The Devices, Experimental Scaffolds, and BiomaterialsOntology (DEB): A Tool for Mapping, Annotation, and Analysis of Biomaterials' Data

Article *in* Advanced Functional Materials · February 2020

DOI: 10.1002/adfm.201909910

Osnat Hakimi
Universitat Politècnica de Catalunya

The Devices, Experimental Scaffolds and Biomaterials Ontology (DEB): a Tool for Mapping, Annotation and Analysis of Biomaterials' Data

Osnat Hakimi, Josep Luis Gelpi, Martin Krallinger, Fabio Curi, Dmitry Repchevsky and Maria-Pau Ginebra*

Submitted:27/11/2019
Accepted: 31/1/2020
Available online:26/2/2020
DOI:10.1002/adfm.201909910
Link: https://onlinelibrary.wiley.com/doi/10.1002/adfm.201909910

Dr O.H. Author 1, Prof. M.P.G Author 6
Universitat Politècnica de Catalunya, Biomaterials, Biomechanics and Tissue Engineering Group, Dept. Materials Science and Metallurgical Engineering, Av. Eduard Maristany 16, edif i4.5, 08930 Barcelona, Spain
Email: osnat.hakimi@upc.edu

Prof. M.P.G Author 6
Institute for Bioengineering of Catalonia, Barcelona Institute of Technology, 08028 Barcelona, Spain

Prof. J.L.G Author 2
Dept. Biochemistry and Molecular Biomedicine, University of Barcelona, Diagonal, 643
08028 Barcelona, Spain

M.K. Author 3, F.C. Author 4, Dr. D.R. Author 3
Life Sciences - INB Computational Node 2, Barcelona Supercomputing Center, C/ Jordi Girona 29, 08034-Barcelona, Spain

Keywords: (biomaterials, text mining, ontology, database, annotation)

Abstract

The size and complexity of the biomaterials literature makes systematic data analysis an excruciating manual task. A practical solution is creating databases and information resources. Implant design and biomaterials research will greatly benefit from an open database for systematic data retrieval. Ontologies are pivotal to knowledge base creation, serving to represent and organize

domain knowledge. To name but two examples, GO, the Gene Ontology, and CheBI, Chemical Entities of Biological Interest ontology and their associated databases are central resources to their respective research communities. This study describes the creation of the Devices, Experimental scaffolds and Biomaterials Ontology (DEB), an open resource for organizing information about biomaterials, their design, manufacture and biological testing. It was developed using text analysis for identifying ontology terms from a biomaterials gold standard corpus, systematically curated to represent the domain's lexicon. Topics covered were validated by members of the biomaterials research community. The ontology may be used for searching terms, performing annotations for machine learning applications, standardized meta-data indexing and other cross-disciplinary data exploitation. We encourage and welcome the input of the biomaterials community to this effort to create data-driven open-access research tools.

1. Introduction

Incentives to publish scientific results as often as possible have led to a dramatic hike in the numbers of scientific publications in all areas of research[1], and functional materials such as biomaterials are not an exception. On November 2019, a search of the MEDLINE library with the search query ('biomaterials' or 'cell scaffolds') returns >190,000 abstracts (see **Table_S1**), with over 10,000 articles published yearly since 2014. A more extensive search combining keywords and medical subject headings (MeSH terms, see **Table 1**) returns >600,000 abstracts in the last 20 years. Because of this high volume, but also heterogeneous nature and complexity of the literature in the field[2], systematically collecting and synthesizing knowledge about biomaterials is an excruciating manual task.

A practical solution is the creation of databases and information resources to organize knowledge and enable more sophisticated data analysis. An open database from which data about relevant materials can be systematically retrieved and summarized could speed and improve implant design

and biomaterials research. To date, no such database exists. In certain sub-domains of tissue engineering, scientists have taken it upon themselves to manually curate data into databases, but these normally cover limited aspects of the materials, and are rarely maintained over long time[3,4]. Other initiatives include biomaterials' data sets repositories, but these will mostly serve for deposition of newly acquired data , rather than the extraction of historic data[5,6].

A major challenge remains how to extract and organize biomaterials information from text (eg: scientific articles, patents, abstracts) which are the main form of communication within the biomaterials community. Thus, this study initiates the effort to automatically mine information from the biomedical literature into an open-access database of biomaterials and their biological effect.

Ontologies are pivotal to the creation of knowledge bases, as they serve to represent and organize shared understanding, and as conceptual frameworks for modeling domain knowledge[7]. An ontology of a specific domain contains a set of concepts and categories (classes) as well as their properties and the relations between them. In the context of extracting information from text, ontologies serve to identify terms and the categories they belong too (much like a taxonomy or a vocabulary), but also to recognize hierarchies, rules and relations between concepts, as well text understanding [8]. To name but a few examples, the gene ontology (GO)[9], and the Chemical Entities of Biological Interest (ChEBI)[10] ontologies and their associated databases have all become a central resource to their respective research community and the development of numerous informatics tools.

To the best of our knowledge, only one study to date have developed an open ontology in the biomaterials domain. Viti *et al* (2014), have designed a Bone and Cartilage Tissue Engineering Ontology (BCTEO) as part of an initiative to create a set of guidelines for the minimum information necessary when describing an experimental study in the domain[11]. Their approach to the design of the ontology was to target concepts from the classical bone tissue engineering workflow, where an

experiment starts with stem cell expansion and ends at tissue formation. Moreover, they have chosen a top-down design, relying on 12 experts in the field[11].

The scope of the present study was to develop a biomaterials ontology which applies for the whole domain, including medical devices and clinical evaluations. Moreover, instead of relying on selected experts, this ontology aimed to use text analysis methods and machine learning to discover topics and terms, and crowd-source validation and input from the wider biomaterials community.

In addition, the ontology was developed to solve two specific and critical problems. First, manufactured biomaterials, scaffolds and devices have no taxonomy or naming system, and researchers may arbitrarily choose which features and aspects to emphasize in the name of the object. Thus, automatic information curation and meta-data analysis of testing data is impossible. Second, the field is highly multidisciplinary, and combines terms from material science, biology, engineering and clinical research, some of which are already in existing vocabularies, thus requiring the design of a distinctive approach.

With these challenges in mind, the development of the ontology was directed towards the logical identification, extraction and cataloging of components, properties and associated features of experimental scaffolds and commercial implants. Biomaterials of interest for the purpose of this ontology were defined as all manufactured objects used for biological and medical applications, including surgery, tissue engineering, cell expansion, drug delivery and antimicrobial protection. To this end, a mixed bottom-up and top-down approach was used, where a conceptual map was created in parallel to the creation of a gold-standard literature set to represent the biomaterials domain. This gold standard set was validated by a combination of machine learning and crowd-sourced feedback from over 70 biomaterials scientists. Finally, text analysis was used to extract significant terms and place them into the ontology, which was adapted over several iterations to accommodate the discovered terms.

The resulting ontology, entitled: 'Devices, Experimental scaffolds and Biomaterials Ontology' (DEB) is now deposited in two open repositories: Bioportal[12] and ProjectDebbie's GitHub, and it is expected to continuously expand. The ontology is available for download, and users are encouraged to contribute to its expansion and report bugs via GitHub's issues. It is already implemented as part of an experimental annotation pipeline for the curation of data into the open biomaterials database (DEBBIE), but several other applications are foreseen for this tool. For example, good quality annotated data can be used to train machine-learning algorithms for tasks such as summarization and predictions. It is also a valuable asset for standardized meta-data indexing, resource-linking and cross-disciplinary data exploitation. Finally, proper identification of used biomaterials in experiments can improve the adoption of FAIR principals (findable, accessible, interoperable and reusable data [13]) in biomaterials' research and automatic knowledge discovery in the field.

2. Results

2.1 Gold standard creation

For the purpose of this study, biomaterials research articles were defined as those where manufactured materials, either experimental or commercial, were tested in any biological system (ie: *in vitro*, *in vivo*, or clinically). The ontology was developed as part of the creation of 'DEBBIE- a database of experimental biomaterials and their biological effect', to be used within an automated annotation pipeline. Therefore, articles where no biological system was used in the evaluation, such as those describing only the manufacture or mechanical evaluation were excluded at this stage. In order to cover as broad a field as possible, a combination of medical subject headings (MeSH terms) and key words were identified using the NIH's MeSH Browser (https://meshb.nlm.nih.gov/search) and advanced PubMed searches. After excluding reviews and non-English abstracts, this combination of terms captured 693K records in a PubMed Search spanning the last 20 years (See Table 1).

Table 1. The list of MeSH terms and keywords used to generate a broad set of biomaterials references as a starting corpus for selecting the representative set.

	Search term	Mesh Term/keywords	Number of records (1999-2018)
1	Biomaterials	keyword	142,873
2	Cell scaffolds	keyword	29,892
3	Biomedical and dental materials	MeSH term	378,827
4	Prostheses and implants	MeSH term	302,692
5	Materials testing	MeSH term	57,386
6	Tissue engineering	MeSH term	31261
7	Tissue scaffolds	MeSH term	18,443
8	Equipment safety	MeSH term	6,747
9	Medical device recalls	MeSH term	165
	Total (excluding duplicates and non-english text)		**693,392**

In order to create a gold standard set that represents the biomaterials literature as correctly as possible, several manual and automatic curation steps were combined as explained in the methodology section. In parallel, a random set representing the general biomedical literature but in PubMed (called the 'random set' for the purpose of this study) was created to be used for downstream comparison and analysis. The process of creating these two corpora is shown in **Figure 1**. The final random set contained 1201 records and the gold standard set 1173 records. The distribution of both sets by year of publication is in **Figure_S1**, showing some bias in the biomaterials set towards the period from 2009 onwards, and particularly towards 2018/2019, most likely because of the order by which the triage tool, MedlineRanker, organizes highly ranked entries. However, given that the ontology was created for ongoing and future information extraction, some bias towards recent terms was seen as acceptable, perhaps even desired.

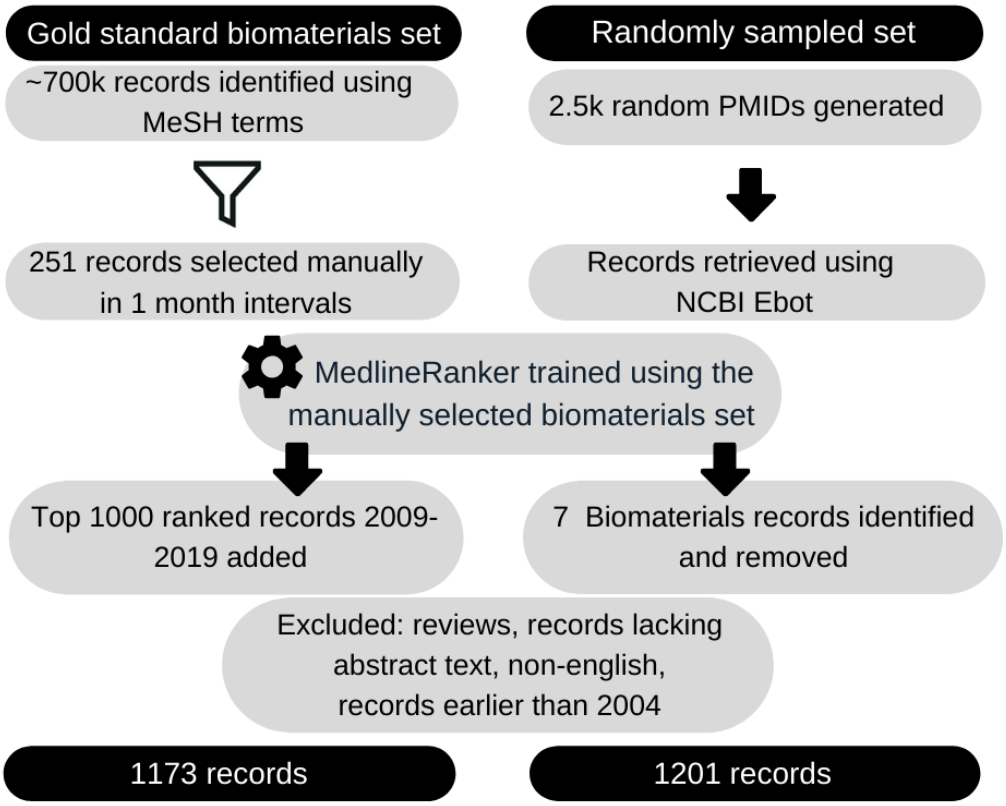


Figure 1. The steps taken to curate the gold standard set, which is a set of abstracts aiming to represent the biomaterials field. Parallel 'random set' pseudo-randomly sampled from the general abstracts archived in PubMed was created for all the comparative analyses.

2.2 Ontology approach and schema

The main issue the ontology was aiming to tackle is the lack of an accepted taxonomy for manufactured biomaterials. In the absence of such a taxonomy, novel scaffolds are named freely, with no restrictions or rules, emphasizing features which are considered of interest. The base assumption of the ontology was therefore that experimental scaffolds, implants and medical devices are all 'manufactured objects' with some fixed features that have undergone testing in a biological environment. Thus, the ontology associates the key features used in the naming of the manufactured object, and curate them for the purpose of identification and classification. As the ontology was created as part of an information extraction effort in the biomaterials domain and

towards the creation of an open-access biomaterials database, it was of great interest that the terms most commonly used to describe these objects are curated.

Figure 2 shows the process developed for the ontology curation. The process started with the design of a logical schema, using similar steps to Viti *et al*[11], as originally outlined by Noy and McGuinness[14], namely: defining the domain, clarifying the aims and creating a top-down schema of main classes. However, beyond ontology definition, concepts and terms were curated using a systematic approach, rather than expert suggestions. Using the process shown in Figure 2, terms were identified through bag-of-words analysis of the gold standard set (as explained in section) and thereafter were used to populate the ontology as classes or data properties. This in turn led to the discovery of additional classes and the modification of the ontology in parallel to the addition of more terms. Thus, the approach took numerous bottom up and top down iterations, aiming to produce a well-formed and coherent ontology. The resulting concept map of the ontology (**Figure 3**) is the result of this cyclic process. The concept map shows the main superclasses (ie: the highest classes in the hierarchy) of the ontology, which are defined in **Table 2**. 'Manufactured object', which can be a medical device, experimental scaffold or a biomaterial manufactured for a specific purpose, is the centre of the ontology schema, and therefore has all features associated to it through pre-defined relationships (or object properties) elaborated in **Table 3**.

It was of interest that the ontology reflects the complexity of how scaffolds and devices are described and named. For example, a scaffold may be named with few or a many technical details in the name, with emphasis on materials used (biomaterials or biologically active substances) engineered features (such as structure, architecture or a processing technique), its medical application or a biological aspect of its performance, depending on the author's preference. Examples of manufactured objects' naming in the gold standard set include: 'Electrospun polycaprolactone nanofibres decorated by drug loaded chitosan nano-reservoirs for antibacterial treatments' [15], which gives information about manufacture technique, materials, structures and

medical application, but also: 'a silver-releasing foam dressing in venous leg ulcer healing' [16], naming the medical application, structure and the bioactive substance, or alternatively: 'A bioartificial dermal regeneration template' [17] , which only names the manufactured object and its medical application. All these articles go on to describe the manufactured object in more detail, often naming it using different variations. For example, in the latest example, the dermal template is also called 'platelet-rich plasma-collagen sponge' and 'collagenase-containing platelet-rich plasma-collagen sponge' in the abstract text [17]. Using the DEB ontology, this associated information can be catalogued into a class and linked to the manufactured object via the designated object properties. Thus, a single manufactured object will be defined by multiple, separate features, and data from testing it can be compared based on those features. This is quite different from the BCTEO ontology[11] that have a biomaterial super-class under which scaffolds are classified into ceramic, composite, gel, metal, polymer or other, but not classify them by other features. Finally, to ensure DEB is well connected to existing ontologies, the relevant classes were linked to reference ontologies (Table 2). Those ontologies, such as BCTEO, often include some superclasses (such as biomaterials), but rarely the terms we sought to curate. More details about the referenced ontologies linked to DEB are in **Table_S2**.

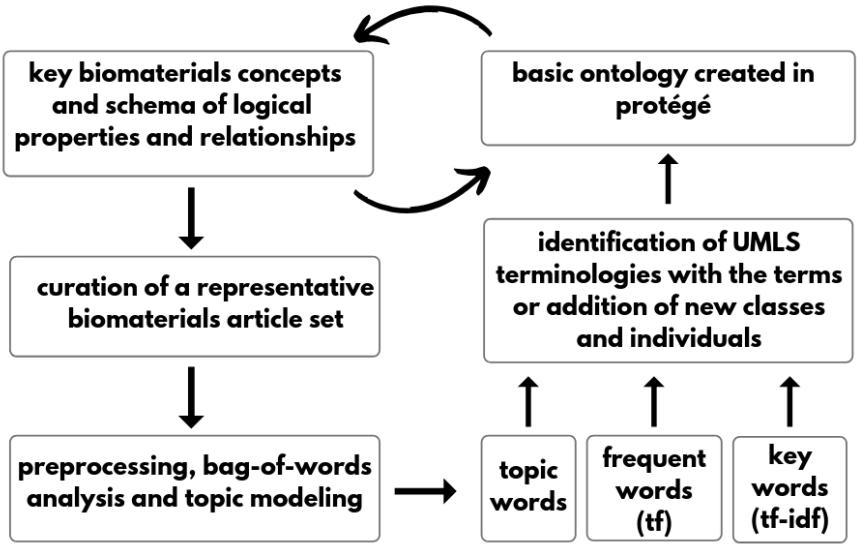


Figure 2. the approach used to create the DEB ontology resulted in a cyclic process, where identified terms were fed into a schema which was in turn adapted to link all the key terms logically.

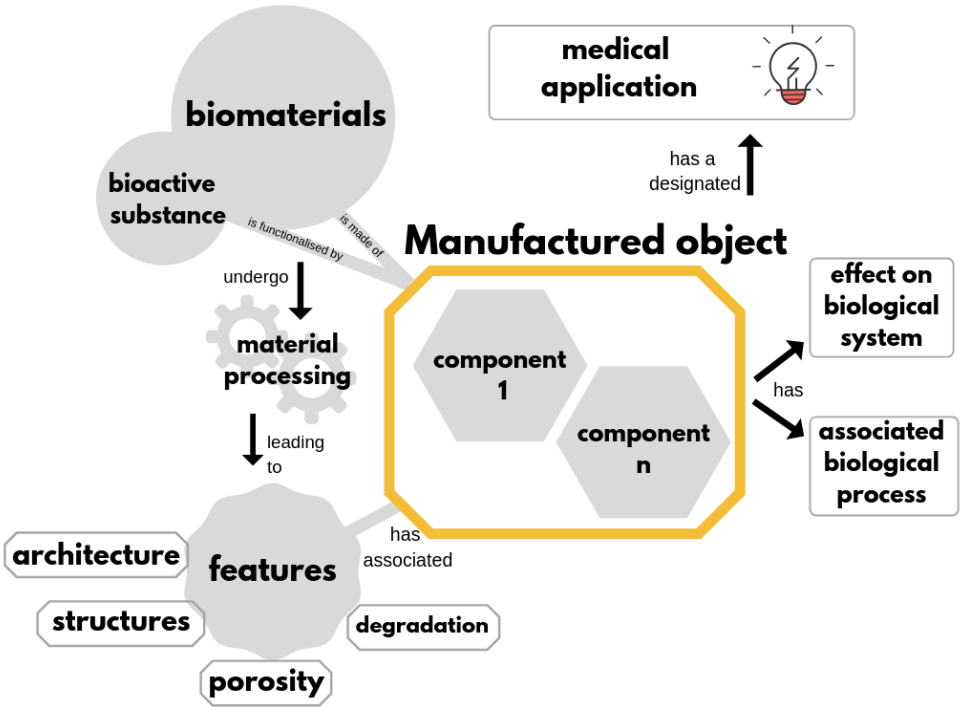


Figure 3. A map of the key concepts directly related to medical devices/experimental scaffolds. Each concept in the map is defined below as a class in the ontology. Object properties linking classes are adjacent to the arrows (schema produced using Canva).

Table 2. The superclasses used in the ontology, their definition, example of subclasses and references to other terminologies or ontologies

Class	Definition	Example of subclasses	Reference ontology
Manufactured object	A physical object created by hand or machine.	Medical device Experimental scaffold	STY
Manufactured object component	A part, region or component referred to as a distinct unit.	Core Layer Surface	
Biomaterial	A non-drug raw material or substance suitable for inclusion in systems which augment or replace the function of bodily	Titanium Alginate Fibroin	BCTEO

	tissues or organs.	PCL	
Biomaterial Type	Classification or nature of biomaterials	Alloy Blend Ceramic	
Biologically active substance	Substance, often a peptide or protein included in a manufactured object in order to impart a biological activity	BMP RGD Collagen	STY
Manufactured object features	Characteristics inherent or given during processing to a manufactured object or its component	Structure Shape Mechanical property Degradation features	
Material processing	A planned process which results in physical changes in a specified input material	Electrospinning Weaving Coating Etching	OBI CHMO
Associated biological process	A cellular or biological process that the manufactured object is designed to cause or support, or is measured to affect	Adipogenesis Angiogenesis Cell attachment	NCIT
Effect on biological system	Biological effect associated with the manufactured object in a specific test system (cells, tissue or organism)	Adverse effects Biocompatibility Bioinertness	
Medical application	Intended use of the manufactured object	Artificial organs Drug delivery Surgical Encapsulation	

Table 3. Object properties (relations) created in the ontology to define the relationships between the classes. Domain and range refer to the classes connected via the defined relation.

Object properties	Domain	Range	Example
create structure	Material processing	Structure Architectural organisation Structural details	Electrospinning[material processing] create structure Aligned[Architectural organisation]
functionalized by	Biomaterials Manufactured object component Manufactured object	Biologically active substance	Titanium[Biomaterial] functionalized by RGD[Bioactive substance]

has associated biological process	Biomaterials Manufactured object component Manufactured object	Associated biological process	Experimental scaffold[Manufactured object] has associated biological process Angiogenesis[Associated biological process]
has biological effect	Biomaterials Manufactured object component Manufactured object	Effect on biological system	Alginate[Biomaterial] has biological effect cytocompatibility [Effect on biological system]
has biomaterial	Manufactured object component Manufactured object	Biomaterial	Experimental scaffold[Manufactured object] has biomaterial Alginate[Biomaterial]
has component	Manufactured object	Manufactured object component	Experimental scaffold[Manufactured object] has component Layer[Manufactured object component]
has designated application	Manufactured object	Medical application	Experimental scaffold[Manufactured object] has designated application Tissue reconstruction [Medical application]
has features	Manufactured object component Manufactured object	Manufactured object features	Layer[Manufactured object component] has features Biodegradability [Manufactured object features]
Is incorporated via	Biologically active substance	Material processing	RGD[Bioactive substance] is incorporated via Deposition[material processing]
Is processed via	Biomaterials Manufactured object component Manufactured object	Material processing	PLLA[Biomaterial] is processed by 3D printing [material processing

2.3 Terms and topics discovered automatically in the set and crowd-sourced validation

Text analysis (bag-of-words) was used to confirm the content of the semi-automatically curated gold standard. Moreover, the analysis of the gold standard generated lists of candidate terms for the ontology. The top 4000 terms in each list were scrutinized for inclusion by checking their presence in alternative terminologies (UMLS terminology services[18], bioportal repository). If terms appeared in a relevant existing terminology they were excluded, but the terminology was noted as an associated vocabulary for future annotations. Terms which could not be found or appeared but with an alternative meaning in reference ontologies were included (ex: 'layered', 'multilayered', which may refer to anatomical structure or surgical procedure rather than a manufactured object). This bottom-up, statistical approach using term frequency (tf) and term frequency inverted document frequency (tf_idf) enabled a minimally biased detection of relevant terms to populate the ontology, ensuring both common (tf) and less common but key (tf_idf) terms were selected.

Figure 4A and **B** show the top 10 most common terms (tf) and literature-specific terms (tf_idf) in the biomaterials gold standard compared to the random set. Predictably, some frequent terms were generic scientific words, with overlap to the general literature (ex: 'cell', 'patient', 'study'), but overall the biomaterials gold standard set had a distinct vocabulary. The tf_idf metric highlighted terms which are especially over-represented in the gold standard set compared to the random set (ex: pcl, nanofibre, osteogenic). For completeness, both of these categories were of interest for the purpose of the ontology.

Many of the discovered terms that were not found in UMLS/Bioportal were related to manufacturing techniques, material structures and associated features. In order to accommodate those terms the ontology was dynamically adapted, creating new classes, such as the subclass 'shape' (example: 'cube', 'cylinder') and the subclass 'architectural organization' (example: 'layered', 'oriented'). The significant number of discovered structures and material processing terms can be seen in **Figure 5**.

Hierarchical Latent Dirichlet Allocation (hLDA) model allows mining topics from a large amount of discrete data and organizes them hierarchically. Here, topic discovery was used to enable validation for the gold standard set, which was expected to have a certain degree of bias due to the initial selection of abstracts which was done manually. In total, 26 topics were discovered, with five words representing each topic (**Figure 4C**). The three largest topics, with over 100 abstracts each, were all related either to bone or cartilage (or both). Other tissues mentioned included nerve, tendon, skin, vascular, ligament, corneal, lung and periodontal. As terms were discovered based on probability, they belonged to varied categories ranging from material feature to cell type. Whilst topic mining showed variation in tissues, materials and study approaches within the gold standard set, further crowd-sourced validation was sought from scientists researching biomaterials in order to ensure the gold standard is sufficiently representative of research themes in the domain. The survey is still live, and continued feedback is expected to be used for further iterations of the gold standard set and the ontology.

At the time of writing, over 70 scientists from 19 countries responded to our request to take a survey about their topic of research, replying to the questions outlined in section. A summary of the results of survey questions 1 – 4 and 6 are presented in **Figure_S2**. 2% of survey respondents were researchers in industry, 10% were researchers in both industry and academia, and the rest (>87%) were academic researchers. 88.9% of researchers marked ‘biomaterials and tissue engineering’ as relevant to their research interests. When asked to select one or more topics relevant to their research, all topics but one (caso, shockwave, radial, fibrocartilage, csgel) were selected by at least one researcher as relevant, and 14 topics were selected by 10 or more researchers as relevant. 72% marked the topic ‘bone, osteogenic, differentiation, surface, cell’ as relevant to their area of research, showing a bias towards bone tissue engineering. This bias appears to be built into the literature, as a simple PubMed search of the top journals in the field shows a significantly higher studies mentioning ‘bone’ as keywords compared to all other searched tissues (see **Table_S3**).

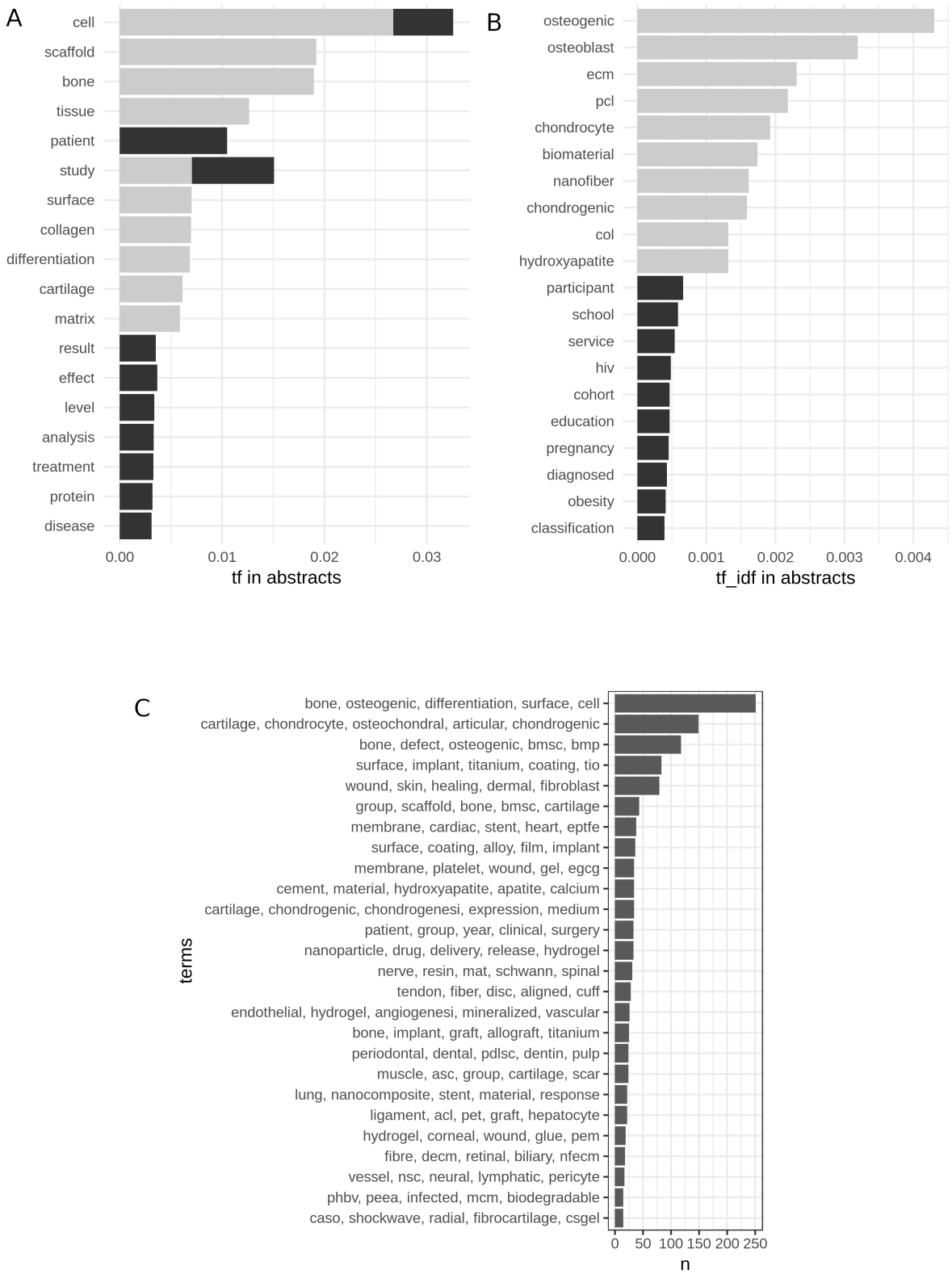


Figure 4. Text analysis results A) The ten most frequent terms (*tf*) and B) the ten words most important terms (*tf-ifd*) in the gold standard (light gray) and random sets (black). Comparing the

gold standard biomaterials article set to the randomly sampled set enabled identification of key terms in the domain. C) The 5 terms associated with each of the 26 sub-topics identified in the Biomaterials gold standard set using hLDA. 'n' is the number of abstracts belonging to each discovered topic. Both tissues and materials appear as topic terms, and bone and cartilage are the dominant tissues in the set.

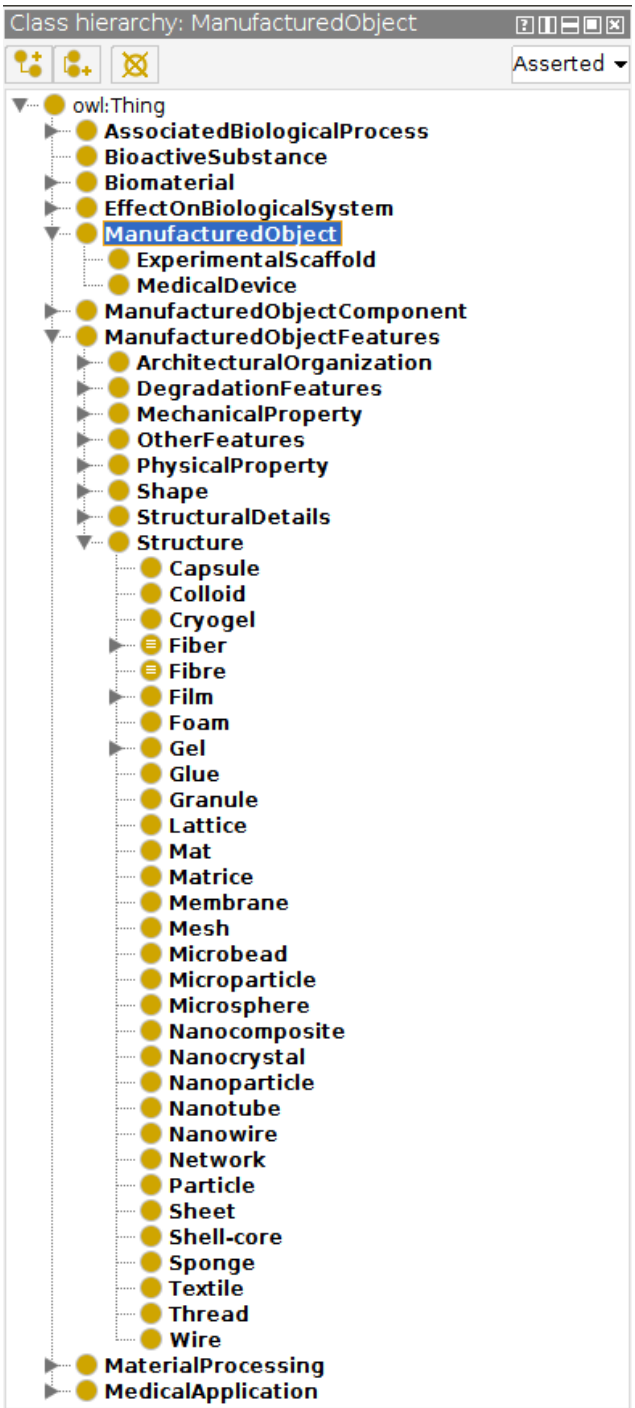
Another interesting finding was the high number of scientists selecting 'nanoparticles', although only ~50 articles were found to belong to that topic. That could be explained by a recent surge of interest in nanoparticles, or a bias in the group that responded to the survey.

Only 4 researchers did not find any relevant topic to their research in the list. The terms they proposed to represent their area of research are in Figure_S2, and are center around encapsulation of beta-islets, plasma medicine, and embryonic development. To address this, the gold standard has been searched for these suggested terms. The term 'plasma treatment' was mentioned 9 times and the term 'plasma' 104 times. The term embryonic development was only mentioned twice, and thus, to strengthen the set, three additional abstracts describing biomaterials used for the investigation of embryonic development were manually selected and added to the gold standard list. Finally, the terms 'beta cells' and 'islets' were completely missing from the set, and thus 5 abstracts were manually identified to add to the gold standard.

2.4 Overview and availability of the ontology

The resulting owl2 ontology contained 244 classes, 11 object properties and 95 data properties. A beta version of the DEB ontology was deposited in the open repository NCBO-bioportal [19] and is available for all to download and use (https://bioportal.bioontology.org/ontologies/DEB). The ontology may be viewed dynamically or statically online (https://projectdebbie.github.io/), and a snap shot of the dynamic visualisation using the VOWL plug-in is in **Figure_S3**.

A



B

Figure 5. Overviews of the class hierarchy of the ontology in Protégé, showing the subclasses of ManufacturedObjectFeatures >> Structure (A) and MaterialsProcessing (B), which were the

classes with the least available terminologies. A yellow circle marked with white lines (ex: fibre, fiber) denotes a synonym. The black triangle indicates the existence of subclasses.

As can be seen in Figure 5, Features related to the manufacture and material properties (and in particular properties related to cell-material interaction and biocompatibility), such as 'manufactured object features' and its subclass 'structure' have extensive lists of subclasses. This was intentional, as these terms are very important for the extraction and interpretation of biomaterials information, but to date they have not been organized in a structured vocabulary.

3. Discussion

Algorithms used in machine learning systems and artificial intelligence (AI) can only be as good as the data used for their development[20], and getting good quality annotated data is a well-known bottleneck for applying those techniques. With knowledge in the biomaterials domain continuously expanding, efforts to mine data from published studies are already being made in sub-areas such as inorganic materials [21,22] and polymers [23], and the potential value of text mining in the biomaterial domain is clear. Thus, the aim of this study was to develop a biomaterials ontology to facilitate information extraction in the domain. To fulfill that aim, a new systematic method was developed to create and validate two important open-access lexical assets: a gold standard biomaterials literature set and the biomaterials ontology (DEB). The method described here relied on both automated tools and human experts' input. Importantly, an effort was made to work systematically and rely on statistical measures, using human input for validation tasks. We believe this approach is both robust and economic.

The first resource, a gold standard biomaterials literature set, was developed using a combination of manual and automated steps. It was created to enable relevance classification as well as be used as a representative set of the biomaterials field. Metrics of the gold standard are presented in this article, and it is openly available from the project's GitHub account (https://github.com/ProjectDebbie).

Because the first step in the curation process (selecting key records) was manual, it was especially important to validate it is inclusive of as many topics as possible. Validation was therefore crowd-sourced via an online survey of wet-bench biomaterials scientists using professional networks, conferences and academic meetings. Out of approximately 120 scientists contacted, over 70 responded to the survey to date (the survey remains open), providing feedback to the topics discovered in the set. This validation led to the expansion of the set to include a few additional abstracts in areas that were not initially covered, and it ensured that the set is representative of sufficient topics.

The second asset created here is the Devices, Experimental scaffolds and Biomaterials Ontology (DEB). Using a bottom-up approach and statistical text analysis, the ontology was populated with frequent and significant terms in the domain, rather than relying on the opinion of selected experts [11]. Also, in contrast to the BCTEO ontology, DEB aims to cover all materials tested in a biological system, rather than focus on a specific tissue. We believe this approach reduces bias, giving a more complete representation of the field's terminology. Operating within an interdisciplinary domain, the ontology was created to complement rather than overlap with existing vocabularies in the medical and cell biology domains. This partial reliance on external terminologies (such as UMLS) exploits prior knowledge and ensure good interoperability. Nevertheless, existing have clear limitations and only capture partial knowledge, and therefore identifying terms in the contemporary literature and using a data-driven approach to extract directly from text helps better reflect the key technical terms used by the biomaterials research community.

Although ontologies are often created to normalize the use of terms within a defined domain and encourage their consistent use[24], this was not the main aim here. Below are the main purposes seen for the DEB ontology:

1. An annotation resource: DEB was primarily created to facilitate Named Entity Recognition (NER) in the biomaterials literature and enable pooling and comparisons of information relating to manufactured biomaterials as part of the creation of a biomaterials database.
2. Overcome the lack of an accepted taxonomy: DEB defined the classes of terms which play a role in the naming and description of scaffolds. Thus, it aspired to describe as many attributed features as possible (such as manufacturing methods, structures, physical properties) in order to enable accurate comparison of reported results, making it highly useful for the recognition of distinct scaffolds with their associated features beyond the materials they are made of.
3. Improve discoverability of terms: one advantage of the methodology developed here is that terms were also selected based on their frequency in the raw text, making navigation through the ontology more tightly linked to the accepted language in the field.
4. Indexing, retrieval and knowledge completion: whilst these are all secondary uses, an ontology can be exploited to search and index documents as well as for logical reasoning. Where it can be applied for completing partial information.

A point worth making is that although the state of the art in text mining is extremely advanced, its applications in many scientific and interdisciplinary domains remain limited or non-existent, because they require specific efforts tailored to domain-specific issues and the investment of experts' time and effort. In order to move beyond the 'proof of concept' studies to actually use existing data, mining efforts must be directed by the end users, which are the domains' experts. Therefore, the inclusion of wet-bench scientists in the process of developing the assets described here was seen to have additional value to mere validation. Also for that reason, both the ontology and the gold standard set are free to download from open repositories, and are thus available for the scrutiny, use and improvement of the research community, in line with open science initiatives and contributing to FAIR biomaterials data.

The assets described here are currently being tested as components/resources of an automated information extraction pipeline for database curation of biomaterials information (see an annotated abstract in **Figure_S4**). Ongoing work with these resources means they are constantly expanded and improved, and future iterations are expected to be larger, more complex and ultimately better. Future work will also need to address several limitations of the ontology. One is the constant growth in terms and techniques, and the process of expanding the ontology automatically. A second important issue is resolving ambiguous terms in the multidisciplinary space. These are terms that may have parallel, different meaning in different disciplines (example: medicine vs. material science). Resolving such ambiguities will require additional resources, beyond the scope of the

$$tf_{(word)} = \frac{n_{appearances\,of\,the\,word}}{n_{total\,number\,of\,all\,words}}$$

ontology developed here. The expansion plan for the ontology includes internal input from its application to specific sub-domains as case studies as well as external, from continuous monitoring of the survey's results and any user feedback received via the open GitHub repository. The gold

$$idf_{(word)} = \ln\left(\frac{n_{number\,of\,documents}}{n_{number\,of\,documents\,containing\,the\,word}}\right)$$

standard set is also expected to be updated using user feedback and case studies.

Conclusions

We describe here a systematic development process and validation of two important lexical assets: a

$$tf-df_{(word)} = tf\left(word\right) \times idf\left(word\right)$$

gold standard literature set and an ontology for the biomaterials domain. Such open access resources stand to benefit researchers by enabling faster and more efficient access to valuable information. Whilst text mining toolkits are revolutionizing research and development in domains such as genetics and chemistry, introducing similar efforts in the biomaterials domain requires community participation in the form of validation, trying new tools, sharing data and supporting

domain-specific efforts. Ultimately, such efforts have the potential to build on the vast amount of existing information for better, more effective advanced biomaterials and medical implants.

4. Methodology

4.1 Biomaterials corpus definition and gold standard set

To create the gold standard set, articles were identified using the MeSH terms and keywords list in Table 1 and downloaded in .csv format from the PubMed database. Thereafter, relevant articles were manually selected by scanning the spreadsheet manually and selecting relevant records in approximately one-month intervals. Special effort was made to select articles with as many varying topics as possible. After a preliminary set of 251 relevant abstracts was selected manually, their PubMed ID's (PMIDs) were used to train the MedlineRanker classifier in order to rank all abstracts from the past 10 years by their similarity to the relevant set. MedlineRanker is an open access University of Mainz (Mainz, Germany) webserver offering a Naïve Bayse classification algorithm to directly rank PubMed abstracts [25]. The top 1000 ranked records, which are those considered to be most related to the topic of interest, the were manually scanned to remove reviews and added to the Biomaterials set. The final filtration step was the exclusion of records with no abstract, non-English records and records published earlier than 2004. The final biomaterials set contained 1173 references.

A second, randomly sampled set (called 'random set' here) was created to enable a comparative analysis of the biomaterials set against the general literature, as detailed in the results. To ensure maximal random distribution of the abstracts from the PubMed database, PMIDs were generated as pseudo-random numbers (within the range of PMIDs in the years 1999-2018), using python's random package (Python Software Foundation, version 3.6, Available at http://www.python.org). These PMIDs were used to retrieve full citations via the PubMed eBot tool, which is one of the educational resources offered by the National Center for Biotechnology Information (NCBI)[26]. To ensure the random set does not contain any biomaterials articles, it was also ranked in the

MedlineRanker, and the top 200 ranked records were manually scanned, of which 7 records were deemed relevant to the biomaterials set and were therefore removed. Reviews, records with no abstracts, non-english records and records published earlier than 2004 were also excluded.

4.2 Ontology creation

The ontology was developed using Protégé(19), which is a free and open source ontology development tool(20). The methodology developed to curate terms into the ontology is described in detail in results' section . To avoid the curation of terms already well-organized in other lexical sources, the existence of key terms in other terminologies and ontologies were searched using the Unified Medical Language System (UMLS) Metathesaurus Browser (uts.nlm.nih.gov/metathesaurus.html), the bioportal ontology recommender[27], the OBO Foundry (http://obofoundry.org/) and the ontobee server (http://www.ontobee.org/). The key ontologies and semantic types identified for linking with the biomaterials ontology are summarized in Table_S2.

4.3 Text analysis using the bag-of-words model

To validate the semi-manually curated gold standard set, as well as populate the ontology, a bag-of-words analysis was carried out on the two groups of abstracts (gold standard set vs random set). The bag-of-words model is a representation used in natural language processing (NLP) where a text (such as a sentence or a document) is represented as the bag (multi-set) of its words, disregarding grammar and word order[28]. This approach generates two commonly used metrics[29], calculated here for the purpose of selecting terms for the ontology: term frequency (tf) and term frequency inverse document frequency (tf-idf). The formulas to calculate these metrics are in Equation 1, Equation 2 and Equation 3

Equation 1: Term frequency (tf)

Equation 2: Inverse document frequency (idf)

Equation 3: Term frequency-inverse document frequency (tf-idf)

tf enables identification of the most common terms, whilst *tf-idf* allows identification of terms which are significantly important in the corpus of interest compared to a control corpus (in this case biomaterials against the randomly sampled PubMed records). For the text pre-processing, R packages (dplyr, tidytext, tidyr) were used to tokenize words, remove English stop words, digits and 1-2 letters expressions. The pluralize package (github repository, "hrbrmstr/pluralize") was then used to turn plural words to singular. The total unique word count after cleaning was 13,878 in the biomaterials set and 11,586 in the random set. In addition to single term expressions, n-grams (or a sequence of n terms, in this case n=2) were also generated, and similar values (*tf*, *tf-idf*) were used to order them by importance.

4.4 Topic discovery using hLDA

To characterize and validate the gold standard set, which was expected to be biased by the manual selection of the seed 251 records, Hierarchical Latent Dirichlet Allocation (hLDA) (python package github repository, "joewandy/hlda") was used to create a hierarchical LDA object and find topics within the corpus.

Historically, Latent Dirichtlet Allocation (LDA), which is a generative probabilistic model of a corpus, has been a common procedure in topic modelling [30]. In LDA, documents are represented as

random mixtures over latent topics, where each topic is characterized by a distribution of words. The model can be trained to fit text features, such as *tf* and *tf-idf*. However, one question that has arisen in the use of models such as LDA is how many topics a given set of texts has, which is a dimension that should be set ahead of topic discovery. This is an essential issue, given that data sets often grow over time, and as they do, new entities and structures are added[31].

To alleviate the requirement of setting a number of topics, Hierarchical Latent Dirichlet Allocation (hLDA) addresses the problem by learning topic hierarchies from data. This methodology generates a tree of classes, where each branch is a topic, and deeper levels inherit from upper branches. The model relies on a non-parametric prior called the nested Chinese restaurant process, or CRP [31], which allows for arbitrarily large branching factors and readily accommodates growing data collections. The hLDA model combines this prior with a likelihood that is based on a hierarchical variant of LDAs. The main requirement in this model is to specify the tree depth (or number of levels) through which it will iteratively look for subtopics.

Here, hLDA was applied to the gold standard set, English stopwords and words <3 letters were removed before *tf* scores were calculated for each remaining word in the vocabulary. 500 iterations and 3 levels of hierarchy were used to generate a list of the most probable 5 words per topic and the number of abstracts belonging to each topic. These topics were then validated for their relevance as described in section .

4.5 Crowd-sourced validation of the gold standard set

To validate the gold standard set, feedback from the research community was crowd-sourced using an online questionnaire on the typeform platform (https://www.typeform.com/). Over 120 scientists were requested to respond to the survey as well as forward it to colleagues in the area. At the time of writing, 72 scientists have answered the questionnaire, which contains 5 multiple choice and one open question as follows:

1. Which of the following best describes what you do (academic research/Research in both academia and industry/R&D in a company/None)

2. Choose all the fields that are relevant to your research interests (Biomaterials and tissue engineering/Material science and engineering/Biocompatibility/Bioactive substances for tissue engineering/Implant design/Clinical trials/Other)

3. Are any of the following 5-term sequences partly or fully relevant to your area of research? (Here options were all the topics discovered in the gold standard set, see Figure 4C)

4. If you did not find any relevant terms, could you please type 3-5 terms that best describe your area of work

5. Given an open-access biomaterials database, which 2-3 terms are you most likely to search?

The questionnaire was used to identify sub-topics that were missing in the the gold standard set and if necessary add abstracts covering these for a more complete terminology, and to correct for the expected bias in the gold standard set. The questionnaire remains open to enable continuous expansion of the validation and improvement of the set.

4.6 Figures and diagrams preparation

Bar charts were created in R ggplot2 package [32]. Diagrams and mind maps were created in Canva (www.canva.com). Ontology views were created in Protégé either by taking direct screen shots of the class hierarchy or using the VOWL plug-in[33] to generate a graphical depiction.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgments

The authors acknowledge the financial support of the the European Union Horizon 2020 research and innovation programme under the Marie Skodowska-Curie (MSCA-IF) grant agreement

DEBBIE, project number: 751277. MPG acknowledge the Generalitat de Catalunya for funding through project 2017SGR-1165 and the ICREA Academia award. The authors thank José María Fernández and Javier Corvi for their technical help.

[1] S. Rawat, S. Meena, *J. Res. Med. Sci.* **2014**, *19*, 87.

[2] N. Groen, M. Guvendiren, H. Rabitz, W. J. Welsh, J. Kohn, J. de Boer, *Acta Biomater.* **2016**, *34*, 133.

[3] E. N. Vasina, E. Paszek, D. V. Nicolau, D. V. Nicolau, *Lab Chip* **2009**, *9*, 891.

[4] K. S. Subia B, Mukherjee S, Bahadur RP, Vitor M Correlo, Reis Rui L, Sabarinathan R, Sekar K, *Open Tissue Eng. Regen. Med. J.* **2012**, *5*, 59.

[5] D. G. A. J. Hebels, A. Carlier, M. L. J. Coonen, D. H. Theunissen, J. de Boer, *Biomaterials* **2017**, *149*, 88.

[6] B. Blaiszik, K. Chard, J. Pruyne, R. Ananthakrishnan, S. Tuecke, I. Foster, *JOM* **2016**, *68*, 2045.

[7] M. Uschold, M. Gruninger, M. Uschold, M. Gruninger, **1996**.

[8] S. Hassanpour, A. K. Das, *CEUR Workshop Proc.* **2011**, *774*, 40.

[9] Gene Ontology Consortium, *Nucleic Acids Res.* **2004**.

[10] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcantara, M. Darsow, M. Guedj, M. Ashburner, *Nucleic Acids Res.* **2007**, *36*, D344.

[11] F. Viti, S. Scaglione, A. Orro, L. Milanesi, *BMC Bioinformatics* **2014**, *15*, S14.

[12] P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, M. A. Musen, *Nucleic Acids Res.* **2011**.

[13] M. D. Wilkinson, M. Dumontier, Ij. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S. A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. Van Der Lei, E. Van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, *Sci. Data* **2016**.

[14] N. Noy, D. Mcguinness, *Med. Informatics* **2011**, 1.

[15] V. Guarino, I. Cruz-Maya, R. Altobelli, W. K. Abdul Khodir, L. Ambrosio, M. A. Alvarez Pèrez, A. A. Flores, *Nanotechnology* **2017**, *28*, 505103.

[16] P. Senet, R. Bause, B. Jørgensen, K. Fogh, *Int. Wound J.* **2014**, *11*, 649.

[17] P. Chang, B. Guo, Q. Hui, X. Liu, K. Tao, *Oncotarget* **2017**, *8*, 25226.

[18] O. Bodenreider, *Nucleic Acids Res.* **2004**, *32*, D267.

[19] M. A. Musen, N. F. Noy, N. H. Shah, P. L. Whetzel, C. G. Chute, M.-A. Story, B. Smith, NCBO team, *J. Am. Med. Informatics Assoc.* **2012**, *19*, 190.

[20] FRA, *Fra – Eur. Union Agency Fundam. Rights* **2019**.

[21] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, A. Jain, Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **2019**.

[22] O. Kononova, H. Huo, T. He, W. Sun, Z. Rong, T. Botari, V. Tshitoyan, G. Ceder, *Prep.* **2019**, 1.

[23] R. B. Tchoua, A. Ajith, Z. Hong, L. T. Ward, K. Chard, A. Belikov, D. J. Audus, S. Patel, J. J. de Pablo, I. T. Foster, In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; 2019.

[24] S. Seppälä, A. Ruttenberg, B. Smith, *Cienc. da Inf.* **2017**.

[25] J.-F. Fontaine, A. Barbosa-Silva, M. Schaefer, M. R. Huska, E. M. Muro, M. A. Andrade-Navarro, *Nucleic Acids Res.* **2009**, *37*, W141.

[26] P. S. Cooper, D. Lipshultz, W. T. Matten, S. D. McGinnis, S. Pechous, M. L. Romiti, T. Tao, M. Valjavec-Gratian, E. W. Sayers, *Brief. Bioinform.* **2010**, *11*, 563.

[27] M. Martínez-Romero, C. Jonquet, M. J. O'Connor, J. Graybeal, A. Pazos, M. A. Musen, *J. Biomed. Semantics* **2017**.

[28] S. Deepu, R. Pethuru, S. Rajaraajeswari, *Int. J. Adv. Netw. Appl.* **2016**.

[29] J. Silge, D. Robinson, *J. Open Source Softw.* **2016**.

[30] D. M. Blei, A. Y. Ng, M. I. Jordan, *J. Mach. Learn. Res.* **2003**.

[31] D. M. Blei, T. L. Griffiths, M. I. Jordan, J. B. Tenenbaum, In *Advances in Neural Information Processing Systems*; 2004.

[32] H. Wickham, *Ggplot2 : elegant graphics for data analysis*.

[33] S. Lohmann, S. Negru, F. Haag, T. Ertl, *Semant. Web* **2016**, *7*, 399.

[34] M. Ortiz, R. Rosales-Ibáñez, A. Pozos-Guillén, C. De Bien, D. Toye, H. Flores, C. Grandfils, *J. Biomed. Mater. Res. B. Appl. Biomater.* **2017**, *105*, 785.

List of Figures (in order of appearance in the text)

Table 1. The list of MeSH terms and keywords used to generate a broad set of biomaterials references as a starting corpus for selecting the representative set.

Figure 1. The steps taken to curate the gold standard set, which is a set of abstracts aiming to represent the biomaterials field. Parallel 'random set' pseudo-randomly sampled from the general abstracts archived in PubMed was created for all the comparative analyses.

Figure 2. the approach used to create the DEB ontology resulted in a cyclic process, where identified terms were fed into a schema which was in turn adapted to link all the key terms logically.

Figure 3. A map of the key concepts directly related to medical devices/experimental scaffolds. Each concept in the map is defined below as a class in the ontology. Object properties linking classes are adjacent to the arrows (schema produced using Canva).

Table 2. The superclasses used in the ontology, their definition, example of subclasses and references to other terminologies or ontologies

Table 3. Object properties (relations) created in the ontology to define the relationships between the classes. Domain and range refer to the classes connected via the defined relation.

Figure 4. Text analysis results A) The ten most frequent terms (*tf*) and B) the ten words most important terms (*tf-ifd*) in the gold standard (light gray) and random sets (black). Comparing the gold standard biomaterials article set to the randomly sampled set enabled identification of key terms in the domain. C) The 5 terms associated with each of the 26 sub-topics identified in the Biomaterials gold standard set using hLDA. 'n' is the number of abstracts belonging to each discovered topic. Both tissues and materials appear as topic terms, and bone and cartilage are the

dominant tissues in the set. tissues and materials appear as topic terms, and bone and cartilage are the dominant tissues in the set.

Figure 5. Overviews of the class hierarchy of the ontology in Protégé, showing the subclasses of ManufacturedObjectFeatures >> Structure (A) and MaterialsProcessing (B), which were the classes with the least available terminologies. A yellow circle marked with white lines (ex: fibre, fiber) denotes a synonym. The black triangle indicates the existence of subclasses.

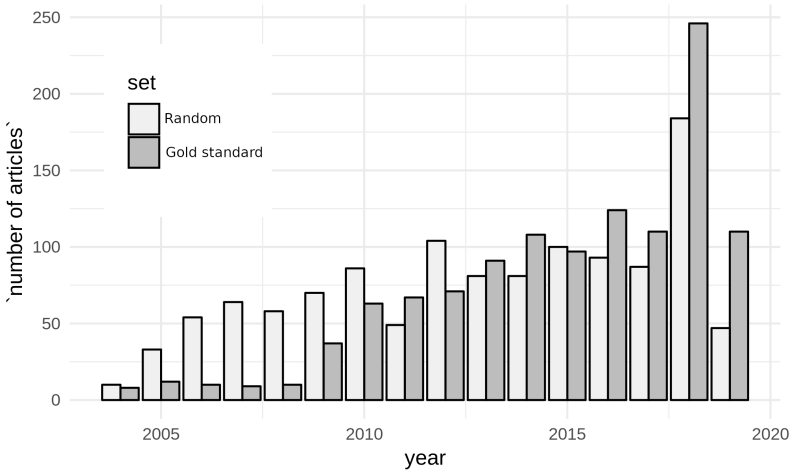
Supporting Information

Title: The Devices, Experimental Scaffolds and Biomaterials Ontology (DEB): a Tool for Mapping, Annotation and Analysis of Biomaterials' Data

Osnat Hakimi, Josep Luis Gelpi, Martin Krallinger, Fabio Curi, Dmitry Repchevski and Maria-Pau Ginebra*

Table_S1. Search results of various scientific documents databases in November 2019 using the query 'biomaterials OR cell scaffolds' applied to all fields

Indexing service	Number of records
MEDLINE (PubMed)	194,941
Web of Science	377,741
Scopus	880,416
Google scholar	~1,100,000



Figure_S1. Distribution of the article sets representing the biomaterials field (gold standard) and the general biomedical literature indexed in MEDLINE (random) by year. Bias towards the last 10 years in the gold standard set was a result of using MedlineRanker to expand the set over the maximal period of ten years built into the ranker.

Table_S2. Semantic types and ontologies linked to the biomaterials ontology

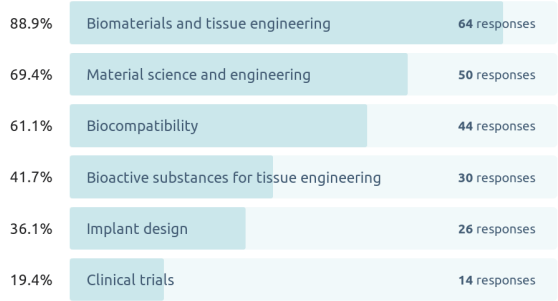
Name of Ontology	Acronym	Description and relevance	Source	Ref
Bone and Cartilage Tissue Engineering Ontology	BCTEO	Ontology that describes the field of Tissue Engineering for what concerns bone and cartilage tissues. Although there is some overlap of terms, most of the overlapping terms were organized differently in DEB. This is mostly because the BCTEO ontology focuses on experimental design rather than information extraction.	NCBO Bioportal	[1]
Semantic Types ontology	STY	A set of broad subject categories, or Semantic Types, that provide a consistent categorization of all concepts represented in the UMLS Metathesaurus. Used to link key objects,	NCBO Bioportal	[3]
National Cancer Institute Thesaurus	NCIT	A vocabulary for clinical care, translational and basic research, and public information and administrative activities. Used to link the list of biomedical and dental terms.	NCBO Bioportal	[4]
Chemical Methods Ontology	CHMO	Describes methods used to collect data in chemical experiments, prepare and separate material for further analysis, and synthesise materials. It also describes the instruments used in these experiments. Most important linked class was materials processing.	OBO Foundry	[5]
Ontology for Biomedical	OBI	Describes investigations, the protocols and instrumentation used, the material used, the data	NCBO Bioportal	[5]

Investigations	generated and the types of analysis performed on it.

Figure_S2. Summary of responses to questions 1-4 of the survey

Q1: Which of the following best describes what you do?

Q2: Choose all the fields that are relevant to your research interests (multiple answers allowed)



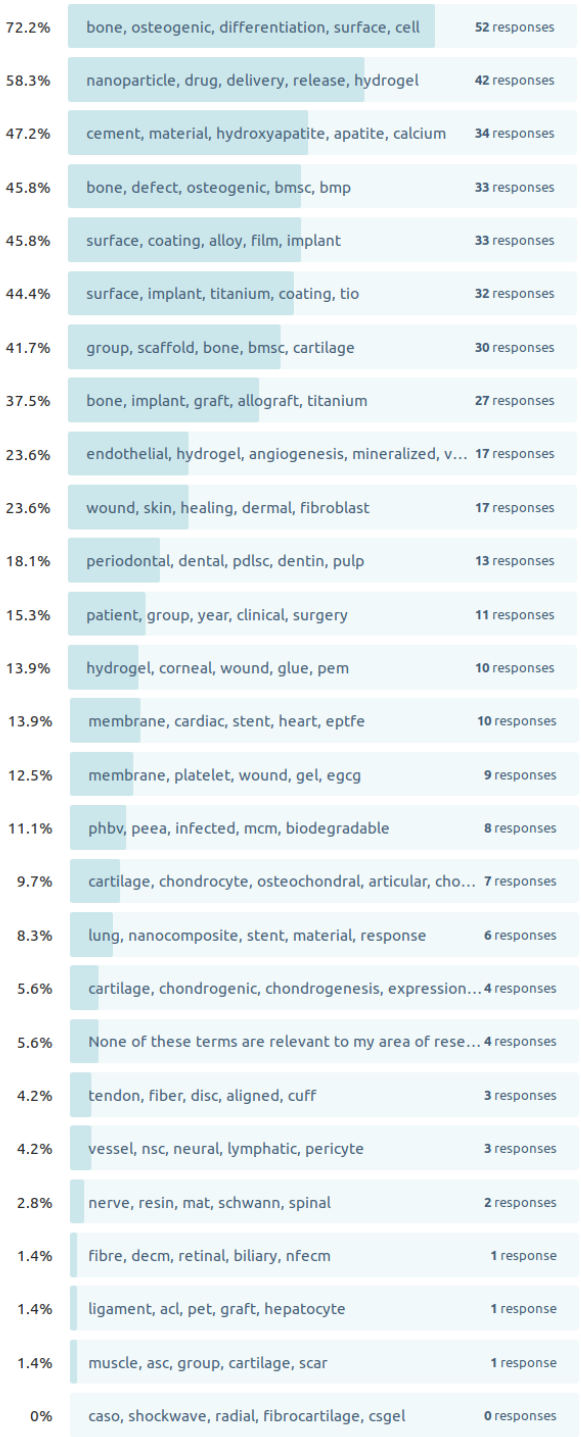
Q4: If you did not find any relevant terms, could you please type 3-5 terms that best describe your area of work

- Diabetes
- immunoprotection
- islets, beta cells
- embryonic development
- polymers
- hydrogels
- polymer characterization
- plasma-liquid interactions
- plasma medicine

Q6: In which country are you currently based?

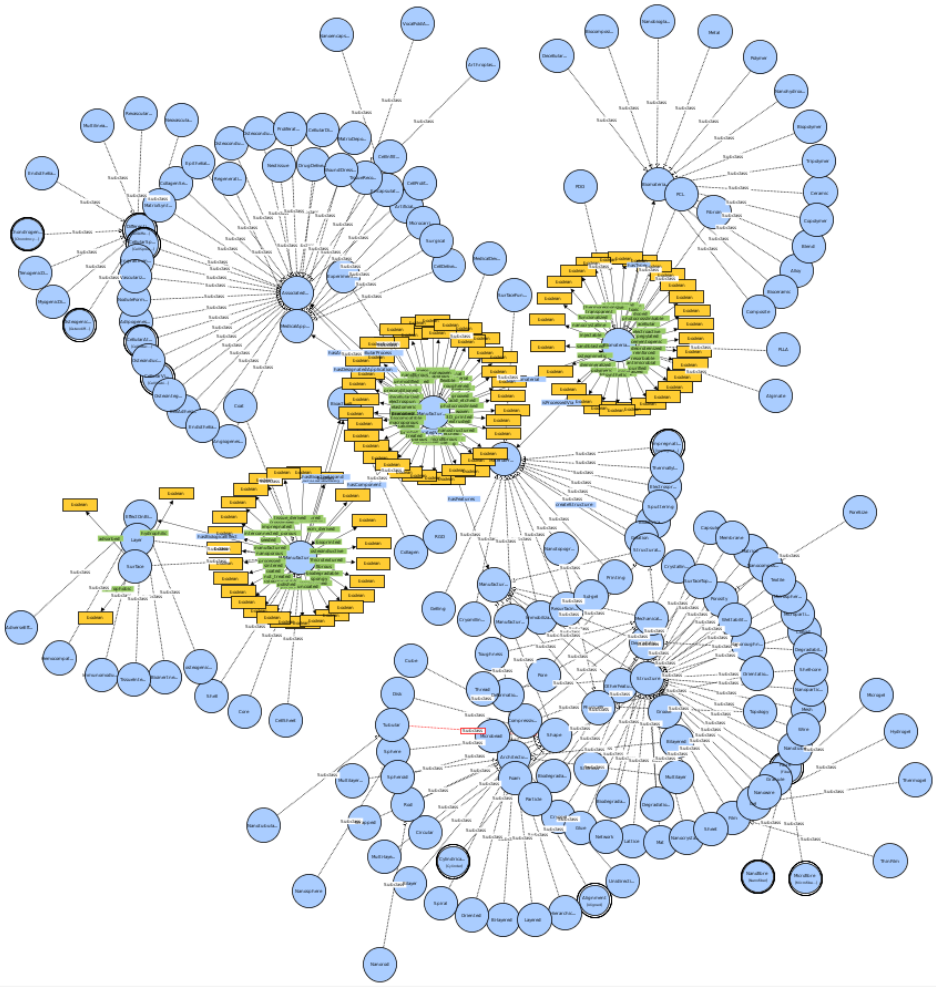
Brazil (1), Czech republic (5), Finland (1), France (1), Germany (4), Israel (2), Italy (3),
Mexico (1), Netherlands (11), Norway (3), Poland (1), Portugal (1), Spain (25),
Sweden (3), UK (2), USA (6), N/A (2).

Q3: Are any of the following 5-term sequences partly or fully relevant to your area of research? (multiple answers allowed)



Table_S3. PubMed searches of keywords and top biomaterials journals as of November 2019

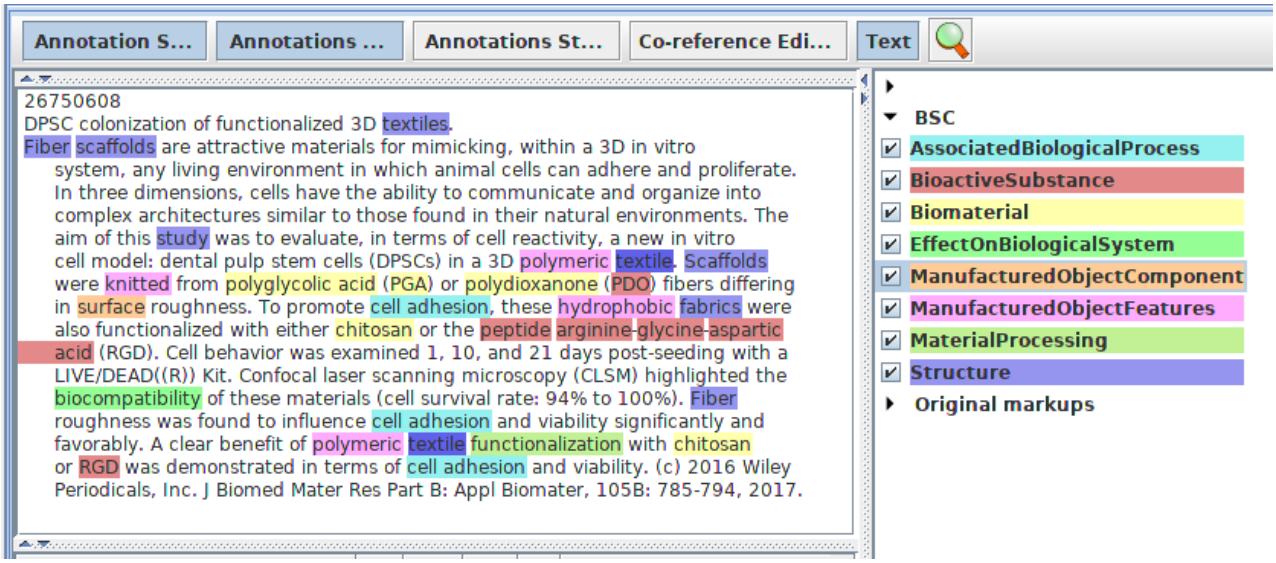
Search term	Tissue	Total entries
(("Biomaterials"[Journal]) OR "Advanced functional materials"[Journal]) OR "Acta biomaterialia"[Journal]	-	23,094
((("Biomaterials"[Journal]) OR "Advanced functional materials"[Journal]) OR "Acta biomaterialia"[Journal])) AND bone	Bone	5049
((("Biomaterials"[Journal]) OR "Advanced functional materials"[Journal]) OR "Acta biomaterialia"[Journal])) AND cartilage	Cartilage	974
Search ((("Biomaterials"[Journal]) OR "Advanced functional materials"[Journal]) OR "Acta biomaterialia"[Journal])) AND tendon	Tendon	233
Search ((("Biomaterials"[Journal]) OR "Advanced functional materials"[Journal]) OR "Acta biomaterialia"[Journal])) AND nerve	Nerve	618



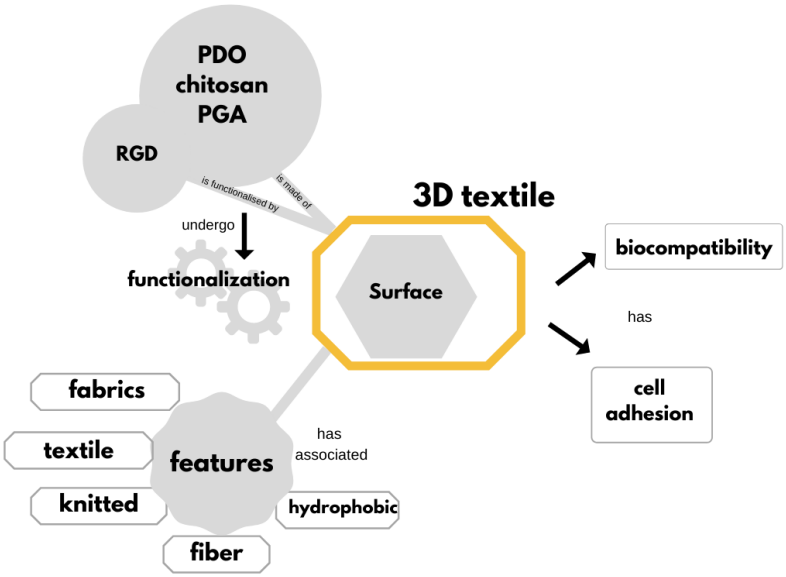
Figure_S3. The appearance of the ontology using VOWL plug-in, showing classes and sub-classes as blue

circles, object properties highlighted in blue, data properties highlighted in green and the latter's data type in yellow rectangles. The ontology may be viewed dynamically or statically online (https://projectdebbie.github.io/)

Figure_S4. A) An example of an abstract [34] annotated by a combination of the ontology in GATE and selected UMLS semantic types and B) Visualization of the identified terms within the ontology.



A



B