UNIVERSITAT JAUME I
Departament de Llenguatges i Sistemes Informàtics

BIOTEA-BIOLINKS: A SEMANTIC INFRASTRUCTURE FOR EXPLORING AND ANALYZING SCIENTIFIC PUBLICATIONS

Ph.D. dissertation

Leyla Jael García Castro

Supervisor
Dr. Rafael Berlanga Llavori

Co-supervisor
Dr. Alexander García

Castellón, March, 2017

A mi mami,
por ser un continuo apoyo
en todo tiempo y lugar.

Statement of Originality

I declare that the work presented in this thesis is, to the best of my knowledge and belief, original and my own work, except as acknowledged in the text. The material (presented as my own) has not been submitted previously, either in whole or in part, for a degree at this or any other institution. Some of the material can have been published as research papers in journals, workshops, book chapters or conferences.

Leyla Jael García Castro

Statement of Contribution Of Others

In those cases in which the work presented in this thesis was the product of collaborative efforts I declare that my contribution was substantial and prominent, involving the development of original ideas as well as the definition and implementation of subsequent work.

Pr. Dr. Rafael Berlanga Llavori
Dr. Alexander Garcia

Leyla Jael García Castro

Abstract

Background. Since the publication of the first journal in 1662, "Philosophical Transactions of the Royal Society of London", scientific publications have been used to communicate scholarly work, including hypotheses, methods, experiments and results. Despite of the availability of electronic formats and advances on information retrieval supported by public repositories such as PubMed, scientific publications remain poorly connected to each other as well as to external resources. In fact, most of the information remains locked up in discrete documents which makes it difficult to integrate it to automatic processes and workflows.

With the continuous growth of scientific publications, more than 1.2 million articles published in PubMed during 2016, benefitting from scientific literature without a machine-processable infrastructure poses a major challenge to researchers. Finding relevant publications for a particular research topic is one of the areas where machine-processable content would make a difference. Although a list of recommended publications –i.e., related regarding their content, is offered by some repositories such as PubMed or Elsevier, no similarity score nor the terms participating in the relation are provided, making it difficult to understand how recommended articles relate to each other.

The Linked Open Data initiative together with semantic technologies provide a connectivity tissue that has not yet been fully used to support the generation of self-describing, semantic and machine-processable documents. The availability of linked data on top of the digital form currently adopted by scientific publications should facilitate knowledge retrieval, making it possible finding out relations and facts otherwise hidden or difficult to grasp. Furthermore, it should facilitate approaches working on full-text rather than just title-and-abstract.

Results. Here we present Biotea, our approach to semantically generate self-describing, machine-processable scholarly documents. We initially define a Resource Description Framework (RDF) model to integrate metadata and content from scientific publications into the Linked Open Data cloud. We enrich this infrastructure with a semantic annotation process, meaning we extract terms and expressions from the documents and connect them to ontological concepts. Our RDF model makes extensive use of existing ontologies and semantic enrichment services. We have applied our model to the full-text, open-access subset of PubMed Central.

Biolinks is built on top of Biotea. We initially propose a reclassification of the Unified Medical Language System (UMLS) semantic groups. Such reclassification is later used to semantically characterize documents as well as relations between scientific publications. A semantic model is defined for both the characterization of the similarity as well as the processes required to apply the Biolinks principles to any publication following the Journal Article Tag

Suite format or the RDF model defined by Biotea. Biolinks has been applied to a subset of documents in the TREC-05 Genomics Track collection, which have been annotated with UMLS concepts. On top of these annotated documents, we have added a distribution score according to semantic profiles.

Our models and processes are open-access and publicly available in GitHub (see https://github.com/biotea and https://github.com/ljgarcia/biotea-biolinks). The data produced by applying Biotea to PubMed Central Open Access is also public (see http://doi.org/10.5281/zenodo.376814) as well as the data generated from applying Biolinks to the TREC-05 Genomics Track Collection (see http://doi.org/10.5281/zenodo.290371).

Conclusions. The semantic processing of the biomedical literature supported by Biotea makes it possible to integrate scholarly communications to the Linked Open Data cloud. Biotea also delivers a flexible and adaptable set of tools for metadata enrichment and semantic processing of scientific publications. In such a way, Biotea provides a semantic-based scaffolding that should make it easier benefiting from the myriad of documents currently published.

Biolinks is an example of the possible benefits opened up thanks to Biotea. With the semantic characterization and similarity scores, Biolinks provides tools that make it easier to researchers to understand the general subject of a publication as well as how it relates to other publications. The weighting and similarity processes can be narrowed to a subset of the semantic groups, enabling researchers to focus on what is more relevant to them. Biolinks also contributes to understanding differences when working with only title-and-abstract versus full-text.

To sum up, Biotea together with Biolinks contribute to enable literature-based knowledge discovery from a semantic perspective.

Resumen

Motivación

Día a día, estudiantes, investigadores y académicos se enfrentan a un alto volumen de producción literaria a nivel científico. Para poder aprovechar esta información, es necesario un acceso efectivo tanto a las publicaciones como a los datos asociados. Actualmente, y desde varios años ya, las publicaciones científicas son distribuidas en repositorios electrónicos asequibles a través de Internet. Sin embargo, la mayor parte de la información contenida en dichas publicaciones, permanece oculta bajo la verbosidad del texto. Adicionalmente, los textos mismos no se encuentran debidamente interconectados entre sí ni a bases de datos especializadas.

La Web Semántica y sus tecnologías de soporte como el Resource Description Framework (RDF) y la iniciativa de Linked Open Data (LOD) ofrecen un tejido conectivo que puede facilitar la interconexión e interoperabilidad de las publicaciones científicas. El primer paso hacia publicaciones semánticas consiste en extraer los datos embebidos en el texto y hacerlos asequibles en formatos que puedan ser fácilmente procesados por computadores.

Dicho formato no debería limitarse a los metadatos como el título, la revista de publicación, los autores, entre otros, sino que debería también cubrir el contenido mismo. En particular no referimos a los términos o expresiones embebidos en el texto y su relación con conceptos ontológicos y registros en bases de datos públicas. Esta extracción de conocimiento no debería limitarse a título y resumen sino que debería ser extendida a las diferentes secciones en las cuales se describen procedimientos, métodos, resultados y discusiones. Una vez se haya construido una infraestructura semántica para publicaciones científicas, entonces será posible soportar nuevas alternativas para acceder y recuperar los datos y hechos consignados en los textos, lo cual posibilitaría la construcción y avance del conocimiento basado en literatura.

Esta tesis busca definir los formalismos y servicios necesarios para construir dicha infraestructura y posteriormente aprovecharla de tal forma que sea más sencillo encontrar literatura relacionada y relevante dentro de una investigación en particular. Esta tesis se centra en el caso de publicaciones científicas en el dominio de Ciencias Naturales. El enfoque propuesto cubre tres tópicos principales: (i) estructuración semántica de publicaciones científicas, (ii) categorización y comparación de publicaciones semánticamente estructuradas y (iii) construcción de servicios que soporten la construcción y aprovechamiento de la infraestructura definida.

Hipótesis y objetivos

La hipótesis principal bajo la cual se ha desarrollado esta tesis plantea que una infraestructura semánticamente enriquecida para publicaciones científicas mejora el acceso a publicaciones y registros tanto relacionados como relevantes.

En particular, el acceso a publicaciones similares permitiría optimizar sistemas de recuperación y recomendación lo cual sería un paso adelante hacia la construcción y avance del conocimiento basado en literatura.

Con el fin de comprobar nuestra hipótesis y desarrollar nuestro trabajo, hemos definido los siguientes objetivos:

1. Integrar las publicaciones científicas en la nube de LOD; para esto es necesario
   - Especificar un modelo semántico para representar metadatos y contenido, incluyendo secciones, subsecciones y párrafos.
   - Especificar un modelo semántico para representar términos y expresiones embebidas en el texto y relacionarlos a conceptos ontológicos.
   - Crear un proceso para convertir el texto de publicaciones científicas a los modelos semánticos definidos.
2. Clasificar publicaciones de acuerdo con un conjunto de categorías predefinidas, para ello es necesario
   - Definir un conjunto de categorías semánticas que se ajusten al dominio de Ciencias Naturales.
   - Definir un proceso que permita asignar categorías a una publicación semánticamente estructurada.
3. Encontrar similitudes entre pares de publicaciones, incluyendo
   - Definir un modelo de similitud semántico.
   - Utilizar categorías semánticas para acotar el cálculo de la similitud de acuerdo con las preferencias de los investigadores.
4. Construir un escenario donde de la categorización y similitud semántica sean utilizadas para navegar un conjunto de publicaciones previamente agrupadas por temas.

Metodología

Con el propósito de alcanzar los objetivos definidos y probar nuestra hipótesis, hemos definido el siguiente marco metodológico:

1. Inicialmente se realizará un análisis exhaustivo de las ontologías existentes para representar datos bibliográficos y contenido de publicaciones científicas. Los modelos que se definan con el propósito de estructurar las publicaciones deben reutilizar tanto como sea posible las ontologías ya existentes y deben ajustarse a los principios definido por la comunidad dentro del LOD.
2. Una vez se haya definido el modelo semántico para metadatos, contenido y datos embebidos en el texto, el siguiente paso debe ser la definición y desarrollo de un proceso automático que permita

semánticamente estructurar y enriquecer publicaciones. Este proceso permitirá construir la infraestructura sobre la cual se soportarán los siguientes pasos.

3. Para definir una caracterización semántica apropiada en el ámbito de Ciencias Naturales, se requiere analizar las formas actuales de agrupar conceptos en este dominio. En caso de ser necesario, se puede proponer una alternativa a las formas existentes. Es necesario tener en cuenta que las categorías se utilizarán para describir a un alto nivel qué tema o subdominio caracteriza la publicación; por ejemplo genética o anatomía.
4. Una vez se hayan definido las categorías semánticas, se debe construir un proceso que permita automáticamente asignar un peso a todas las categorías presentes en una publicación. Las categorías asignadas se deben utilizar además para entender mejor las diferencias de trabajar sólo con título y resumen vs. todo el texto.
5. En cuanto a la medida de similitud semántica, el primer paso consiste en analizar métricas existentes como el coseno, Best Matching y Artículos Relacionados en PubMed, y determinar cuál se adapta mejor al tipo de documentos que estamos trabajando – publicaciones científicas en Ciencias Naturales anotadas con conceptos ontológicos. Posteriormente, se utilizarán las categorías semánticas y los pesos correspondientes, pasos 3 y 4, para acotar la medida de similitud a una selección de una o más categorías semánticas. Igual que antes, se analizará la diferencia entre sólo título y resumen vs. todo el texto.
6. Con el propósito de facilitar el análisis de categorías y similitudes semánticas, se deberá definir un conjunto de herramientas visuales.
7. Finalmente, tanto la infraestructura como los servicios construidos sobre la misma deberán ser expuestos en una pequeña aplicación que permita la navegación de publicaciones.

Adicionalmente, la reproducibilidad debe ser siempre tenida en cuenta. Para ello, todos los algoritmos y modelos deben ser de acceso público.

Contribuciones

Una de las principales contribuciones de esta tesis en la definición de un modelo semántico para estructurar y enriquecer publicaciones científicas desde el punto de vista semántico, de tal forma que las publicaciones entren fácilmente a ser parte de la nube del LOD. Además del modelo, se definieron todos los algoritmos necesarios para procesar y estructurar publicaciones en lote a partir del modelo Journal Article Tag Suite. El modelo y los algoritmos[1] constituyen la base del proyecto llamado Biotea [1]. La infraestructura semántica es

[1] Mayor información sobre el modelo y los algoritmos se puede encontrar en el repositorio público GitHub (https://github.com/biotea)

considerada como el principal aporte de esta tesis ya que sirve como cimiento para el aprovechamiento semántico de publicaciones científicas. El modelo planteado puede incluso utilizarse en dominios diferentes a las Ciencias Naturales.

El proyecto Biotea fue inicialmente concebido por Alexander García Castro. Fue presentado por primera vez al público como uno de los proyectos participantes del Elsevier Grand Challenge en el año 2008 bajo el título "La historia de dos ciudades en la tierra de la serendipia: La web semántica y social se unen para dar paso a un documento vivo en Ciencias Naturales" ("A tale of two cities in the land of serendipity: The semantic web and the social web heading towards a living document in life sciences") [2]. Como parte del proyecto Biotea se procesaron artículos disponibles en el conjunto de publicaciones conocido como PubMed Central Open Access[2]. Tanto metadatos como entidades biológicas presentes en el texto fueron transformados a RDF/XML; en total se procesaron 270.834 artículos [3].

Como resultado del análisis de ontologías existentes para representar conceptos asociados a términos y expresiones embebidas en textos científicos, se llevó a cabo una colaboración con el proyecto Annotation Ontology [4]. Este proyecto ha sido liderado desde sus inicios por Paolo Ciccarese y Timothy Clark de la Escuela de Medicina de la Universidad de Harvard y el Hospital General de Massachusetts. El principal objetivo del proyecto fue la definición de una ontología para marcar anotaciones dentro de un texto. Una anotación es una porción dentro del texto posiblemente con un comentario adicional y una conexión a un concepto ontológico. Más adelante se participó en otra colaboración con el proyecto Open Annotation Data Model [5], cuyo objetivo fue agrupar y armonizar iniciativas similares al proyecto Annotation Ontology.

Con el objetivo de mostrar uno de los posibles usos de la plataforma generada gracias a Biotea, definimos y desarrollamos el proyecto Biolinks [6, 7]. Biolinks ofrece modelos y servicios de categorización y similitud semántica. La conceptualización detrás de Biolinks puede ser aplicada a cualquier dominio, sin embargo, las fórmulas han sido optimizadas para su uso en Ciencias Naturales[3]. Inicialmente analizamos las medidas de similitud existentes [8], particularmente aquellas basadas en la distribución de Poisson, Best Matching 25 y Coseno. Biolinks inicialmente adapta las categorías definidas por el Unified Medical Language System (UMLS). Dichas categorías son luego utilizadas para caracterizar y comparar publicaciones desde un punto de vista semántico.

Finalmente, realizamos un análisis sobre la colección de publicaciones en genómica TREC-05 [9, 10] y observamos las diferencias al utilizar Biolinks en sólo título y resumen versus todo el texto. Una de las principales conclusiones es que tanto la categorización como la similitud semántica pueden divergir

[2] Los archivos originales se pueden descargar de http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/ mientras que los archivos RDF generados se encuentran en http://doi.org/10.5281/zenodo.376814

[3] Los algoritmos y modelos se encuentran disponibles en https://github.com/ljgarcia/biotea-biolinks. Una versión congelada también está disponible en http://doi.org/10.5281/zenodo.290371

substancialmente de uno caso al otro; es por esto que recomendamos utilizar texto completo siempre que sea posible. PubMed, uno de los repositorios más extensos y más usados en Ciencias Naturales, utiliza sólo el título y resumen para recomendar publicaciones similares. La baja disponibilidad de texto completo junto con el alto costo y tiempo de su procesamiento son razones comúnmente citadas en contra de análisis sobre texto completo. Sin embargo, existe un movimiento dentro de la comunidad científica hacia una apertura total, donde el texto completo juega un papel protagónico. Esta apertura en conjunción con la capacidad de procesamiento y rapidez siempre en aumento de los computadores, ofrecen un panorama positivo para los análisis basados en texto completo. Más aún, amplían las posibilidades de avance del conocimiento basado en literatura.

Conclusiones y trabajo futuro

Biotea define una estructura semántica de publicaciones científicas incluyendo metadatos, contenido y datos extraídos del texto. Esta estructura constituye la base sobre la cual se hace posible la interconectividad e interoperabilidad no sólo entre publicaciones sino también de recursos, ontologías y bases de datos públicas. De esta forma Biotea aumenta el valor de la literatura científica ya que las publicaciones se hacen más útiles al estar interconectadas La disponibilidad electrónica que actualmente soportan la mayoría, si no todos, los repositorios de publicaciones científicas, debería facilitar no sólo la recuperación de información sino también la recuperación de relaciones y hechos consignados en la literatura [11]; haciendo de esta forma posible el avance del conocimiento a través de métodos basados en literatura. Biotea es nuestro grano de arena.

Biotea es el primer paso hacia un acceso semánticamente normalizado en cuanto a publicaciones científicas se refiere. Los siguientes pasos deberían utilizar y aprovechar la estructura semántica provista por Biotea. Al respecto, nuestra contribución es Biolinks. Biolinks define formulas, modelos, procedimientos y herramientas visuales para categorizar y comparar publicaciones. Biolinks reorganiza los grupos semánticos de UMLS; sin embargo, diferentes reagrupamientos se pueden configurar sin necesidad de cambiar los algoritmos de categorización y comparación semántica.

Un posible escenario en el cual se podrían utilizar Biotea y Biolinks consiste en ofrecer una experiencia enriquecida al usuario de la interfaz web del repositorio PubMed. En particular nos referimos a las listas de artículos recomendados. Estas listas enumeran los artículos de acuerdo con su similitud con respecto al documento desplegado para lectura. La similitud se calcula con base en palabras encontradas en el título y resumen así como también términos del vocabulario Medical Subject Headings (MeSH) asignados al documento. La similitud semántica calculada por Biolinks se utilizaría para reorganizar la lista de acuerdo con términos identificados en todo el texto. La categorización semántica se utilizaría para presentar una idea global de los temas más representativos dentro de la lista de artículos similares. De esta forma, los investigadores y académicos se pueden concentrar en aquellos temas de su interés que cuenten con una buena representación dentro de la colección de publicaciones analizada.

Las anotaciones identificadas gracias a Biotea se pueden utilizar también de forma independiente. Diferentes autores en el dominio biomédico se han enfocado en encontrar relaciones implícitas a partir de anotaciones semánticas como las soportadas por Biotea. Por ejemplo, se han utilizado términos acuñados en MeSH para identificar patrones que permitan seleccionar candidatos para nuevas asociaciones entre medicamentos y enfermedades [12]. En un estudio similar, se identificaron términos de la Gene Ontology (GO) que aparecieran en el texto en forma conjunta con nombres de genes [13]. Las anotaciones con conceptos de la GO que son compartidas por varios genes también se han utilizado para identificar posibles relaciones entre dichos genes [14, 15].

Nuestros resultados muestran la utilidad al usar anotaciones semánticas para caracterizar un conjunto predefinido de artículos, como lo son los artículos de TREC-05. Esta caracterización puede ser utilizada para analizar similitud entre documentos desde una perspectiva global o acotada a ciertas categorías. De esta forma, los investigadores se pueden enfocar en aquellos artículos que son más similares dentro de los temas más relevantes para su investigación en particular.

Dentro de nuestro trabajo futuro se encuentra un análisis en profundidad de la variación de la similitud cuando se acota a categorías específicas. Con este propósito, proponemos trabajar con un experto de dominio en el campo de protocolos de laboratorio, posiblemente con una ontología específica para este tipo de documentos.

También dentro del trabajo futuro se encuentra la extensión de nuestras herramientas de visualización y análisis y la construcción de un prototipo más amplio y más robusto. Nuestro objetivo final es contribuir con un navegador basado en conceptos para publicaciones científicas, donde los investigadores puedan encontrar el texto pero también los datos y hechos detrás del mismo. De esta forma, queremos enriquecer la experiencia de lectura y facilitar el análisis, exploración y avance del conocimiento basado en literatura.

Acknowledgements

First of all, I would like to express my deepest gratitude to my supervisor Rafael Berlanga Llavori who accepted me as an external PhD student. He has been always very supportive and patient and has kindly and assertively contributed and supervised my scientific work in order to make it better and stronger. Despite the distance, he always found the time to guide me. I truly appreciate the opportunity he offered me when he accepted me in his group.

My deepest gratitude as well to Alexander García, my brother, who introduced me to the semantic web world. He has co-supervised my research since before it turned out to be a PhD quest. It has not been easy for him as our characters cannot be more different. However, he has taken the time and care to revise my work and contribute with valuables ideas and comments. Thanks to his critics my papers have become into a more readable shape.

I want to thank to Dietrich Rebholz-Schuhmann who introduced me to Rafael Berlanga. He showed me a window when a door had just closed. My sincere gratitude to my colleagues in the TKBG group; although we did not interact much due to the distance, I benefited from their previous work and experience. I hope current and future members could benefit as well from the outcomes of my research. I also want to thank to people in the e-Business group at Universität der Bundeswehr who introduced me to SPARQL and other Semantic Web technologies, as well as to the iDigInfo group at The Florida State University who initially hosted the first version of Biotea dataset. Many thanks as well to the experts who review my thesis, Dietrich Rebholz-Schuhmann and José Francisco Aldana Montes, they contributed with valuable comments to improve this document.

Finally, thanks to my mother, who always supports and encourages me. She knows how much I have struggled to get here. She has patiently listened every time even when I have lost my temper (so sorry about it!). She is a rock, the greatest supporter I could ever have.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

1.1 Motivation

Scientists face a continuously increasing amount of scientific production; therefore, effective ways to access scientific publications and related data plays a major role in the scientific process. Scholarly web-based repositories have succeed as a dissemination platform for scientific publications; however, most of the information remains locked up in discrete documents, which are poorly interconnected to each other as well as to relevant databases.

The Semantic Web (SW) and its technologies, e.g., the Resource Description Framework (RDF) and the Linked Open Data (LOD) initiative, offer a connectivity tissue that could make it easier for scientific publication to become interconnected and interoperable. The first step towards semantically aware publications consists in providing the data embedded in the text in a machine-processable format. Such a format should cover not only metadata such as title, authors and journals, but also the content itself as well as references to database entries and recognized ontological entities. Furthermore, embedded data should be processed in full-text, rather than be limited to the title-and-abstract only. Once a highly interconnected infrastructure for scientific publications has been built, new alternatives regarding information retrieval become possible, leading to literature-based knowledge discovery [16, 17].

This thesis aims to introduce formalisms and services to facilitate the integration of the scientific literature into the Life Sciences (LS) information structure as well as scenarios where this integration plays a major role. Thus, this thesis focuses on three main topics: (i) structuring scientific publications from a semantic perspective, (ii) categorization and comparison of structured scientific publications, and (iii) definition of recommendation and retrieval services on top of such an infrastructure.

1.2 Hypothesis and objectives

The main hypothesis underlying this thesis states that a semantically enriched literature infrastructure can improve access to relevant and related data. Particularly, some of that data can be in the form of additional literature, making it possible enhanced recommendation systems and thus enabling literature-based knowledge discovery.

In order to test such a hypothesis, we have defined the following goals and sub-goals.

1. Integrating scientific publications into the LOD cloud
    - Specifying a semantic model to represent metadata and content for scientific publications
    - Specifying a semantic model to represent database entries and ontological entities embedded in the text of scientific publications
    - Creating a pipeline to integrate semantically structured publications into the LOD cloud
2. Categorizing scientific publications according to the some pre-defined semantic categories
    - Defining a set of semantic categories
    - Defining a category weighting model for semantically structured publications
    - Automatically assigning categories to semantically structured publications
3. Finding similarities across publications
    - Defining a similarity scoring model for semantically structured publications
    - Using semantic categories to narrow such a similarity model according to possible users' research interests
4. Showcasing a scenario where semantic categories and similarity scores are used to facilitate the navigation across sets of related articles.
5. Evaluating the defined set of semantic categories, as well as the category weighting and similarity scoring models using metrics defined in the literature.

1.3 Methodology

Aligned with the hypothesis and objectives of this thesis, the operational methodology can be summarized in the following points.

1. It all starts with a thorough analysis on existing ontologies to represent publications and their content. Such analysis should also consider interconnecting publications to each other as well as to external resources. As much as possible, existing vocabularies should be used and LOD principles should be observed.
2. Once the semantic model has been defined, the next step is designing and developing a pipeline to process publications in batch so a semantic infrastructure will be created. This infrastructure will be the scaffolding for the next steps. In this way, we aim to integrate

the scientific publications into the LOD, not just metadata but also content and entities recognized within the text.

3. Semantic categories aim to assign a topic to a publication based on the data embedded in the text. In this regards, an analysis on current trends used in the LS domain should be performed. If necessary, alternatives should be proposed. Proposed alternatives will aim a better granularity and semantic coherence. An evaluation against a baseline using an Information Gain (IG) approach will be performed. Differences between using only title-and-abstract against using the full-content should be analyzed.
4. Once the semantic categories have been coined, it is necessary to define formulas to weight the categories present in a particular publication. Formulas should use the data provided by the scientific publications semantic infrastructure. The category weighting model will be evaluated against a baseline; the model will be tuned so that all of the possible categories will be given a fair chance. In this way, generic categories will not have a much higher impact than categories with specialized terms.
5. The weighting formulas defined in the previous step, will be automated via algorithms that will allow batch processing so documents in a large corpus can be categorized according to the categories defined on the 4th step of this list.
6. In order to find similarities across publications, a detailed analysis of similarity metrics is required. Such analysis should evaluate what the best approach for the data extracted from the publications is. Well known similarity metrics such as Best Matching 25 (BM25), Cosine similarity and PubMed Related Articles (PMRA) will be used for comparison and tuning. The similarity approach should take advantage of the semantic categories so publications can be compared to each other regarding all –global similarity, or just some pre-selected categories –group-narrowed similarity. Again, differences regarding title-and-abstract vs. full content should be carefully analyzed.
7. The global and group-narrowed similarity formulas will be integrated to the batch process. The same corpus of documents categorized according to our model, will all be processed with the similarity formulas. Rather than comparing all against all, we will use predefined subsets of documents sharing some characteristics. For instance, a subset could refer to "Huntington's disease" while another to "Mad cow disease". Subsets will be defined taking into account the analyzed corpus.
8. In order to facilitate analysis on the assigned categories and similarity scores, a visual representation should be defined. Visual components should observe some good practices for web-based visualizations.

9. After creating a semantic infrastructure for a publicly available set of publications, evaluated and tuned models –semantic groups, category weighting model and similarity scoring model, should be showcased. This showcase aims to be a proof of concept supporting the hypothesis stated in this thesis.

Additional to the listed steps, reproducibility should be always considered. Therefore, algorithms and dataset should be provided in a public and persistent space.

1.4 Contributions

One of the main contributions of this thesis is the definition of a semantic model to represent metadata and content for scientific publications, making it easier to include publication in the LOD cloud. Additional to the model, algorithms to transform publications available in the Journal Article Tag Suite (JATS) to RDF were also defined. These models and algorithms[4] constitute to the first part of this thesis and correspond to the project named Biotea.[1]. Biotea was initially conceived by Alexander Garcia Castro and was first introduced as one of the projects participating in the Elsevier Grand Challenge in 2008 –"A tale of two cities in the land of serendipity: The semantic web and the social web heading towards a living document in life sciences" [2].

Biotea includes not only metadata such as authors, journal and publication dates, but also the full-text as well as entities extracted from it. The extracted entities are mapped to well-known ontologies in the Life Sciences domain and modelled as annotations. In order to semantically represent such annotations, we collaborate with the Annotation Ontology project [4]. This project has been led by Paolo Ciccarese and Timothy Clark from the Medicine School in the Harvard University and the Massachusetts General Hospital. The Annotation Ontology defines a model to represent annotations that can be or not semantic, i.e., linked to ontological concepts. Later, we also collaborate to the Open Annotation Data Model [5] whose purpose was harmonizing and merging together different efforts to model annotations.

The semantic infrastructure for PubMed Central Open Access (PMC-OA) together with the enriched content in form of annotations[5] is probably the most important contribution of this thesis as it acts as a scaffolding required to improve access to relevant and related data. A total of 270834 publications were processed and transformed to RDF/XML [3]. Our RDFization process transcends the LS domain as it can be applied to scientific publications on any domain. In order to show how the access to relevant data can be improved

[4] More information regarding the models and algorithms are available in GitHub (https://github.com/biotea)

[5] Open Access files provided by PubMed Central can be downloaded at http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/ while the corresponding RDF/XML files can be found at http://doi.org/10.5281/zenodo.376814

once the publications are part of the LOD, we focused on finding additional literature related to an already known publication, a reference article. Biolinks is our contribution in this regard [6, 7], which constitutes the second part of this thesis.

Biolinks provides models and algorithms[6] making it possible to (i) semantically categorize a publication with respect to a predefined set of semantic categories or groups, and (ii) find the semantic similarity between two publications annotated with the Biotea model. Biolinks defines a set of semantic groups based on those provided by the Unified Medical Language System (UMLS). In order to regroup the UMLS semantic groups, we analyze the semantic coherence regarding the **is-a** hierarchy and the UMLS groups co-occurrence in a well-known collection in Life Sciences, the Genomic Track collection from the 2005 Text Retrieval Conference (TREC-05).

Biolinks proposes a formula to weight the presence of the predefined set of semantic groups in a publication, making it possible to get an overview of the groups' distribution. Biolinks also provides a similarity formula. The similarity between two publications can be found based on all the annotations or can be narrowed to a subset of the predefined semantic groups, making it easier to focus only on those groups well-represented in the publications or those more interested for a researcher. Different similarity formulas were initially assessed, including Poisson distribution, Best Matching and Cosine [8].

Finally, we showcase the use of Biolinks group-base distribution and similarity formulas on the TREC-05 collection [9, 10]. The main contribution of Biolinks is the analysis of the differences when working only with title-and-abstract versus full-text. Our results indicate that both the group-based distribution and the semantic similarity vary, therefore we recommend using full-text as much as possible. We recognize the difficulty of finding and processing full-text; however, more and more open access to full-text is being promoted by researchers, publishers and governments. This openness together with the continuous improvement of storage and processing capabilities facilitate working with full-text rather than just title-and-abstract, which constitutes a step forward to the literature-based knowledge discovery.

1.5 Outline of this thesis

The research context is presented in Chapter 2, there we cover topics corresponding to the general background used in our investigation. Concepts such as ontologies, Semantic Web, Linked Open Data, and scientific literature in the LS domain are introduced.

The first part, namely Biotea, is covered in Chapter 3 and Chapter 4. Biotea relates to formal models for representing metadata, content, and annotations

[6] Algorithms and models are available at https://github.com/ljgarcia/biotea-biolinks. A frozen version can also be found at http://doi.org/10.5281/zenodo.290371

on scientific publications. Initially, in Chapter 3, we present the models to structure metadata and content following well-known ontologies for bibliographic metadata representation as well as others of general purpose. In Chapter 4, we present the enrichment process by identifying relevant terms linked to biological entities in different bio-medical ontologies. The aim of our annotation model is to facilitate the interoperability across publications and external related databases.

The second part, namely Biolinks, is developed in Chapter 5 to Chapter 8. Biolinks uses structured knowledge from scientific publications in order to support semantic literature exploration. We first create a Biotea-like semantic infrastructure for datasets extracted from the TREC-05 collection, using UMLS to annotate the publications contained in the collection. We then use this dataset to propose a variation to the UMLS semantic groups, a similarity metric and a group-based categorization for scientific publications. In this part we also include a showcase where we demonstrate how the semantic infrastructure plus categorization and similarity services can be used to improve recommendation systems.

We finalize this document with conclusions and future work. To ensure the reproducibility of science, we envision publications providing access to raw data as well as to machine-processable descriptions of methodologies, experimental protocols, results, etc. In such a way, it becomes possible for scientific publication semantically link to each other as well as to other resources; thus, facilitating the interoperability and interconnectedness by having literature data fully immersed in the LOD cloud.

Chapter 2

Research context

2.1 Ontologies

The word “Ontology” has its roots in philosophy where it refers to the study of those attributes and characteristics that constitute the very nature of a thing that is. The term was first used in Computer Science in the 1970’s within the Artificial Intelligence community. Ontologies were proposed as computational models enabling automated reasoning. Later, in the 1990’s, ontology became a fashion word inside the Knowledge Engineering community when building ontologies was considered an innovative idea [18]. The research community realized about the possibilities offered by ontologies and the machine-enable-understanding of them; then, communities of ontology users, domain experts, and ontology engineers appeared.

Around 1993 Gruber defined an ontology as an “explicit specification of a conceptualization” [19]. In 1997, Borst added a key word “shared” to the definition, stating that ontologies are “formal specification of a shared conceptualization” [20]. Both definitions were later merged by so the definition became “an ontology is a formal, explicit specification of a shared conceptualization” [21]. An explanation on the different key points in these definitions, i.e., conceptualization, formal and explicit, and shared, were offered in detail by Guarino, Oberle & Staab [22]. At that point, Ontology Engineering was an incipient scientific discipline looking for “agreements upon standards for protocols to achieve the unified and consistent progression of innovation and knowledge” [23, 24]. Those agreements arose with the appearance and use of methodologies and tools, and the active participation of the community involved so the specifications were indeed shared.

Conceptualizations can be used, for instance, to describe the academic aspect of universities, e.g., lecturers, students, courses and interactions between them. A conceptualization is particular to a domain, e.g., management, and to a purpose. For instance, **"lecturer"** , **"student"** , and **"course"** would be entities observed in a university while **"lecturer" → teaches → "course"** and **"student" → attends → "course"** would be relations, i.e., possible interactions between entities. Instances, i.e., individuals belonging to a particular type of entity, are also possible –e.g., **"Andreas" → is-a → "lecturer"** . Both domain and purpose can be broader or narrower depending on what a community wants to achieve with such a specification.

Different levels of semantics are also possible, from informal to formal models, from weak to strong semantics (see Figure 1). At the beginning of the spectrum we have glossaries and data dictionaries while at the end we have ontologies [25]. Ontologies have strong semantics as they are expressed by using "logical languages that allow specifying rigorously formalized logical theories" [22]. Those logical languages make it easier to specify axioms stating symmetry and transitivity for the relations between entities. For instance, if "is-sibling-of" is stated as symmetric and transitive, asserting that "*Eloisa*" → is-sibling-of → "*Esteban*" and "*Esteban*" → is-sibling-of → "*Nicholas*" would be enough to infer that "*Esteban*" → is-sibling-of → "*Eloisa*" due to symmetry, and that "*Eloisa*" → is-sibling-of → "*Nicholas*" due to transitivity. Thanks to the reasoning and inferring possibilities behind ontologies, they became a cornerstone for the Semantic Web.

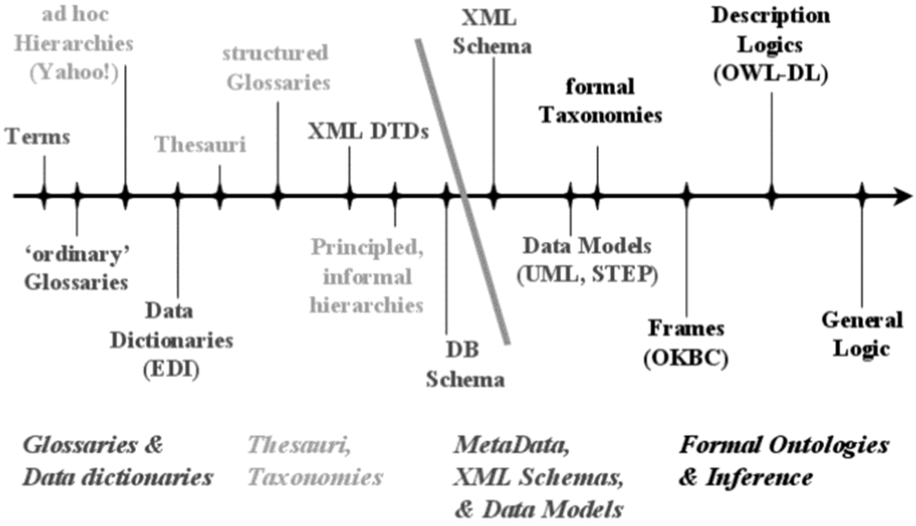


Figure 1. Different levels of semantics, from weak to strong, from informal to formal. Taken from [25].

2.2 Semantic Web

Less than 30 years ago, in 1989, Sir Tim Berners-Lee invented the World Wide Web (WWW). Also known as the Web 1.0, it introduced hyperlinks so documents could link to each other and users could navigate from document to document regardless where they were actually stored. About 10 years later, the Web 2.0 emerged, the Web was no longer about just document but people generating and sharing content. Two years later, in 2001, Sir Berners-Lee introduced the concept of the Semantic Web (SW), referred as well as sometimes as the Web 3.0, the Web of Data or the Web of Things. The SW is an extension of the current Web where data embedded in it is well-defined and linked to each other, in such a way that machines can understand it and process it so automated services can be built on top of it [26].

The SW takes advantage of structured collections of data together with a set of reasoning and inference rules on top of that data. However, it should not be dependent of the structure; it is the meaning, the semantics, the key issue here.

As stated by Sir Berners-Lee, a vast amount of information can be expressed along the lines of "a hex-head bolt is a type of machine bolt"; however, the SW technologies should enable for more complex relationships. Features such as reasoning and inference should facilitate software to move create logic paths based on descriptive information. Along with ontologies, some other technologies are fundamental to the SW: eXtensible markup language (XML), the Resource Description Framework (RDF) [27] and the Web Ontology Language (OWL).

XML makes it possible for users to add a structure layer to describe data, for instance **<Student id=" 1234" ><Name>Esteban</Name></Student>**. However it does not allow us to express what such a structure means, i.e., no reasoning or inference is yet possible. RDF is described as a "standard model for data interchange on the Web" holding a stronger semantic representation than other standards such as XML. It comprises a set of specifications that make it easier to model resources on the Web. Particularly, it provides a data model for representing graph data structures, making it ideal to express statements in the form subject-predicate/property-object, e.g., **"Ivan → is son of → Jazmin"** .

Data described following the RDF model can be serialized in XML but also in some other flavors such as Turtle, N-triples and JavaScript Object Notation-Linked Data (JSON-LD). Subject and object are identified by Internationalized Resource Identifiers (IRIs), usually in the form of Universal Resource Identifiers (URIs). Identical entities, either subject or object, can be given different identifiers in different data model representations. In order to make it possible for machines to realize that two entities with different identifiers refer to the same thing, we need ontologies. OWL comprises a set of knowledge representation languages for representing ontologies. OWL languages are characterized by formal semantics, i.e., they allow reasoning and inference.

SW technologies offer a connectivity tissue that facilitates data integration and interoperability across multiple sources. However, such connectivity is only possible when resources link to each other in a decentralized, open and shareable way.

2.3 Linked Data

While the Web 1.0 is about documents and links connecting them, all of it intended to be consumed by humans, the Web 3.0 is about data to be processed by machines. However isolated data offers no added value, in order to facilitate knowledge discovery, data should be connected. This basic notion is the idea behind Linked Data (LD) [28]. In 2006, Sir Berners-Lee defined the four principles that LD should follow in order to make the most of it. These four principles are:

- Use URIs as names for things

- Use Hypertext Transfer Protocol (HTTP) URIs so that people can look up those names
- When someone looks up a URI, provide useful information, using the standards such as RDF and the SPARQL Protocol and RDF Query Language (SPARQL)
- Include links to other URIs, so that they can discover more things.

While RDF allows representing the semantics behind data by using a direct and labeled graph, SPARQL allows querying that data. SPARQL makes it possible to query for triple patterns that can be defined by using disjunction, conjunctions, and optional patterns. The current version is SPARQL 1.1., which introduced federated queries, i.e., queries that access, join and retrieve data from multiple data sources [29].

LD particularly refers to data published on the Web, with its meaning explicitly defined and linked to other external datasets [30]. All that inter-linked data constitutes what is known as the Web of Data –although sometimes also used as a synonym of SW. In the Web of Data, there are no restrictions regarding the type of data or the authoring of such data. Data publishers can select those vocabularies, i.e., ontologies, which better suit their requirements. As graphs representing different datasets are linked to each other, a global graph emerges, linking datasets and making it possible to discover new sources of data.

The LOD initiative adds a fifth principle: release your data under an open license so it can be reused for free. In 2010, the five-star rating system was proposed by Sir Berners-Lee, in which LOD should aim to fulfill the five stars:

1. ★ Available on the web (whatever format) but with an open licence, to be Open Data
2. ★★ Available as machine-readable structured data (e.g. spreadsheets instead of image scan of a table)
3. ★★★ As (2) plus non-proprietary format (e.g. CSV instead of spreadsheets)
4. ★★★★ All the above plus using open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff
5. ★★★★★ All the above plus linking your data to other people's data to provide context.

Given the advantages delivered by open shared data, the LOD was initially adopted by researchers, developers in university labs and small companies [30]. Since its inception, the project has grown considerably (see Figure 2).

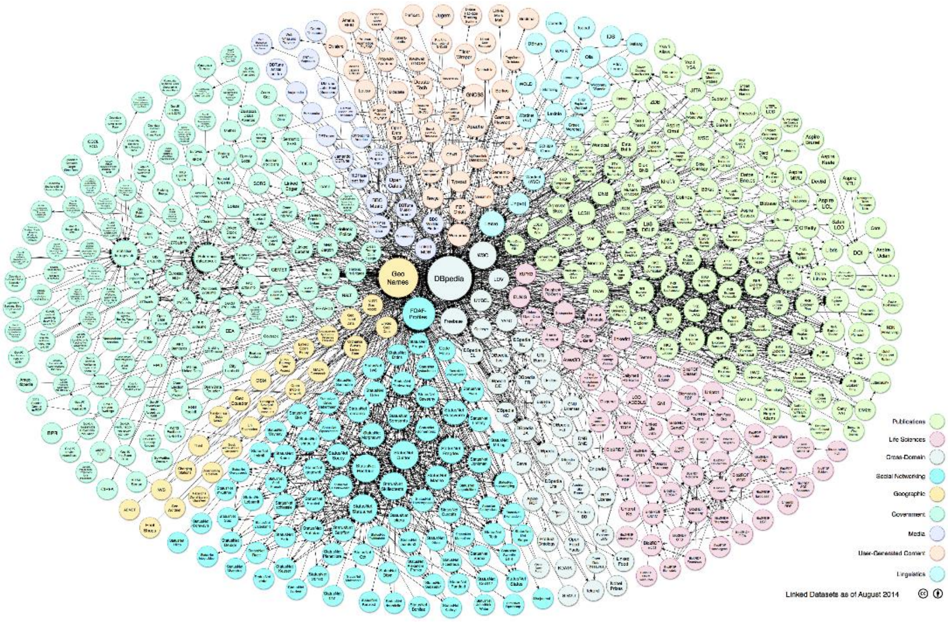


Figure 2. LOD cloud as for August 2014. Taken from [31].

2.4 Linked Open Data in Life Sciences

LS and health care domains comprise more than 1500 public databases potentially overlapping to each other [32]. Data in the biomedical domain has continuously increased in the last years partially due to the advent of omics studies, e.g., genomics, proteomics and metabolomics, which in turn have been boosted thanks to the improvement of high-throughput gene sequencing technologies [33]. As observed in Figure 2, a considerable portion of the LOD cloud corresponds to LS data. Some of the most significant contributors include Bio2RDF [34-36], Linked Life Data [37] and the EBI-RDF platform [38].

Due to the complexity and descriptive nature of LS and health care data as well as the cross-referencing mechanisms commonly used in this domain, SW technologies provide an ideal mechanism for the representation and integration of LS data [32]. In fact, LS data becomes a natural test-bed for SW technologies such as RDF, SPARQL and OWL [39]. Research in LS frequently require understanding data at different levels, e.g., from cells to biological systems, potentially involving different species, e.g. orthologs, under diverse experimental conditions. The biology behind research questions is intrinsically connected; however, as data is distributed across multiple databases, multiple queries should be placed and results should be collated in a coherent manner [38]. SW technologies such as LOD could facilitate both queries and collation, thus contribute to improve biological research.

Although RDF enables interoperability across heterogeneous databases, in order to take advantage of the LOD in LS, biologists, bioinformaticians, health care professionals and researchers need to learn SPARQL. Furthermore, they need to understand the data models behind the different datasets they might need for their research. In order to overcome this practical difficulty, different efforts have been developed in recent years. For instance, SPARQLGraph [40] builds on top of Bio2RDF and the EBI-RDF platform introducing a graphical query builder. The visual graph is then translated into a query and executed on a public SPARQL endpoint. In a similar vein, cMapper [41] uses the EBI-RDF data. It offers a graphical gene-centric interface that enable users to find connections to genes or small molecules relevant to their research. Different from SPARQLGraph, cMapper relies on a relational data model rather than SPARQL endpoints.

Rather than facilitating the access to LOD, other semantic projects in LS aim to target a specific population, offering tailored datasets. For instance, DisGeNET-RDF [42] supports knowledge related to the genetic basis of human diseases. PubChem [43] provides a public repository for information on chemical substances and their biological activities while kPath [44] integrates information related to metabolic pathways. kPath also provides an interface for browsing pathways and building metabolomics networks.

2.5 Scientific literature in Life Sciences

LS have developed into a data driven science where research is grounded on knowledge provided by fact repositories and databases such as GenBank [45], the Universal Protein Resource Knowledge Base (UniProtKB) [46], or the Kyoto Encyclopedia of Genes and Genomes (KEGG) [47]. Data hosted by those repositories is used not only to assess and verify researchers' domain knowledge but also to propose and develop novel hypotheses. Research processes and findings are commonly reported in scientific publications, which in turn should nurture back fact repositories and databases with novel and up-to-date knowledge. However, due to the fast pacing in scientific literature production, facts extraction from text-based publications is not a fully automated process.

LS databases commonly include curated knowledge, i.e., facts carefully reviewed and assessed by a group of domain experts as an effort to produce high quality and standardized data. Automated processes producing data based on heuristics and statistics are also incorporated in order to keep up with the myriad of data produced from omics studies. Efforts related to text mining and automatic entity extraction from scientific publications aim to fill the gap between the knowledge reported as text, a semi-structured and non-directly machine-processable format, and the structured data used recorded in scientific databases.

Text-mining and entity extraction are challenging processes when used in the LS domain literature. On one hand, the language used in scientific literature is not necessarily aligned with database standards. On the other hand, databases are not complete as new knowledge is continuously produced. For instance, it is possible to find novel results reported in literature regarding rare genetic mutations or new gene candidates.

Once structured data has been extracted from the literature, it is still necessary to follow an entity disambiguation process. Extracted entities should be semantically aligned to concepts coined in ontologies and supported in databases. Entities extracted and disambiguated from publications include genes, proteins, chemical entities, gene ontology terms, drugs, diseases, phenotypes, tissues, amongst others. Well-known ontologies exist for most, if not all, of those terms in LS. As ontologies are a fundamental part of the backbone supporting SW and LOD, once the literature-based data has been disambiguated it becomes possible to integrate it into the LOD cloud. However, in order to achieve this integration, metadata in scientific publications and extracted entities need to be put together in a semantic-based infrastructure. Such an infrastructure is the first subject to be addressed by this thesis.

PART I - BIOTEA

SEMANTIC SCAFFOLDING FOR
SCIENTIFIC PUBLICATIONS

Biotea delivers a semantically enriched
infrastructure aiming to facilitate
literature-base knowledge discovery.

Chapter 3

A semantic infrastructure for scientific publications

3.1 Background

Over 350 years ago, the first scholarly journals were published: Philosophical Transactions of the Royal Society (of London) and the Journal de Sçavans. Since then, scientific papers have been used to communicate scientific activities, from hypotheses to novel protocols and developments [48]. Technological advances have made it possible to move from paper-based dissemination channels to electronic formats, giving way to digital libraries and repositories providing search, retrieval and filter capabilities.

In the Life Sciences (LS) domain, PubMed is one of the biggest public repositories for scientific publications. PubMed is a service of the US National Library of Medicine providing free access to abstracts in a variety of fields including medicine, nursery, dentistry, veterinary, chemistry, bioinformatics and health care. It was developed by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM). New citations are added on a daily basis; for instance, on the 9th of September of 2016, 2190 publications were added. PubMed Central Data (PMC) [49] is a free archive of LS journal literature, with a particular focus on biomedicine, supported by the U.S. National Institutes of Health's National Library of Medicine (NIH/NLM). The PubMed Central Open Access (PMC-OA) is a subset of PMC including about 1 million articles distributed across 2700 journals, as of 2015. PMC-OA includes full-text open-access articles, which are still protected by copyright but are also available under the Creative Commons license, i.e., a more liberal redistribution is allowed. Articles are available as eXtensible Markup Language (XML) files downloadable via File Transfer Protocol (FTP).

Supporting XML, a semi-structure format, is a step forward in digital libraries as it facilitates automated processes on the text; however it still lacks two of the five stars suggested by the Linked Open Data (LOD) principles. It does not use W3C standards such as the Resource Description Framework (RDF) and SPARQL Protocol and RDF Query Language (SPARQL) to identify things and it does not support connectivity between scientific publications or to other data. Semantic Digital Libraries (SDLs) use semantic technologies in order to provide uniform access to metadata as well as machine-processable content; in such a way, SDLs intend to better support information retrieval and classification tasks [50, 51]. Within the context of SDLs, ontologies can be used to: (i) organize bibliographic descriptions, (ii) represent and expose document contents, and (iii) share knowledge amongst users [51].

For instance, the Bricks project [52] aims to integrate existing digital resources into a shared digital memory. It relies on Web ontology language-Description Login (OWL-DL) in order to support, organize, and manage metadata. Other projects supporting content representation and classification tasks will be discussed in the next chapter.

In an effort to add value to the content of scientific publications, publishers are actively improving programmatic access to their products. For instance, in 2012, the Nature Publishing Group (NPG) RDF released more than 270 million of RDF statements [53], covering primary metadata for more than 450,000 articles published by NPG since 1869. In that first release, the dataset included basic citation information (title, author, publication date, etc.), identifiers, and Medical Subject Headings (MeSH) terms. NPG RDF data model makes use of vocabularies such as the Bibliographic Ontology (BIBO) [54], Dublin Core Metadata Initiative elements (DCMI-elements) [55, 56], Friend of a Friend (FOAF) [57, 58], and the Publishing Requirements for Industry Standard Metadata (PRISM) [59] as well as ontologies that are specific to NPG [53]. PRISM is an XML specification that defines a controlled vocabulary for managing, aggregating, and publishing content. The data model for the NPG RDF data[7] [60] is presented in Figure 3; as for September 2016, the latest available release corresponded to the 24th of August on 2015.

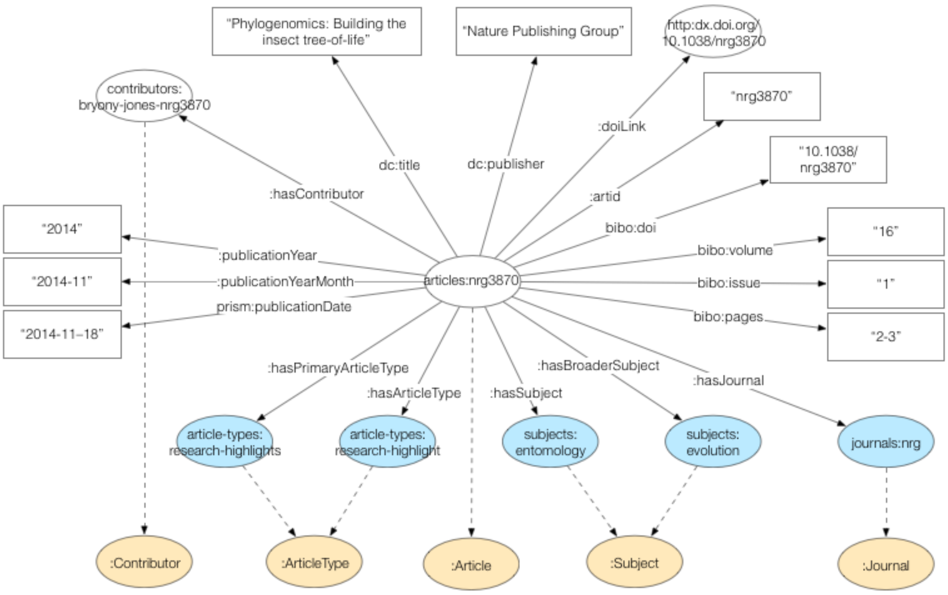


Figure 3. Data model for NPG RDF[8].

In this chapter, we present our knowledge model for metadata in scientific publications as well as our RDFization process. We aim at delivering

[7] NPG RDF data is freely available at http://www.nature.com/ontologies

[8] Figure taken from http://www.nature.com/ontologies/datasets/articles.

interoperable, interlinked, and self-describing documents in the biomedical domain. We applied our approach to the full-text, open-access subset of PMC, i.e., PMC-OA. In our approach, existing ontologies are brought together in order to facilitate the representation of sections in scientific literature.

3.2 Methodological approach

3.2.1 Model

Following the LOD principles, our model reuses as much as possible existing vocabularies. We mainly use the Bibliographic Ontology (BIBO) and the Dublin Core Metadata Initiative terms (DCMI-terms) to model bibliographic metadata. BIBO reuses concepts from DCMI-terms and PRISM –the latter one of the ontologies used as well in the NPG RDF dataset. We also use the Provenance Ontology (PROV-O) [61], mainly for attribution to original sources as well as serialization agents and dates. PROV-O is a specification from the World Wide Web Consortium (W3C); W3C is the main international organization for standards in the World Wide Web (WWW). PROV-O provides classes and properties to represent and interchange provenance data. We also use the Vocabulary of Interlinked Datasets (VoID) [62], a vocabulary for expressing metadata about RDF datasets. By using VoID, we can provide information about our dataset such as version and provenance.

BIBO reuses DCMI-terms in two ways. It includes some properties similar to those found in DCMI-terms but for some other properties it inherits from DCMI-terms. We use bibo:pmid and bibo:doi as publication identifiers. We also use dcterms:identifier because, being a domain-independent property, it is widely used by other existing RDF datasets –e.g., NPG RDF dataset; thus, it facilitates compatibility. Dcterms:identifier is also used for the PMC identifier as BIBO only support PubMed identifier (PMID) and the Digital Object Identifier (DOI). PROV-O also includes some similar properties to DCMI-terms. For instance, prov:wasAttributedTo is similar to dcterms:creator, and prov:generatedAtTime is similar to dcterms:created. For the compatibility reasons already stated, we include both PROV-O and DCMI-terms properties.

In BIBO, authors are modeled as items in either an rdf:List or an rdf:Seq, depending whether or not the order should be kept. In both cases, the range is defined as an rdfs:Resource [63]. Modeling authors as items makes it possible to represent single authors as resources rather than plain text. We then use FOAF [58] to model authors as well as their affiliations as resources. Particularly, we use foaf:Person and foaf:Organization. FOAF provides a set of classes and properties to represent people and their connections to other people, organizations, and other resources, e.g., publications. FOAF integrates information related to social networks. Such networking is also identifiable in publications; authors collaborate with co-authors and are affiliated to organizations. Although authors could be represented as dcterms:Agent, we have chosen FOAF as it is more detailed and explicit. It includes elements such as first and last name, institution to which authors belong, personal homepage and email account.

With BIBO, PROV-O, DCMI-terms and FOAF we are covering the metadata describing the paper, including the title-and-abstract as well as references. PMC-OA articles are modeled as bibo:AcademicArticle. Whenever a bibliographic reference includes a PMID, we use again bibo:AcademicArticle. Patents, books, book chapters and other are also specified by BIBO. However, when it is not possible to identify a particular type of publication, we use bibo:Document, a broader classification for publications. dcterms:publisher is used to link to the publisher while BIBO is used to reference both the journal and the publisher. The journal is modeled as a bibo:journal and the publisher as a foaf:Organization. Data corresponding to the International Standard Serial Number (ISSN), volume, issue, and starting and ending pages are also recorded. Authors are modeled as a bibo:authorList, where each member is a foaf:Person. Abstract and sections are modeled as a doco:Section while the actual text is modeled as an rdf:value. The references are modeled as bibo:Document; the relations used from main article to references is bibo:cites while bibo:citedBy is used in the opposite direction . References are available at both document and section levels. A summary of our model is presented in Figure 4.

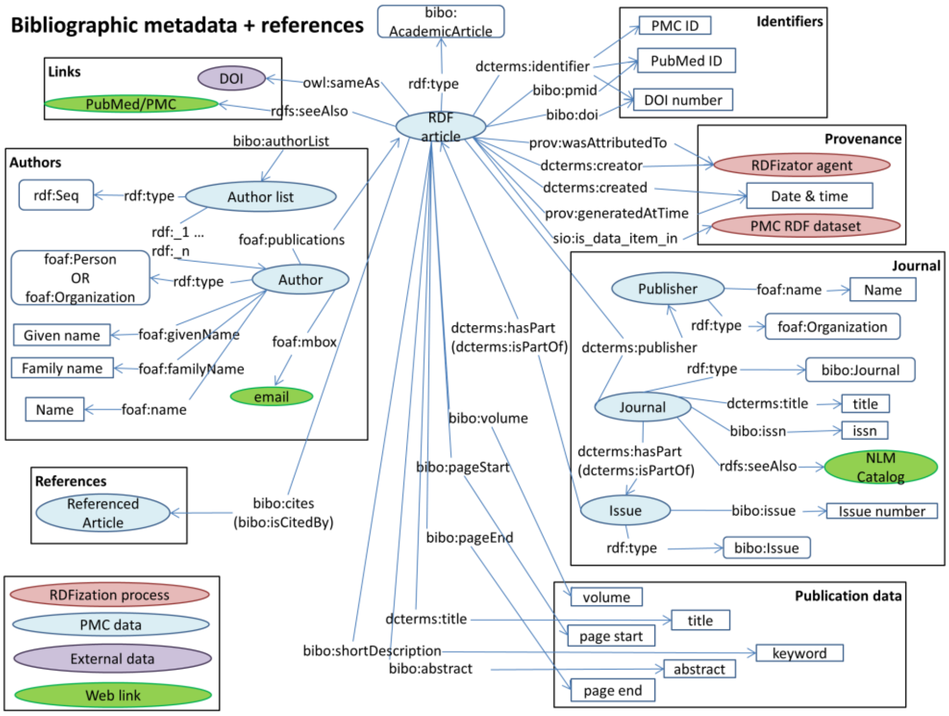


Figure 4. Our RDF data model for metadata, title, abstract and references. Figure adapted from [1].

As full-text is available for PMC-OA articles, we also provide a model supporting the representation of the content itself. Having the content in RDF as well makes it easier to refer to particular sections or even paragraphs, thus opening broader text-based analyses possibilities. We use Document Components Ontology (DoCO) [64] to explicitly identify sections and paragraphs and rdf:value to link paragraphs identifiers to the actual text. DoCO provides a structured vocabulary to represent document components. It covers both a structural perspective (e.g., block, inline, paragraph, section, and chapter) and a rhetorical one (e.g., introduction, discussion, acknowledgements, reference list, figure, and appendix). For those articles with full-text available but not under a license allowing redistribution or change in formats, it is possible to just represent the skeleton, i.e., sections. Figure 5 shows a summary of our model for structure and full-text.

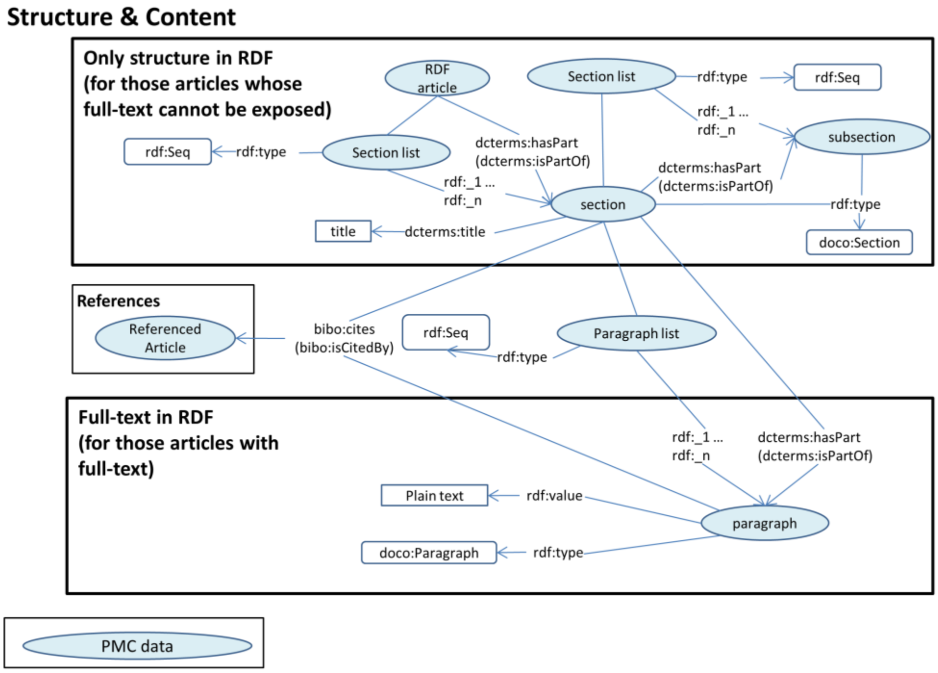


Figure 5. Our RDF data model for full-text.

3.2.2 Integration with Bio2RDF

The Bio2RDF project [34-36] aims to provide biomedical following principles from the SW; it relies on technologies such as RDF and SPARQL. Bio2RDF brings together information from diverse public databases such as KEGG, Protein Databank (PDB), UniProt, National Cancer Institute thesaurus (NCIt), and PubMed, amongst others. Both Bio2RDF and Biotea aim to support biological knowledge discovery, the former by providing a single access point to several biomedical data sources, and the latter by delivering a semantically enriched infrastructure for scientific literature, particularly full-text articles like those in PMC-OA.

In order to make Biotea compatible with Bio2RDF, we follow their guidelines regarding Uniform Resource Identifiers (URIs), metadata for datasets identification and use of the Semantic science Integrated Ontology (SIO) [65] entities. In order to use the SIO entities, we provide mappings between the ontological terms used in our model and those specified by SIO (see Table 1 and Table 2). These mappings are required as Bio2RDF mainly relies on generic ontologies such as the RDF model, the RDF schema (RDFS), DCMI-terms, Prov-O and SIO. Datatype properties in Bio2RDF required a more verbose approach (see Table 3).

In SIO, there is only one datatype property, "has value". In order to distinguish all the different values, it is necessary to select a SIO class representing the datatype range, instantiate such a class and then link it from the datatype property domain as well as to the original datatype value via the "has value" SIO property. For instance, representing a link between a publication and a DOI identifier is done via bibo:doi in Biotea. In Bio2RDF, it is necessary to find a class for DOI, we have selected "identifier". We link the publication with an identifier via the "has identifier" object property, then we link the "identifier" to the DOI value via the "has value" datatype property.

Mappings are collected in a Java properties file and a default Bio2RDF mapping is provided. Mapping to other models is possible by creating a new mapping file; however, reification is allowed only for datatype properties[9].

Table 1. Mappings between SIO classes and our metadata and content models.

Classes used in Biotea	Classes in SIO
bibo:Article	SIO_000154 (article)
bibo:AudioDocument	SIO_001168 (audio recording)
bibo:Book	SIO_000106 (book)
bibo:Chapter	SIO_000107 (chapter)
biotea:ElementSelector	SIO_000602 (computational entity)
bibo:Proceedings	SIO_000157 (conference proceedings)
bibo:Bill, bibo:Brief, bibo:CollectedDocument, bibo:Document, , bibo:LegalCaseDocument, bibo:LegalDecision, bibo:LegalDocument, bibo:Legislation, bibo:PersonalCommunicationDocument, bibo:ReferenceSource, bibo:Statute	SIO_000148 (document)
bibo:DocumentPart, bibo:Slide	SIO_000171 (document component)
bibo:BookSection, doco:Section	SIO_000111 (document section)
bibo:EditedBook	SIO_000159 (edited publication)
bibo:Email	SIO_000304 (email)
bibo:Excerpt	SIO_000298 (excerpt)
bibo:Issue	SIO_001169 (issue)
bibo:Journal	SIO_000160 (journal)

[9] More information about how to use a customized mapping file is found on the GitHub project https://github.com/biotea/biotea-rdfization

bibo:Letter	SIO_000306 (letter)
bibo:Manual	SIO_000161 (manual)
bibo:Manuscript	SIO_000151 (manuscript)
bibo:AudioVisualDocument, bibo:Film	SIO_000297 (movie)
bibo:Note	SIO_000152 (note)
bibo:Patent	SIO_000153 (patent)
bibo:AcademicArticle	SIO_001029 (peer reviewed article)
bibo:Quote	SIO_000299 (quote)
bibo:Report	SIO_001026 (report)
bibo:SlideShow	SIO_001170 (slideshow)
bibo:Standard	SIO_000618 (standard)
bibo:Thesis	SIO_000165 (thesis)
bibo:WebPage	SIO_000302 (webpage)
foaf:Agent	SIO_000776 (object)
foaf:Organization	SIO_000012 (organization)
foaf:Person	SIO_000498 (person)
foaf:OnlineAccount	SIO_001165 (user account)
doco:Figure	SIO_000080 (figure)
doco:Paragraph	SIO_000110 (paragraph)
doco:Section	SIO_000111 (document section)
doco:Table	SIO_000419 (table)

Table 2. Mappings between SIO object properties and our metadata and content model.

Object properties used in Biotea	Object properties in SIO
bibo:cites	SIO_000277 (cites)
bibo:authorList, bibo:editorList, dcterms:publisher, foaf:account, foaf:publications	SIO_000008 (has attribute)
dcterms:creator, prov:wasAttributedTo	SIO_000364 (has creator)
dcterms:hasPart	SIO_000028 (has part)
dcterms:isFormatOf, dcterms:references, dcterms:source, prov:wasDerivedFrom	SIO_000253 (has source)
bibo:citedBy	SIO_000278 (is cited by)
dcterms:isPartOf	SIO_000068 (is part of)
dcterms:hasFormat	SIO_000219 (is source of)

Table 3. Mappings between datatype properties used in Biotea and SIO.

Datatype property used in Biotea	Reification to object property, class and datatype property in SIO

bibo:doi, bibo:pmid, bibo:issn, dcterms:identifier	SIO_000671 (has identifier)	SIO_000115 (identifier)
bibo:abstract	SIO_000008 (has attribute)	SIO_000188 (abstract)
bibo:edition		SIO_000308 (edition number)
bibo:issue		SIO_000942 (numerical label)
bibo:numPages		SIO_000178 (page total)
bibo:pageEnd		SIO_000953 (page ends)
bibo:pageStart		SIO_000943 (page starts)
bibo:volume		SIO_000309 (volume)
bibo:shortDescription, dcterms:description		SIO_000136 (description)
dcterms:created, dcterms:issued, prov:generatedAtTime		SIO_001314 (date of issue)
dcterms:title		SIO_000175 (title)
foaf:accountName		SIO_000116 (name)
foaf:familyName		SIO_000182 (last name)
foaf:givenName		SIO_000181 (first name)
foaf:name		SIO_000183 (personal name)
rdf:value	SIO_000628 (refers to)	SIO_001073 (text span)

3.2.3 RDFization batch process

We define RDFize as a verb, meaning (i) to generate an RDF representation of something that was originally in a different format and (ii) to convert or transform to RDF [1]. The workflow followed to RDFize metadata and content for PMC articles is depicted in Figure 6. We produce two files per publication; for example, for *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2971111* we produce one RDF file for metadata, abstract, tittle and references and another one for sections, paragraphs and content. Previous to the RDFization batch process, it is necessary to download the Journal Article Tag Suite (JATS) files for PMC-OA; they constitute the main input for our process.

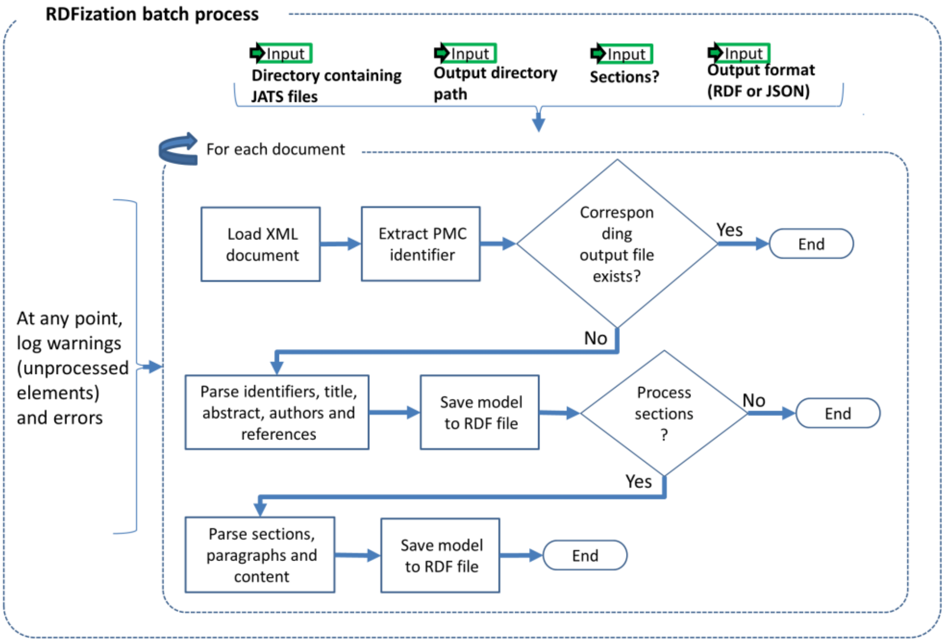
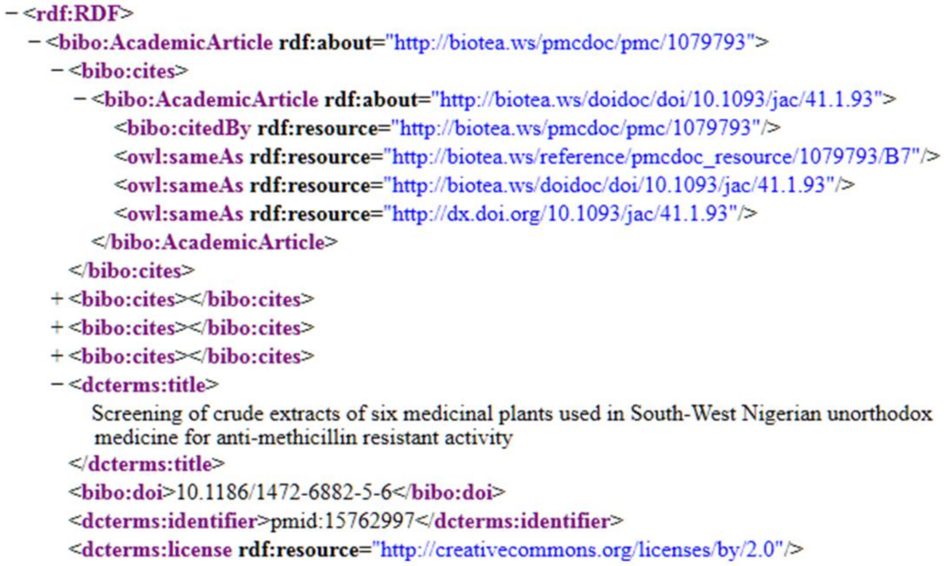

RDFization batch process
Input
Directory containing JATS files
Input
Output directory path
Input
Sections?
Input
Output format (RDF or JSON)
For each document
At any point, log warnings (unprocessed elements) and errors
Load XML document
Extract PMC identifier
Correspon ding output file exists?
Yes
End
No
Parse identifiers, title, abstract, authors and references
Save model to RDF file
Process sections ?
No
End
Yes
Parse sections, paragraphs and content
Save model to RDF file
End

Figure 6. RDFization batch process.

Before running the batch process, some configuration properties in the config.properties files should be modified in order to assure that the output follows the expected format and style. For instance, it is possible to specify the prefix; in the case of PMC-OA it must be *pmc*. The domain used in the Unified Resource Locators (URLs) can also be specified as well as the identification of the dataset. There are also some development dependencies that should be configured before running the process[10].

From the input directory, we process all the files with extension *nxml*, i.e., JATS files. We parse each article, starting with metadata and then moving to references, up to that point, we save the corresponding RDF file to an output directory. The output file name is given in the form PMC<id number>.rdf or PMC<id number>.json depending the output format specified in the program arguments, <id number> corresponds to the PMC identifier for the RDFized article. If the output file already exists in the output directory, the input file is not even parsed but just dismissed. The next step consists in RDFizing the sections, paragraphs and content. Figures, tables, footnotes and formulas are omitted. As a rule of thumb, our process attempts to RDFize as much as possible; however, if any XML element cannot be parsed, then it is omitted and the process continues with the rest. An excerpt of the metadata corresponding to the article PMC1079793 is provided in Figure 7.

[10] Further documentation as well as the algorithms are available in GitHub at https://github.com/biotea/biotea-rdfization



```
– <rdf:RDF>
  – <bibo:AcademicArticle rdf:about="http://biotea.ws/pmcdoc/pmc/1079793">
    – <bibo:cites>
      – <bibo:AcademicArticle rdf:about="http://biotea.ws/doidoc/doi/10.1093/jac/41.1.93">
          <bibo:citedBy rdf:resource="http://biotea.ws/pmcdoc/pmc/1079793"/>
          <owl:sameAs rdf:resource="http://biotea.ws/reference/pmcdoc_resource/1079793/B7"/>
          <owl:sameAs rdf:resource="http://biotea.ws/doidoc/doi/10.1093/jac/41.1.93"/>
          <owl:sameAs rdf:resource="http://dx.doi.org/10.1093/jac/41.1.93"/>
        </bibo:AcademicArticle>
      </bibo:cites>
    + <bibo:cites></bibo:cites>
    + <bibo:cites></bibo:cites>
    + <bibo:cites></bibo:cites>
    – <dcterms:title>
        Screening of crude extracts of six medicinal plants used in South-West Nigerian unorthodox
        medicine for anti-methicillin resistant activity
      </dcterms:title>
      <bibo:doi>10.1186/1472-6882-5-6</bibo:doi>
      <dcterms:identifier>pmid:15762997</dcterms:identifier>
      <dcterms:license rdf:resource="http://creativecommons.org/licenses/by/2.0"/>
```

Figure 7. An excerpt of an RDF file for metadata, title, abstract and references following the Bio2RDF URLs style.

3.3 Results

The Biotea RDFization project defines an RDF model for scientific publications that can be used beyond the LS domain. It also makes available the algorithms required to RDFize articles available on PMC-OA. A first version of PMC-OA RDFized articles following the Biotea model was released in 2012 [3]. A second version is under preparation at the time of writing, i.e., March 2017. Future releases will be coordinated to be synchronized with the Bio2RDF releases, usually once or twice a year.

Here we present the results of the RDFization process carried on 2012. The released dataset comprised 270,834 articles from PMC-OA, distributed across 2,401 journals. About 40% of these articles correspond to 20 journals (see Figure 8).

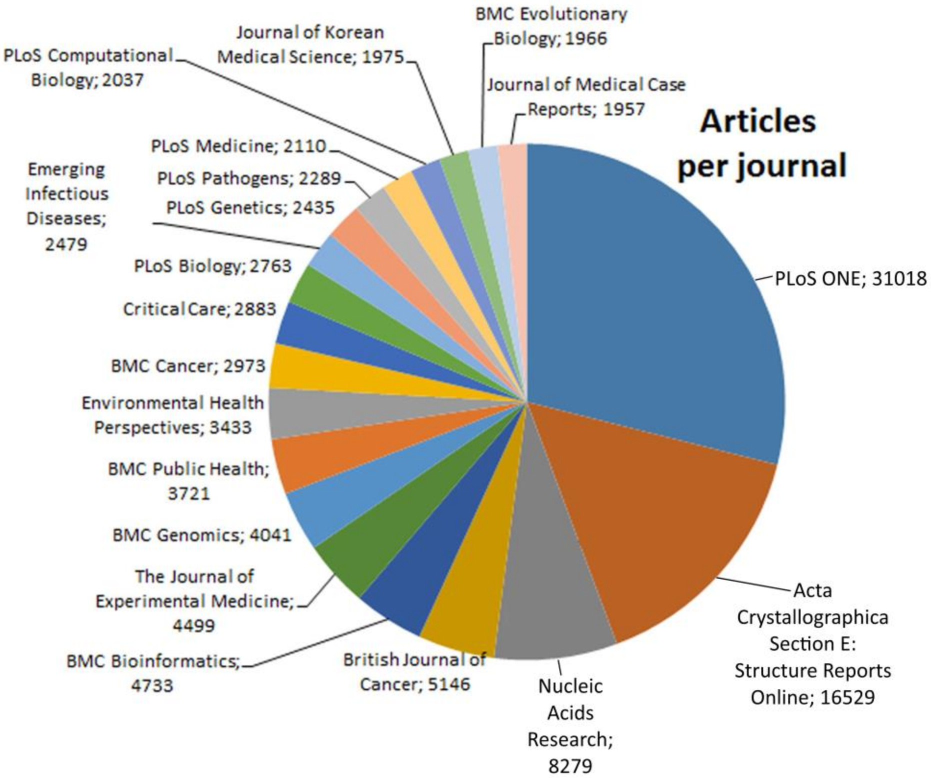


Figure 8. Coverage per journal. Only those journals corresponding to the 40% of the articles are represented in the figure. Figure taken from [1].

Our model and RDFization process follow four of the five principles proposed by Tim Berners-Lee for publishing Linked Data [28]: (i) providing machine-readable structured data , (ii) using a non-proprietary format, (iii) using W3C standards –RDF, and (iv) inter-linking data. The fifth principle, supporting an open license cannot be guarantee as it depends on the original articles licenses. Our model uses BIBO and DCMI-terms to represent the bibliographic metadata, DoCO to explicitly identify sections, and FOAF to identify authors and organizations. Relations to other resources representing the same entity are included as owl:sameAs while relations to web pages are included as rdfs:seeAlso. References include metadata similar to that of the main article.

As we are RDFizing the article structure and content, it is possible to use our model in order to find articles with terms present in specific sections of the document. This is an improvement over traditional keyword-based search and retrieval tools, since such tools currently search for keywords either in the title, abstract, or entire text without the possibility of specifying a particular section. An example SPARQL query is provided in the following code excerpt –taken from [1]. This query retrieves the PubMed identifier, article title, section title, and paragraphs for those articles containing the term “cancer” in any section whose title includes “introduction”.

```
SELECT ?pmid ?title ?secTitle ?text WHERE {
  ?article a bibo:Document ;
    bibo:pmid ?pmid ;
    dcterms:title ?title .
  ?section a doco:Section ;
    dcterms:isPartOf ?article ;
    dcterms:title ?secTitle .
  FILTER (
    regex(str(?secTitle), "introduction", "i")
  ).
  ?para a doco:Paragraph ;
    dcterms:isPartOf ?section ;
    rdf:value ?text .
  FILTER (
    regex(str(?text), "cancer", "i")
  ).
} LIMIT 50
```

3.4 Discussion

We have defined a model to represent scientific publications and developed the algorithms required to apply such a model to PMC-OA. Although it has not been fully tested, the implementation allows as well RDFizing PubMed articles with only title-and-abstract. Other formats apart from JATS can be supported by defining classes similar to the one under the package named ws.biotea.ld2rdf.rdfGeneration.pmcoa. Such a class should implement a predefined interface in order to be compatible with the rest of the process.

Similar to NPG RDF, we use BIBO; however, rather than DCMI-elements, we use DCMI-terms. DCMI-elements comprise a set of 15 descriptive semantic definitions. Elements are expressed as simple attribute-value pairs without further restrictions regarding domains or ranged of the offered properties. We prefer DCMI-terms as it uses additional qualifiers such as property ranges to further refine the meaning of a resource, increasing specificity and precision of the metadata [55].

Some of the identified limitations or our approach are: (i) at least four different forms are used to model references in JATS files, (ii) author names are represented with initials and last name, making it difficult to disambiguate them, (iii) FOAF for authors and institutions are not provided thus we have to make up URLs for them despite the fact they could have been already defined somewhere on Internet. In order to deal with different reference styles, we create specific methods for each identified form and transform them into a common RDF model following the recommendations from BIBO.

Our model along with the RDFization process makes it possible to generate an interoperable semantic dataset. Within the owl:sameAs links, we include PubMed, DOI and identifiers.org. A DOI is a unique alphanumeric identifier assigned to electronically available scientific articles providing a persistent link to its location on the World Wide Web (WWW). Identifiers.org provides resolvable persistent URIs used to identify data common across the LS domain such as scientific publications, genes, proteins, and chemicals. Similar to the NPG experience, we also rely on ontologies such as BIBO in order to model metadata. Since we are targeting only open-access documents within PMC, we also include the content of the document.

Chapter 4

An enriched infrastructure based on semantic annotations

4.1 Background

In the previous chapter we introduced a semantic model to represent metadata, references and content in scientific publications. Although it constitutes a step forward to literature-based knowledge discovery as it makes publications semantic and machine-processable, reported facts remain locked up behind the free text. Adding annotations to scientific articles is a common practice while reading text and analyzing the content. Annotations are simple yet effective as they provide a lightweight mechanism to add value to the content embedded on the free text.

Sematic Digital Libraries (SDLs) use ontologies to: (i) organize bibliographic descriptions, (ii) represent and expose document contents, and (iii) share knowledge amongst users. In the previous chapter we use ontologies to organize bibliographic descriptions; in this chapter we will use them to represent and expose enriched content extracted from publications. We will use semantic annotations in order to model the extracted content. A sematic annotation is an annotation that relates the annotated content to unique concept in an ontology. In the past years, semantic annotation support has been added to SDLs. For instance, JeromeDL [50] allows users to semantically annotate books, papers, and resources. In the same vein, DOMEO [66] and the Living Document [67] support semantic annotations for publications in the biomedical domain. Both DOME and the Living Document offer a web-based interface where users can read and annotate articles.

DOMEO uses web-based components developed with the Google Web Toolkit. Such components allow users to manually create unstructured –i.e., free text, or semi-structured –i.e., related to a unique concept in an ontology, annotations that can be private, shared within selected groups, or made public. Semi-structured annotations are assisted by a semi-automatic process. The Living Document follows a similar approach although annotations are mainly kept private within the user's space. UTOPIA [68], a desktop tool rather than a digital library, provides a semantic and social layer on top of a Portable Document Format (PDF) reader. Similar to DOMEO and the Living Document, UTOPIA also aims to improve interoperability and user experience.

Similar to DOMEO and the Living Document, Biotea relies on text-mining and data-extraction tools to automatically identify expressions within the text with a biological significance. Such expressions can be related to ontological terms corresponding to proteins, chemicals, drugs, or diseases, among other biological concepts. Rather than providing an end-user environment, Biotea aims to enrich Resource Description Framework (RDF) scientific publications by extracting biological relevant content annotated with semantic annotations. Biotea delivers a semantic model to represent such annotations as well as algorithms to create the corresponding RDF files. In such a way, Biotea aims to support interoperability as publications become linked to each other and to biological resources. Same as for the RDFization of metadata, references and content, we still focus on PubMed Central Open Articles (PMC-OA) articles.

4.1.1 The Annotation Ontology

The Annotation Ontology (AO) [69] defines a model to represent annotations as well as semantic annotations. AO is built on top of the Annotea Project [70]. A basic annotation represents an expression –one or more consecutive words, found on a document; the corresponding model according to AO is shown on the top part of Figure 9. In order to relate this annotation to a unique concept in an ontology, just one more triplet is required (see bottom of Figure 9). AO also contemplates the provenance of the annotation. In order to model such information, it uses the Provenance, Authoring and Versioning (PAV) ontology [71] which corresponds to a specialization of the Provenance Ontology (PROV-O) for the biomedical domain. The PAV approach is compatible with the provenance model used for the RDFization of metadata, content and references. AO is split in different modules that will be explained in the paragraphs below Figure 9. Those modules include the core –namespace ao, the topics –namespace aot, and the selectors –namespace aos. In the latest version of AO, all the modules are grouped under the same namespace ao.

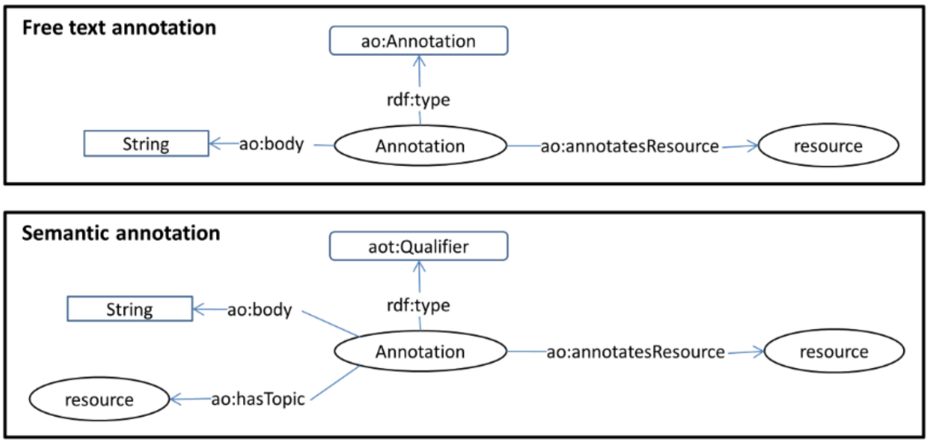


Figure 9. AO basic schema for free text and semantic annotations.

The AO core defines the minimum set of classes and properties required to model a simple, i.e., free text, annotation. A different module in AO, AO-topic, is used to define elements required to model semantic annotations. The AO-topic defines a class aot:Qualifier, i.e., annotations associate to a controlled vocabulary entity; and a property aot:hasTopic to record the association between the annotation and the ontological entity. Depending on the type of association, different qualifiers are defined, all of the compatible with the Simple Knowledge Organization System (SKOS) model [72]. SKOS is a common data model for sharing and linking knowledge organization systems such as thesauri, taxonomies, classification schemes and subject heading systems. Table 4 shows the type of associations supported by AO-topic and their corresponding mapping to SKOS properties.

Table 4. AO and SKOS.

Qualifier	skos:relatedMatch	Relationship between the object of the annotation and a well-defined semantic entity.
ExactQualifier	skos:exactMatch	The object of the relationship aot:hasTopic (the semantic entity) exactly represents the portion of the annotated document.
CloseQualifier	skos:closeMatch	The object of the relationship aot:hasTopic closely but not exactly represents the portion of the annotated document.
BroaderQualifier	skos:broaderMatch	The object of the relationship aot:hasTopic broadly represents the portion of the annotated document.
NarrowerQualifier	skos:narrowerMatch	The object of the relationship aot:hasTopic represents a more specific entity than the portion of the annotated document.

AO also considers atomic annotations, *i.e.* annotations that refer to a particular part within a document. Atomic annotations in AO are represented as selectors, i.e., classes that model a portion of the document. The AO-selector set contains the definitions for the different type of selectors natively supported by AO. This set includes:

- aos:XPointerSelector used to identify any element in an XML document,
- aos:ImageSelector for portions of an image,
- aos:AudioSelector for portions of audio files, and
- aos:TextSelector for portions of text in a document.

New selectors can be easily added by inheritance mechanisms. Such inheritance mechanism is used by AO-selector in order to specify specializations of the main selectors. Particularly for the aos:TextSelector, two specializations are considered:

- aos:OffsetRangeTextSelector uses two numbers to indicate the position regarding the first character on a document –offset, and the number of annotated characters –range,
- aos: PrefixPostfixTextSelector uses some character before the annotated text –prefix, and some more characters after the annotated text –postfix.

4.1.2 The Open Annotation Ontology

The Open Annotation Ontology (OA) is a community effort to collate different ontologies into a single one aiming to represent annotations on resources [5]. OA evolved into the Web Annotation Vocabulary [73], currently a World Wide Web Consortium (W3C) candidate recommendation. Although the name has changed, we still will refer to this vocabulary as OA as that is the namespace defined for the model. Same as AO, OA includes classes and properties to annotate with free text but also with ontological terms. OA defines 3 basic concepts: an annotation –oa:Annotation containing a body –oa:hasBody, that is somehow related to a target –oa:hasTarget. The basic model to represent free text as well as semantic annotations is presented in Figure 10.

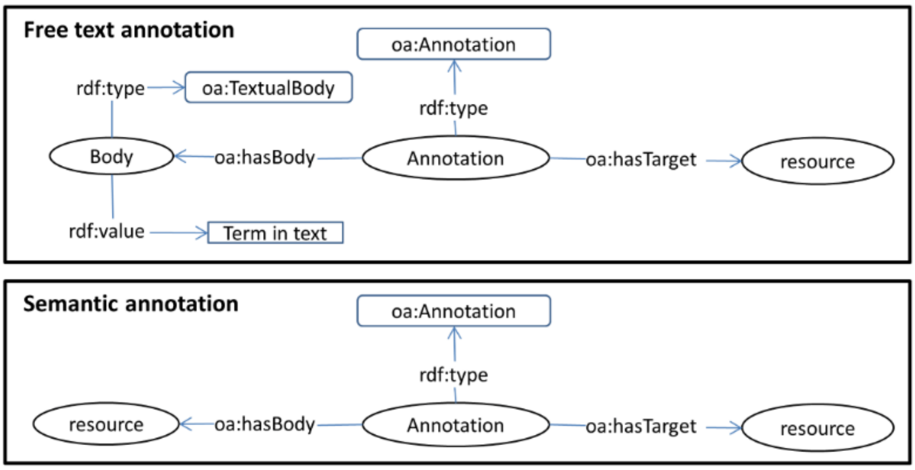


Figure 10. OA basic schema for free text and semantic annotations.

OA also considers atomic annotations. These are the selectors used to refer different portions within a resource:

- oa:FragmentSelector used to identify segments within a resource, i.e., URLs including anchors;
- oa:RangeSelector used to identify portions of a text or records within a dataset, it includes
  - oa:TextPositionSelector to specify the start and end positions of the annotations text,

    - oa:TextQuoteSelector to specify an excerpt of the text, and
    - oa:DataPositionSelector to specify start and end positions of a record in a database; and finally,
- oa:AreaSelector used to defined polygon-based selections on a media type

4.2 Methodological approach

4.2.1 Model

Our model to represent semantic annotations supports both AO and OA. Following the AO, we represent an annotation as an ExactQualifier as our annotations always refer to a concept in an ontology. The chunk of text identified in a publication and associated to a unique concept in an ontology is linked to via the ao:body property. The concept itself is linked to via the ao:hasTopic property. The publication is linked to by using the ao:annotatesResource property. Provenance regarding the serialization agent, the annotator agent and the time of the annotation is also recorded. Different from the metadata model, we use PAV, a specialization of PROV-O for the biomedical domain. Our model goes further as it includes information regarding the section and paragraph where the chunk of text was identified. In order to link to a section containing a paragraph, we use ao:context. AO provides contexts related to positions within the text but it does not provide a context to select an RDF element.

In order to link to paragraphs as well as to provide information related to the term frequency (TF) and the inverse document frequency (IDF) of an ontological term in the whole publication, we have created some customized properties. Biotea properties and classes are put together in the form of an ontology[11]. Here we detail those bits used for the annotation model:

- biotea:ElementSelector, a class representing an AO context related to an RDF element;
- biotea:tf, a datatype property to represent the TF, and
- biotea:idf, a datatype property to represent the IDF.

In Figure 11, we provide an overview of our annotation model for publications using AO.

[11] Biotea ontology is available at https://github.com/biotea/biotea-ontololgy

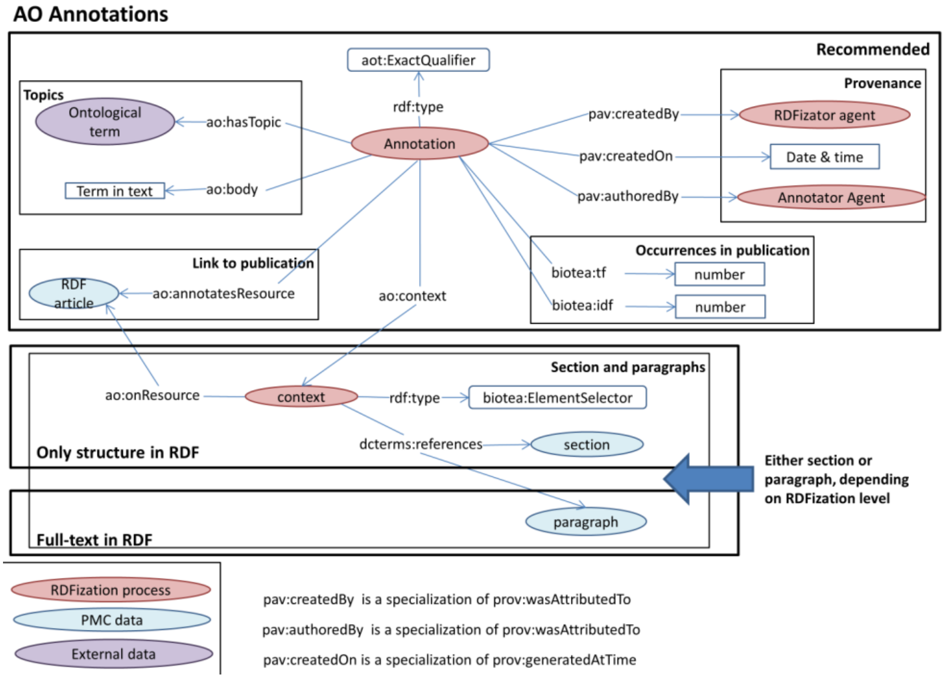


Figure 11. Our RDF data model for annotations on publications using AO. Figure adapted from [1].

We also offer support for OA. Annotations are modeled as an oa:Annotation with multiple bodies, linked to via the oa:hasBody property. A body can be an ontological term or a chunk of text identified in a publication. Chunks of text correspond to an oa:TextualBody where the actual text is linked to via rdf_value property. Both the publication and the sections within are represented by using the property oa:hasTarget. Whenever a section is used as a target, a link to the publication will be added, oa:hasSource, to indicate that the section belongs to a bigger entity corresponding to the entity being annotated. The bits related to provenance, TF and IDF remain the same. In Figure 12, we present our model for annotations on publications using OA. Main differences with respect to AO are highlighted with a grey background.

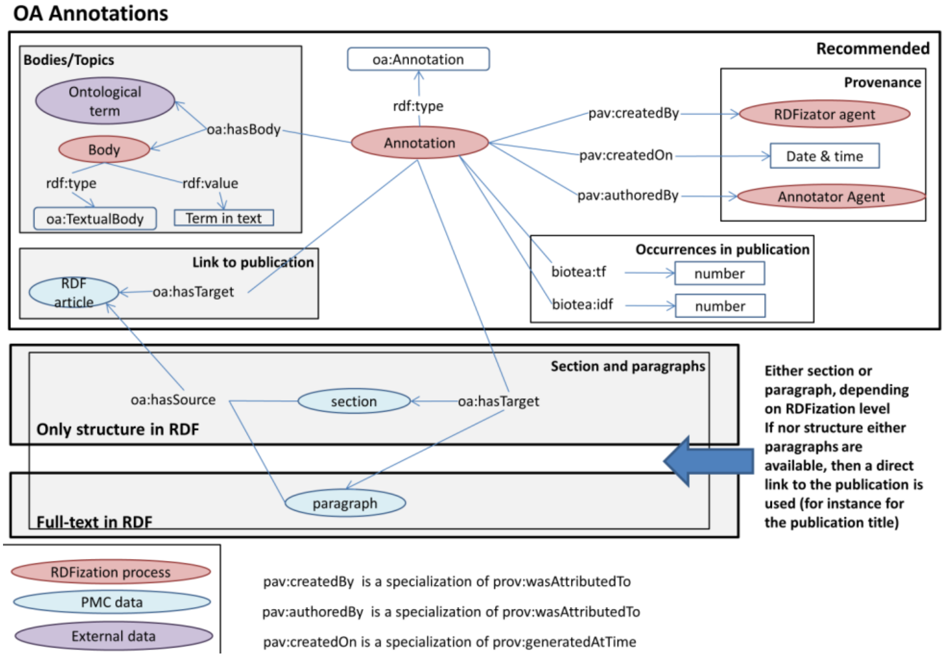


Figure 12. Our RDF data model for annotations on publications using OA. A soft grey background is used to highlight differences regarding the model using AO.

4.2.2 Integration with Bio2RDF

Our annotation model is also compatible with Bio2RDF. Again, it is necessary to provide a mapping between classes and properties used to represent annotations and terms in the Semantic science Integrated Ontology (SIO). A summary of such a mapping is provided in Table 5.

Table 5. Mappings between SIO classes and properties and our annotations model.

	Terms used in Biotea	Terms in SIO
Classes	ao:Annotation, aot:ExactQualifier, oa:Annotation	SIO_001166 (annotation)
	oa:TextualBody	SIO_001073 (text span)
Object Properties	pav:authoredBy, pav:createdBy	SIO_000364 (has creator)
	oa:hasSource	SIO_000253 (has source)
	ao:onResource	SIO_000332 (is about)
	ao:annotatesResource, oa:hasTarget	SIO_000254 (is annotation of)
	ao:context, ao:hasTopic, oa:hasBody	SIO_000628 (refers to)
Datatype properties (using reification)	ao:body	SIO_000628 (refers to) - SIO_001073 (text span) - SIO_000300 (has value)

	biotea:idf	SIO_000900 (has frequency) - SIO_001018 (ratio) - SIO_000300 (has value)
	biotea:tf	SIO_000900 (has frequency) - SIO_000794 (count) - SIO_000300 (has value)
	prov:generatedAtTime	SIO_000008 (has attribute) - SIO_001314 (date of issue) - SIO_000300 (has value)

4.2.3 Semantic enrichment batch process

We process titles, sections and paragraphs with two text-mining tools: Whatizit [74, 75] and the National Center for Biomedical Ontology (NCBO) Annotator [76]. Both annotators are based on exact string matching as well as pre-defined dictionaries. Whatizit is based on monq.jfa [77], an open source Java library for text-mining. Regular expressions are defined and bound to actions; whenever a matching expression is found in the text, the action is executed. In Whatizit those actions produce an XML element representing the expression found and the associated ontological term.

We use Whatizit to identify chemical entities in ChEBI, protein accessions in the Universal Protein Resource (UniProt) and Gene Ontology (GO) identifiers. The NCBO Annotator is built upon Mgrep [78]. The process is similar to that one performed in Whatizit, i.e., expressions in the text are associated to ontological terms. With the NCBO Annotator is possible to cover more ontologies. It also provides more configuration options, for instance, the words that should be excluded as well as the minimum length of an expression to be taken into account in the matching process.

The order of usage of the annotations services has no impact as a different RDF file is produced per annotator. If the same entity is recognized by more than one annotator, we will keep all of the generated annotations. The same entity can also be associated to more than one concept in one or more ontologies. This case will produce only one annotation but multiple ontological associations. Another case corresponds to an entity covering multiple words, for instance "mad cow" and another entity covering only one of them, for instance "cow". In this case, the all of the entities recognized by the annotators will be transformed into annotations, one annotation per recognition. In the example, we would have one annotation for "mad cow" and another one for "cow".

At the time of writing, Whatizit was no longer provided as a public service thus only the NCBO Annotator is currently used in Biotea. The annotations can be created from JATS files or Biotea RDF files. Both the input and output format, either RDF/eXtensible Markup Language (RDF/XML) or JavaScript Object Notation-Linked Data (JSON-LD), can be specified via execution parameters. The annotator and the annotation style, either AO or OA, can also be specified. Comprehensive usage information is provided[12].

From the input directory, we process all the files with extension *nxml* for Journal Article Tag Suite (JATS) files or with extension *rdf* for RDF/XML. The RDF file that should be used is the one containing sections and paragraphs. We parse each article, first the main title, then the section titles and then paragraph by paragraph. Each paragraph is sent to the annotation services. The annotation service then replies with an XML. That XML is processed so we extract the expressions found in the text together with their associations to ontological terms. One file is produced per annotator. A summary of the semantic enrichment process is presented in Figure 13.

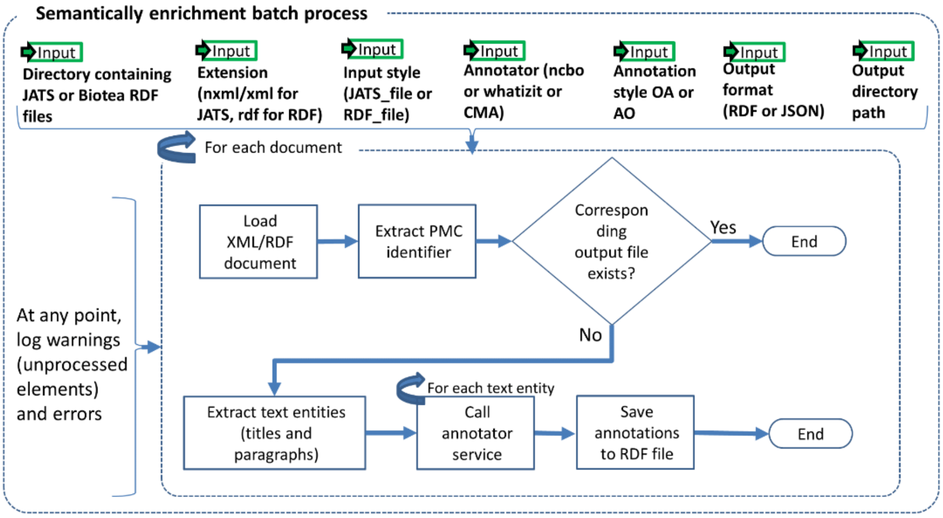

Semantically enrichment batch process
Input
Directory containing JATS or Biotea RDF files
Input
Extension (nxml/xml for JATS, rdf for RDF)
Input
Input style (JATS_file or RDF_file)
Input
Annotator (ncbo or whatizit or CMA)
Input
Annotation style OA or AO
Input
Output format (RDF or JSON)
Input
Output directory path
For each document
At any point, log warnings (unprocessed elements) and errors
Load XML/RDF document
Extract PMC identifier
Correspon ding output file exists?
Yes
End
No
Extract text entities (titles and paragraphs)
For each text entity
Call annotator service
Save annotations to RDF file
End

Figure 13. Semantic enrichment batch process.

The NCBO Annotator is used for the following ontologies:

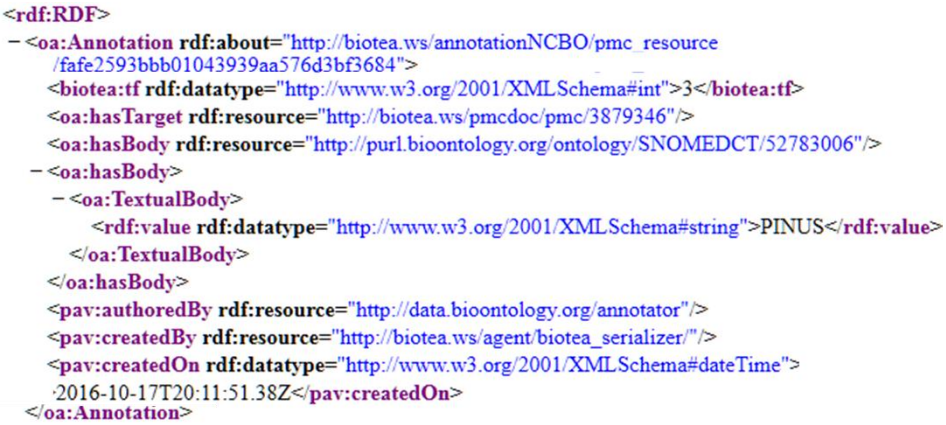
- ChEBI for chemicals;
- Pathway, and Functional Genomics Data Society (MGED) for genes and proteins;
- Master Drug Data Base (MDDB), National Drug Data File (NDDF), and the National Drug File - Reference Terminology (NDF-RT) for drugs;
- SNOMED Clinical Terms (SNOMED-CT), Symptom Ontology (SYMP), Medical Dictionary for Regulatory Activities (MedDRA),

[12] Further documentation as well as the algorithms are available in GitHub at https://github.com/biotea/biotea-annotation

MeSH, MedlinePlus Health Topics (MedlinePlus), Online Mendelian Inheritance in Man (OMIM), Foundational Model of Anatomy Ontology (FMA), International Statistical Classification of Diseases vr. 10 (ICD10), and Ontology for Biomedical Investigations (OBI) for diseases and medical terms;

- Plant Ontology for plants; and
- MeSH, SNOMED-CT, and NCIt for general terms.

Whatizit is used for GO, UniProt proteins, UniProt Taxonomy, and diseases mapped to UMLS concepts. ChEBI, GO, and organisms are supported by both NCBO and Whatizit. It is possible to keep them both or to exclude one of them. Ontologies to be excluded during the NCBO annotation are configured with the property ncbo.annotator.exclude in the config.properties file. For Whatizit, two exclusion properties exist, one for ChEBI, whatizit.CHEBI, and one for GO, whatizit.GO. In addition to the ontology links provided by the annotators, we also included links to Bio2RDF for ChEBI, GO, MeSH, NCIt, UniProt proteins, UniProt Taxonomy, and NCBI Taxon. An excerpt of the annotations produced for the article PMC3879346 using the OA model is presented in Figure 14.



```
<rdf:RDF>
 - <oa:Annotation rdf:about="http://biotea.ws/annotationNCBO/pmc_resource
       /fafe2593bbb01043939aa576d3bf3684">
     <biotea:tf rdf:datatype="http://www.w3.org/2001/XMLSchema#int">3</biotea:tf>
     <oa:hasTarget rdf:resource="http://biotea.ws/pmcdoc/pmc/3879346"/>
     <oa:hasBody rdf:resource="http://purl.bioontology.org/ontology/SNOMEDCT/52783006"/>
   - <oa:hasBody>
     - <oa:TextualBody>
         <rdf:value rdf:datatype="http://www.w3.org/2001/XMLSchema#string">PINUS</rdf:value>
       </oa:TextualBody>
     </oa:hasBody>
     <pav:authoredBy rdf:resource="http://data.bioontology.org/annotator"/>
     <pav:createdBy rdf:resource="http://biotea.ws/agent/biotea_serializer/"/>
     <pav:createdOn rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">
      2016-10-17T20:11:51.38Z</pav:createdOn>
   </oa:Annotation>
```

Figure 14. An excerpt of an RDF file for annotations following the OA model.

4.3 Results

4.3.1 The Biotea Ontology

Initially, the Biotea ontology[13] was created with only two properties, biotea:idf to represent IDF and biotea:tf to represent TF. At the time of defining the initial Biotea algorithms, not suitable properties were found thus we created them. Biotea ontology later evolved to include a representation for semantic

[13] Biotea ontology is available at https://github.com/biotea/biotea-ontololgy

similarity and semantic group-based categorization. The Biotea ontology will be further discussed in the following chapters.

4.3.2 Enriched content for PMC-OA

The first Biotea version, released in 2012, included annotations from both Whatizit and NCBO Annotator following the AO model. As Biotea is not the only effort RDFizing scientific publication, although it was one of the pioneers in the field, we include owl:sameAs links to Bio2RDF and identifiers.org [79]. We also include rdf:seeAlso links to the same resource in different formats such as PubMed, PMC-OA and the Digital Object Identifier (DOI). We also use owl:sameAs for biological entities covered by either Bio2RDF or identifiers.org such as MeSH and National Center for Biotechnology Information (NCBI) Taxonomy. At the time of writing, a new version of Biotea is under production. Options such as Neji [80] are being considered as an alternative in order to cover vocabularies not supported by the NCBO Annotator such as UniProt.

Annotations covered by the RDFization process are distributed across 18 different controlled vocabularies. Thirteen of them are covered by the NCBO Annotator and the other five by Whatizit; seven of all those vocabularies are also part of Bio2RDF. In Figure 15, we present a summary of the coverage per vocabulary. None of them was particularly higher than others as observed in the left side of the figure. On the right side of the figure, we present the number of terms recognized in the articles in comparison to the number of entities coined in the vocabularies.

A term can be expressed in multiple ways, for instance either “human” or “homo sapiens” will be recognized and associated to the same ontological entity. Therefore, the number of terms is usually higher than the number of entities. This is not true for the case of UniProt as Whatizit can associate more than one protein to the same term. For instance, for PMC:1043860 "Scr" and "lacZ" were recognized by Whatizit as proteins; the first term was associated to two proteins –Q93CH6 and P09077, the second term was associated to seventeen proteins –Q59750, P30812, P23989, P06219, Q48727, Q56307, P70753, P26257, P81650, P0C1Y0, P77989, Q9K9C6, Q59140, O33815, P00722, Q47077, and Q1G9Z4.

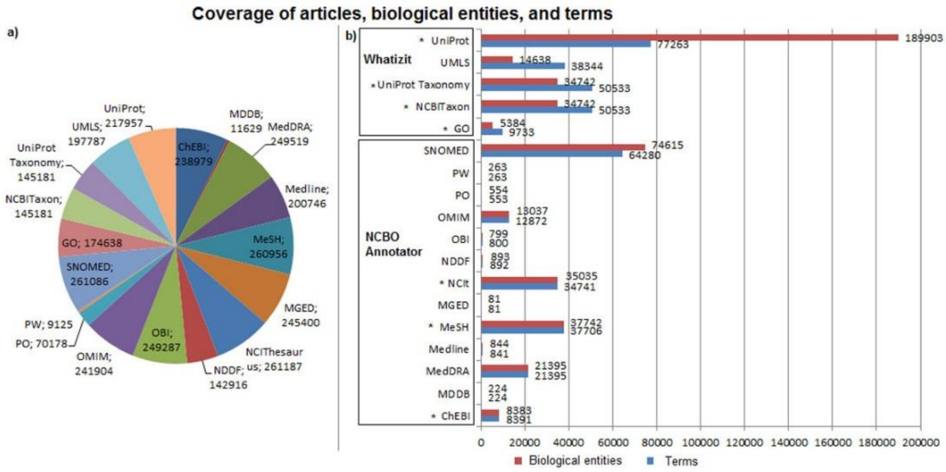


Figure 15. a) Number of articles covered per controlled vocabulary. b) Number of biological entities and terms covered per vocabulary and annotator. Vocabularies that are part of Bio2RDF are indicated with a star (*). Taken from [1].

Article retrieval based on annotations becomes possible thanks to RDF and SPARQL. In the following excerpt, extracted from [1], we present a query retrieving all PubMed identifiers for articles with annotations associated to a particular chemical entity identified as CHEBI:60004 (“mixture”).

```
SELECT DISTINCT ?pmid
  WHERE {
    ?article a bibo:AcademicArticle ;
      bibo:pmid ?pmid .
    ?annotation a aot:ExactQualifier ;
      ao:annotatesResource ?article ;
      ao:hasTopic
<http://purl.obolibrary.org/obo/CHEBI_60004> .
}
```

4.4 Discussion

One of the main challenges faced during the enrichment process was the availability of the annotation services. Initially, we sent the full text to the annotators and the parse the complete response. This approach put some stress on the annotation services thus they continuously timed out. We changed our strategy in order to send small chunks of text, corresponding to either titles or paragraphs. This approach worked better for the annotation services and also allowed us to parse the response more easily while at the same time identifying the specific sections and paragraphs associated to them. However, the enrichment process remains significantly heavier than the metadata and content

RDFization process. While processing metadata and content for about one million articles would take a couple of weeks, annotating the content takes a couple of months. Installing a local version of the annotation services could improve this time; unfortunately, we did not count with the resources required for this.

The resulting annotated dataset is semantically richer than that one provided by the Nature Publishing Group (NPG). NPG does not semantically link to ontologies. It does include MeSH terms but they are included as plain literals, thus information retrieval keeps being a keyword experience rather than a semantic one. We do not only link to the ontological entities but add owl:sameAs and rdf:seeAlso relations whenever possible.

Semantically enriched datasets, as the one delivered by Biotea, facilitate literature-base knowledge discovery as semantic annotations make it possible to find hidden relations [12, 14, 15]. For instance, patterns across the MeSH terms have been used to identify potential new associations between drugs and diseases [12]. Also, annotations shared by a group of genes have contributed to identify possible relationships between these genes [14, 15]. Broader possibilities become possible in Biotea. For instance, as sections are explicitly identified, it is possible to deliver analyses focusing on a particular section such as "Materials and methods" and "Results", leaving aside the "Introduction" or "Background".

Adding new annotators is possible by adding a new class implementing the *AnnotatorParser* in the package *ws.biotea.ld2rdf.annotation.parser*. The batch process should also be modified so the new annotator can be specified within the execution options. Supporting new vocabularies for the existing annotators is easier as they are configured in property files. Furthermore, our algorithms are open source thus they can be reused, modified and extended.

Part II - BIOLINKS

SEMANTIC GROUPS, CATEGORIZATION
AND COMPARISON FOR SCIENTIFIC
PUBLICATIONS

Biolinks is a project using the enriched infrastructure provided by Biotea in order to semantically categorize and compare scientific publications.

Chapter 5

Biolinks semantic groups

5.1 Background

Biotea, introduced in Part I of this document, delivers a semantically enriched infrastructure. Biotea aims to provide a scaffolding to facilitate the development of semantic information retrieval as well as literature-based knowledge discovery tools. Knowledge discovery becomes possible mainly because of the semantic annotation added on top of the text. As the annotations relate specific portions of the text with specific concepts in controlled vocabularies, fine-grained analyses, involving properties across concepts and vocabularies, become possible. Although such fine-grained analyses are both interesting and useful, our attention draws to more high-level analyses, regarding semantic groups rather than semantic concepts. In this Part of the document, we introduce Biolinks, a project using the enriched infrastructure provided by Biotea in order to semantically categorize and compare scientific publications. While the categorization focuses on annotations at a high-level, the comparison works with both fine-grained and high-level conceptualizations.

The Unified Medical Language System (UMLS) is a meta-thesaurus gathering together concepts from multiple controlled vocabularies within the biomedical domain [81-83]. Concepts in UMLS are identified with a Concept Unique Identifier (CUI). Different from other controlled vocabularies, UMLS does not only provide concepts but also a semantic grouping at two different levels. UMLS concepts are attributed to Semantic Types (STY), which are in turn categorized into Semantic Groups (SGRs). UMLS is released twice a year. The current version of Biolinks uses the UMLS version 2012AB; although updating to a more recent version is also possible. UMLS makes use of 15 SGRs assigned to about 99% of the UMLS concepts. SGRs can be seen as a partition over STYs, i.e., one STY maps to just one semantic group. The SGRs have been defined for organizational reasons in order to better manage the conceptual complexity of STYs [84].

Biolinks proposes an alternative grouping for UMLS STYs which will be later used to categorize scientific publications as well as calculate and narrow the semantic similarity between publications. The purpose of the Biolinks groups is to achieve a better semantic coherence taking into account (i) the distribution of the different UMLS concepts along the is-a relations of themselves and their corresponding STYs, and (ii) the distribution of the UMLS concepts along the used test collection. Semantic groups support the identification and access of more relevant and related data as researchers could use these groups as a parameter to narrow the publications they are interested in.

Alternative organizations of the is-a hierarchy for STYs and their grouping into SRGs have been suggested. For instance, visual approaches have been used to analyze SGRs and then assessing their semantic coherence regarding the semantic relations among the STYs belonging to each group [85]. A simplification into broader groups useful for biomedical text-mining has been also proposed [86]. Microorganisms, biological function together with genes and proteins were identified as groups that could benefit from a reclassification. Biolinks focuses on a division of those SGRs where more than one main branch is observed in the hierarchy, is-a, tree. For instance, two branches are clearly observed in the Living Beings (LIVB) UMLS group, one starting with the STY "Organism" and the other one with the STY "Group". In order to assess our approach, we analyze the semantic coherence and the Information Gain (IG) [87] regarding a particular collection, namely the Text Retrieval Conference 2005 Genomics Track collection (TREC-05) [9, 10].

5.2 Methodological approach

5.2.1 TREC-05 collection

The TREC-05 collection comprises a ten-year subset of MEDLINE articles, all of them identified with a PubMed identifier (PMID). This collection contains 34633 articles distributed across 50 topics corresponding to different information needs. A domain-expert group of people assigned one of three relevance judgements to PMIDs in each topic. The possible relevance judgments were: relevant, partially relevant and non-relevant. In Biolinks, we base our analyses on a subset corresponding to relevant and partially relevant articles for which title-and-abstract were available via National Center for Biotechnology Information (NCBI) Entrez Programming Utilities (e-utils*)* [88]. Such a subset includes 4240 articles. E-utils are a set of web services supporting programmatic access to different databases hosted at NCBI such as PubMed and PubMed Central Open Access (PMC-OA). A summary of the TREC-05 topics with more than 100 relevant and partially relevant articles is presented in Table 6.

Table 6. TREC-05 topics with more than 100 relevant and partially relevant articles. Taken from [89].

Topic code	Description	Non-relevant articles	Partially relevant articles	Relevant articles	Relevant and partially relevant articles with data for title-

					and-abstract
117	Role of the gene Apolipoprotein E (ApoE) in the disease Alzheimer's Disease	385	182	527	653
146	Mutations of hypocretin receptor 2 and its/their role in narcolepsy	388	67	370	421
114	Role of the gene APC (adenomatous polyposis coli) in the disease Colon Cancer	375	169	210	346
120	Role of the gene nucleoside diphosphate kinase (NM23) in the process of tumor progression	182	122	223	331
126	Role of the gene P53 in the process of apoptosis	1013	117	190	307
142	Sonic hedgehog mutations and its/their role in developmental disorders	257	120	151	263
108	Procedure or methods for identifying in vivo protein-protein interactions in time and space in the living cell	889	127	76	191
107	Procedure or methods for normalization procedures that are used for microarray data	294	114	76	189
111	Role of the gene PRNP in the disease Mad Cow Disease	473	93	109	185
109	Procedure or methods for fluorogenic 5'-nuclease assay	210	14	165	175
106	Procedure or methods for chromatin IP (Immuno Precipitations) to isolate proteins that are bound to DNA in order to precipitate the proteins out of the DNA	1061	125	44	158

5.2.2 Annotation with UMLS concepts

In Biolinks we use a semantic annotator supporting UMLS concepts, the Concept Mapping Annotator (CMA) [90, 91]. CMA follows a mapping process: text chunks in the text are judged relevant if a corresponding CUI exists in UMLS. Similar to other semantic annotators, CMA allows tuning the mapping process by adjusting a required level of confidence. In Biolinks, we use a low setting in order to favor a high recall. CMA does not only provide the corresponding CUI but also the STY and the SGR. Having the STY allows users to easily categorize terms according to an alternative semantic grouping, offering an advantage for Biolinks purposes. We use CMA rather than NCBO Annotator or Whatizit because CMA supports UMLS semantic types and groups as well as customized versions of them.

Additional to the CUI, STY and SGR, CMA semantic annotations contain as well the initial and length of the mapped text, a confidence score and the inverse document frequency (IDF) associated to the corresponding CUI. The annotations obtained from the 4240 relevant and partially relevant articles with title-and-abstract conform our TA-dataset. One more dataset was created from the TA-dataset, the TAFT-dataset, containing annotations on title-and-abstract for the 62 articles with full-text retrieved. CMA was also used over the full-text of these 62 articles, conforming our FT-dataset. The TA-dataset is used for the analyses in this and the subsequent chapters while the TAFT and FT datasets are not used in the chapter but later in this document. A summary of the datasets extraction is illustrated in Figure 16.

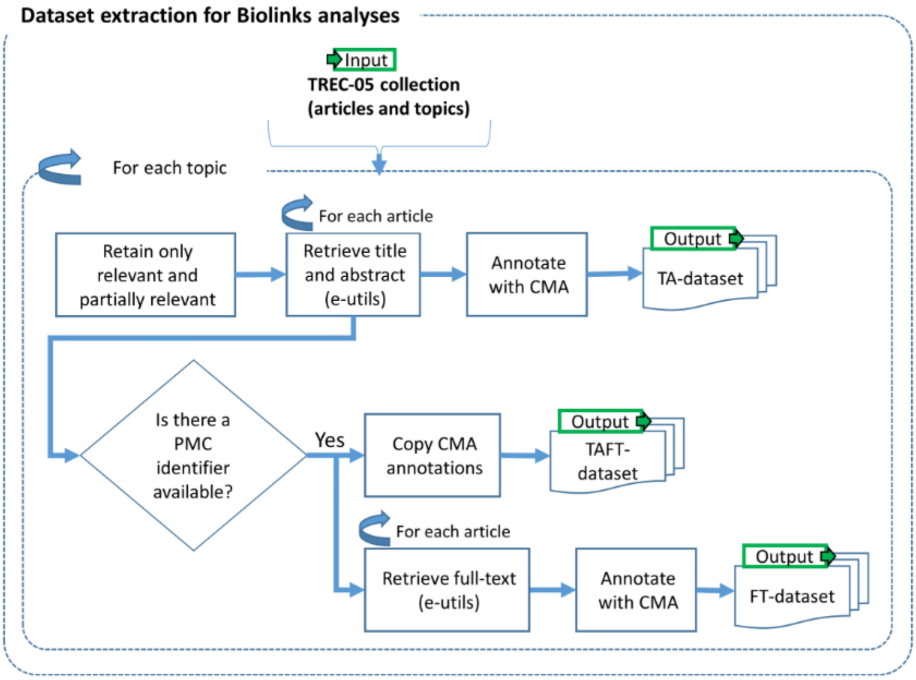

Dataset extraction for Biolinks analyses
Input
TREC-05 collection (articles and topics)
For each topic
For each article
Retain only relevant and partially relevant
Retrieve title and abstract (e-utils)
Annotate with CMA
Output
TA-dataset
Is there a PMC identifier available?
Yes
Copy CMA annotations
Output
TAFT-dataset
For each article
Retrieve full-text (e-utils)
Annotate with CMA
Output
FT-dataset

Figure 16. Annotation process with CMA annotator on the TREC-05 relevant and partially relevant articles.

5.2.3 Semantic groups reorganization

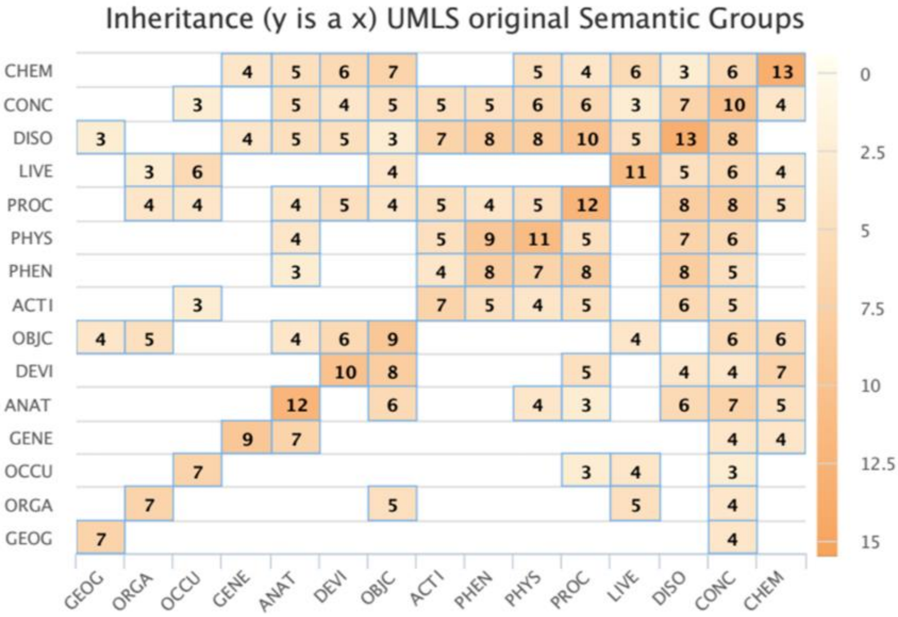
As previously mentioned, in order to configure the Biolinks semantic groups, we focus on those SRGs where more than one cluster is clearly exhibited. Particularly we focused on:

- LIVB displays two clusters, one corresponding to organisms (STY T001) and another to groups of people (STY T096);
- Disorders (DISO) displays 4 clusters, the first one covering "anatomical abnormalities" (STY T190), the second one "injuries or poisoning" (STY T037), the third one pathological functions (STY T046), and the forth one findings (STY T033);
- Physiology (PHYS) displays two clusters, one corresponding to physiological functions (STY T039) and another to organism attributes (STY T032); and
- Chemical (CHEM) displays three clusters, the first one corresponding to chemicals viewed structurally (STY 104), the second one to chemicals viewed functionally (T120), and the third one to clinical drugs (STY T200).

In order to evaluate the Biolinks group configuration, we assess the semantic coherence with respect to the **is-a** relations of CUIs and their STYs. Additionally, we use the TA-dataset together with the corresponding TREC-05 topics in order to analyze the IG regarding the UMLS SGRs and the Biolinks configuration. Our goal is finding a configuration maximizing both semantic coherence and IG.

For the semantic coherence, we define a coherence matrix of semantic groups, each cell displaying the number of times that each pair of semantic groups relates via **is-a** at a CUI level. The higher the number of **is-a** relations within concepts of the same group, the better the configuration. We built a baseline matrix based on the current SGRs in UMLS. In order to do so, we used the Metathesaurus relationships file (MRREL). The entries from this file considered in Biolinks correspond to **is-a** relations only. An **is-a** relation looks like **C0330206 is-a C0330148**. Both C0330206 and C0330148 are CUIs, all of them associated to a STY that it is in turn associated to a SGR. The initial relations of the type **CUI1 is-a CUI2** are translated as **STY(CUI1) is-a STY(CUI2)**, which are again translated as **SGR(CUI1) is-a SGR(CUI2)**. Those **is-a** relations between SGRs are used to build an upper-level inheritance matrix.

Our baseline matrix is shown in Figure 17. The number displayed in the different cells corresponds to the logarithm of the number of concepts in the y-axis related via **is-a** to concepts in the x-axis group. For instance, the first group on the y-axis, CHEM, has a value of 4 regarding inheritance from group GENE; this means that about 10000 concepts in CHEM inherit from concepts in GENE. The maximum scores for each pair are placed in the diagonal, indicating a high semantic coherence. However, there are other cells with high scores that would indicate potential incoherent assignments, like the cell PHYS-PHEN.



	GEOG	ORGA	OCCU	GENE	ANAT	DEVI	OBJC	ACTI	PHEN	PHYS	PROC	LIVE	DISO	CONC	CHEM
CHEM				4	5	6	7			5	4	6	3	6	13
CONC			3		5	4	5	5	5	6	6	3	7	10	4
DISO	3			4	5	5	3	7	8	8	10	5	13	8	
LIVE		3	6				4					11	5	6	4
PROC		4	4		4	5	4	5	4	5	12		8	8	5
PHYS					4			5	9	11	5		7	6	
PHEN					3			4	8	7	8		8	5	
ACTI			3					7	5	4	5		6	5	
OBJC	4	5			4	6	9					4		6	6
DEVI						10	8				5		4	4	7
ANAT					12		6			4	3		6	7	5
GENE				9	7									4	4
OCCU			7								3	4		3	
ORGA		7					5					5		4	
GEOG	7													4	

Figure 17. is-a inheritance matrix for UMLS SGRs. The higher the concentration in the diagonal, the better the semantic coherence. Taken from [6].

We use the TA-dataset and the UMLS SGRs in order to obtain a baseline IG. To do so, we first calculate the total TF-IDF per group and article, as well as per group and TREC-05 topic, then we get the entropy per TREC-05 topic as well as per group and TREC-05 topic, and, finally, we obtain the IG per group and TREC-05 topic. All the different steps are detailed in the Equation 1. All the TREC-05 topics are represented by $T=\{t_1, \cdots, t_M\}$ while all the groups are represented by $G=\{g_1, \cdots, g_N\}$. As observed in the equation, we first calculate the IG for a specific TREC-05 topic regarding all the groups $IG(t_i, G)$ and then use these partial IG to calculate the IG regarding all TREC-05 topics and semantic groups, i.e., $IG(T, G)$. The calculation of the IG requires to calculate the probability of a specific TREC-05 topic given a specific semantic group, $P(t_i, g_j)$; this probability is based on the summation of all the TF-IDF for those concepts belonging to the group g_j in documents contained in the TREC-05 topic t_i.

Equation 1. IG calculation for documents in TREC-05 topics. Taken from [6].

$t_1 \cdots t_M$ represents TREC-05 topics, $g_1 \cdots g_N$ represents semantic groups and $d_1 \cdots d_P$ represents documents for a given TREC-05 topic

$$IG(t_i, G) = H(t_i) - H(t_i \mid G)$$

As we are distributing across groups, we have

$$H(t_i) = -\sum_{j=1}^{N} P(t_i, g_j) \quad \times \log_2(P(t_i, g_j))$$

$$as\ well\ as\ H(t_i \mid G) = \sum_{j=1}^{N} prob(g = g_j) \times H(t_i \mid g = g_j)$$

The final IG for the dataset and the list of groups is given as

$$IG(T, G) = \sum_{i=1}^{M} IG(t_i, G)$$

In order to calculate $P(t_i,g_j)$, we first calculated the **total_tf_idf** per and group, thus, given a topic t_i and a group g_j, we have

$$total\_tf\_idf(t_i, g_j) = \sum_{k=1}^{P} total\_tf\_idf(d_k, g_j) \quad with\ d_k\ a\ document\ in\ t_i$$

$$then\ P(t_i, g_j) = \frac{total_{tf_{idf(t_i,g_j)}}}{\sum_{m=1}^{M} \sum_{n=1}^{N} total_{tf_{idf(t_m,g_n)}}}$$

With the probability for a given group g_j defined as

$$prob(g = g_j) = \frac{\sum_{m=1}^{M} total\_tf\_idf(t_m, g_j)}{\sum_{m=1}^{M} \sum_{n=1}^{N} total\_tf\_idf(t_m, g_n)}$$

5.3 Results

5.3.1 Annotation with UMLS concepts

Our TA-dataset comprises 4240 articles with 260666 annotations corresponding to 146353 different concepts, giving us an average of 61 annotations and 34 concepts per article. Regarding the 62 full-text articles, a total of 5376 annotations and 3071 concepts were identified in our TAFT-dataset, while a total of 83937 annotations and 23926 concepts in the TA-dataset. The ratio between concepts versus annotations, 0.56, is the same across the title-and-abstract datasets. That ratio decreases to 0.28 for the FT-dataset meaning a higher repetition of the same concept across the collection.

In Figure 18, we present the distribution regarding the different journals existing in the full-text subset extracted from the TREC-05 collection. The "Journal of Cell Biology" is the best represented one with about 40% of the articles, followed by "Genome Biology" and "BMC Bioinformatics".

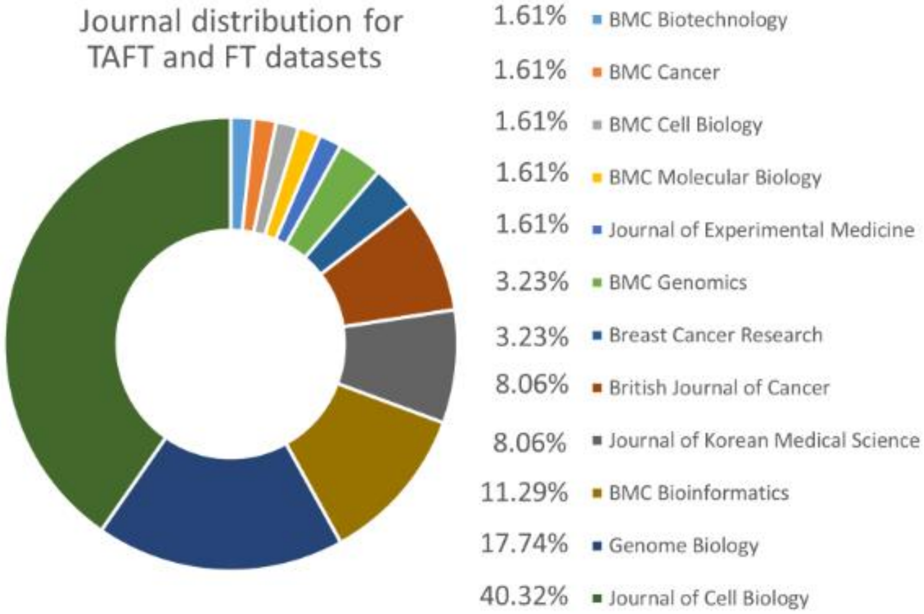
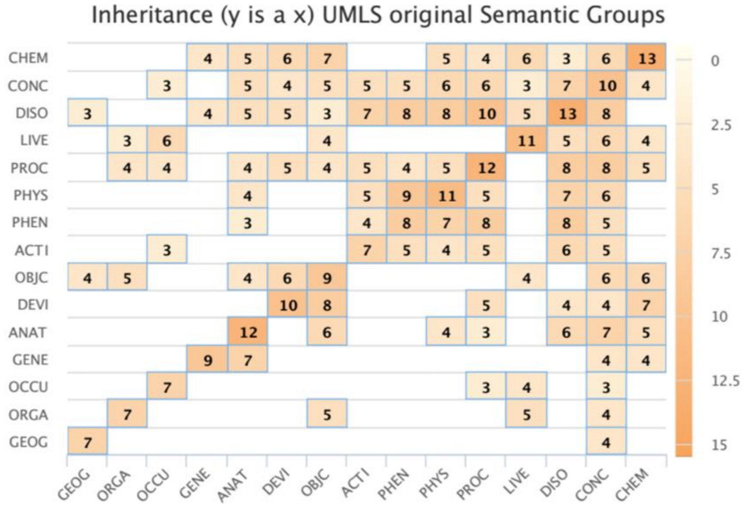


Figure 18. Journal distribution for the TAFT and FT datasets. Taken from [6].

5.3.2 Biolinks semantic groups

We analyzed about three possible redistributions of the UMLS SGRs LIVB, DISO, PHYS and CHEM and one more including the SGR Concepts & Ideas (CONC). The selected configuration was that one maximizing both the IG and the semantic coherence regarding the is-a hierarchy matrix. . The IG for UMLS SGRs and the TREC-05 topics was 7.521818 while the IG for Biolinks groups was 7.92894, representing an enhancement of 5.41%. In Figure 19 we show again the inheritance matrix for the UMLS original SGRs along with the matrix for the Biolinks groups. As observed, there is an increment of values in the diagonal, showing a higher semantic coherence.

Matrix A



Matrix B

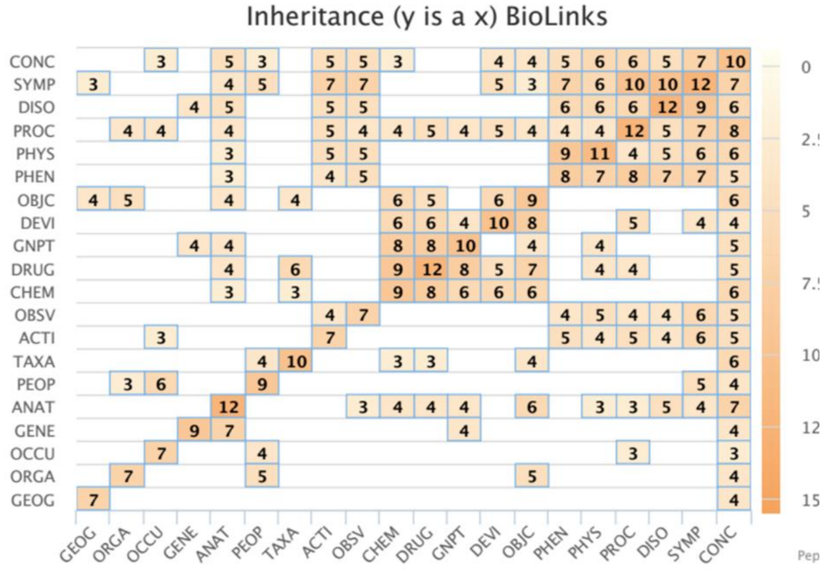


Figure 19. Inheritance matrices for UMLS original SGRs –matrix A, and Biolinks –matrix B. Taken from [6].

The UMLS group LIVB was split in two new groups: TAXA for all of those STYs extending from “Organism” and PEOP for all of those extending from “Group” (see Figure 20).

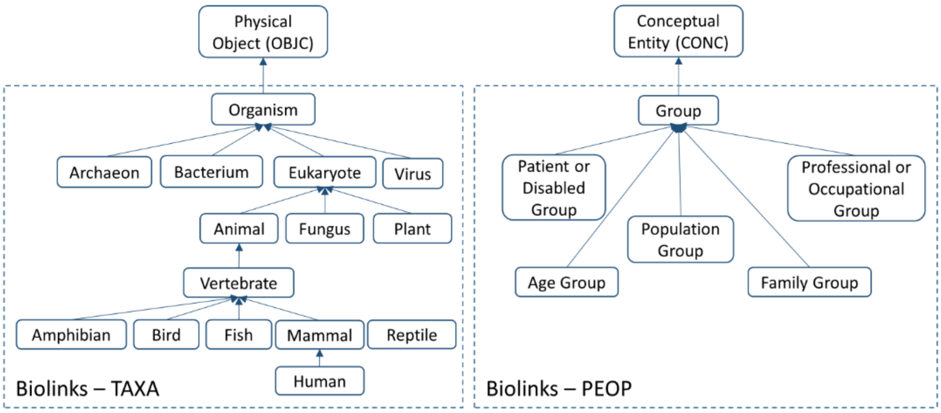


Figure 20. Former group LIVB split in TAXA and PEOP.

The UMLS group DISO was broken down in two groups, one keeping the group name DISO' and another for Symptoms (SYMP) (see Figure 21). DISO' contains those types related to disorders and diseases, all of them extending from either a type in the SGR "Anatomy" (ANAT) or types in the SGR Phenomenon (PHEN), and SYMP for those related to symptoms or findings in patients, extending from a type in the SGR CONC.

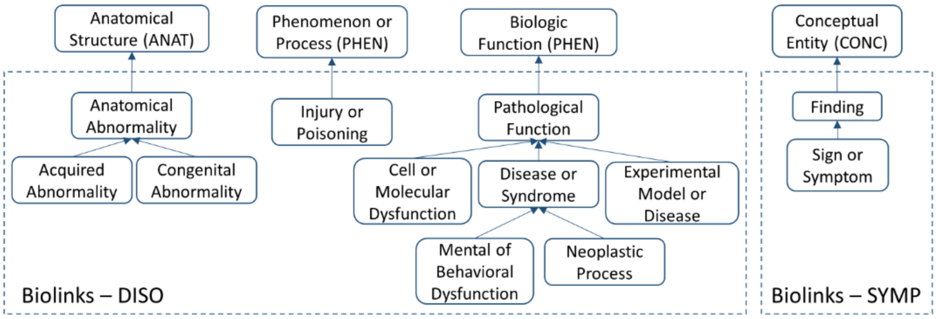


Figure 21. Former group DISO split in DISO' and SYMP.

The UMLS group PHYS was split in two groups, namely: one retains the group name PHYS', and the other is named Observations (OBSV) (see Figure 22). PHYS' contains types "Physiological Function" and its descendants while OBSV contains types "Organism Attribute" as well as its descendants.

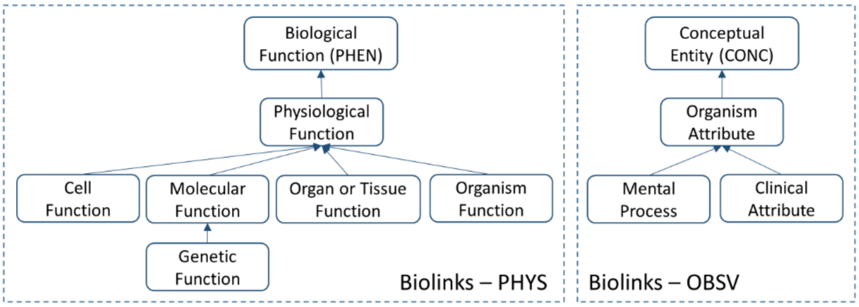


Figure 22. Former group PHYS split in PHYS' and OBSV.

Finally the UMLS group CHEM was broken down in three groups, CHEM', Genes and Proteins (GNPT), and DRUGs (see Figure 23). GNPT contains four types highly related to either nucleic or amino acid molecules, DRUG contains three types related to drugs or pharmaceutical substances, and CHEM' containing the rest of the types originally in CHEM. The remaining SGRs were not modified: activities and behaviors (ACTI), ANAT, CONC, devices (DEVI), genes and molecular sequences (GENE), geographic areas (GEOG), objects (OBJC), occupations (OCCU), organizations (ORGA), PHEN, and procedures (PROC).

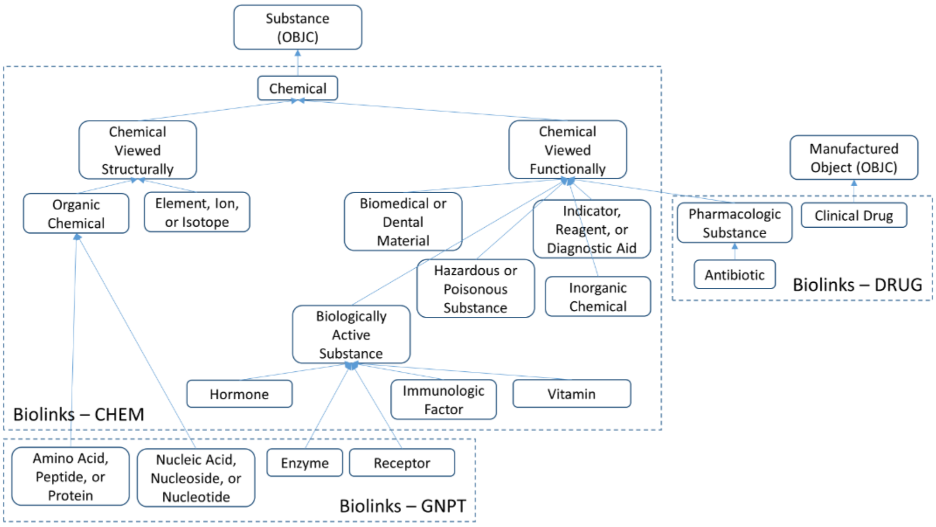

Substance (OBJC)
Chemical
Chemical Viewed Structurally
Chemical Viewed Functionally
Organic Chemical
Element, Ion, or Isotope
Biomedical or Dental Material
Indicator, Reagent, or Diagnostic Aid
Hazardous or Poisonous Substance
Inorganic Chemical
Biologically Active Substance
Hormone
Immunologic Factor
Vitamin
Biolinks – CHEM
Amino Acid, Peptide, or Protein
Nucleic Acid, Nucleoside, or Nucleotide
Enzyme
Receptor
Biolinks – GNPT
Manufactured Object (OBJC)
Pharmacologic Substance
Clinical Drug
Antibiotic
Biolinks – DRUG

Figure 23. Former group CHEM split in CHEM', GNPT and DRUG.

5.4 Discussion

Our initial approach to reconfiguring the UMLS SGRs is similar to a previously visual approach proposal [85]. Alike such a proposal, we started our analysis from the STYs hierarchy tree, focusing on the semantic coherence for the SRGs. Although different methods were used, our final distribution is similar to the possible partitions reported by the authors [85]. Bodenreider & McCray report a high cohesion for OCCU and ORGA while a large dispersion for CONC, OBJC and PROC. We considered all of them wide-spectrum groups and such we did not find any added value in reconfiguring them. We actually tried a configuration splitting CONC in two groups, one for concepts and another one for ideas; neither the IG nor the semantic coherence exhibited an improvement therefore we left those groups untouched. GEOG and DEVI are both small groups so we did not perform any partition on them either.

In [85], the authors find ACTI, LIVB and CHEM being organized around several distinctive poles. In the ACTI case, this is clear from the group name itself, "Activities and behaviors". Only one STY corresponds to behaviors thus we did not split the group. In the case of LIVB, there are two clear and separated branches, one for types related to taxonomical entities and another one related to groups of people. CHEM is a more complex case; there are three clear branches, "chemical viewed structurally", "chemicals viewed functionally" and "clinical drug". In this case we focused more on the STYs definitions and affinities amongst them.

DISO, PHYS and PHEN are reported as essentially cohesive groups with one isolated branch each. The isolated branch in DISO corresponds to the types "Finding" and its descendant "Sign or Symptom"; we assigned a new group SYMP to them. In PHYS there are two main types: "Physiological Function" and "Organism Attribute". The former extends from a type "Biological Function" belonging to the group PHEN while the latter extends from a type "Conceptual Entity" belonging to the group CONC. We kept the groups PHYS' for the first branch while created a new group OBSV for the second one. In the case of PHEN, only one type is isolated, "Laboratory or Test Result"; same as for ANAT, we kept these groups together in order to avoid groups with only one STY.

The re-grouping proposed by Biolinks results in an increment of the IG for the TREC-05 collection as well as an improved semantic coherence. This semantic coherence takes into account not only the TREC-05 collection but also the distribution of the different UMLS concepts along the **is-a** relations of themselves and their corresponding STYs. Although the Biolinks groups are promising and have been successfully used in our research, further re-organizations should be explored. We are especially interested in analysis regarding the semantic re-configuration at the type level, i.e., STY, so the coherence can be further improved, particularly for those noisy groups observed in the Biolinks matrix in Figure 19 –e.g., CONCepts and Ideas, which remains a rather generic group.

Chapter 6

Semantic group-based distribution for scientific publications

6.1 Background

Scientific publications are naturally related one to each other. Such relations come immediately from shared authors as well as bibliographic references. Co-citation occurs when at the same article, two other articles are cited; those two articles are said to be co-cited. Co-citation is recognized as a measure of relatedness across documents. The more time two articles are co-cited, the higher the co-citation factor is, indicating a likeness of relation between them [92, 93]. Although useful, co-citation analyses lack of a semantic layer. Such a layer becomes possible when connectivity is analyzed from the ontological concepts shared across documents. Despite the existence of widely used vocabularies in the biomedical domain such as Medical subject headings (MeSH) [94], SNOMED [95], and Unified Medical Language System (UMLS) [83], concept-based connectivity across scientific publications in the biomedical domain remains underexploited.

Analyzing the concepts shared between two publications could lead to a better understanding of how they are related. However, annotations are not all necessarily at the same level, some could be judged more relevant than others. Relevance varies from researcher to researcher. Some researcher might be interested in algorithms or methods used to achieve a particular purpose while some other could be more focused on enzymes used to accelerate a particular reaction. Annotations associated to expressions such as "Markov Chain", "probabilistic regression", "protein signature" would be more interested for the former researcher while "enzyme", "growth factor", "catalysis", "antigen" would be more attractive to the latter.

The UMLS semantic Groups (SGRs) have been defined for organizational reasons in order to better manage the conceptual complexity of semantic types (STYs) [84]. In a similar vein, scientific publications can be analyzed and categorized based on the SGRs assigned to those annotations extracted from either the title-and-abstract or full-text. Understanding how groups distributed along a document would be a first step in order to profit from the concept-base connectivity across publications. It could, for instance, facilitate analyses based on similarity. A similarity network could be navigated according to the most representative or relevant semantic group identified from the annotations in the text. This would lead to better recommendation systems, thus facilitating literature-based knowledge discovery.

Here we present a semantic group-based distribution formula. Such a formula weights the presence of each Biolinks semantic group in a publication annotated with UMLS concepts. We also analyze differences regarding annotations on only title-and-abstract versus full-text.

6.2 Methodological approach

6.2.1 Data

We use the TA, TAFT and FT datasets obtained from the TREC-05 collection, see 5.2.1 –TREC-05 collection, and 5.2.2 –Annotation with UMLS concepts. The TA-dataset is used to estimate a parameter in the distribution formula while the TAFT and FT datasets are used to analyze similarities and differences when applying the proposed formula on title-and-abstract and full-text documents.

6.2.2 Semantic group-based distribution

We propose a formula to weight Biolinks groups in a document annotated with UMLS concepts. Rather than classifying a group according to its more relevant group, i.e., the group with the highest weight, we assign a weight to each group with at least one annotation. Our formula takes as input a set of groups $G=\{g_1, \cdots, g_N\}$ and a document d with annotations $A=\{a_1, \cdots, a_S\}$; all annotations are as well distributed along the groups. An estimated parameter, $\gamma(g_j)$, is used to take into account the group distribution across a corpus. In our case, $\gamma(g_j)$ is estimated from the IG for the Biolinks groups identified in the TA-dataset. The full set of weights gives us the semantic group-based distribution for a publication. Our formula is presented in Equation 2. The group-based distribution formula for a document d regarding a group g_j is expressed as a ratio between the weight for all the annotations belonging to the semantic group g_j and the summation of the weights for all the semantic groups; with the weight being smoothed by an estimated parameter $\gamma(g_j)$. The weight is defined as the summation of the TF-IDF for all the annotations belonging to g_j.

Equation 2. Group-based distribution formula.Taken from [6].

We have a document **d** with annotations $a_1, \cdots, a_S$ distributed in groups $g_1, \cdots, g_N$.

$$dist(d \mid g_j) = \frac{\gamma(g_j) \times weight(d, g_j)}{\sum_{m=1}^{M} \gamma(g_m) \times weight(d, g_m)}$$

$$With\ weight\ defined\ as$$

$$weight(d, g_j) = \sum_{k=1}^{S} tf\_idf(a_k)$$

where a_k corresponds to a concept in g_j

From the Equation 1 we have

$$H(t_i \mid G) = \sum_{j=1}^{N} prob(g = g_j) \times H(t_i \mid g = g_j)$$

Thus a partial

$$H(t_i \mid g_j) = prob(g = g_j) \times H(t \mid g = g_j)$$

And finally we have the estimated parameter γ defined as

$$\gamma(g_j) = \sum_{i=1}^{M} partial\_H(t_i \mid g_j)\ with\ t_i\ a\ TREC\_05\ topic$$

6.3 Results

6.3.1 Group-based distribution for TAFT and FT datasets

Concepts & Ideas (CONC) and Genes and Proteins (GNTP) were the groups most observed in both the TAFT and TA datasets. In about 53% of the 62 full-text articles in the TAFT-dataset, CONC was the most dominant group. The same group was the most dominant one in about 46% of the articles in the FT-dataset. GNPT presence got 32% in the TAFT-dataset while a slightly higher number, 38%, was exhibited in the FT-dataset. In Table 7 we present a summary of the most representative groups across TREC-05 topics for articles with full-text.

Table 7. Most dominant groups per TREC-05 topic in the TAFT and FT datasets. The total number of articles corresponding to dominant group are shown before the group acronym. A dominant group is the one with the greatest score in a document. Taken from [6].

Topic	TREC-05 topic description	Total articles in topic	TAFT-dataset Biolinks groups	FT-dataset Biolinks groups
100	PROC/METH how to "open up" a cell through a process called "electroporation."	1	1 GNPT	1 CONC
103	PROC/METH green fluorescent protein (GFP) tagged proteins to do experiments with tagged proteins	1	1 CONC	1 CONC
106	PROC/METH chromatin IP (Immuno Precipitations) to isolate proteins that are bound to DNA in order to precipitate the proteins out of the DNA	1	1 GNPT	1 GNPT
107	PROC/METH normalization procedures that are used for microarray data	19	18 CONC 1 PROC	19 CONC
108	PROC/METH identifying in vivo protein-protein interactions in time and space in the living cell	6	4 CONC 2 GNPT	1 CONC 5 GNPT
111	Role of gene PRNP in the disease Mad Cow Disease	1	1 GNPT	1 GNPT
114	Role of gene APC (adenomatous polyposis coli) in the disease Colon Cancer	3	1 DISO 2 GNPT	1 DISO 2 GNPT
119	Role of gene GSTM1 in the disease Breast Cancer	3	1 CONC 2 GNPT	2 CONC 1 GNPT
120	Role of gene nucleoside diphosphate kinase (NM23) in the process of tumor progression	5	4 CONC 1 GENE	4 CONC 1 GENE
121	Role of gene BARD1 in the process of BRCA1 regulation	2	2 GNPT	1 GENE 1 GNPT
122	Role of gene APC (adenomatous polyposis coli) in the process of actin assembly	5	1 ANAT 1 CONC 2 DISO 1 GNPT	1 ANAT 2 DISO 2 GNPT
123	Role of gene COP2 in the process of transport of CFTR out of the endoplasmic reticulum	2	1 GNPT 1 PHYS	2 GNPT
124	Role of gene casein kinase II in the process of ribosome assembly	1	1 GNPT	1 GNPT
126	Role of gene P53 in the process of apoptosis	4	1 CONC 3 GNPT	4 GNPT
132	About genes APC (adenomatous polyposis coli) and wnt in colon cancer	1	1 GNPT	1 DISO
139	About genes Ret and GDNF in kidney development	2	1 GENE 1 GNPT	1 GENE 1 GNPT
141	About Huntingtin mutations and its/their role in Huntington's Disease	1	1 CONC	1 GNPT
146	About Mutations of presenilin-1 gene and its/their biological impact in Alzheimer's disease	4	3 CONC 1 GNPT	1 CONC 3 GNPT

Figure 26 makes it easier to grasp the differences regarding title-and-abstract against full-text when analyzing the group-based distribution. As already observed in Table 7, the most dominant groups are CONC and GNPT. CONC was expected to have a high representation because CONC is a rather generic group. GNPT is no surprise either as the TREC-05 collection is about genomics. Groups no longer conserve the same positions in both TAFT and FT datasets from the third position. For instance, Physiology (PHYS) and Procedures (PROC) present opposite tendencies. PHYS occupies the third position in the TAFT-dataset while the fifth in the FT-dataset. PROC, instead, rises from the sixth position to the third. Other groups such as Chemicals (CHEM), devices (DEVI) and DRUG are also better represented in the full-text than in title-and-abstract.

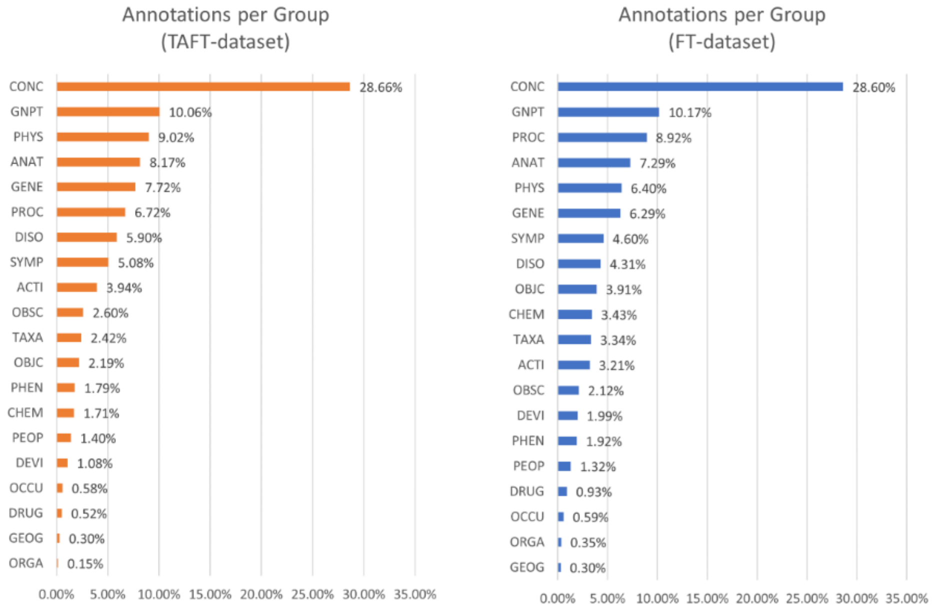


Figure 24. Biolinks group distribution across TAFT and FT datasets. Taken from [6].

We further analyzed the differences in the distribution when using annotations from title-and-abstract against full-text. In Figure 25 we present the Biolinks group distribution for four of the TREC-05 topics with more full-text articles. Each matrix represents the distribution of the Biolinks groups for a particular topic regarding the annotations observed in a particular dataset. The columns correspond to the articles while the rows correspond to the Biolinks groups; due to space limitations only the most relevant groups are explicitly named in the figure. It is possible to observe across the matrixes how the most dominant groups are usually conserved regardless the Text Retrieval Conference 2005 Genomics Track Collection (TREC-05) topics and the datasets. However, the FT-dataset has a better coverage of groups: some groups are observed in the full-text but not in the title-and-abstract. Additionally, some groups become more representative when moving from title-and-abstract to full-text. Such is the case of Topic 107 "PROC/METH normalization procedures that are used for microarray data" where GNPT is better represented in the full-text. A representation of this group based only on title and abstract could give the impression of poor information regarding genomics and proteomics; however the panorama changes when the full-text is taken into account.

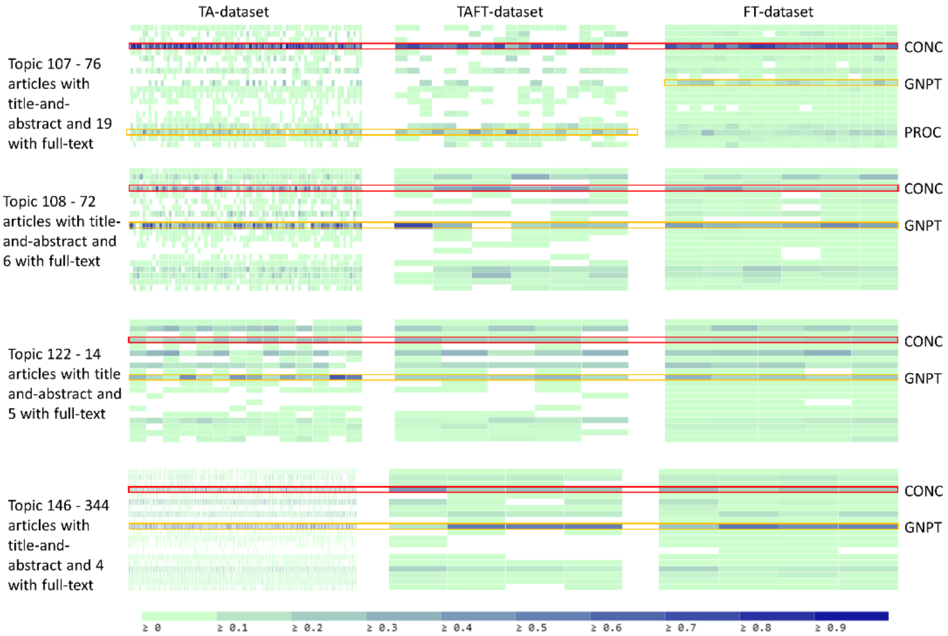


Figure 25. Group-based distribution for four selected TREC-05 topics. The two more representative groups are highlighted. Taken from [6].

6.3.2 Semantic model

Here we present our model[14] to represent group-based distribution in RDF (see Figure 26). We have defined some custom elements under the namespace "biotea". We also use the PAV ontology [71] to trace provenance and the Resource Description Framework Schema (RDFS) for general purposes.

The central object is a biotea:TopicDistribution, a topic in Biolinks refers to a semantic group. A distribution regards to, biotea:onDocument, an annotated document whose annotations are attributed, biotea:annotator, to an annotator and have been serialized by, pav:createdBy, a software agent on a particular date, pav:createdOn. A link to the model describing the semantic groups used is also recorded, biotea:hasModel. The total number of term occurrences is stored as an integer. Finally, each topic, i.e., semantic group, is available via biotea:hasTopic, the label corresponds to the name of the topic and the score to the weight within the distribution.

[14] Biotea ontology, including the group-based distribution model is available at https://github.com/biotea/biotea-ontololgy

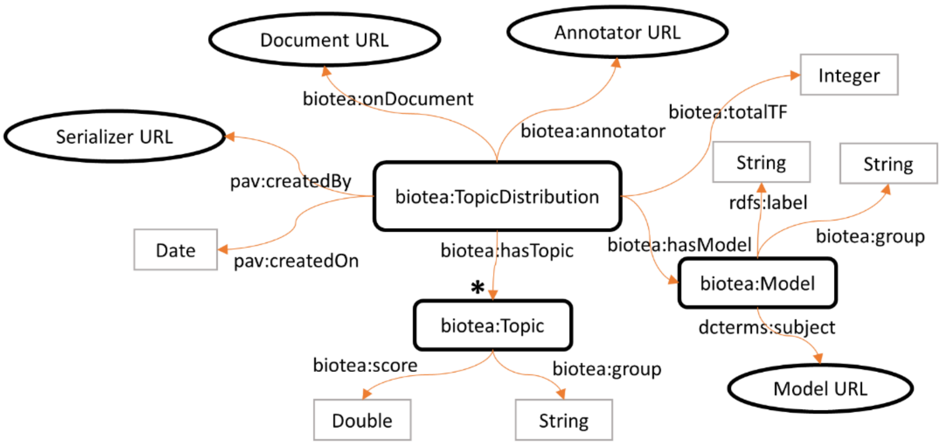


Figure 26. RDF model for group-based distribution.

6.4 Discussion

Analyses performed on the TAFT and FT datasets show that the group-based distribution does not significantly vary from one dataset to another. Due to the difficulties on freely obtaining full-text plus the extra time and machine-power required to process it, text-mining approaches on scientific literature have been traditionally performed on title-and-abstract. It has been argued that most representative terms will be contained there, thus it is fair to use them as a reasonable representation of the whole paper for term-based analyses. When it becomes to group-based distribution, the TREC-05 collection annotated with UMLS concepts seems indeed to support this claim. However, this might not necessarily be true if other collection or annotators are used.

Our distribution formula includes estimated parameters per group. We have calculated such parameters based on the IG formula applied to annotations and groups identified in the TA-dataset. As these parameters could be not the best estimation for other datasets, we calculated and analyzed the group-based distribution with a $\gamma(g_j)=1$ for all the Biolinks groups. The distribution scores did not exhibit big changes; however, the most representative group did vary in about 30% of the documents. Whenever calculating the estimated parameters become difficult for a particular corpus, we recommend using $\gamma(g_j)=1$, giving all groups the same probability regardless their general distribution across the whole corpus.

Although both Biotea and Biolinks focus mainly on full-text articles, the group-based distribution formula can be applied to either title-and-abstract or full-text articles. Furthermore, the distribution results could be used to cluster articles together, in a similar way as it is done in topic modeling. In the following chapter we will present how to use the distribution results in order to assess the similarity between two documents, narrowing it to groups of interest.

Chapter 7

A semantic similarity metric based on annotations

7.1 Background

Nowadays, scientific articles are made available in electronic formats. This not only improves dissemination but also makes it easier for repositories to offer functionalities on top of basic key-based search and retrieval. One of the additional features offered by PubMed consists of a list or similar or recommended articles to the one actually displayed on their web-based interface. Articles in this list are presented in descending ordered regarding the similarity score, from the most similar to the least.

The PubMed Related Articles (PMRA) metric [96] is used to calculate the similarity; scores are computed based on (i) word stems identified in title-and-abstract as well as (ii) Medical Subject Headings (MeSH) terms assigned to the reference document, i.e., the article being read. Neither the score nor the words participating in the similarity are provided, making it difficult for readers to realize how close or far apart two contiguous articles are in the similarity list. Although evidence suggests that this list of similar articles is a useful feature [96], from this list alone is not currently possible to understand how publications relate to each other.

Other similarity metrics, such as Best Matching 25 (BM25) [97-99], and Cosine Similarity [100, 101], could also be used to find similar articles. A benchmarking using related and partially related articles in the Text Retrieval Conference 2005 Genomics Track Collection (TREC-05), particularly the TA-dataset, was conducted in order to assess the performance of PMRA, BM25 and Cosine [89]. The purpose of that benchmarking was to determine how these three metrics behaved when applied to semantic annotations. PMRA outperformed the others. The evaluation considered correlation, precision, recall and F1 statistics to select the most adequate metric for semantic annotations.

Despite efforts to automatically recognize ontological entities in scientific publications, e.g., Whatizit [74, 75], Concept Mapping Annotator (CMA) [90, 91], MetaMap [102] and the National Center for Biomedical Ontology (NCBO) Annotator [76], only MeSH terms are currently used to calculate similarity between articles in PubMed. PMRA could be adapted to work with semantic annotations as those provided by Biotea. In this chapter, we extend Biolinks for it to include a semantic similarity metric based on PMRA. Furthermore, we take advantage of the group-based distribution so that similarity can be narrowed to groups of interest. In this way Biolinks makes it easier to understand where the similarity between two articles relies regarding semantic annotations and how similarity varies when narrowing annotations to some particular groups. We also analyze how similarity differs from annotations only in the title-and-abstract against those in the full-text.

7.2 Methodological Approach

7.2.1 Data

We use the TA, TAFT and FT datasets obtained from the TREC-05 collection, see 5.2.1 –TREC-05 collection, and 5.2.2 –Annotation with UMLS concepts. We apply our similarity formula to articles in all of the datasets. The TAFT and FT datasets are also used to analyze similarities and differences when applying the proposed formula on title-and-abstract versus full-text documents.

7.2.2 A semantic similarity metric

The similarity between two articles, (c, r), is defined as the probability of being interested in compared article c given a known interest in reference article r. The similarity score is calculated based on the terms identified in both articles c and r [96]. A term is either a single word like "phosphorylation" or several words associated with a single idea like "adenosine triphosphate (ATP)". PMRA is a Poisson-based probabilistic approach that considers the length of the document, $l(d)$, as well as the term frequency (TF) and inverse document frequency (IDF) for the identified word stems and MeSH terms. It also includes two parametric constants μ and λ. We use here the optimal values reported in [96], i.e., ., $\mu=0.013$ and $\lambda=0.022$. In our case, rather than using words or terms, we use concepts as identified by semantic annotators; thus, the length of the document is calculated from the total count of annotations, i.e., $l(d) = total\_ann(d)$. In particular, we have used annotations obtained from CMA applied to TREC-05 collection.

A similarity score involves a numerator including information from both the compared and the reference articles as well as a denominator including information about the reference article only. As PMRA is used to present the similar articles list in PubMed, where the ranking but not the final score is important, only the numerator is taken into account. In Equation 3 we present the original PMRA formula as well as its adaptation as a semantic similarity metric. In order to take into account the semantics behind the text, we use semantic annotations rather than word stems. The formulas presented in Equation 3 use all the annotations present in the compared and reference articles, regardless the groups they belong to; thus, it is referred to as a global semantic similarity metric.

Equation 3. Global **s**emantic similarity metric based on PMRA. Taken from [89]

Here we have the similarity between **c** and **r** with annotations $a_1,\cdots,a_N$ present in document **r** (but not necessarily in document c)

$$pmra\_ranking(c,r) = \sum_{t=1}^{N} w(a_t,c) \times w(a_t,r)$$

$$with\, w(a,d) = (1 + (\frac{\mu}{\lambda})^{tf\_idf(a,d)-1} \times\; e^{-(\mu-\lambda)\times l(d)})^{-1} \times \sqrt{idf(a)}$$

As 0≤similarity≤1.0, and similarity(r,r) = 1, we define the similarity metric as a normalized ranking

$$pmra\_sim(c,r) = \frac{\sum_{t=1}^{N} w(a_t,c) \times w(a_t,r)}{\sum_{t=1}^{N} w(a_t,r) \times w(a_t,r)}$$

The group-based distribution analyses carried on in the previous chapter showed that some groups dominate over others. This of course depends on the particular article under consideration. Additionally, researchers' interests vary from one to another. In order to take into account group-based distribution and particular interests, we propose a group-narrowed similarity formula (see Equation 4). One or more groups, noted by **G**, can be considered at once. In order to avoid multiplications by 0, only groups with at least one annotation present in the reference document are included. Remember that all groups are represented by **G**.

Initially, the PMRA semantic similarity is modified so that only the annotations belonging to the groups of interest are included. Then, a group-weight factor is added to the equation. Without this weight factor, underrepresented groups, i.e., groups with just a few annotations, would result in similarities close to 1. The group-weight factor involves the compared article **c** as well as the reference article **r**, **dist(c, r, G)**, and corresponds to a smoothened version of the group-based distribution score for each article **d**, **dist(d, G)**. A smooth factor, **alpha**, is incorporated into the formula to give more relevance to the group-based distribution in the reference article.

In order to estimate **alpha**, multiple values from 0.1 to 1.0 were contemplated. The goal was getting an alpha in such a way that the group-narrowed similarity over all groups had the best correlation with the global similarity. After analyzing all the possible values, **alpha=1** was selected.

Equation 4. Group-narrowed semantic similarity metric based on PMRA. Taken from [6].

Taking into account only annotations $\mathsf{a_t}$ in r belonging to groups $\mathsf{G'}$, we have

$$pmra\_sim(c,r,G') = dist(c,r,G') \times \frac{\sum_{t=1}^{N} w(a_t,c,G) \times w(a_t,r,G')}{\sum_{t=1}^{N} w(a_t,r,G') \times w(a_t,r,G')}$$

$$w(a,d,G') = (1 + (\frac{\mu}{\lambda})^{tf\_idf(a,d)-1} \times e^{-(\mu-\lambda)\times l(d,G\prime)})^{-1} \times \sqrt{idf(a)}$$

$$dist(c,r,G') = alpha \times dist(r,G') + (1 - alpha) \times dist(c,G')$$

$$With\ alpha = 1, we\ have\ dist(c,r,G') = dist(r,G')$$

$$dist(r,G') = \sum_{J=1}^{N} dist(r,g_j)\ with\ g_j \in G'$$

7.3 Results

7.3.1 Global similarity

We calculated the global similarity for all of our three datasets; however, we focused our analyses on TAFT and FT dataset so we could understand how similarity scores vary when calculated on annotations only on title-and-abstract against those in the full-text. In this regard, results on group-based distribution showed no significant variations, the most dominant groups in title-and-abstract could be the same as well in the full-text. Although it was not possible to identify a variation pattern, our analyses on similarity score reveal indeed a variation. We found cases were the similarity in title-and-abstract was close to that one in the full-text, but also cases were they were higher or lower.

As an example, in Figure 27, we present a scattered plot for a randomly selected article belonging to the TREC-05 topic “Procedures and methods for normalization procedures that are used for microarray data”. This topic is the one with more full-text articles, a total of 19 articles from the 62 articles with full-text. The reference article corresponds to PMID:12049663 “How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach" [103]. The scattered plot shows similarity values against all of the other 61 articles in the TAFT and FT datasets. We use squares to distinguish articles in the same topic and circles for articles in any other topic.

As observed in Figure 27, regardless the topic, similarity scores regarding title-and-abstract, y-axis, vary with respect to those for full-text, x-axis. However, no definitive pattern is perceived. We selected a second article in the same TREC-05 topic, PubMed identifier (PMID) 12537560 "Bayesian analysis of gene expression levels: statistical quantification of relative mRNA level across multiple strains or treatments" [104]. Both articles were published in 2002 in the journal Genome Biology but in different volumes; they do not cite each other. The similarity between these two articles is greater when considering title-and-abstract, 0.88, than for full-text, 0.69.

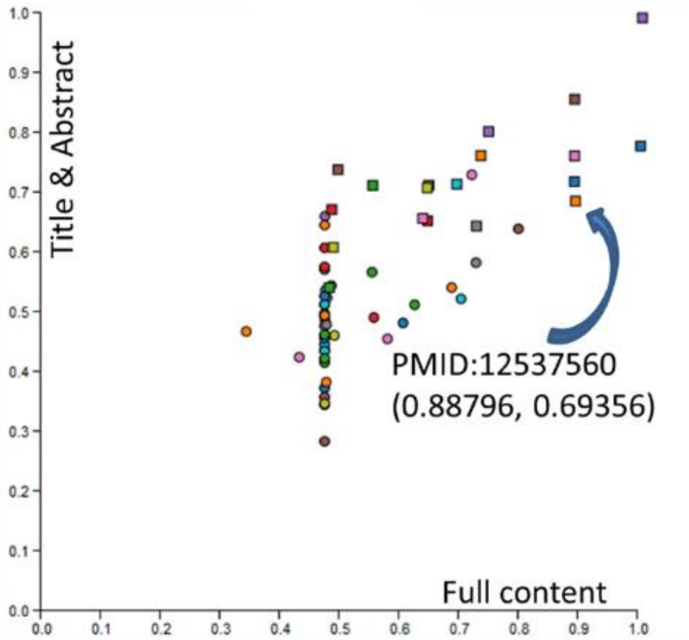


Figure 27. Title-and-abstract versus full-text annotations for PMID:12537560 given an interest in PMID:12049663. Adapted from [89]

Similar analyses were carried out for different pairs of the articles. In order to show a summary of the variations, we calculated the mean, median, and standard deviation for the similarities calculated on title-and-abstract as well as on full-text (see Figure 28). In the TAFT-dataset, media and mean are close to each other, suggesting a symmetric distribution. Although similarity scores concentrated between 0.4 and 0.6, multiple outliers are observed. The FT-dataset shows lower similarity scores, concentrated between 0.2 and 0.4. This suggests that a pair of articles in the TAFT-dataset that could be considered similar, with a score above 0.5, would probably show a different tendency in the FA-dataset, meaning a score below 0.5. No outliers are observed for the FT-dataset, which could be attributed to the increase of the amount of information when working with full-text.

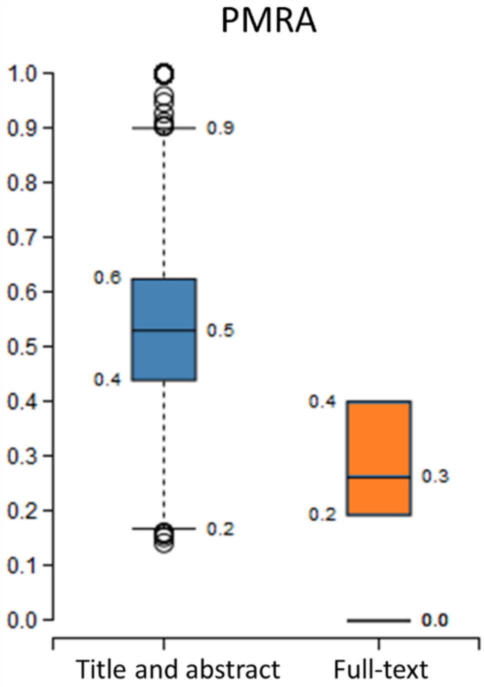


Figure 28. Box-plots for mean, median, and standard deviation for TAFT and FT datasets. Adapted from [89]

TAFT and FT datasets comprise eighteen TREC-05 topics, only five of them with at least four articles. This set of five topics is actually reduced to four as full-text in Topic 120 "Role of gene nucleoside diphosphate kinase (NM23) in the process of tumor progression" corresponds to scanned images, thus not machine-processable. The other four topics with at least four articles are: Topic 107 "PROC/METH normalization procedures that are used for microarray data" with 19 articles, Topic 108 "PROC/METH identifying in vivo protein-protein interactions in time and space in the living cell" with 6 articles, Topic 122 "Role of gene APC (adenomatous polyposis coli) in the process of actin assembly" with 5 articles, and Topic 146 "About Mutations of presenilin-1 gene and its/their biological impact in Alzheimer's disease" with 4 articles.

Global similarity for three topics is shown in Figure 29. In this graphic is again possible to observe how similarity scores vary from title-and-abstract to full-text. Scores are usually greater for full-text. The similarity order, from high to low scores, is not conserved either. A similar behavior is observed for Topic 107, which was excluded from the figure as it would become difficult to read for 19 articles.

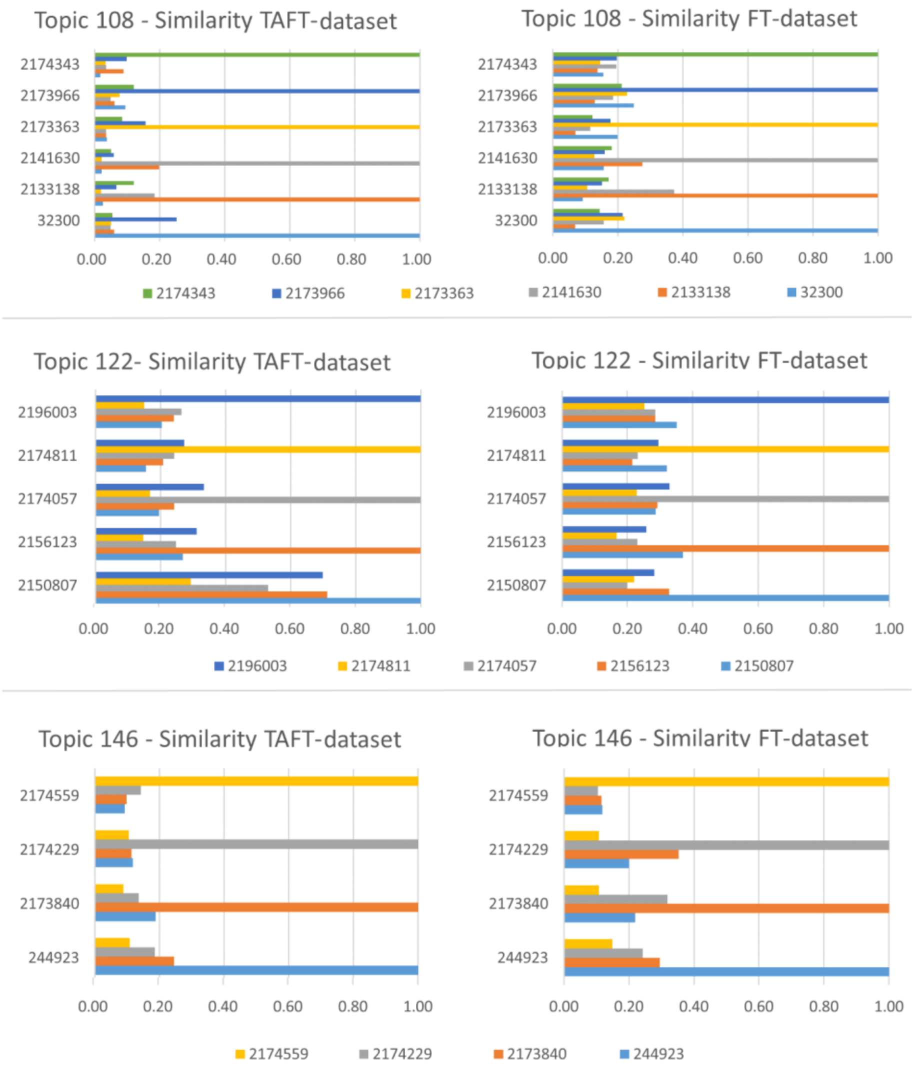


Figure 29. Global similarity for TAFT and FT datasets for some selected TREC-05 topics. Taken from [6].

7.3.2 Group-narrowed similarity

The same four TREC-05 topics with at least four articles were used to analyze variations when narrowing the similarity to some particular groups. We first analyze variations regarding global versus group-narrowed similarities. For the group-narrowed similarities, we selected the most representative groups according to the group-based distribution: Concepts & Ideas (CONC) and Genes and Proteins (GNPT). In Figure 30, we present the global similarity as well as the one narrowed to CONC+GNPT; only the TAFT dataset is considered in this figure. The similarity scores show a similar tendency regardless being global or group-narrowed. Neither the similarity values nor the ranking are conserved.

Topic 146 presents interesting insights. GNPT is the most representative group while CONC is not particularly strong. The weakness of the CONC group is also reflected in the similarity narrowed to this group. CONC-narrowed similarity is very poor compared to GNPT-narrowed scores. Furthermore, the GNPT-narrowed similarity is close to the global one, showing that the most representative group is the most relevant when it comes to narrow similarity by groups.

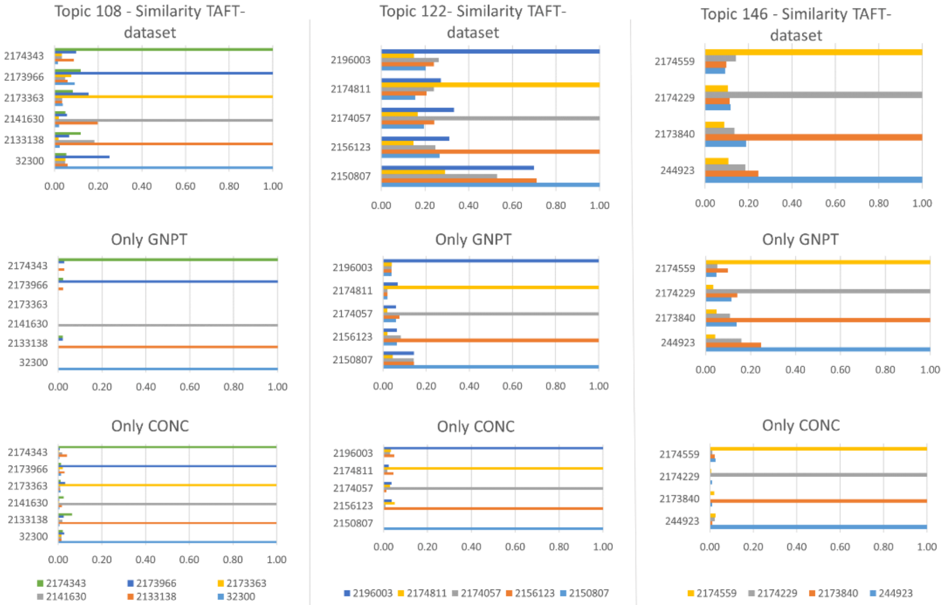


Figure 30. Group-narrowed similarity for some selected TREC-05 topics in the TAFT-dataset. Taken from [6].

In Figure 31, we present results for the same TREC-05 topics but this time for the FT-dataset. Once again, both the similarity values and the ranking order vary. Topic 122 does not exhibit a dominant group. Although CONC and GNPT are the most representative ones, they are closely followed by other groups. A similar behavior is exhibited by Topic 108. Topic 146 does have a strong group, GNPT. In Figure 31, it is possible to observe how neither CONC nor GNPT are strong enough to significantly influence in the group-narrowed similarity for topics 108 and 122.

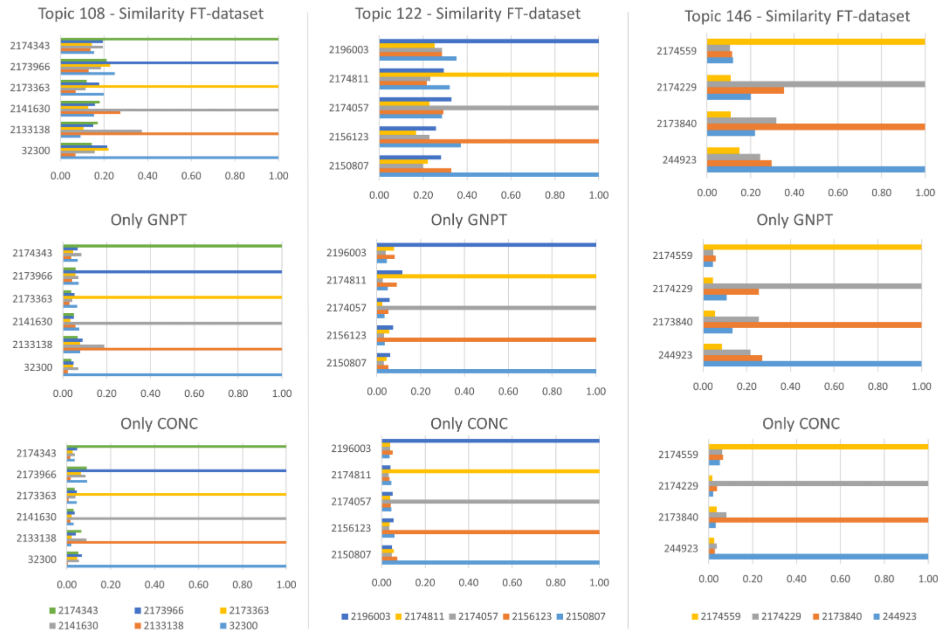


Figure 31. Group-narrowed similarity for some selected TREC-05 topics in the FT-dataset. Taken from [6].

7.3.3 Semantic Model

Biolinks provides a model corresponding to the semantic similarity, either global or group-narrowed (see Figure 32). Provenance, annotator and serializer agent information is similar to the one included in the group-based distribution model. Two documents are required, the reference document, a.k.a. query document, linked via biotea:onQueryDocument, and the compared document, linked via biotea:onRelatedDocument. The similarity score is stored as a double and bond, biotea:score, from the central class, biotea:Biolink.

Additional information such as the annotations contributing to the similarity are reached by biotea:link. Links to concepts as well as details regarding TF and IDF values are also kept. Whenever the group-narrowed option is used, the model describing the groups are saved, biotea:hasModel, as well as labels identifying the specified groups, biotea:group. We use a star symbol * whenever properties allow multiple values, otherwise it should be only one value. Classes and propertied inside the dotted box are used only if one or more particular groups were used during the similarity calculation.

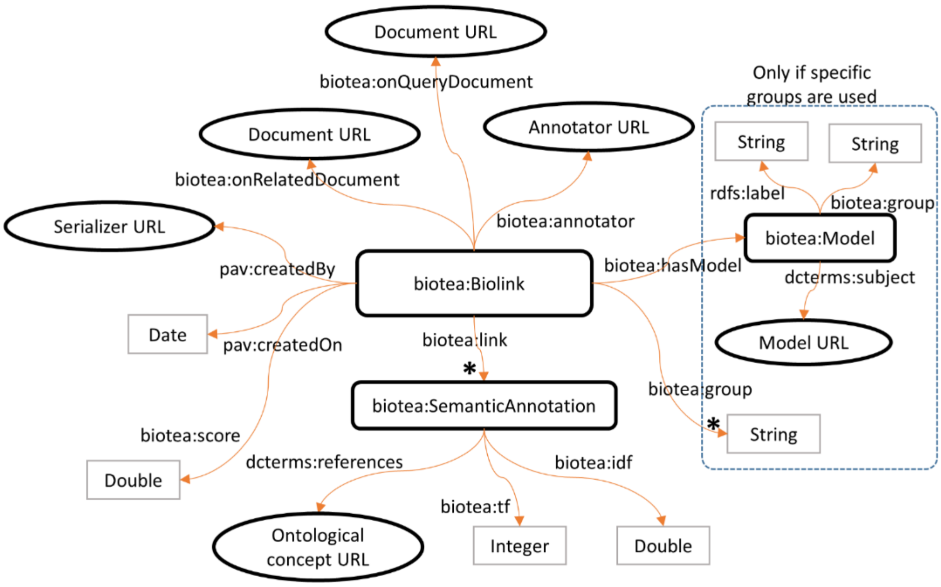


Figure 32. RDF model for semantic similarity between two articles.

7.4 Discussion

Scientific publications are nowadays commonly available in electronic formats. Furthermore, repositories such as PubMed offer web data services supporting machine-processable formats such as XML and JavaScript Object Notation-Linked Data (JSON-LD). However, most of the text available corresponds to title-and-abstract. Although the machine-processing capabilities get better and better and data move from megabytes to terabytes and petabytes, access to full-text publications remains limited. While PubMed covers more than 26 million articles in the biomedical domains, PubMed Central Open Access (PMC-OA) only covers over a million of articles, i.e., less than 5% compared to PubMed.

In the last years, with the aim to improve transparency and reproducibility, more and more, journals encourage sharing not only full-text but also code and data supporting processes and findings. Thus, scientific literature is slowly moving to a fully open-access model. As full-text become available, approaches aiming to provide a list of similar articles should get ready to take advantage of it. Furthermore, such approaches should also make use of entity recognition supported by efforts such as Whatizit [74, 75], CMA [90, 91], MetaMap [102] and the NCBO Annotator [76].

Despite requiring higher machine-processing as well as storage capabilities, using full-text offers some advantages in comparison to using only title-and-abstract. Although Shah and colleagues [105] found that abstracts contain the best ratio of keywords per total words, they also stressed the potential opened up by analyzing full-text. As the distribution of keywords is heterogeneous along sections, some biologically relevant keywords would be present outside the abstract. Sections other than the abstract could also contribute with specific data such as gene names, anatomical terms and organisms. Although we have not considered variations in similarity along different sections, our findings show that the similarity score indeed varies when considering full-text in contrast to only title-and-abstract, supporting the heterogeneous distribution of keywords along text; keywords corresponding to ontological concepts associated to terms identified in the text in our case. Our scattered plot analyses show that in about 50% of the 62 full-text articles, the global similarity based on title-and-abstract are close to 0 while they are above 0.5 when based on full-text.

Evidence suggests that the related article search in PubMed is a useful feature [96]; however the group-narrowed variation of the our semantic similarity formula makes it possible to go further. Group-based distribution provides insights on what the main subject –semantic group, is in a document. Similar or recommended article lists present a descending list from most related to least; such a list can be modified based on the groups more relevant to both, the articles in the list and the researchers' interests. Our analyses show that the similarity varies when it is narrowed to some particular groups. If a group is not representative, similarity regarding this group will be not significant, exhibiting values close to 0.

The PMRA similarity formula is not symmetric thus pair-wise analysis of articles in a group becomes difficult. By analyzing the formula itself, it is possible to observe how annotations only present in the reference document have a low impact in the overall result. Modifying the similarity formula by restricting it to concepts present in both documents would make it symmetric. This would make it simpler to analyze at once all document groups together, for instance groups as a TREC-05 topic, as a list of relevant citations for a particular research of as a list of recommended articles. Network-based visualizations would help to get a quick idea about closeness between articles in an all-versus-all comparison graph.

Chapter 8

A prototype to analyze groups of semantically enriched publications

8.1 Background

As publications in biomedicine increase day by day, so do the connections across them. Research articles are immediate connected by co-authorship and citations; however, this interconnectedness becomes richer when taking into account the actual terms found in the documents. A semantic dataset as the one obtained from Biotea together with the categorization and similarity techniques proposed by Biolinks make an ideal scenario to semantically navigate clusters of documents.

Once an article has been selected to read, the PubMed web interface presents on the right hand side a list of similar articles. Only about the first five titles are initially displayed. A "see all" link takes the user to a result set displaying information such as title, authors, journal, and publication dates are displayed. From this list and based on the information presented, users can easily navigate to any of the similar articles. Users are somehow blind regarding the terms and topics expected on those similar articles.

Here we present a prototype to navigate clusters of articles. Starting with a set of already known related articles, our prototype displays the group-based distribution, allowing users, for instance, to focus on those articles with terms within the groups more relevant to them or with a better representation. Once an article has been selected as the reference one, the similarity network regarding the rest of articles is presented as well as the cloud of annotations. From here, users can select any other article in the network so its annotations will be also displayed. Our goal is to present users with a summarized information related to the groups and terms existing in the articles as well as the similarity scores between them. By doing so we aim to facilitate users to make an informed decision when choosing one article or another to navigate to.

8.2 Methodological approach

8.2.1 Visualization components

Data visualization has become a common practice as a mean to present summarized data in a compact and intuitive way, facilitating data analysis and interpretation. Web-based components make it easier to increase the accessibility, dissemination and usage of data visualization as embedding them into web pages is simple and straightforward. BioJS [106, 107] is an open-source collection of web-based components for biological data visualization and manipulation. It encompasses a set of principles so components are discoverable, reusable and shareable. BioJS also defines guidelines in order to assure modularity and interoperability with other web-based components.

We have developed five visual and one parser component following BioJS principles and guidelines. Our components are written in JavaScript version ECMAScript5 and make use of well-known libraries such as jQuery [108], Data-Driven Documents (D3) [109] and Underscore [110]. jQuery facilitates traversing and manipulating Hypertext Markup Language (HTML) documents as well as handling events and retrieving data, D3 is used to manipulate and visualize documents based on data, while Underscore provides basic and advance functionality to traverse collections and arrays. We use jQuery to retrieve JavaScript Object Notation-Linked Data (JSON-LD) files containing Biotea annotations, D3 to render data extracted from those annotations in a graphical way, and Underscore to handle collections. Our data collections refer to annotations mainly but also to data derived from them, in particular group-based distribution as well as global and group-narrowed similarity.

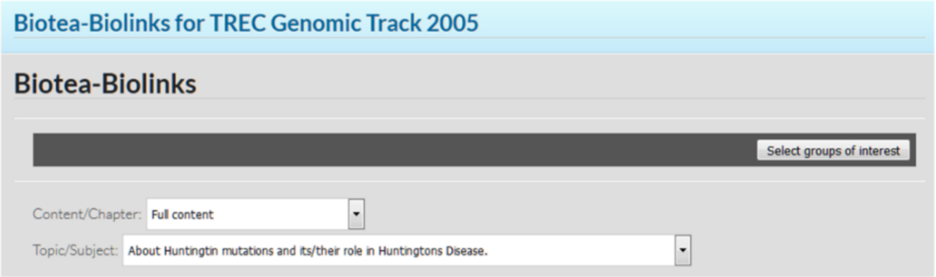
8.2.2 Showcase data

We use the TA, TAFT and FT datasets obtained from the TREC-05 collection, see 5.2.1 –TREC-05 collection, and 5.2.2 –Annotation with UMLS concepts. We apply our similarity formula to articles in all of the datasets. For all the datasets and articles is possible to display group-based distribution, similarity scores and cloud of annotations.

8.3 Results

We have created a mini-application showcasing the usage of Biolinks on top of relevant and partially relevant articles for the TREC-05 collection[15]. As shown in Figure 33, by default, the FT-dataset and the first alphabetical TREC-05 topic for such dataset –"About Huntingtin mutations and its/their role in Huntington's Disease", are selected. The "Select groups of interested" button is used to select some groups in order to narrow the group-based distribution, similarity and annotations to only those groups selected.

[15] The showcase code is available at https://github.com/ljgarcia/biotea-biolinks/tree/gh-pages and can be seen in action at http://ljgarcia.github.io/biotea-biolinks/biolinks4TREC.html



Biotea-Biolinks for TREC Genomic Track 2005

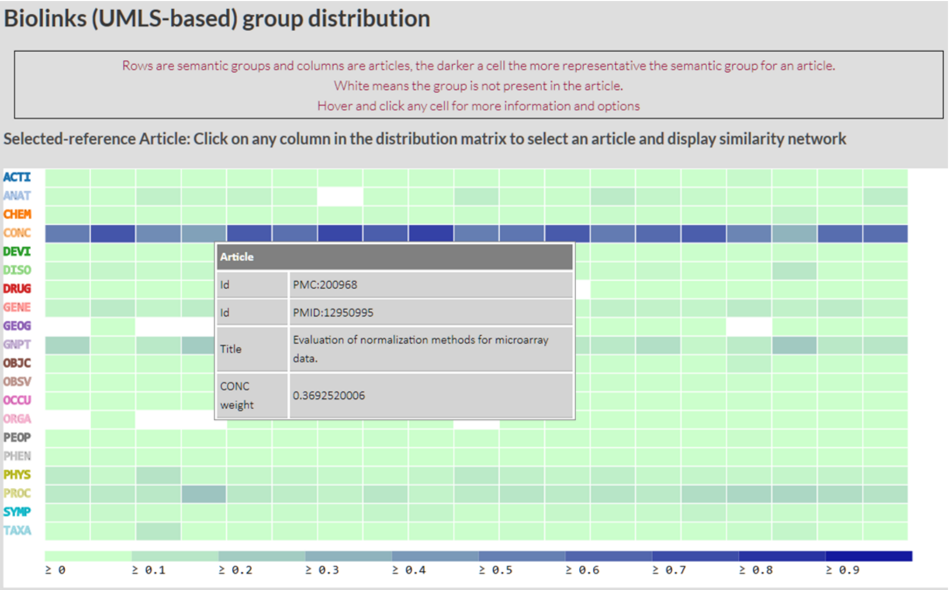
Biotea-Biolinks

Content/Chapter: Full content

Topic/Subject: About Huntingtin mutations and its/their role in Huntingtons Disease.

Figure 33. TREC-05 Biolinks showcase

Once a TREC-05 topic is selected, the group-based distribution for this topic is displayed. A heat-map is used to show what groups are more representative for each article on the selected TREC-05 topic. For instance, it is possible to observe how CONCepts & Ideas is the most representative semantic group for the TREC-05 topic "PROC/METH normalization procedures that are used for microarray data" in the FT-dataset (see Figure 34). A color scale is used to show how the group-based distribution varies: the darker blue the more representative, the lighter green, the less. On hovering an article, further information such as identifiers, tittle and distribution score regarding the group on the left side are displayed.



Biolinks (UMLS-based) group distribution

Rows are semantic groups and columns are articles, the darker a cell the more representative the semantic group for an article.
White means the group is not present in the article.
Hover and click any cell for more information and options

Selected-reference Article: Click on any column in the distribution matrix to select an article and display similarity network

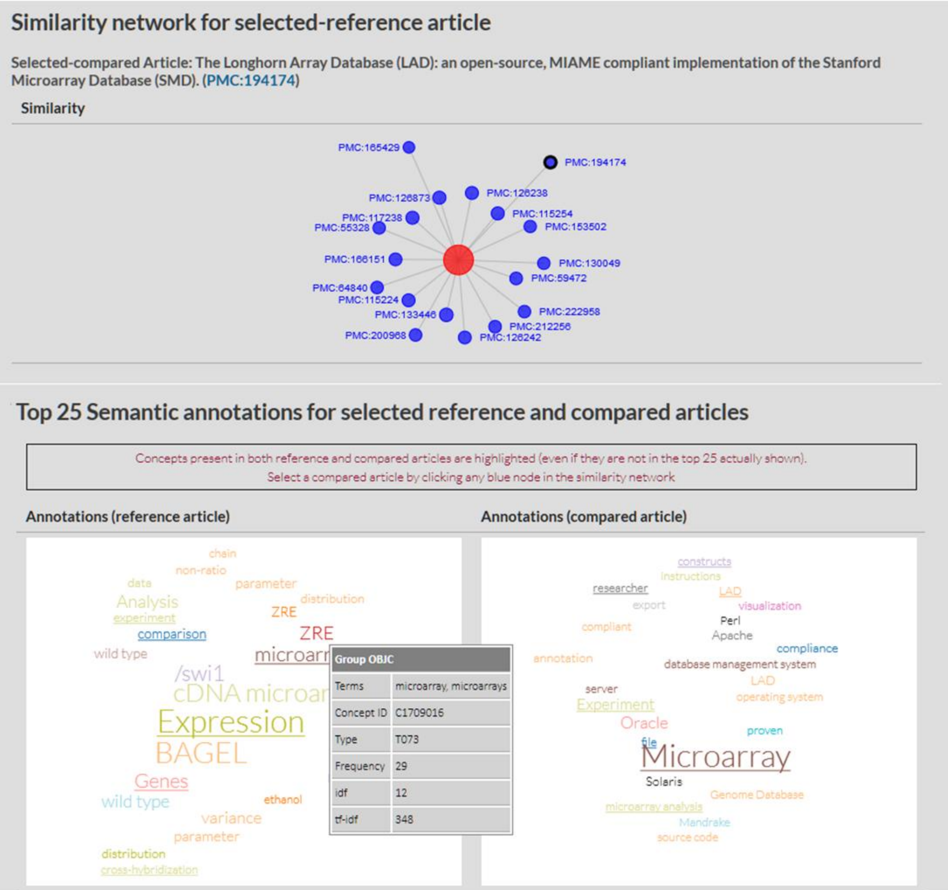
Article	
Id	PMC:200968
Id	PMID:12950995
Title	Evaluation of normalization methods for microarray data.
CONC weight	0.3692520006

Figure 34. Group-based distribution example

Any article can be selected by simply clicking on it. Once an article is selected, this article is set as the reference article, i.e., the one the reader is currently interested in. The reference article is highlighted with a black rectangle in the group-based distribution heat-map matrix. After selection, the similarity scores regarding all of the other articles are displayed as a similarity network;

the central node corresponds to the reference article. Additionally, a cloud of tags representing the 25-top annotations for the reference article is also displayed –a top annotation is that one with a high TF-IDF.

Whenever an article in the similarity network is selected, its cloud of tags is also displayed and the compared-to selected article is highlighted with a black circumference. The elements underlined in the cloud of annotations correspond to terms found in both articles. As before, additional information is displayed on hovering an element (see example in Figure 35). In this example the reference article corresponds to "Bayesian analysis of gene expression levels: statistical quantification of relative mRNA level across multiple strains or treatments. (PMC:151173)". The similarity network shows how much close are the rest of the articles in the TREC-05 topic "PROC/METH normalization procedures that are used for microarray data". One of the compared articles has been selected, the one with a black border, this compared article corresponds to "The Longhorn Array Database (LAD): an open-source, MIAME compliant implementation of the Stanford Microarray Database (SMD). (PMC:194174)". The cloud of annotations is displayed for both articles, the reference article and the selected compared one.



Similarity network for selected-reference article

Selected-compared Article: The Longhorn Array Database (LAD): an open-source, MIAME compliant implementation of the Stanford Microarray Database (SMD). (PMC:194174)

Similarity

Top 25 Semantic annotations for selected reference and compared articles

Concepts present in both reference and compared articles are highlighted (even if they are not in the top 25 actually shown).
Select a compared article by clicking any blue node in the similarity network

Annotations (reference article)

Group OBJC	
Terms	microarray, microarrays
Concept ID	C1709016
Type	T073
Frequency	29
idf	12
tf-idf	348

Annotations (compared article)

Figure 35. Similarity network and cloud of annotations

8.4 Discussion

We argue that using a semantic similarity approach could guide the navigation across the related articles list as it can provide information about the nature of the relation itself. Our results regarding Biolinks indicate that it is possible to use semantic annotations in order to characterize a predefined set of related documents. This characterization is used to analyze similarity between documents from a global perspective or narrowed to semantic groups of interest, enabling researchers to move to similar documents depending on what is more relevant to their investigation.

As mentioned previously, the similarity formula could be modified so only common concepts are taken into account. This would make the formula symmetric so the similarity score between articles R and C would be exactly the same as between C and R. The star-like graph would become a graph with no predefined shape where all the nodes would be connected. As a graph can become easily busy and difficult to grasp, other alternatives such as matrixes, Manhattan plot or scattered plots should be considered.

Chapter 9

Final remarks

9.1 Conclusions

Although scientific publications are nowadays available in electronic formats, the data embedded in the text remains locked up behind the scenes. In order to reduce the gap from discrete documents to fully machine-processable content, we have developed two complimentary projects: Biotea and Biolinks. They both aim to make it easier to semantically represent, structure and benefit from data and facts extracted from the scientific literature. Furthermore, they facilitate the integration and processing of literature-based data within the Linked Open Data (LOD) cloud.

Biotea semantically structures scientific publications including metadata, content and data extracted from the text. Such a structure constitutes a semantic scaffold facilitating interconnectivity and interoperability across not only publications but also online resources. Biotea adds value to scholarly publications as such documents become more useful when they are interconnected rather than independent [111]. The availability of linked data on top of the digital form currently adopted by scientific publications should move literature information retrieval from finding documents to finding relations, facts, and actionable intelligence [11]; it should enable literature-based knowledge discovery from a semantic perspective. Biotea is our contribution to this aim.

Biotea is a first step towards accessing semantically normalized information provided by scientific publications. Following steps should make use of the data and semantics provided by Biotea. Our contribution in this regards is Biolinks. Biolinks provides formulas, algorithms and visual tools to semantically categorize and compare scientific publications. It also defines a semantic model to represent the group-based categorization as well as global and group-narrowed similarity. Biolinks is built on top of a tailored redefinition of the Unified Medical Language System (UMLS) semantic groups; however, the algorithms can be easily configured in order to work with UMLS groups or any other semantic arrangement of the UMLS semantic types.

Biotea and Biolinks can work together in different ways in order to facilitate literature-based knowledge discovery. Our showcase scenario is just a glimpse of what is possible. Providing users with a rich experience built on top of the PubMed similar article list is another possible scenario that could benefit from our approach. The similar articles list ranks publications from most to least similar according to an analysis performed on word-stems in title-and-abstract and Medical subject headings (MeSH) terms assigned to the currently loaded article, i.e., the reference article. The global similarity score would be used to reorganize the compared articles taking into account concepts extracted from the full-text. Concepts would be not limited to MeSH but extended to other vocabularies covered by UMLS. The group-based distribution would be used to present an overview of the most representative subjects in the listed articles, making it easier for researchers to get a quick idea regarding the aboutness of the documents. This distribution can be used to narrow the similarity to those more relevant to them and more representative in the listed articles.

Annotations provided by Biotea can also be further used on their own. Different efforts in the biomedical domain aim to find hidden relations from semantic annotations. For instance, MeSH terms have been used in pattern identification with the purpose of finding candidates for new associations between drugs and diseases [12]. In a similar vein, the identification of Gene Ontology (GO) terms co-occurring with human genes have been used to suggest new GO annotations for those genes [13]. Shared GO annotations across different genes have also been used to identify possible relations between those genes [14, 15].

Our results indicate that it is helpful to use semantic annotations in order to characterize a predefined set of related documents. This characterization is used to analyze similarity between documents from a global perspective or narrowed to semantic groups of interest, enabling researchers to move to similar documents depending on what is more relevant to their investigations.

9.2 Future work

As part of our future work, we want to explore probabilistic topic model approaches [112, 113] such as Latent Dirichlet Allocation (LDA). Our group-based distribution uses a predefined set of categories, i.e., semantic groups, while topic model approaches find out a set of possible topic from the analysis of co-occurrence of different words present in the texts. This could give us new insights regarding the Biolinks groups.

We also want to further analyze variations in similarity when scores are narrowed to particular groups. In order to do so, we will work together with a domain expert and set of full-text laboratory protocols annotated with a tailored ontology. For such a collection we will use the National Center for Biomedical Ontology (NCBO) annotator [76] as it can be easily configured to work with any ontology.

Our next step will be to extend our prototype. We envision a recommendation system where, from a list of preferred publications, semantic annotations are generated and compared against annotations corresponding to the latest publications in PubMed or other repositories. Automatically generated annotations would be also complemented by manual annotations carried out by the researchers themselves. Semantic categorization and similarity would be used to generate a list containing the best matches between the preferred and the new publications.

Our ultimate goal is to provide a concept-based browser for scientific publications where researchers find the text but also the data and the facts behind it. In such a way, we aim to facilitate literature analysis, exploration and knowledge discovery.

9.3 List of contributions

- Semantic model to represent metadata, structures, content and entities recognized in the full-text.
- RDFization algorithms to process in batch a corpus complaint with the Journal Article Suite (JATS) format.
- RDFization of PubMed Central Open Access
- Alternative distribution of UMLS semantic types (STYs) into groups, i.e., Biolinks groups.
- Semantic model to represent group-based distribution for scientific publications.
- Algorithms to weight a predefined set of categories in a publication so the group-based distribution is calculated; batch process for a corpus is supported.
- Semantic model to represent similarity and group-narrowed similarity between a pair of publications.
- Algorithms to calculate global and group-narrowed similarity for a couple of publications; batch process for a corpus is supported.
- Showcase for TREC-05 collections showing annotations, group-based distribution and similarity.

9.4 List of publications

9.4.1 Journal papers

- Ciccarese, P., Ocana, M., Garcia Castro, L.J., Sudeshna, D., Clark, T. An open annotation ontology for science on web 3.0. Journal of Biomedical Semantics, 2011. 2(Suppl 2): p. S4. http://jbiomedsem.biomedcentral.com/articles/10.1186/2041-1480-2-S2-S4
- Garcia Castro, L.J., McLaughlin, C., Garcia, A. Biotea: RDFizing PubMed Central in Support for the Paper as an Interface to the Web of Data. Biomedical semantics, 2013. 4 Suppl 1: p. S5. http://jbiomedsem.biomedcentral.com/articles/10.1186/2041-1480-4-S1-S5
- Garcia Castro, L.J., Berlanga, R. and Garcia, A. In the pursuit of a semantic similarity metric based on UMLS annotations for articles in

PubMed Central Open Access. Journal of Biomedical Informatics, 2015. 57: p. 204-218. http://www.sciencedirect.com/science/article/pii/S1532046415001550

- Garcia Castro, L. J., Berlanga, R., Garcia, A. (Submitted February 2017). "Biolinks, semantic-based distribution and similarity for scientific publications." Journal of Biomedical Semantics.

9.4.2 Conference papers and posters

- Garcia Castro, L. J., Garcia, A., Gómez, J. (2010). Conceptual Exploration of Documents and Digital Libraries in the Biomedical Domain. SWAT4LS, Paris, France. http://ceur-ws.org/Vol-952/paper_40.pdf
- Garcia Castro, L. J., Berlanga, R., Rebholz-Schuhmann, D., Garcia, A. (2013). Connections across scientific publications based on semantic annotations. SePublica, Motpellier, France. http://ceur-ws.org/Vol-994/paper-05.pdf

9.4.3 Datasets

- Leyla Jael Garcia Castro, Olga Giraldo, Casey McLaughlin, & Alexander Garcia. (2012). Biotea dataset (vr. July 2012) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.376814
- Garcia Castro, L.J., Berlanga, R., Garcia, A. dataset used to determine a semantic similarity metric based on UMLS for PMC-OA, 2014: Zenodo. http://doi.org/10.5281/zenodo.13323
- Garcia Castro, L.J., Berlanga, R., Garcia, A., Biolinks, datasets and algorithms supporting semantic-based distribution and similarity for scientific publications, 2017: Zenodo. http://doi.org/10.5281/zenodo.290371

9.4.4 Software repositories and ontologies

- https://github.com/biotea/biotea-ontololgy
- https://github.com/biotea/biotea-utilities
- https://github.com/biotea/biotea-ao
- https://github.com/biotea/biotea-rdfization
- https://github.com/biotea/biotea-annotation
- https://github.com/ljgarcia/biotea-biolinks
- https://github.com/ljgarcia/biotea-io-parser
- https://github.com/ljgarcia/biotea-vis-annotation
- https://github.com/ljgarcia/biotea-vis-tooltip

- https://github.com/ljgarcia/biotea-vis-topicDistribution
- https://github.com/ljgarcia/biotea-vis-similarity
- https://github.com/ljgarcia/biotea-vis-biolinks

APPENDICES

A. Ontologies used in the RDFization process

Table 8. Ontologies used in RDFization of metadata, structure and content.

Ontology	Purpose	Main elements used in Biotea
Bibliographic ontology	Metadata	bibo:AcademicArticle, bibo:Document, bibo:doi, bibo:identifier, bibo:issn, bibo:Issue, bibo:issue, bibo:Journal, bibo:numPages, bibo:pageEnd, bibo:pageStart, bibo:pmid, bibo:shortDescription, bibo:volume
	References	bibo:AcademicArticle, bibo:Book, bibo:Chapter, bibo:citedBy, bibo:cites bibo:Document bibo:Proceedings
Biotea	Metadata (list of elements)	biotea:authorList
	Structure (list of elements)	biotea:paragraphList, biotea:sectionList
Document ontology	Structure and content	doco:Figure, doco:Section, doco:Paragraph, doco:Table
Dublin core terms	Metadata	dcterms:description, dcterms:issued, dcterms:publisher, dcterms:title
	Provenance	dcterms:creator, dcterms:hasFormat, dcterms:isFormatOf, dcterms:references, dcterms:source
Friend of a friend ontology	Metadata	foaf:familyName, foaf:givenName, foaf:name, foaf:OnlineAccount, foaf:Organization, foaf:Person, foaf:publications
	References	foaf:familyName, foaf:givenName, foaf:name, foaf:OnlineAccount, foaf:Organization, foaf:Person, foaf:publications
OWL	Link to other semantic representations	owl:sameAs
Provenance ontology	Provenance	prov:generatedAtTime, prov:wasAttributedTo, prov:wasDerivedFrom
RDF	Content (text in paragraphs)	rdf:value
RDFS	Link to related web pages	rdfs:seeAlso
Semantic science integrated ontology	Provenance	sio:is_data_item_in

Table 9. Ontologies used in RDFization of annotations

Ontology	Purpose	Main elements used in Biotea
Annotation ontology	Annotation	ao:Annotation, aot:ExactQualifier, ao:body
	Link to biomedical ontologies	ao:hasTopic
	Link to RDFized publication	ao:annotatesResource, ao:context, ao:onResource
Biotea	Frequency	biotea:idf, biotea:tf
Open Annotation	Annotation	oa:Annotation, oa:hasBody (with a oa:TextualBody)
	Link to biomedical ontologies	oa:hasBody (with a direct link to the ontological concept)
	Link to RDFized publication	oa:hasSource, oa:hasTarget
Provenance, authoring and versioning ontology	Provenance	pav:authoredBy, pav:createdBy
Provenance ontology	Provenance	prov:generatedAtTime

B. Algorithms

B.1 RDFization packages for scientific publications

Algorithms required to RDFize PMC-OA articles have been developed in Java 7 and are publicly available at GitHub, they are provided as they are under the Apache License, version 2.0 of January 2004. Two repositories are required: https://github.com/biotea/biotea-utilities and https://github.com/biotea/biotea-rdfization. biotea-utilities provides generic classes and files mainly for configuration purposes. The following list describes its main packages.

- elsevier.jaxb.math.mathml and pubmed.openAccess.jaxb.generated. They contain the Java model representing JATS files, including some auxiliary classes. These classes were automatically generated from a sample of files using the Java Architecture for XML Binding (JAXB). JAXB binds XML schemata to Java classes.
- ws.biotea.ld2rdf. It contains some generic exceptions, the URL style definition, and utility classes taking care of configuration property files as well as namespaces and base URLs for the ontologies used in the RDFization process. The URL style defines patterns followed by the URLs assigned to each RDFized element. Two URL patterns are

supported: a customized style defined for Biotea and Bio2RDF as defined at its RDFization guide[16].

The module biotea-rdfization corresponds to the RDFization process; it locally depends on biotea-utilities so it should be within the classpath scope. Most of the dependencies are configured via Maven; however, some few that were not available on Maven repositories are offered on the lib directory. Vocabularies used to represent metadata and content in RDF were mapped to Java classes using RDFReactor [114]. Log output can be configured via log4j.properties file. The main packages are:

- ws.biotea.ld2rdf.rdf.model. It contains the ontology mapping from BIBO and DOCO to Java classes.
- ws.biotea.ld2rdf.rdfGeneration. It contains generic handlers used by the batch process.
- ws.biotea.ld2rdf.rdfGeneration.pmcoa. It contains the class that actually parses a JATS file and creates an RDF representation.
- ws.biotea.ld2rdf.rdfGeneration.batch. It contains the main class PMCOABatchApplication in charge of RDFize JATS files within an input directory. Either an RDF/XML or a JSON-LD file is produced per each article and is stored in an output directory. The RDFization takes place only if the output file does not exist. By default it creates 10 threads in order to parallelize the process.

B.2 RDFization packages for enriched content

Algorithms used to annotate scientific publications and RDFize those annotations have been developed in Java 7 and are publicly available at GitHub. These algorithms are covered under the Apache License (version 2.0 of January 2004). Three repositories are required: https://github.com/biotea/biotea-utilities, https://github.com/biotea/biotea-ao and https://github.com/biotea/biotea-annotation.

Before running the batch process, some configuration is required. In the apikey.properties file, a valid Application Programming Interface (API) key for the NCBO Annotator should be provided. In the config.properties the directory where the Whatizit Web Service Definition Language (WSDL) file is located should be specified. Other properties such as stop words and URLs for the annotation services can be configured but it is highly recommended not to touch them.

The module biotea-utilities provide generic classes and files mainly for configuration purposes. Some of its characteristics were stated in the previous chapter. Here we add some more information related to the annotation process.

[16] https://github.com/bio2rdf/bio2rdf-scripts/wiki/RDFization-Guide

- The package ws.biotea.ld2rdf.util.annotation contains utility classes used during the annotation process.
- The file ontology.properties describes all the ontologies used to annotate with NCBO Annotator. It includes the virtual NCBO identifier, description, namespace, URL and NCBO name. For ontologies not supported by NCBO such as the UniProt Taxonomy, only the namespace and URL are required.
- In the file config.properties, it is possible to configure the URL to the annotators, add or remove stop words for NCBO and Whatizit, and specify the ontologies to be excluded for NCBO and Whatizit.
- The apikey.properties should be modified so it contains a valid API key to the NCBO Annotator

Biotea-ao was initially created as a Java representation for AO; however, it was later extended to represent as well OA. It contains a package ws.biotea.ld2rdf.rdf.model with a detailed model for AO annotations. In Biotea, annotations are always represented following the AO model but can be serialized as AO or OA. The serialization to RDF is defined in the package ws.biotea.ld2rdf.rdf.persistence.

Biotea-annotation corresponds to the RDFization process for annotations. It locally depends on biotea-utilities and biotea-ao so they both should be accessible within the classpath scope. All of the other dependencies are configured with Maven except for a dependency to Protégé [115] (version 05.08.2009). Protégé dependency is available in the lib directory. Similar to biotea-rdfization, the log output can be configured via log4j.properties file. The main packages in biotea-annotation are:

- ws.biotea.ld2rdf.annotation.parser. It contains the parser for the annotators. Initially it had parsers for both NCBO and Whatizit; however only NCBO is currently supported. Once a new public version of Whatizit becomes available, we will integrate it to our annotation process.
- ws.biotea.ld2rdf.annotation.batch. It contains the main class BatchApplication which is in charge of processing the input directory, invoke the annotators and produce the output. The input directory can contain either JATS files or articles RDFized with Biotea. Either an RDF/XML or a JSON-LD file is produced per each article and is stored in an output directory. The RDFization takes place only if the output file does not exist. By default it creates 10 threads in order to parallelize the process.
- ws.biotea.ld2rdf.annotation.ws.controller. It offers a basic web-service support so annotations can be retrieved on the fly. These web services are built on top of Spring Boot [116]. This is not the preferred annotation method thus is not as well supported as the batch process.

B.3 Group-based distribution packages

Biolinks delivers algorithms to calculate the group-based distribution for articles annotated with UMLS concepts. These algorithms have been developed in Java 7 and are publicly available at GitHub[17]. Algorithms are covered under the Apache License (version 2.0 of January 2004). Biolinks algorithms work on Resource Description Framework / eXtensible Markup Language (RDF/XML) annotation files following the model proposed by Biotea. If no annotation files exist, Biolinks depends on Biotea annotation and utility projects[18]. In particular, the file biolinks.properties in biotea-utilities contains Biolinks groups and their corresponding STYs as well as the estimated group parameters.

The module biotea-biolinks performs the calculations necessary to find the group-based distribution in a document. Such a distribution is saved in RDF following a model proposed in Biolinks. Similar to Biotea projects, dependencies are configured via Maven while logs output can be configured via log4j.properties file. The main packages in biotea-biolinks regarding the group-based distribution are:

- biolinks.parser, which contains a parser in charge of applying the group-based distribution to a document. An interface TopicDistributionParser is implemented by CMADistributionParser so annotations with an STY assigned can be processed. It uses the biolinks.properties in order to (i) find out the group assigned to an annotation, and (ii) read the estimated $\gamma(g_j)$ parameters.
- biolinks.persistence, which contains the RDFization process to translate the distribution to the RDF models defined by Biolinks.
- biolinks.batch, which contains the main class BatchApplication which is in charge of processing the input directory, invoke the annotators (if needed), calculate the distribution and produce the output. The input directory should contain all RDF/XML annotation files to be processed. By default, it creates 10 threads in order to parallelize the process.

B.4 Semantic similarity packages

Similar to the algorithms delivered to calculate the group-based distribution for articles annotated with Unified Medical Language System (UMLs) concepts, Biolinks also delivers algorithms regarding the global and group-narrowed

[17] https://github.com/ljgarcia/biotea-biolinks

[18] https://github.com/biotea/biotea-ao and https://github.com/biotea/biotea-annotation are used to produce the annotations while https://github.com/biotea/biotea-utilities is used for generic configuration purposes

similarity[19]. Biotea-biolinks performs the required calculations to find either global or group-narrowed similarity scores between any pair of articles annotated with UMLS concepts. Information regarding the participant articles and groups, if applicable, together with the similarity score are saved in the Resource Description Framework (RDF). The main packages in biotea-biolinks regarding the similarity are:

- biolinks.parser, which contains a parser in charge of applying the similarity formula to a pair of documents. The main class in this package is PMRASimilarityParser
- biolinks.persistence, which contains the RDFization process to translate the distribution to the RDF models defined by Biolinks.
- biolinks.batch, which contains the main class BatchApplication which is in charge of processing the input directory, invoke the annotators (if needed), calculate the similarity score and produce the output. The input directory should contain all RDF/ eXtensible Markup Language (RDF/XML) annotation files to be processed. By default, it creates 10 threads in order to parallelize the process.

B.5 Visualization components

Biotea-io-parser is the starting point. It takes Biotea JSON-LD annotation files and parses them in order to calculate group-based distribution and similarity scores. Biotea-vis-tooltip is a generic-purpose component. Biotea-vis-annotation, biotea-vis-topicDistribution and biotea-vis-similarity are the core visualization components while biotea-vis-biolinks mix them together in order to provide a single access to a corpus of documents previously organized in clusters. Biotea-vis-biolinks is used to showcase Biolinks capabilities applied to the relevant and partially relevant documents of the Text Retrieval Conference 2005 Genomics Track Collection (TREC-05).

Biotea-io-parser. This component[20] provides the core configuration for Biolinks including semantic groups and constant values used for group-based distribution and similarity calculations. Such constant values were estimated while tuning distribution and similarity formulas for Biolinks. This component retrieves Biotea annotations from JSON-LD and transforms them into a simple JSON object used by the core visualization components. It also delivers functions to calculate group-based distribution for one document, global similarity between two documents and group-narrowed similarity as well.

[19] The GitHub repository used for distribution and similarity is the same, https://github.com/ljgarcia/biotea-biolinks

[20] Biotea-io-parser is publicly available at https://github.com/ljgarcia/biotea-io-parser

Although its main purpose is preparing the data used by the visual components, it can also be used standalone.

Biotea-vis-tooltip. This component[21] creates a tooltip that is rendered in a table-like style. This generic component is used by all of the other visual components but can also be used standalone.

Biotea-vis-annotation. This component[22] displays a cloud of semantic annotations. A semantic annotation corresponds to one ontological concept, one or more terms found in the text, the term frequency (TF) and the inverse document frequency (IDF). The rendered text corresponds to the first term collected for an annotation. On mouse-over, a tooltip is displayed so semantic group, the concept identifier, and frequency related information are shown. By default it render the 100-top concepts regarding their TF-IDF; however, this number can be configured to a smaller and more visually appealing one. This component has two Biolinks dependencies: biotea-io-parser and biotea-vis-tooltip.

Bioteas-vis-topicDistribution. This component[23] renders the group-based distribution for one or more documents as a heat-map matrix. It aims to present an overview of the most representative groups for a cluster of documents. A heat-map uses different hues of color to show variation within a range; in this case from 0 to 1. In the matrix, rows correspond to Biolinks semantic groups while columns correspond to the different articles in the cluster. A cell displays the color corresponding to the group-based distribution score. Cell colors vary from green to blue, the closer to 0 the lighter green, the closer to darker blue. On mouse over, a tooltip displays the PubMed and PubMed Central Data (PMC) identifiers, the title, and the distribution score for the corresponding article and group. This component has two Biolinks dependencies: biotea-io-parser and biotea-vis-tooltip.

Biotea-vis-similarity. This component[24] renders a star-like graph representing a similarity network. It requires two or more articles, a reference article which is the central node of the network, and the compared articles which are shown around. The higher the similarity score, the bigger the circle representing the compared article and the shorter the arc to the reference article. On mouse over nodes, a tooltip showing information related to the corresponding article as

[21] Biotea-vis-tooltips is publicly available at https://github.com/ljgarcia/biotea-vis-tooltip

[22] Biotea-vis-annotations is publicly available at https://github.com/ljgarcia/biotea-vis-annotation

[23] Biotea-vis-topicDistribution is publicly available at https://github.com/ljgarcia/biotea-vis-topicDistribution

[24] Biotea-vis-similarity is publicly available at https://github.com/ljgarcia/biotea-vis-similarity

well as the similarity score is displayed. This component has two Biolinks dependencies: biotea-io-parser and biotea-vis-tooltip.

Biotea-vis-biolinks. This component[25] merges together the previous components making it easier to build a showcase based on a larger number of articles. It takes a list of clusters of articles and displays them as a two-level list; the first list shows the cluster titles while the second one shows the article titles for the selected cluster. It also allows selecting one or more groups of interested so the similarity and annotations are narrowed to those selected groups. This component has four Biolinks dependencies: biotea-io-parser, biotea-vis-tooltip, biotea-vis-topicDistribution and biotea-vis-similarity.

C. List of Abbreviations

Table 10. List of abbreviations (abbreviations are introduced in each chapter).

Abbreviation	Full text
ACTI	Activities & Behaviors
ANAT	Anatomy
AO	Annotation Ontology
API	Application Programming Interface
BIBO	Bibliographic ontology
CHEM (in Biolinks)	Chemicals
CHEM (in UMLS)	Chemicals & Drugs
CMA	Concept Mapping Annotator
CONC	Concepts & Ideas
CUI	Concept Unique Identifier
D3	Data-Driven Documents
DC	Dublin Core
DCMI	Dublin Core metadata initiative
DEVI	Devices
DISO	Disorders
DL	Digital library
DoCO	Document Components Ontology
DOI	Digital Object Identifier
FMA	Foundational Model of Anatomy Ontology
FOAF	Friend of a friend ontology
FTP	File protocol transfer
GENE	Genes & Molecular Sequences

[25] Biotea-vis-biolinks is publicly available at https://github.com/ljgarcia/biotea-vis-biolinks

GEOG	Geographic Areas
GNPT	Genes and Proteins
GO	Gene Ontology
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
ICD10	International Statistical Classification of Diseases vr. 10
IDF	Inverse document frequency
IG	Information Gain
IRI	Internationalized Resource Identifier
ISSN	International Standard Serial Number
JATS	Journal Article Tag Suite
JAXB	Java Architecture for XML Binding
JSON-LD	JavaScript Object Notation-Linked Data
KEGG	Kyoto Encyclopedia of Genes and Genomes
LD	Linked Data
LIVB	Living Beings
LOD	Linked open data
LS	Life Sciences
MDDB	Master Drug Data Base
MedDRA	Medical Dictionary for Regulatory Activities
MeSH	Medical subject headings
MRREL	Metathesaurus relationships file
MRREL	Metathesaurus relationships file
NCBI	National Center for Biotechnology Information
NCBO	National Center for Biomedical Ontology
NCIt	National Cancer Institute thesaurus
NDDF	National Drug Data File
NDF-RT	National Drug File - Reference Terminology
NIH/NLM	U.S. National Institutes of Health's National Library of Medicine
NLM	National Library of Medicine
NPG	Nature Publishing Group
OA	Open Annotation Ontology
OBI	Ontology for Biomedical Investigations
OBJC	Objects
OBSV	Observations
OCCU	Occupations
OMIM	Online Mendelian Inheritance in Man
ORGA	Organizations
OWL	Web ontology language
OWL-DL	Web ontology language - Description logic

PAV	Provenance, Authoring and Versioning
PDF	Portable Document Format
PEOP	People and groups of people
PHEN	Phenomena
PHYS	Physiology
PMC	PubMed Central Data
PMC-OA	PubMed Central Open Access
PMID	PubMed Identifier
PRISM	Publishing Requirements for Industry Standard Metadata
PROC	Procedures
PROV-O	Provenance Ontology
RDF	Resource description framework
RDFS	RDF schema
SDL	Semantic digital library
SGR	Semantic Group
SIO	Semantic science Integrated Ontology
SKOS	Simple Knowledge Organization System
SNOMED-CT	SNOMED Clinical Terms
SPARQL	SPARQL Protocol and RDF Query Language
STY	Semantic Type
SW	Semantic Web
SYMP	Symptom Ontology
SYMP (in Biolinks)	Symptoms
TAXA	Taxonomy
TF	Term frequency
TREC-05	Text Retrieval Conference 2005 Genomics Track Collection
UMLS	Unified Medical Language System
UniProtKB	Universal Protein Resource Knowledge Base
URI	Universal Resource Identifier
URL	Universal Resource Locator
VoID	Vocabulary of Interlinked Datasets
W3C	World Wide Web Consortium
WSDL	Web Service Definition Language
WWW	World Wide Web
XML	eXtensible markup language

BIBLIOGRAPHY

1. Garcia Castro, L.J., C. McLaughlin, and A. Garcia, *Biotea: RDFizing PubMed Central in Support for the Paper as an Interface to the Web of Data.* Biomedical semantics, 2013. **4 Suppl 1**: p. S5.
2. Garcia, A., et al., *The Semantic Web and the Social Web heading towards a Living Document in life sciences.* Journal of the Semantic Web., 2009(Special issue " The ELSEVIER Grand challenge").
3. Garcia Castro, L.J., et al., *Biotea dataset (vr. July 2012)*, 2012: Zenodo.
4. Ciccarese, P., et al., *An open annotation ontology for science on web 3.0.* Journal of Biomedical Semantics, 2011. **2**(Suppl 2): p. S4.
5. Sanderson, R., P. Ciccarese, and H. Van de Sompel. *Open Annotation Data Model [http://www.openannotation.org/spec/core/].* 2013.
6. Garcia Castro, L.J., R. Berlanga, and A. Garcia, *Biolinks, semantic-based distribution and similarity for scientific publications.* Journal of Biomedical Semantics, Submitted (February 2017).
7. Garcia Castro, L.J., R. Berlanga, and A. Garcia Castro, *Biolinks, datasets and algorithms supporting semantic-based distribution and similarity for scientific publications*, 2017: Zenodo.
8. Garcia Castro, L.J., R. Berlanga, and A. Garcia Castro, *A dataset used to determine a semantic similarity metric based on UMLS for PMC-OA*, 2014: Zenodo.
9. Text Retrieval Conference 2005 - Genomics Track. *TREC-05 Genomics Track ad hoc relevance judgement.* 2005 [cited 2016 23rd August]; Available from: http://trec.nist.gov/data/genomics/05/genomics.qrels.large.txt.
10. Hersh, W., et al. *TREC 2005 Genomics Track Overview.* in *Text Retrieval Conference.* 2005.
11. Sheth, A., I.B. Arpinar, and V. Kashyap, *Relationships at the Heart of Semantic Web: Modeling, Discovering, and Exploiting Complex Semantic Relationships*, in *Enhancing the Power of the Internet.* 2003, Springer Berlin Heidelberg. p. 63-94.
12. Srinivasan, P., B. Libbus, and A.K. Sehgal. *Mining MEDLINE: Postulating a Beneficial Role for Curcumin Longa in Retinal Diseases.* in *Workshop BioLINK, Linking Biological Literature, Ontologies and Databases at HLT NAACL.* 2004. Boston, Massachusetts, USA.
13. Good, B. and A.I. Su. *Mining Gene Ontology Annotations From Hyperlinks in the Gene Wiki.* in *Translational Bioinformatics Conference.* 2011. Washington, D.C.
14. Saha, B., et al. *Dense subgraphs with restrictions and applications to gene annotation graphs.* in *14th Annual international conference on Research in Computational Molecular Biology.* 2010. Lisbon, Portugal: Springer-Verlag.

15. Thor, A., et al. *Link prediction for annotation graphs using graph summarization*. in *International Conference on the Semantic Web*. 2011. Bonn, Germany: Springer-Verlag.

16. Harrow, I., et al., *Towards Virtual Knowledge Broker services for semantic integration of life science literature and data sources*. Drug Discovery Today, 2012. **in press**.

17. Rebholz-Schuhmann, D., et al., *A case study: semantic integration of gene–disease associations for type 2 diabetes mellitus from literature and biomedical data resources*. Drug Discovery Today, 2014. **19**(7): p. 882-889.

18. Corcho, O., M. Fernández-López, and A. Gómez-Pérez, *Methodologies, tools and languages for building ontologies: where is their meeting point?* Data & Knowledge Engineering, 2003. **46**(1): p. 41-64.

19. Gruber, T., *Toward Principles for the Design of Ontologies Used for Knowledge Sharing*. International Journal Human-Computer Studies, 1995. **43**(5-6): p. 907-928.

20. Borst, W., *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*, in *Centre for Telematics and Information Technology*1997, University of Twente.

21. Studer, R., R. Benjamin, and D. Fensel, *Knowledge Engineering: Principles and Methods*, in *Knowledge engineering: principles and methods*. 1998, Elsevier Science Publishers. p. 161-197.

22. Guarino, N., D. Oberle, and S. Staab, *What Is an Ontology?*, in *Handbook on Ontologies*, S. Staab and R. Studer, Editors. 2009, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 1-17.

23. Garcia, A., et al., *Developing Ontologies within Decentralised Settings*, in *Semantic e-Science*, H. Chen, Y. Wang, and K.-H. Cheung, Editors. 2010, Springer US. p. 99-139.

24. Garcia, A., et al., *The melting point, communities of practice developing ontologies*, in *Semantic e-Science, Annals of Information System*, H. Chen, Y. Wan, and K. Cheung, Editors. 2009, Springer.

25. Uschold, M. and M. Gruninger, *Ontologies and semantics for seamless connectivity*. SIGMOD Rec., 2004. **33**(4): p. 58-64.

26. Berners-Lee, T., J. Hendler, and O. Lassila, *The Semantic Web*. Scientific American, 2001.

27. World Wide Web Consortium, *RDF 1.1 Primer*, G. Schreiber, et al., Editors. 2014: Internet (http://www.w3.org/TR/rdf11-primer/).

28. Berners-Lee, T., *Design issues: Linked data*, 2009, http://www.w3.org/DesignIssues/LinkedData.

29. Consortium, W.W.W., *SPARQL 1.1 Overview*, 2013: https://www.w3.org/TR/sparql11-overview/.

30. Bizer, C., T. Heath, and T. Berners-Lee, *Linked Data - The Story So Far*. International Journal on Semantic Web and Information Systems, 2009. **5**(3): p. 1-22.

31. Cyganiak, R. and A. Jentzsch, *The Linking Open Data cloud diagram*, 2014: http://lod-cloud.net/.

32. Hu, W., H. Qiu, and M. Dumontier, *Link Analysis of Life Science Linked Data*, in *The Semantic Web - ISWC 2015: 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part II*, M. Arenas, et al., Editors. 2015, Springer International Publishing: Cham. p. 446-462.
33. Hasnain, A., et al., *A Roadmap for Navigating the Life Sciences Linked Open Data Cloud*, in *Semantic Technology: 4th Joint International Conference, JIST 2014, Chiang Mai, Thailand, November 9-11, 2014. Revised Selected Papers*, T. Supnithi, et al., Editors. 2015, Springer International Publishing: Cham. p. 97-112.
34. Dumontier, M., et al., *Bio2RDF release 3: a larger connected network of linked data for the life sciences*, in *Proceedings of the 2014 International Conference on Posters &#38; Demonstrations Track - Volume 1272*2014, CEUR-WS.org: Riva del Garda, Italy. p. 401-404.
35. Belleau, F., et al., *Bio2RDF: Towards a mashup to build bioinformatics knowledge systems.* Journal of Biomedical Informatics, 2008. **41**(5): p. 706-716.
36. Callahan, A., et al., *Improved dataset coverage and interoperability with Bio2RDF Release 2*, in *Semantic Web Applications and Tools for Life Sciences*2012: Paris, France.
37. Ontotext, *Linked Life Data*, Ontotext, Editor 2014, Ontotext.
38. Jupp, S., et al., *The EBI RDF platform: linked open data for the life sciences.* Bioinformatics, 2014. **30**(9): p. 1338-1339.
39. Egaña Aranguren, M., J.T. Fernández-Breis, and M. Dumontier, *Special issue on Linked Data for Health Care and the Life Sciences.* Semant. web, 2014. **5**(2): p. 99-100.
40. Schweiger, D., Z. Trajanoski, and S. Pabinger, *SPARQLGraph: a web-based platform for graphically querying biological Semantic Web databases.* BMC Bioinformatics, 2014. **15**(1): p. 1-5.
41. Shoaib, M., A.A. Ansari, and S.-m. Ahn, *cMapper: Gene-Centric Connectivity Mapper for EBI-RDF Platform.* Bioinformatics, 2016.
42. Queralt-Rosinach, N., et al., *DisGeNET-RDF: harnessing the innovative power of the Semantic Web to explore the genetic basis of diseases.* Bioinformatics, 2016. **32**(14): p. 2236-2238.
43. Kim, S., et al., *PubChem Substance and Compound databases.* Nucleic Acids Research, 2016. **44**(D1): p. D1202-D1213.
44. Navas-Delgado, I., et al., *kpath: integration of metabolic pathway linked data.* Database, 2015. **2015**.
45. Benson, D.A., et al., *GenBank.* Nucleic Acids Research, 2013. **41**(D1): p. D36-D42.
46. , T.U.C., *UniProt: a hub for protein information.* Nucleic Acids Research, 2015. **43**(D1): p. D204-D212.
47. Kanehisa, M., et al., *KEGG as a reference resource for gene and protein annotation.* Nucleic Acids Research, 2016. **44**(D1): p. D457-D462.
48. Swan, A., *Overview of scholarly communication*, in *Open Access: Key Strategic, Technical and Economic Aspects*, N. Jacobs, Editor. 2006, Chandos.

49. U.S. National Institutes of Health's National Library of Medicine, *PubMed Central*, 2000: Internet (http://www.ncbi.nlm.nih.gov/pmc/).
50. Kruk, S.R., et al. *JeromeDL - a Semantic Digital Library*. in *International Semantic Web Conference - Semantic Web Challenge*. 2007. Busan, Korea.
51. Kruk, S., et al. *The Role of Ontologies in Semantic Digital Libraries*. in *European Networked Knowledge Organization Systems (NKOS) Workshop*. 2006. Alicante, Spain.
52. *BRICKS Project: Building Resources for Integrated Cultural Knowledge Services*, [http://www.brickscommunity.org/].
53. Carpenter, P., *Nature Publishing Group releases linked data platform*, 2012, (http://www.nature.com/press_releases/linkeddata.html).
54. D'Arcus, B. and F. Giasson. *Bibliographic Ontology Specification*. 2009.
55. Dublin Core Metadata Initiative. *DCMI Metadata Terms* 2012; Available from: http://dublincore.org.
56. Weibel, S., *The Dublin core metadata initiative - The Frankfurt focus and the year 2000*. Zeitschrift Fur Bibliothekswesen Und Bibliographie, 2000. **47**(1): p. 3-13.
57. FOAF. *The Friend of a Friend (FOAF) project*. 2008; Available from: http://www.foaf-project.org/.
58. Brickley, D. and L. Miller, *FOAF Vocabulary Specification 0.98*, 2010, (http://xmlns.com/foaf/spec/).
59. IDEAlliance. *Publishing Requirements for Industry Standard Metadata*.
60. Hammond, T. and M. Pasin, *The nature. com ontologies portal*, in *Workshop on Linked Science 2015, colocated with International Semantic Web Conference 2015*2015: Bethlehem, USA.
61. Belhajjame, K., et al., *The PROV ontology*, 2012, [http://www.w3.org/TR/prov-o/].
62. Alexander, K., et al. *Describing Linked Datasets with the VoID Vocabulary [http://www.w3.org/TR/void/]*. 2011.
63. Brickley, D. and R.V. Guha, *RDF Vocabulary Description Language 1.0: RDF Schema*, 2004, World Wide Web Consortium W3C: (http://www.w3.org/TR/rdf-schema/).
64. Shoton, D. and S. Peroni, *DoCO, the Document Components Ontology*, 2011, [http://purl.org/spar/doco].
65. Dumontier, M., et al., *The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery*. Journal of Biomedical Semantics, 2014. **5**(1): p. 1-11.
66. Ciccarese, P., M. Ocana, and T. Clark. *DOMEO: a web-based tool for semantic annotation of online documents*. in *Bio-Ontologies*. 2011. Vienna, Austria.
67. Garcia-Castro, A., et al., *Semantic Web and Social Web heading towards Living Documents in the Life Sciences*. Web Semantics: Science, Services and Agents on the World Wide Web, 2010. **8**(2-3): p. 155-162.
68. Attwood, T.K., et al., *Utopia Documents and The Semantic Biochemical Journal experiment*. EMBNet News, 2010. **15**(4).

69. Ciccarese, P., et al. *AO: An Open Annotation Ontology for Science on the Web*. in *Bio-ontologies*. 2010. Boston MA, USA.

70. Koivunen, M.-R. *Annotea Project*. 2010.

71. Ciccarese, P., et al., *PAV ontology: provenance, authoring and versioning*. Journal of Biomedical Semantics, 2013. **4**: p. 37-37.

72. Miles, A. and S. Bechhofer. *SKOS Simple Knowledge Organization System Reference*. 2009 [cited 2016 18.Sep.]; Available from: https://www.w3.org/TR/skos-reference/.

73. Sanderson, R., P. Ciccarese, and B. Young. *Web Annotation Data Model*. 2016 [cited 2016 18.Sep.]; Available from: https://www.w3.org/TR/annotation-model/.

74. Rebholz-Schuhmann, D., et al., *Text processing through Web Services: Calling Whatizit*. Bioinformatics, 2007. **24**(2).

75. Kirsch, H. and D. Rebholz-Schuhmann, *Distributed modules for text annotation and IE applied to the biomedical domain*, in *International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*2004: Geneva, Switzerland. p. 50-53.

76. Jonquet, C., et al. *NCBO Annotator: Semantic Annotation of Biomedical Data*. in *International Semantic Web Conference, Poster and Demo session*. 2009.

77. *Monq.jfa*, [http://freecode.com/projects/monq_jfa].

78. Dai, M., et al. *An Efficient Solution for Mapping Free Text to Ontology Terms*. in *AMIA Summit on Translational Bioinformatics*. 2008. San Francisco, CA, U.S.A.

79. Juty, N., N. Le Novère, and C. Laibe, *Identifiers.org and MIRIAM Registry: community resources to provide persistent identification*. Nucleic Acids Research, 2012. **40**(D1): p. D580-D586.

80. Campos, D., S. Matos, and J.L. Oliveira, *A modular framework for biomedical concept recognition*. BMC Bioinformatics, 2013. **14**(1): p. 281.

81. *2. Metathesaurus*, in *UMLS® Reference Manual [Internet]*2009, National Library of Medicine (US): Bethesda (MD).

82. NIH U.S National Library of Medicine. *The UMLS Semantic Network ([Updated: February 11, 2011])*. 2011 [cited 2013 27 Oct].

83. Bodenreider, O., *The Unified Medical Language System (UMLS): integrating biomedical terminology*. Nucleic Acids Research, 2004. **32**(suppl 1): p. D267-D270.

84. McCray, A.T., A. Burgun, and O. Bodenreider, *Aggregating UMLS semantic types for reducing conceptual complexity*. Proceedings of Medinfo, 2001. **10**(Pt 1): p. 216-220.

85. Bodenreider, O. and A.T. McCray, *Exploring semantic groups through visual approaches*. Journal of biomedical informatics, 2003. **36**(6): p. 414-432.

86. Fan, J.-W. and C. Friedman, *Semantic reclassification of the UMLS concepts*. Bioinformatics, 2008. **24**(17): p. 1971-1973.

87. Bird, S., E. Klein, and E. Loper, *4.1 Entropy and Information Gain*, in *Natural Language Processing with Python*, T. Apandi, Editor. 2009, O'Reilly Media, Inc.

88. Sayers, E., *E-utilities Quick Start (2008 Dec 12 [Updated 2013 Aug 9])*, in *Entrez Programming Utilities Help [Internet]*. 2008, Bethesda (MD): National Center for Biotechnology Information (US); 2010-. Available from: http://www.ncbi.nlm.nih.gov/books/NBK25500/.

89. Garcia Castro, L.J., R. Berlanga, and A. Garcia, *In the pursuit of a semantic similarity metric based on UMLS annotations for articles in PubMed Central Open Access.* Journal of Biomedical Informatics, 2015. **57**: p. 204-218.

90. Berlanga, R., V. Nebot, and E. Jimenez-Ruiz, *Semantic annotation of biomedical texts through concept retrieval.* Procesamiento de Lenguaje Natural, 2010. **45**: p. 247-250.

91. Berlanga, R., V. Nebot, and M. Pérez, *Tailored semantic annotation for semantic search.* Web Semantics: Science, Services and Agents on the World Wide Web, 2015. **30**: p. 69-81.

92. Small, H., *Co-citation in the scientific literature: A new measure of the relationship between two documents.* Journal of the American Society for Information Science, 1973. **24**(4): p. 265-269.

93. Hummon, N.P. and P. Dereian, *Connectivity in a citation network: The development of DNA theory.* Social Networks, 1989. **11**(1): p. 39-63.

94. Rogers, F., *Medical subject headings.* Bulletin of the Medical Library Association, 1963. **51**: p. 114-116.

95. Cornet, R. and N. de Keizer, *Forty years of SNOMED: a literature review.* BMC Medical Informatics and Decision Making, 2008. **8 Suppl 1**: p. S2.

96. Lin, J. and W.J. Wilbur, *PubMed related articles: a probabilistic topic-based model for content similarity.* BMC Bioinformatics, 2007. **8**(1): p. 423.

97. Robertson, S.E., C.J.v. Rijsbergen, and M.F. Porter, *Probabilistic models of indexing and searching*, in *Proceedings of the 3rd annual ACM conference on Research and development in information retrieval*1981, Butterworth & Co.: Cambridge, England. p. 35-56.

98. Robertson, S.E. and S. Walker, *Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval*, in *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*1994, Springer-Verlag New York, Inc.: Dublin, Ireland. p. 232-241.

99. van Rijsbergen, C.J., S.E. Robertson, and M.F. Porter, *New models in probabilistic information retrieval*, in *British Library Research and Development Report, no. 5587*1979, British Library: London.

100. Jannach, D., et al., *The cosine similarity measure*, in *Recommender Systems: An Introduction*. 2010, Cambridge University Press. p. 360.

101. Armstrong, J., *Cosine similarity: the similarity of two weighted vectors*, in *Programming Erlang, 2nd edition*. 2013, The Pragmatic Programmers. p. 548.

102. Aronson, A.R. and F.-M. Lang, *An overview of MetaMap: historical perspective and recent advances.* Journal of the American Medical Informatics Association, 2010. **17**(3): p. 229-236.

103. Pan, W., J. Lin, and C. Le, *How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach.* Genome Biology, 2002. **3**(5): p. research0022.1 - research0022.10.

104. Townsend, J. and D. Hartl, *Bayesian analysis of gene expression levels: statistical quantification of relative mRNA level across multiple strains or treatments.* Genome Biology, 2002. **3**(12): p. research0071.1 - research0071.16.

105. Shah, P., et al., *Information extraction from full text scientific articles: Where are the keywords?* BMC Bioinformatics, 2003. **4**(1): p. 20.

106. Gómez, J., et al., *BioJS: An Open Source JavaScript Framework for Biological Data Visualization.* Bioinformatics, 2013.

107. Corpas, M., et al., *BioJS: an open source standard for biological visualisation – its status in 2014.* F1000Research, 2014. **3**: p. 55.

108. The jQuery Foundation. *jQuery.* [cited 2016 13.Dec]; Available from: https://jquery.com/.

109. Bostock, M. *Data-Driven Documents.* 2015 [cited 2016 13.Dec]; Available from: https://d3js.org/.

110. Ashkenas, J. *Underscore.* [cited 2016 13.Dec]; Available from: http://underscorejs.org/.

111. Borgman, C.L., *Scholarship in the Digital Age: Information, Infrastructure, and the Internet.* 2007, Cambridge, MA, and London: MIT Press. 336.

112. Blei, D.M., *Probabilistic topic models.* Communications of the ACM, 2012. **55**(4): p. 77-84.

113. Yao, L., D. Mimno, and A. McCallum, *Efficient methods for topic model inference on streaming document collections*, in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*2009, ACM: Paris, France. p. 937-946.

114. Völkel, M. *RDFReactor - From Ontologies to Programatic Data Access.* in *Jena User Conference.* 2006. Bristol, UK.

115. Gennari, J., et al., *The evolution of Protégé: an environment for knowledge-based systems development.* International Journal of Human-Computer Studies, 2003. **58**(1): p. 89 - 123.

116. Spring. *Srping Boot.* [cited 2016 18.Oct.]; Spring Boot makes it easy to create stand-alone, production-grade Spring based Applications that you can "just run"]. Available from: https://projects.spring.io/spring-boot/.