varela, Matthias Seidel, Albert Deler-Hernández, Viktor Senderov, and Martin Fikácek. 2017. “A review of the Cercyon Leach (Coleoptera, Hydrophilidae, Sphaeridiinae) of the Greater Antilles”. *ZooKeys* 681 (): 39–

93. ISSN: 1313-2970, 1313-2989. doi:10.3897/zookeys.681.12522. https://zookeys.pensoft.net/articles.php?id=12522. Индексация в WoS IF 1.079, Q3 SCOPUS, SJR 0.533.

8. Viktor Senderov, Kiril Simov, Nico Franz, Pavel Stoev, Terry Catapano, Donat Agosti, Guido Sautter, Robert A. Morris, and Lyubomir Penev. 2018. "OpenBiodiv-O: ontology of the OpenBiodiv knowledge management system". *Journal of Biomedical Semantics* 9, no. 5 (). ISSN: 2041-1480. doi:10.1186/s13326-017-0174-5. https://jbiomedsem.biomedcentral.com/articles/10.1186/s13326-017-0174-5 Индексация в WoS IF 1.6, Q3 SCOPUS, SJR 0.952. Намерени 3 уникални цитата от Michel et al., 2018, Page, 2018b и Page, 2018a.

9. OpenBiodiv-LOD: Populating the OpenBiodiv Ontology. Непубликувана статия, представена в списанието Cybernetics and Information Technologies.

Представен е списък с публикации, свързани с дисертацията. Изброените статии са публикувани без изключение в четири международни научни списания: пет статии в Research Ideas and Outcomes, една статия в ZooKeys (WoS IF 1.079, Q3 SCOPUS, SJR 0.533), една статия в Biodiversity Data Journal (WoS SCOPUS, SJR 0.465) и една статия в Journal of Biomedical Semantics (WoS IF 1.6, Q3 SCOPUS, SJR 0.952). Общият брой намерени цитати, които кандидатът е натрупал, изключвайки само-цитати е 20, като конкретните цитиращи статии са посочени в списъка по горе. Общият брой намерени цитати, включително само-цитати и цитати на работа извън обхвата на дисертацията, е 48 (Google Наука).

[1] е ранна версия на въведението, както и на глава 1 и съдържа труд по Задача 2 (Архитектура). Текстовете на публикациите [2, 3, 5, 6, 7] не са част от текста на дисертацията дословно, но съдържат труд по Задача 5 (Методи за работа). Представените в тези публикации идеи са в голяма степен включени в глава 5, чиито текстови гръбнак се формира от публикация [4]. [8] е най-важната публикация в рамките на тази дисертация и е публикувана във реномираното списание Journal of Biomedical Semantics. [8] съставлява съдържанието на глава 2 и е основната част от работата, изпълняваща Задача 1 (Онтология). Статията беше на заглавната страница на JBS в продължение на няколко месеца (фиг. 6.2). Глава 3 и глава 4, които формират Задачи 3, 4, се подготвят като ръкописи в международни списания. Освен това софтуерната библиотека RDF4R, описана в глава 4, се подготвя да бъде изпратена в хранилището с отворен код rOpenSci[1].

Апробация на резултатите

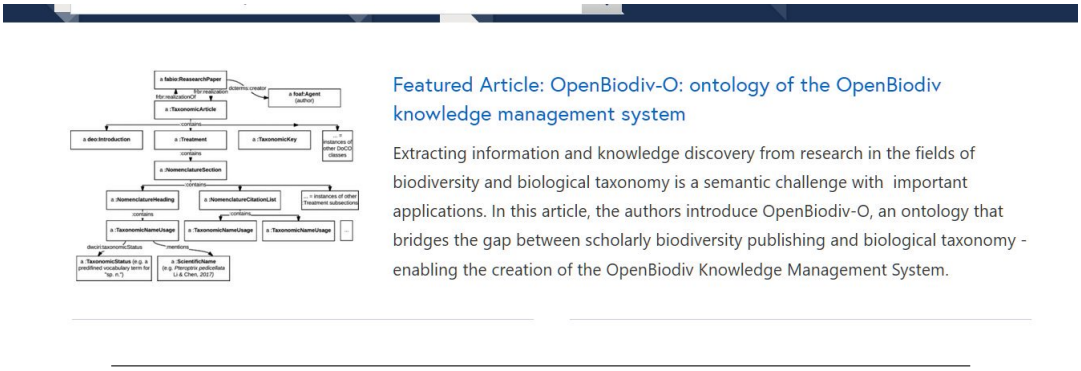
Доклади пред научен семинар на ПНЗ

1. Доклад пред научен семинар на ИБЕИ на БАН на 26.10.2015 г. ("Публикуване, визуализация и разпространение на първични и геномни данни за биологичното разнообразие на основата на открита система за управление на информацията").

2. Доклад пред научен семинар в ИИКТ на БАН на 31.03.2016 г. (Open Biodiversity Knowledge Management System)

3. Доклад пред научен семинар на ИИКТ на БАН за 23.03.2018 г. (OpenBiodiv: a knowledge-based system of biodiversity information)

[1] "We build software with a community of users and developers, and educate scientists about transparent research practices." https://ropensci.org/



Фигура 6.2: Статията относно OpenBiodiv-O е на началната страница на Journal of Biomedical Semantics.

Доклади пред научно мероприятие в чужбина или пред международно научно мероприятие у нас

1. Доклад пред международния симпозиум EU BON в София на 23.03.2016 г. (The Data Publishing Toolkit at EU BON: Automated creation of data papers, data and text integrated publishing via the ARPHA Publishing Platform.)

2. Доклад по време на работната среща на BIG4 в Хавраники, Чехия на 03.06.2016 г. (Project Progress Report (OBKMS))

3. Доклад по време на работната среща на BIG4 в Хавраники, Чехия на 03.06.2016 г. (Modern Methods of Systematic Research and the BOLD Algorithm)

4. Уеб-базиран доклад (уебинар) пред международна аудитория в рамките на семинар на iDigBio на 16.07.2016 г. (Online direct import of specimen records from iDigBio instrastructure into taxonomic manuscripts)

5. Доклад по време на работната среща на BIG4 в Копенхаген на 14.10.2016 г. (Midterm Progress Report)

6. Доклад на международния симпозиум TDWG 2016 в Санта Клара де Сан Карлос от 5. до 9.12.2016 г. (Streamlining the Flow of Taxon Occurrence Data Between a Manuscript and Biological Databases)

7. Доклад на международния симпозиум TDWG 2016 в Санта Клара де Сан Карлос от 5. до 9.12.2016 г. (The Open Biodiversity Knowledge Management System: A Semantic Suite Running on top of the Biodiversity Knowledge Graph)

8. Доклад на международния симпозиум TDWG 2016 в Санта Клара де Сан Карлос от 5. до 9.12.2016 г. (Demonstrating the Prototype of the Open Biodiversity Knowledge Management System)

9. Доклад на международния симпозиум TDWG 2016 в Санта Клара де Сан Карлос от 5. до 9.12.2016 г. (Creation of Data Paper Manuscripts from Ecological Metadata Language (EML))

10. Уеб-базиран доклад пред международния семинар на работната група по семантични технология към Университета Вандербилт (Тенеси, САЩ) на 20.02.2017 г. (Open Biodiversity Knowledge Management System)

11. Доклад на европейската конференция на биосистематиците, BioSyst.eu 2016 на 15.08.2017 г. (The OpenBiodiv Knowledge System: The Future of Access to Biodiversity Knowledge)

12. Доклад на международния симпозиум TDWG 2017 в Отава, Канада от 1. до 6.10.2017 г. (OpenBiodiv Computer Demo: an Implementation of a Semantic System Running on top of the Biodiversity Knowledge Graph)

13. Доклад на международния симпозиум TDWG 2017 в Отава, Канада от 1. до 6.10.2017 г. (OpenBiodiv: an Implementation of a Semantic System Running on top of the Biodiversity Knowledge Graph)

14. Постер на международния симпозиум TDWG 2017 в Отава, Канада от 1. до 6.10.2017 г. (OpenBiodiv: an Implementation of a Semantic System Running on top of the Biodiversity Knowledge Graph)

15. Доклад по време на работната среща на BIG4 в Ла Палма, Испания от 30. окт. до 3 ноем. 2017 г. (Midterm Progress Report)

16. Доклад пред научен семинар на групата по биоинформатика (група Ронкуист) в Кралския природо-научен музей в Стокхолм на 29.11.2017 г.

Bibliography

Agosti, D., C. Klingenberg, N. Johnson, C. Stephenson, and C. Catapano. 2007. *Why not let the computer save you time by reading the taxonomic papers for you?* Tech. rep. Zenodo. doi:10.5281/zenodo.15584.

Agosti, Donat. 2006. "Biodiversity data are out of local taxonomists' reach". *Nature* 439, no. 7075 (): 392–392. ISSN: 0028-0836, 1476-4687. doi:10.1038/439392a. http://www.nature.com/articles/439392a.

Arriaga-Varela, Emmanuel, Matthias Seidel, Albert Deler-Hernández, Viktor Senderov, and Martin Fikácek. 2017. "A review of the Cercyon Leach (Coleoptera, Hydrophilidae, Sphaeridiinae) of the Greater Antilles". *ZooKeys* 681 (): 39–93. ISSN: 1313-2970, 1313-2989. doi:10.3897/zookeys.681.12522. https://zookeys.pensoft.net/articles.php?id=12522.

Auer, Sören, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. "DBpedia: A Nucleus for a Web of Open Data". In *The Semantic Web*, ed. by David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, 4825:722–735. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-540-76297-3 978-3-540-76298-0. doi:10.1007/978-3-540-76298-0_52. http://link.springer.com/10.1007/978-3-540-76298-0_52.

Bachman, Steven P, Eimear M Nic Lughadha, and Malin C Rivers. 2018. "Quantifying progress toward a conservation assessment for all plants". *Conservation Biology* 32 (3): 516–524.

Barrasa, Jesús. 2017. *RDF Triple Stores vs. Labeled Property Graphs: What's the Difference?* Published online. Visited on 01/11/2019. https://neo4j.com/blog/rdf-triple-store-vs-labeled-property-graph-difference/.

Baskauf, Steve, and Campbell O. Webb. 2016. "Darwin-SW: Darwin Core-based terms for expressing biodiversity data as RDF". *Semantic Web Journal* 7 (6): 629–643. doi:10.3233/SW-150203.

Beck, Kent, Mike Beedle, Arie van Bennekum, Alistair Cockburn, Ward Cunningham, Martin Fowler, James Grenning, Jim Highsmith, Andrew Hunt, Ron Jeffries, Jon Kern, Brian Marick, Robert C. Martin, Steve Mellor, Ken Schwaber, Jeff Sutherland, and Dave Thomas. 2001. *Manifesto for Agile Software Development.* Published online. Visited on 01/11/2019. http://www.agilemanifesto.org/.

Berendsohn, Walter G. 1995. "The concept of"potential taxa"in databases". *Taxon*: 207–212. Visited on 07/22/2017. http://www.jstor.org/stable/1222443.

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. "The Semantic Web". *Scientific American* 284, no. 5 (): 34–43. ISSN: 0036-8733, visited on 04/13/2018. doi:10.1038/scientificamerican0501-34. http://www.nature.com/doifinder/10.1038/scientificamerican0501-34.

Blomquist, HL. 1948. *The grasses of North Carolina.* Duke Universtiy Press.

Boettiger, Carl, Scott Chamberlain, Edmund Hart, and Karthik Ram. 2015. "Building software, building community: lessons from the rOpenSci project". *Journal of Open Research Software* 3 (1).

Bollacker, Kurt, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. "Freebase: a collaboratively created graph database for structuring human knowledge", 1247. ACM Press. ISBN: 978-1-60558-102-6, visited on 07/22/2018. doi:10.1145/1376616.1376746. http://portal.acm.org/citation.cfm?doid=1376616.1376746.

Burt, Trevor, and Ximo Mengual. 2017. "Origin and diversification of hoverflies: a revision of the genera Asarkina and Allobaccha–A BIG4 Consortium PhD project". *Research Ideas and Outcomes* 3:e19860.

Cardoso, Pedro, Pavel Stoev, Teodor Georgiev, Viktor Senderov, and Lyubomir Penev. 2016. "Species Conservation Profiles compliant with the IUCN Red List of Threatened Species". *Biodiversity Data Journal* 4 (): e10356. ISSN: 1314-2828, 1314-2836. doi:10.3897/BDJ.4.e10356. http://bdj.pensoft.net/articles.php?id=10356.

Catapano, Terence. 2010. "TaxPub: an extension of the NLM/NCBI journal publishing DTD for taxonomic descriptions". Published online. Visited on 07/11/2017. https://www.ncbi.nlm.nih.gov/books/NBK47081/.

Challenges in irreproducible research. 2010. Published online. Visited on 04/15/2018. https://www.nature.com/collections/prbfkwmwvz.

Chavan, Vishwas, and Lyubomir Penev. 2011. "The data paper: a mechanism to incentivize data publishing in biodiversity science". *BMC Bioinformatics* 12 (Suppl 15): S2. ISSN: 1471-2105, visited on 02/13/2018. doi:10.1186/1471-2105-12-S15-S2. http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-S15-S2.

Chen, Mingmin, Shizhuo Yu, Nico Franz, Shawn Bowers, and others. 2014. "Euler/X: a toolkit for logic-based taxonomy integration". Published online, *arXiv preprint arXiv:1402.1992.* Visited on 08/11/2017. https://arxiv.org/abs/1402.1992.

Claerbout, Jon F., and Martin Karrenbach. 1992. "Electronic documents give reproducible research a new meaning", 601–604. Society of Exploration Geophysicists. Visited on 04/15/2018. doi:10.1190/1.1822162. http://library.seg.org/doi/abs/10.1190/1.1822162.

Constantin, Alexandru, Silvio Peroni, Steve Pettifer, David Shotton, and Fabio Vitali. 2016. "The Document Components Ontology (DoCO)". Ed. by Oscar Corcho. *Semantic Web* 7, no. 2 (): 167–181. ISSN: 22104968, 15700844, visited on 08/13/2017. doi:10.3233/SW-150177. http://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/SW-150177.

Deans, Andrew R., Matthew J. Yoder, and James P. Balhoff. 2012. "Time to change how we describe biodiversity". *Trends in Ecology & Evolution* 27, no. 2 (): 78–84. ISSN: 01695347, visited on 07/11/2017. doi:10.1016/j.tree.2011.11.007. http://linkinghub.elsevier.com/retrieve/pii/S0169534711003302.

Dmitriev, D.A., and M. Yoder. 2017. *NOMEN*. Published online. Visited on 07/22/2017. https://github.com/SpeciesFileGroup/nomen.

Egloff, Willi, David Patterson, Donat Agosti, and Gregor Hagedorn. 2014. "Open exchange of scientific knowledge and European copyright: The case of biodiversity information". *ZooKeys* 414 (): 109–135. ISSN: 1313-2970, 1313-2989, visited on 04/13/2018. doi:10.3897/zookeys.414.7717. http://zookeys.pensoft.net/articles.php?id=3830.

Filippova, Nina V., Ilya V. Filippov, Dmitry S. Schigel, Natalia V. Ivanova, and Maxim P. Shashkov. 2017. "Biodiversity informatics: global trends, national perspective and regional progress in Khanty-Mansi Autonomous Okrug". *Environmental Dynamics and Global Climate Change* 8, no. 2 (): 46–56. ISSN: 2541-9307, 2218-4422, visited on 03/15/2018. doi:10.17816/edgcc8246-56. http://journals.eco-vector.com/EDGCC/article/view/7080.

Franz, N.M., and R.K. Peet. 2009. "Perspectives: Towards a language for mapping relationships among taxonomic concepts". *Systematics and Biodiversity* 7, no. 1 (): 5–20. ISSN: 1477-2000, 1478-0933, visited on 07/22/2017. doi:10.1017/S147720000800282X. http://www.tandfonline.com/doi/abs/10.1017/S147720000800282X.

Franz, Nico M, and Beckett W Sterner. 2018. "To increase trust, change the social design behind aggregated biodiversity data". *Database* 2018.

GBIF Secretariat. 2017. "GBIF Backbone Taxonomy. Checklist dataset". Published online. Visited on 06/30/2018. doi:10.15468/39omei. https://doi.org/10.15468/39omei.

Godtsenhoven, Karen, van, Mikael Karstensen Elbaek, Barbara Sierman, Magchiel Bijsterbosch, Patrick Hochstenbach, Rosemary Russell, and Maurice Vanderfeesten. 2009. *Emerging Standards for Enhanced Publications and Repository Technology : Survey on Technology*. Amsterdam: Amsterdam University Press. ISBN: 978-90-8964-189-2, visited on 04/15/2018. doi:10.5117/9789089641892. http://dare.uva.nl/aup/nl/record/316870.

Gruber, Thomas R. 1993. "A translation approach to portable ontology specifications". *Knowledge Acquisition* 5, no. 2 (): 199–220. ISSN: 10428143, visited on 08/07/2017. doi:10.1006/knac.1993.1008. http://linkinghub.elsevier.com/retrieve/pii/S1042814383710083.

Harris, Larry R., Jeffrey M. Hill, Dayton Marcott, and Timothy F. Rochford. 1993. Knowledge base management system. Pat. US5228116A, **patentfiled**July 1993. https://patents.google.com/patent/US5228116A/en.

Heath, Tom, and Christian Bizer. 2011. *Linked data: evolving the web into a global data space*. 1. ed. Synthesis lectures on the semantic web: theory and technology 1. OCLC: 732238828. San Rafael, Calif.: Morgan & Claypool. ISBN: 978-1-60845-430-3 978-1-60845-431-0.

International Commission on Zoological Nomenclature. 1999. *International Code of Zoological Nomenclature*. Fourth Edition. London, UK: The International Trust for Zoological Nomenclature. ISBN: 0 85301 006 4.

— . 2017. *The Official Registry of Zoological Nomenclature*. Publihsed online. Visited on 08/11/2017. http://zoobank.org/.

International code of nomenclature for algae, fungi and plants (Melbourne code): adopted by the Eighteenth International Botanical Congress Melbourne, Australia, July 2011. 2012. Regnum vegetabile v. 154. OCLC: ocn824722354. Königstein, Germany: Koeltz Scientific Books. ISBN: 978-3-87429-425-6.

Jarke, Matthias, Bernd Neumann, Yannis Vassiliou, and Wolfgang Wahlster. 1989. "KBMS Requirements of Knowledge-Based Systems". In *Foundations of Knowledge Base Management*, ed. by Michael L. Brodie, John Mylopoulos, Joachim W. Schmidt, Joachim W. Schmidt, and Constantino Thanos, 381–394. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-642-83399-1 978-3-642-83397-7, visited on 07/22/2018. doi:10.1007/978-3-642-83397-7_17. http://www.springerlink.com/index/10.1007/978-3-642-83397-7_17.

Kennedy, Jessie B., Robert Kukla, and Trevor Paterson. 2005. "Scientific Names Are Ambiguous as Identifiers for Biological Taxa: Their Context and Definition Are Required for Accurate Data Integration". In *Data Integration in the Life Sciences: Second International Workshop, DILS 2005, San Diego, CA, USA, July 20-22, 2005. Proceedings*, ed. by Bertram Ludäscher and Louiqa Raschid, 80–95. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-540-31879-8. doi:10.1007/11530084_8.

Kissling, W Daniel, Ramona Walls, Anne Bowser, Matthew O Jones, Jens Kattge, Donat Agosti, Josep Amengual, Alberto Basset, Peter M Van Bodegom, Johannes HC Cornelissen, and others. 2018. "Towards global data products of Essential Biodiversity Variables on species traits". *Nature ecology & evolution*: 1.

Koperski, M, M Sauer, W Braun, and S Gradstein. 2000. *Referenzliste der Moose Deutschlands*. Vol. 34. Schriftenreihe Vegetationsk.

Kraker, Peter, Derick Leony, Wolfgang Reinhardt, N.A. Gü, and nter Beham. 2011. "The case for an open science in technology enhanced learning". *International Journal of Technology Enhanced Learning* 3 (6): 643. ISSN: 1753-5255, 1753-5263, visited on 04/15/2018. doi:10.1504/IJTEL.2011.045454. http://www.inderscience.com/link.php?id=45454.

Li, Diyan, Tiandong Che, Binlong Chen, Shilin Tian, Xuming Zhou, Guolong Zhang, Miao Li, Uma Gaur, Yan Li, Majing Luo, and others. 2017. "Genomic data for 78 chickens from 14 populations". *GigaScience* 6 (6): 1–5.

Lin, Qiang, Ying Qiu, Ruobo Gu, Meng Xu, Jia Li, Chao Bian, Huixian Zhang, Geng Qin, Yanhong Zhang, Wei Luo, Jieming Chen, Xinxin You, Mingjun Fan, Min Sun, Pao Xu, Byrappa Venkatesh, Junming Xu, Hongtuo Fu, and Qiong Shi. 2017. "Draft genome of the lined seahorse, Hippocampus erectus" [**inlang**en]. *GigaScience* 6, no. 6 (): 1–6. ISSN: 2047-217X, visited on 03/15/2018. doi:10.1093/gigascience/gix030. https://academic.oup.com/gigascience/article-lookup/doi/10.1093/gigascience/gix030.

Linnaeus, Carl von. 1758. *Systema naturae per regna tria naturae: secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis. Volume 1*.

Mallet, James. 2001. "Species, concepts of". In *Encyclopedia of biodiversity*, 5:427–440. Published online. Visited on 07/11/2017. http://tarjomefa.com/wp-content/uploads/2016/02/4420-engilish.pdf.

Mansinghka, Vikash, Richard Tibbetts, Jay Baxter, Pat Shafto, and Baxter Eaves. 2015. "BayesDB: A probabilistic programming system for querying the probable implications of data". ArXiv: 1512.05006, *arXiv:1512.05006 [cs]* (). Visited on 07/22/2018. http://arxiv.org/abs/1512.05006.

Marwick, Ben, and Suzanne Birch. 2017. *A Standard for the Scholarly Citation of Archaeological Data as an Incentive to Data Sharing*. Tech. rep. SocArXiv. Visited on 03/15/2018. doi:10.17605/OSF.IO/PY4HZ. https://osf.io/preprints/socarxiv/py4hz/.

Mathieu, Jérôme. 2018. "EGrowth: A global database on intraspecific body growth variability in earthworm". *Soil Biology and Biochemistry* 122:71–80.

Michel, Franck, Catherine Faron-Zucker, Sandrine Tercerie, and Gargominy Olivier. 2018. "Modelling Biodiversity Linked Data: Pragmatism May Narrow Future Opportunities". *Biodiversity Information Science and Standards* 2:e26235.

Michener, William K., James W. Brunt, John J. Helly, Thomas B. Kirchner, and Susan G. Stafford. 1997. "NONGEOSPATIAL METADATA FOR THE ECOLOGICAL SCIENCES". *Ecological Applications* 7, no. 1 (): 330–342. ISSN: 1051-0761, visited on 02/14/2018. doi:10.1890/1051-0761(1997)007[0330:NMFTES]2.0.CO;2. http://doi.wiley.com/10.1890/1051-0761(1997)007[0330:NMFTES]2.0.CO;2.

Mietchen, Daniel. 2014. "The Transformative Nature of Transparency in Research Funding". *PLoS Biology* 12, no. 12 (): e1002027. ISSN: 1545-7885, visited on 04/15/2018. doi:10.1371/journal.pbio.1002027. http://dx.plos.org/10.1371/journal.pbio.1002027.

Milano, Filippo, Paolo Pantini, Stefano Mammola, and Marco Isaia. 2017. "LA CONSERVAZIONE DELL'ARANEOFAUNA IN ITALIA E IN EUROPA". *ATTI DELL'ACCADEMIA NAZIONALE ITALIANA DI ENTOMOLOGIA. RENDICONTI* 65:91–103.

Miller, Jeremy, Torsten Dikow, Donat Agosti, Guido Sautter, Terry Catapano, Lyubomir Penev, Zhi-Qiang Zhang, Dean Pentcheff, Richard Pyle, Stan Blum, Cynthia Parr, Chris Freeland, Tom Garnett, Linda S Ford, Burgert Muller, Leo Smith, Ginger Strader, Teodor Georgiev, and Laurence Bénichou. 2012. "From taxonomic literature to cybertaxonomic content". *BMC Biology* 10 (1): 87. ISSN: 1741-7007, visited on 04/13/2018. doi:10.1186/1741-7007-10-87. http://bmcbiol.biomedcentral.com/articles/10.1186/1741-7007-10-87.

Miller, Jeremy, Donat Agosti, Lyubomir Penev, Guido Sautter, Teodor Georgiev, Terry Catapano, David Patterson, David King, Serrano Pereira, Rutger Vos, and Soraya Sierra. 2015. "Integrating and visualizing primary data from prospective and legacy taxonomic literature". *Biodiversity Data Journal* 3 (): e5063. ISSN: 1314-2828, 1314-2836, visited on 04/13/2018. doi:10.3897/BDJ.3.e5063. http://bdj.pensoft.net/articles.php?id=5063.

Mindell, David P., Brian L. Fisher, Peter Roopnarine, Jonathan Eisen, Georgina M. Mace, Roderic D. M. Page, and Richard L. Pyle. 2011. "Aggregating, Tagging and Integrating Biodiversity Research". Ed. by Sean A. Rands. *PLoS ONE* 6, no. 8 (): e19491. ISSN: 1932-6203, visited on 04/13/2018. doi:10.1371/journal.pone.0019491. http://dx.plos.org/10.1371/journal.pone.0019491.

Momtchev, Vassil, Deyan Peychev, Todor Primov, and Georgi Georgiev. 2009. "Expanding the pathway and interaction knowledge in linked life data". *Proc. of International Semantic Web Challenge*. Visited on 07/22/2017. http://challenge.semanticweb.org/documents/Linked%20Life%20Data-LLD%20semantic%20web%20challenge%202009.pdf.

Mulsant, E. 1866. "Monographie des Coccinellides". *Mém. L'Acad. Imp. Lyon, Cl. Sci., LMém. L'Acad. Imp. Lyon.* 15:1–112.

Neo4J Developers. 2012. *Neo4J Graph NoSQL Database*. Published online. http://neo4j.com.

Obitko, Marek. 2007. "Translations between ontologies in multi-agent systems". Ph. D. Dissertation, , Czech Technical University, Faculty of Electrical Engineering.

Ontotext. 2018. *GraphDB 8.6*. Published online. Visited on 07/17/2018. http://graphdb.ontotext.com/.

Ordynets, Alexander, Anton Savchenko, Alexander Akulov, Eugene Yurchenko, Vera Malysheva, Urmas Kõljalg, Josef Vlasák, Karl-Henrik Larsson, and Ewald Langer. 2017. "Aphyllophoroid fungi in insular woodlands of eastern Ukraine". *Biodiversity Data Journal* 5 (): e22426. ISSN: 1314-2828, 1314-2836, visited on 03/15/2018. doi:10.3897/BDJ.5.e22426. https://bdj.pensoft.net/articles.php?id=22426.

Page, R. D. M. 2008. "Biodiversity informatics: the challenge of linking data and the role of shared identifiers". *Briefings in Bioinformatics* 9, no. 5 (): 345–354. ISSN: 1467-5463, 1477-4054, visited on 04/13/2018. doi:10.1093/bib/bbn022. https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbn022.

— . 2015. *Putting some bite into the Bouchout Declaration*. Published online. http://iphylo.blogspot.bg/2015/05/putting-some-bite-into-bouchout.html.

— . 2014. *The vision thing - it's all about the links*. Published online. http://iphylo.blogspot.bg/2014/06/the-vision-thing-it-all-about-links.html.

Page, Roderic D. M. 2006. "Taxonomic names, metadata, and the Semantic Web". *Biodiversity Informatics* 3, no. 0 (). ISSN: 15469735, visited on 04/13/2018. doi:10.17161/bi.v3i0.25. https://journals.ku.edu/index.php/jbi/article/view/25.

Page, Roderic DM. 2018a. "Liberating links between datasets using lightweight data publishing: an example using plant names and the taxonomic literature". Published online, *bioRxiv*: 343996.

— . 2018b. "Ozymandias: A biodiversity knowledge graph". Published online, *bioRxiv*: 485854.

Parr, Cynthia S., Robert Guralnick, Nico Cellinese, and Roderic D.M. Page. 2012. "Evolutionary informatics: unifying knowledge about the diversity of life" [**inlang**en]. *Trends in Ecology & Evolution* 27, no. 2 (): 94–103. ISSN: 01695347, visited on 04/13/2018. doi:10.1016/j.tree.2011.11.001. http://linkinghub.elsevier.com/retrieve/pii/S0169534711003247.

Patterson, D.J., J. Cooper, P.M. Kirk, R.L. Pyle, and D.P. Remsen. 2010. "Names are key to the big new biology". *Trends in Ecology & Evolution* 25, no. 12 (): 686–691. ISSN: 01695347, visited on 07/11/2017. doi:10.1016/j.tree.2010.09.004. http://linkinghub.elsevier.com/retrieve/pii/S0169534710002181.

Patterson, David J., David Remsen, William A. Marino, and Cathy Norton. 2006. "Taxonomic Indexing—Extending the Role of Taxonomy". Ed. by Rod Page. *Systematic Biology* 55, no. 3 (): 367–373. ISSN: 1076-836X, 1063-5157, visited on 11/16/2018. doi:10.1080/10635150500541680. https://academic.oup.com/sysbio/article/55/3/367/1667279.

Pellissier Tanon, Thomas, Denny Vrandečić, Sebastian Schaffert, Thomas Steiner, and Lydia Pintscher. 2016. "From Freebase to Wikidata: The Great Migration", 1419–1428. ACM Press. ISBN: 978-1-4503-4143-1, visited on 07/22/2018. doi:10.1145/2872427.2874809. http://dl.acm.org/citation.cfm?doid=2872427.2874809.

Penev, Lyubomir, Teodor Georgiev, Peter Geshev, Seyhan Demirov, Viktor Senderov, Iliyana Kuzmova, Iva Kostadinova, Slavena Peneva, and Pavel Stoev. 2017a. "ARPHA-BioDiv: A toolbox for scholarly publication and dissemination of biodiversity data based on the ARPHA Publishing Platform". *Research Ideas and Outcomes* 3, no. e13088 (). ISSN: 2367-7163. doi:10.3897/rio.3.e13088. http://riojournal.com/articles.php?id=13088.

Penev, Lyubomir, W. John Kress, Sandra Knapp, De-Zhu Li, and Susanne Renner. 2010a. "Fast, linked, and open – the future of taxonomic publishing for plants: launching the journal PhytoKeys". *PhytoKeys* 1, no. 0 (). ISSN: 1314-2003, 1314-2011, visited on 07/22/2017. doi:10.3897/phytokeys.1.642. http://www.pensoft.net/journal_home_page.php?journal_id=3&page=article&type=show&article_id=642&abstract=1.

Penev, Lyubomir, Terence Catapano, Donat Agosti, Teodor Georgiev, Guido Sautter, and Pavel Stoev. 2012. "Implementation of TaxPub, an NLM DTD extension for domain-specific markup in taxonomy, from the experience of a biodiversity publisher". Published online. Visited on 07/11/2017. https://www.ncbi.nlm.nih.gov/books/NBK100351/.

Penev, Lyubomir, Donat Agosti, Teodor Georgiev, Terry Catapano, Jeremy Miller, Vladimir Blagoderov, David Roberts, Vincent Smith, Irina Brake, Simon Ryrcroft, Ben Scott, Norman Johnson, Robert Morris, Guido Sautter, Vishwas Chavan, Tim Robertson, David Remsen, Pavel Stoev, Cynthia Parr, Sandra Knapp, W. John Kress, Frederic Thompson, and Terry Erwin. 2010b. "Semantic tagging of and semantic enhancements to systematics papers: ZooKeys working examples". *ZooKeys* 50 (): 1–16. ISSN: 1313-2970, 1313-2989, visited on 04/15/2018. doi:10.3897/zookeys.50.538. http://zookeys.pensoft.net/articles.php?id=2215.

Penev, Lyubomir, Daniel Mietchen, Vishwas Chavan, Gregor Hagedorn, Vincent Smith, David Shotton, Éamonn Ó Tuama, Viktor Senderov, Teodor Georgiev, Pavel Stoev, Quentin Groom, David Remsen, and Scott Edmunds. 2017b. "Strategies and guidelines for scholarly publishing of biodiversity data". *Research Ideas and Outcomes* 3, no. e12431 (). ISSN: 2367-7163. doi:10.3897/rio.3.e12431. http://riojournal.com/articles.php?id=12431.

Penev, Lyubomir, Christopher Lyal, Anna Weitzman, David Morse, David King, Guido Sautter, Teodor Georgiev, Robert Morris, Terry Catapano, and Donat Agosti. 2011. "XML schemas and mark-up practices of taxonomic literature". *ZooKeys* 150 (): 89–116. ISSN: 1313-2970, 1313-2989, visited on 07/04/2018. doi:10.3897/zookeys.150.2213. http://zookeys.pensoft.net/articles.php?id=3038.

Peroni, Silvio. 2014. "The semantic publishing and referencing ontologies". In *Semantic Web Technologies and Legal Scholarly Publishing*, 1st ed., 15:121–193. Springer. Visited on 07/22/2017.

Peroni, Silvio, and David Shotton. 2012. "FaBiO and CiTO: Ontologies for describing bibliographic resources and citations". *Web Semantics: Science, Services and Agents on the World Wide Web* 17 (): 33–43. ISSN: 15708268, visited on 08/13/2017. doi:10.1016/j.websem.2012.08.001. http://linkinghub.elsevier.com/retrieve/pii/S1570826812000790.

Poorani, J., and Roger Booth. 2016. "Harmonia manillana (Mulsant), a new addition to Indian Coccinellidae, with changes in synonymy". *Biodiversity Data Journal* 4 (): e8030. ISSN: 1314-2828, 1314-2836, visited on 08/13/2017. doi:10.3897/BDJ.4.e8030. http://bdj.pensoft.net/articles.php?id=8030.

Pyle, Richard. 2016. "Towards a Global Names Architecture: The future of indexing scientific names". *ZooKeys* 550 (): 261–281. ISSN: 1313-2970, 1313-2989, visited on 08/12/2017. doi:10.3897/zookeys.550.10009. http://zookeys.pensoft.net/articles.php?id=6241.

Pyron, Robert Alexander. 2018. "A 21st Century Vision for Neotropical Snake Systematics". *Revista Latinoamericana de Herpetología* 1 (1): 58–62.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Published online, Vienna, Austria. https://www.R-project.org/.

Rebholz-Schuhmann, Dietrich, Harald Kirsch, and Francisco Couto. 2005. "Facts from text—is text mining ready to deliver?" *PLoS biology* 3 (2): e65. Visited on 07/22/2017. http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0030065.

Sarah Faulwetter, Evangelos Pafilis, Lucia Fanini, Nicolas Bailly, Donat Agosti, Christos Arvanitidis, Laura Boicenco, Terry Catapano, Simon Claus, Stefanie Dekeyzer, Teodor Georgiev, Aglaia Legaki, Dimitra Mavraki, Anastasis Oulas, Gabriella Papastefanou, Lyubomir Penev, Guido Sautter, Dmitry Schigel, Viktor Senderov, Adrian Teaca, and Marilena Tsompanou. 2016. "EMODnet Workshop on mechanisms and guidelines to mobilise historical data into biogeographic databases". *Research Ideas and Outcomes* 2, no. e10445 (). ISSN: 2367-7163. doi:10.3897/rio.2.e10445. http://rio.pensoft.net/articles.php?id=10445.

Senderov, Viktor, and Lyubomir Penev. 2016. "The Open Biodiversity Knowledge Management System in Scholarly Publishing". *Research Ideas and Outcomes* 2, no. e7757 (). ISSN: 2367-7163. doi:10.3897/rio.2.e7757. http://rio.pensoft.net/articles.php?id=7757.

Senderov, Viktor, Teodor Georgiev, and Lyubomir Penev. 2016. "Online direct import of specimen records into manuscripts and automatic creation of data papers from biological databases". *Research Ideas and Outcomes* 2 (): e10617. ISSN: 2367-7163. doi:10.3897/rio.2.e10617. http://rio.pensoft.net/articles.php?id=10617.

Senderov, Viktor, Kiril Simov, Nico Franz, Pavel Stoev, Terry Catapano, Donat Agosti, Guido Sautter, Robert A. Morris, and Lyubomir Penev. 2018. "OpenBiodiv-O: ontology of the OpenBiodiv knowledge management system". *Journal of Biomedical Semantics* 9, no. 5 (). ISSN: 2041-1480. doi:10.1186/s13326-017-0174-5. https://jbiomedsem.biomedcentral.com/articles/10.1186/s13326-017-0174-5.

Senderov, Viktor, Nico M. Franz, and Kiril Simov. 2017. *OpenBiodiv Ontology and Guide*. Published online. Visited on 08/09/2017. http://openbiodiv.net/ontology.

Shotton, David. 2009. "Semantic publishing: the coming revolution in scientific journal publishing". *Learned Publishing* 22, no. 2 (): 85–94. ISSN: 09531513, visited on 04/15/2018. doi:10.1087/2009202. http://doi.wiley.com/10.1087/2009202.

Singhal, Amit. 2012. *Introducing the knowledge graph: things, not strings.* Published online. Visited on 01/11/2019. https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html.

Staab, Steffen, and Rudi Studer, eds. 2009. *Handbook on Ontologies.* Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-540-70999-2 978-3-540-92673-3, visited on 08/07/2017. doi:10.1007/978-3-540-92673-3. http://link.springer.com/10.1007/978-3-540-92673-3.

Sterner, Beckett, and Nico M. Franz. 2017. "Taxonomy for Humans or Computers? Cognitive Pragmatics for Big Data". *Biological Theory* 12, no. 2 (): 99–111. ISSN: 1555-5542, 1555-5550, visited on 07/11/2017. doi:10.1007/s13752-017-0259-5. http://link.springer.com/10.1007/s13752-017-0259-5.

Taxonomic Names and Concepts Interest Group. 2006. *Taxonomic Concept Transfer Schema (TCS), version 1.01.* Published online. Visited on 01/11/2019. http://www.tdwg.org/standards/117.

Tennant, Jonathan P, Jonathan M Dugan, Daniel Graziotin, Damien C Jacques, François Waldner, Daniel Mietchen, Yehia Elkhatib, Lauren B Collister, Christina K Pikas, Tom Crick, and others. 2017. "A multi-disciplinary perspective on emergent and future innovations in peer review". *F1000Research* 6.

The Bouchout Declaration for Open Biodiversity Knowledge Management. 2014. Published online. Visited on 01/11/2019. http://www.bouchoutdeclaration.org/fileadmin/Dateien/PDF/Bouchout_Declaration_EN.pdf.

Tzitzikas, Yannis, Carlo Allocca, Chryssoula Bekiari, Yannis Marketakis, Pavlos Fafalios, Martin Doerr, Nikos Minadakis, Theodore Patkos, and Leonardo Candela. 2013. "Integrating heterogeneous and distributed information about marine species through a top level ontology". In *Research Conference on Metadata and Semantic Research*, 289–301. Springer. Visited on 07/22/2017. http://link.springer.com/chapter/10.1007/978\-3\-319\-03437\-9_29.

University of Copenhagen, University of Turku, Institute for Systematic Zoology and Evolutionary Biology, Zoologisches Forschungsmuseum Alexander Koenig, Naturhistorisches Museum Wien, Pensoft Publishers company, ERA7 Bioinformatics, Swedish Museum of Natural History, and Decuria IT company. 2014. *BIG4 - Biosystematics, Informatics and Genetics of the big 4 insect groups: training tomorrow's researchers and entrepreneurs.* Published online. Visited on 01/11/2019. http://big4-project.eu.

Vrandečić, Denny, and Markus Krötzsch. 2014. "Wikidata: a free collaborative knowledgebase". *Communications of the ACM* 57, no. 10 (): 78–85. ISSN: 00010782, visited on 07/22/2018. doi:10.1145/2629489. http://dl.acm.org/citation.cfm?doid=2661061.2629489.

Was ist Open Science? Published online. Visited on 01/11/2019. http://openscienceasap.org/open-science/.

What is GBIF? Published online. Visited on 08/12/2017. http://www.gbif.org/what-is-gbif.

Wickham, Hadley. 2017. *httr: Tools for Working with URLs and HTTP.* Published online. Visited on 07/17/2018. https://cran.r-project.org/web/packages/httr/.

Wickham, Hadley, James Hester, and Jeroen Ooms. 2018. *xml2: Parse XML*. Published online. Visited on 07/17/2018. https://cran.r-project.org/web/packages/xml2/index.html.

Wieczorek, John, David Bloom, Robert Guralnick, Stan Blum, Markus Döring, Renato Giovanni, Tim Robertson, and David Vieglais. 2012. "Darwin Core: An Evolving Community-Developed Biodiversity Data Standard". Ed. by Indra Neil Sarkar. *PLoS ONE* 7, no. 1 (): e29715. ISSN: 1932-6203, visited on 07/22/2017. doi:10.1371/journal.pone.0029715. http://dx.plos.org/10.1371/journal.pone.0029715.

Williams, Antony J., Lee Harland, Paul Groth, Stephen Pettifer, Christine Chichester, Egon L. Willighagen, Chris T. Evelo, Niklas Blomberg, Gerhard Ecker, Carole Goble, and Barend Mons. 2012. "Open PHACTS: semantic interoperability for drug discovery". *Drug Discovery Today* 17, **numbers** 21-22 (): 1188–1198. ISSN: 13596446, visited on 07/22/2017. doi:10.1016/j.drudis.2012.05.016. http://linkinghub.elsevier.com/retrieve/pii/S1359644612001936.

Witteveen, Joeri. 2015. "Naming and contingency: the type method of biological taxonomy". *Biology & Philosophy* 30, no. 4 (): 569–586. ISSN: 0169-3867, 1572-8404, visited on 08/13/2017. doi:10.1007/s10539-014-9459-6. http://link.springer.com/10.1007/s10539-014-9459-6.

Wolfram|Alpha, Making the wolrd's knowledge computable. Published online. Wolfram Alpha LLC. Visited on 06/10/2018. https://www.wolframalpha.com/.

pro-iBiosphere project final report. 2014. Published online. Visited on 04/13/2018. http://adm.pro-ibiosphere.eu/getatt.php?filename=oo_4751.pdf.

pro-iBiosphere. Visited on 08/12/2017. http://wiki.pro-ibiosphere.eu/.

Филиппова, НВ, ИВ Филиппов, ДС Щигель, НВ Иванова, and МП Шашков. 2017. "Информатика биоразнообразия: мировые тенденции, состояние дел в России и развитие направления в Ханты-Мансийском Автономном Округе". *Динамика окружающей среды и глобальные изменения климата* 8 (2): 46–56.

Шашков, МП, ИФ Чадин, and НВ Иванова. 2017. "Методические рекомендации по стандартизации данных для публикации через глобальный портал GBIF. ORG и подготовке статьи о данных". *Труды Кольского научного центра РАН*, no. 6-5 (8).

Шашков, МП, and НВ Иванова. 2018. "Стандарты и веб-инструменты для публикации данных через глобальные порталы по биоразнообразию". *Доклады Международной конференции "Математическая биология и биоинформатика"* 7:e98. doi:10.17537/icmbb18.55.