Bioinformatics, 33(22), 2017, 3627–3634
doi: 10.1093/bioinformatics/btx478
Advance Access Publication Date: 2 August 2017
Original Paper

Databases and ontologies

BiobankUniverse: automatic matchmaking between datasets for biobank data discovery and integration

Chao Pang[1,2], Fleur Kelpin[1], David van Enckevort[1], Niina Eklund[3], Kaisa Silander[3], Dennis Hendriksen[1], Mark de Haan[1], Jonathan Jetten[1], Tommy de Boer[1], Bart Charbon[1], Petr Holub[4], Hans Hillege[2] and Morris A. Swertz[1,2,*]

[1]Department of Genetics, Genomics Coordination Center, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands, [2]Department of Epidemiology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands, [3]Department of Public Health Solutions, National Institute for Health and Welfare, Helsinki, Finland and [4]Biobanking and BioMolecular Resources Research Infrastructure (BBMRI-ERIC), Graz, Austria

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on April 21, 2017; revised on July 4, 2017; editorial decision on July 22, 2017; accepted on July 22, 2017

Abstract

Motivation: Biobanks are indispensable for large-scale genetic/epidemiological studies, yet it remains difficult for researchers to determine which biobanks contain data matching their research questions.

Results: To overcome this, we developed a new matching algorithm that identifies pairs of related data elements between biobanks and research variables with high precision and recall. It integrates lexical comparison, Unified Medical Language System ontology tagging and semantic query expansion. The result is BiobankUniverse, a fast matchmaking service for biobanks and researchers. Biobankers upload their data elements and researchers their desired study variables, BiobankUniverse automatically shortlists matching attributes between them. Users can quickly explore matching potential and search for biobanks/data elements matching their research. They can also curate matches and define personalized data-universes.

Availability and implementation: BiobankUniverse is available at http://biobankuniverse.com or can be downloaded as part of the open source MOLGENIS suite at http://github.com/molgenis/molgenis.

Contact: m.a.swertz@rug.nl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The increasing breadth and depth of data in the biological sciences provides many new opportunities to understand the mechanisms that underlie complex diseases and essential background for personalized medicine and health. Much of this data resides in biobanks, which not only store sample collections (urine, blood and DNA) but also large data collections (e.g. history of disease, physical activity, lifestyle and environmental factors) (Scholtens *et al.*, 2015). With so many valuable resources available, one would expect much more scientific output for each biobank at an ever-increasing pace.

However, while working on various biobanking projects over the past five years, we noticed limited biobank reuse. What we observed instead was researchers spending a substantial amount of their time locating, negotiating access to and interoperating biobank data before they could actually study the pooled data. There are useful standards emerging for describing biobank collections such as MIABIS (minimum information about biobank information) (Merino-Martinez *et al.*, 2016), directories that list all available biobanks (Holub *et al.*, 2016), catalogues of biobank data schemas (Maelstrom Research, 2015) and robust integration protocols (Fortier *et al.*, 2010). However, researchers still routinely ask us how to find suitable biobank data collections for their research questions. They also spend many months manually curating and comparing biobank data elements to define integrated datasets because existing tools do not enable automatic matching.

Box 1: Overview of catalogue projects for data discovery

BBMRI-ERIC biobank directory: Main use case is to give an overview of the landscape of biobanks and biobank collections in the BBMRI-ERIC member states.

BBMRI-NL biobank catalogue: Main use case is to advertise all biobank collections available in Netherlands and lead interested researchers to contact these biobanks.

RD-Connect sample catalogue: Main use case is to give a comprehensive overview of the available samples for rare diseases.

LifeLines catalogue: Main use case is to allow the researcher to find and request access to data items of interest.

Maelstrom Research: Main use case is to provide harmonization potential (data attributes) between standard target data schemas and biobank studies.

In our recent experience the process of data harmonization and integration, driven by a research question, typically consists of the following steps (Fortier *et al.*, 2010): (i) find the datasets relevant to the research question; (ii) determine the harmonization potential between the target schema representing the research question and data elements in the relevant dataset; (iii) identify the attribute matches between the target schema and the source data for integration. Through a series of user workshops we listed several use cases in Box 1, based on which we have identified three major user needs in biobank data discovery:

1. Researchers want to **find biobank data collections** that can be potentially useful in terms of relevant data items in order to shortlist biobanks that might be suitable to serve a particular research project.
2. Researchers want to **assess the integration potential** of data collections and their data items (matching research variables) as the basis for data requests and to make decisions about whether it is worthwhile spending time on data integration for pooled analysis.
3. Biobanks (and networks of biobanks) want to **identify attribute matches between similar biobank data collections** to provide integrated datasets as basis for large studies.

In addition, all these use cases needed to be served using only metadata descriptions of the data, as individual level data is typically subject to data access committees because of privacy constraints.

Joining forces with the BBMRI and ELIXIR infrastructures and the CORBEL, ADOPT and RD-connect projects, we have developed BiobankUniverse. BiobankUniverse is an online service that bridges the biobank data discovery gap by (i) enabling users to share data element descriptions of biobank data collections and (ii) providing a new matching score that identifies pairs of related data elements between biobanks and research variables.

2 Materials and methods

In previously published work, we developed BiobankConnect (Pang *et al.*, 2015), a semantic search tool for matching data items between biobank data collections using ontology-based query expansion on top of the information retrieval system Lucene (The Apache Software Foundation, 2006). However, while achieving high precision and recall, BiobankConnect still requires substantial user input. Specifically, each of the desired 'target' attributes needs to be manually annotated with ontology terms before the system can try and find relevant 'source' attributes from biobanks that match this target. This is only feasible if the user wants to compare many 'source' biobanks against one relatively small 'target' set of data items.

To enable pairwise discovery considering all data items of many biobanks without requiring extensive curation we have developed a new algorithm that automatically shortlists matching data items between any two or more collections of data elements (such as data schemas in biobanks). To standardize the terminology throughout this paper, we will use 'attribute' to refer to a variable, data column, data element or data item. We implemented the algorithm as open source in Java and reused data management tools and user interfaces from the MOLGENIS software platform (Swertz *et al.*, 2010).

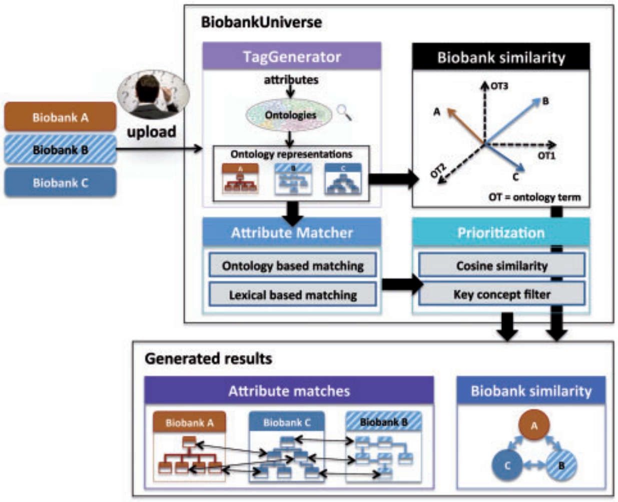


Fig. 1. Overview of the BiobankUniverse system. Users upload/add biobanks attributes to the universe. TagGenerator is automatically triggered to create ontology representations of the uploaded biobank's attributes. These are then used in AttributeMatcher to generate attribute matches with any of the other biobanks. A cosine similarity score is computed for each attribute match pair to prioritize the candidate list, and a strict matching criterion is applied to remove false positives. A biobank similarity is also calculated by computing the cosine angles between the ontology representations of biobanks in the semantic space for each pair

Figure 1 provides an outline of the system, which consists of six key steps: (i) automatic ontology tagging of attributes using lexical matching, (ii) matching pairs of attributes using ontology-based query expansion, (iii) matching pairs of attributes using lexical matching, (iv) prioritizing matches from both lists by calculating a normalized similarity score, (v) filtering irrelevant matches based on key-concepts to improve precision and (vi) calculating semantic similarity scores between biobank pairs. Each step is described in detail below.

2.1 Automatic ontology tagging of attributes using lexical matching

Because of their heterogeneous backgrounds, biobanks often describe their attributes using very different terminologies, which hinders the automatic matching of related or equivalent attributes. To enable matching based on these heterogeneous metadata, we ′tag′ each attribute with one or more groups of ontology terms based on the label + description. For example, ′History of Hypertension′ is tagged with two groups of ontology terms: (History && Hypertension) and (Medical history [synonym: History] && Hypertension). Each group of ontology terms is called a tag group.

With BiobankConnect, users had to do this tagging manually, which was not feasible when matching dozens of biobanks with thousands of attributes. In BiobankUniverse, each attribute is tagged automatically in four steps: (1) Having indexed the Unified Medical Language System (UMLS) ontology (UMLS is a meta-thesaurus that incorporates all major biomedical ontologies such as SNOMED CT, NCI thesaurus and ICD-10), we use the Vector Space Model (VSM) to find potentially relevant ontology terms for each attribute based on its label; (2) We apply a strict matching criterion to remove non-informative ontology terms. Only ontology terms (or synonyms) whose labels (or any their synonyms) can be completely matched to words from the attribute label are considered as tags; (3) We use a cosine-similarity-based string-matching algorithm to compute a similarity score between the attribute and the ontology terms, which we use to order the tags from most relevant to least relevant; (4) We remove non-informative tags. In this step, we use ontology terms with the highest similarity as the initial tag group then prune the rest of the list to see if inclusion of the next ontology terms as the tag group results in an overall improvement of the similarity score. If yes, we keep the new ontology term in the tag group. If no, we remove the term and repeat the same procedure for the next item in the list. The result is a set of ontology term tag groups for each attribute. An example of tagging attribute is shown in Supplementary example S1. In Pang *et al.* (2015), we discussed how to select ontologies for this procedure based on the extent that an ontology covers the data. Based on these experiences, we decided to use UMLS.

2.2 Matching pairs of attributes using ontology based query expansion

The tags established in step 1 are now used to search for semantically matching pairs of attributes between biobanks using semantic query expansion in a manner similar to what we previously described for BiobankConnect (Pang *et al.*, 2015). We have now changed the algorithm to query on terms from both parent and child classes (instead of child only) to ensure that the matches generated by this query expansion are symmetrical. This ensures that queries of more specific biobank attributes will still find matching attributes from another biobank that are tagged with more general ontology terms. An example of matching attributes is provided in Supplementary example S2.

In BiobankUniverse, we have also optimized query execution. In BiobankConnect, we created separate queries for each attribute to match a small number of attributes (<100). This is computationally too expensive for large numbers of biobanks with large numbers of attributes because we have encountered many attribute-matching cases, where more than 100 000 of expanded queries needed to be collected from the UMLS ontology and this process dramatically slowed down the matching process. Thus, in BiobankUniverse, we implemented a more efficient matcher that uses the hierarchical ontology term relations to discover the matching correspondences between those attributes. For example, the concept ‘Vegetables’ is a parent class of the concept ‘Beans’ so inferentially the attributes tagged with ‘Vegetables’ can be concluded as the matches for the attributes tagged with ‘Beans’.

To efficiently compare these hierarchical relationships, we collect all the term paths available for the tagged ontology terms into a list of atom unique identifiers of the current concept and its ancestors. For each attribute, we then check whether this term path or any of its parent term paths overlaps and, if so, we retrieve the corresponding attributes as the candidate match.

For example, the attribute ‘Consumption of Vegetables’ has path ‘A3684559.A3206010.A3314529.A2881738.A3217489.A2887927’ and the attribute ‘Consumption of Beans’ has overlapping path ‘A3684559.A3206010.A3314529.A2881738.A3217489.A2887927.A3189886.A2878987’, so we can conclude that ‘Consumption of Beans’ is a more specific match for ‘Consumption of Vegetables’ based on their paths. To prevent false positive matches based on very general concepts, we decided to limit the upward traversals to stop at level 5 from the root of UMLS after evaluating different cut-offs as discussed in Section 5.4.

2.3 Matching pairs of attributes using lexical matching

We also implemented a lexical matcher that uses standard search functionality from ElasticSearch. Given an attribute label/description from one biobank, the lexical matcher retrieves attributes from another biobank that share at least one word (excluding punctuation marks and stop words). The purpose of this matcher is to retrieve matches where the attribute labels are very similar and to retrieve attributes that have no tags to use for semantic matches. The motivation for this second method is that some of the attributes use terminology not yet defined in any ontology such as the attribute ‘SOKRAS sticker series’ in Finrisk2002 and Finrisk2007. Enabling lexical matching will help capture the matches containing those specific attributes.

2.4 Calculating a normalized similarity score to prioritize matches from both lists

Steps 2 and 3 produce two lists of candidate matches for each attribute based on the lexical matcher and the semantic matcher, respectively. To merge both lists, we calculate a similarity score for each matching pair using the cosine similarity algorithm also used in Lucene (The Apache Software Foundation, 2006). In this score, each ′query′ attribute from one biobank and its candidate matches from another biobank are treated as vectors in a space built of all words derived from all attribute names and descriptions. For each vector, the length of the dimension (word) is calculated by multiplying the word inverse document frequency with the word occurrence in the specific attribute. The vector and similarity score are computed as:

$$\overrightarrow{Vector} = (Word\_1_{tf} \times \ Word\_1_{idf}, \ldots, \ Word\_n_{tf} \times \ Word\_n_{idf})$$

$$\mathrm{Co\,sin\,}e = \frac{\sum_{i=1}^{n} Vector\_target_i \times Vector\_candidate_i}{\sqrt{\sum_{i=1}^{n} Vector\_target_i^{\,2}} \times \sqrt{\sum_{i=1}^{n} Vector\_candidate_i^{\,2}}}$$

It was particularly complicated to generate meaningful scores in cases where a pair of attributes are semantically close but have very different labels. This results in very low cosine similarity scores for matches that an expert user would recognize as a good match, e.g. ‘Consumption of Vegetables’ versus ‘Consumption of Beans’. We therefore also calculate a cosine similarity score based on the ontology terms instead of the attribute labels.

For each pair of attributes, we first retrieve all ontology tags that are either the same or related via parent-child or child-parent. We then replace the relevant substrings of the attribute labels with information from their ontology tags. For example, ‘History of high blood pressure’ and ‘History of hypertension’ are converted to ‘History of hypertension’.

If ontology terms are related via a parent-child or a child-parent relationship, we replace the child ontology terms with the parent terms in the attribute labels. However, these parent/child ontology terms are obviously not equivalent with the attribute label, just of a sub/superclass. We therefore correct their similarity score based on the semantic-relatedness between these parent and child ontology terms (Wu and Palmer, 1994). This correction is only performed on the subscore that is contributed by the relevant substring replaced by the information from ontology tags as follows:

$$\mathrm{Relatedness} = \frac{\mathrm{Level}_{\mathrm{parent}} \times 2}{\mathrm{Level}_{\mathrm{child}} + \mathrm{Level}_{\mathrm{parent}}}$$

$$\mathrm{Score}_{\mathrm{sub}} = \mathrm{Score}_{\mathrm{total}} * \frac{\mathrm{Length}_{\mathrm{replacement}}}{\mathrm{Length}_{\mathrm{total}}}$$

$$\mathrm{Score}_{\mathrm{corrected}} = \mathrm{Score}_{\mathrm{total}} - \mathrm{Score}_{\mathrm{sub}} + \mathrm{Score}_{\mathrm{sub}} \times \mathrm{Relatedness}^2$$

For example, when calculating the similarity score between attribute ‘Consumption of Vegetables’ and attribute ‘Consumption of Beans’, ‘Beans’ (level 8) is replaced with more general term ‘Vegetables’ (level 6). Without correction, the cosine similarity score would be 100% because both attribute labels are the same, which is clearly too high a score because the attributes are of semantically different levels. To correct for this, we first of all calculate the relatedness between ‘Vegetables’ and ‘Beans’,

$$\mathrm{Relatedness} = \frac{6 \times 2}{6 + 8} = 0.857$$

We then calculate the subscore that is contributed by ‘Vegetables’,

$$\mathrm{Score}_{\mathrm{sub}} = 100\% * \frac{10}{23} = 43\%$$

Finally we compute the corrected score,

$$\mathrm{Score}_{\mathrm{corrected}} = 100\% - 43\% + 43\% \times 0.857^2 = 88.6\%$$

After we have calculated all the similarity scores for all the candidate attribute matches, we sort the list based on similarity scores and keep (at most) the first 50 matching pairs (50 is the limit of user-acceptable matches based on BiobankConnect user feedback) (Pang *et al.*, 2015).

2.5 Filter out irrelevant matches based on key concepts to improve precision

The BiobankUniverse search methods are optimized to yield maximum recall. However, not all ontology terms are equally relevant for the research domain, and some may yield false positive matches. To reduce false positives, we enable users to filter results to matches that are based on ‘key concept’ ontology terms such as ‘Hypertension’ while discarding more general ontology terms such as ‘History’. For this we use the ′semantic type′ of UMLS ontology terms that indirectly indicate the importance of these concepts. For example, ontology terms associated with the semantic type ‘Disease or Syndrome’ (e.g. Myocardial infarction) are key concepts while the semantic type ‘Quantitative Concept’ (e.g. Numbers) indicates the common concepts. We used this as basis for the definition of the key concepts and went through the list of all 127 semantic types in UMLS and manually allocated them to the group of key concepts and the group of common concepts that are used in the system to determine the quality of the matched source attributes. Group members of the semantic types can be found in Supplementary Table S3.

Using these key concepts, we apply a lexical matching filter in which all the words from the key concept must be perfectly matched (considering lexical matching methods that allow for stemming etc.). For example, ‘Have you ever had high blood pressure?’ is a good match for ‘history of hypertension’ because both of the attributes are matched on the key concept **hypertension** whereas ‘history of myocardial infarction’ is far less relevant for ‘history of hypertension’ because the matched word **history** is not a key concept.

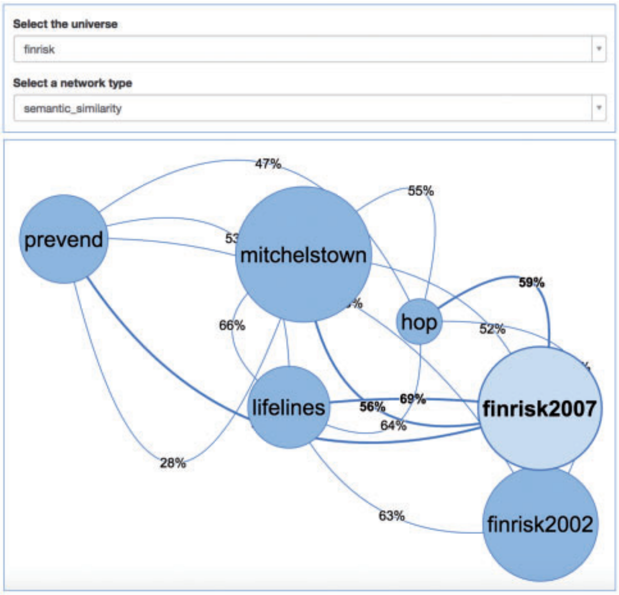
As an additional filter, attributes need to be matched based on words that are not stop words and consist of at least three alphabetic characters. If these two criteria are not met, the matches are treated as false positives and removed from the candidate list.

2.6 Calculate overall semantic similarity between biobanks

Finally, we created a metric to quantify the similarity between two biobank collections. At first we simply calculated the average of the attribute similarity for all of the candidate matches. However, this metric showed bias towards collections that were lexically similar and penalized semantic similarity. For example, the scores of the matches generated between FINRISK2002 and FINRISK2007 are systematically higher than the ones between HOP and Lifelines because FINRISK2002 and FINRISK2007 use very similar attribute labels and descriptions (see description of these biobanks in Section 4). We therefore implemented a metric that uses the semantic tags of the attributes.

Our new metric compares vectors of unique ontology terms derived from the tags of all attributes of both biobanks. Exactly matching terms are given a value of ′1′. Indirectly matching terms (i.e. a parent/child terms) are given a lesser score based on the semantic relatedness (Shima, 2011; Wu and Palmer, 1994). Finally, a cosine similarity is calculated on the vectors for the each biobank pair as described above in Step 4. For example, Biobank A has attributes tagged with the ontology term ‘Vegetables’ and biobank B has attributes tagged with the ontology terms ‘Beans’ and ‘Tomatoes’. When combined, there are three dimensions in their space and the vector representations are:

$$\overrightarrow{\mathrm{Biobank}\ A} = (\mathrm{Vegetables}: \ 1,\ \mathrm{Beans}: \ 0.8, \mathrm{Tomatoes}: \ 0.8)$$



Select the universe: finrisk

Select a network type: semantic_similarity

Fig. 2. User interface for discovering biobanks. Users can choose various network options to visualize the 'universe': the biobank similarity, the number of matches generated by the system or the number of matches curated by the user. The nodes represent biobanks in the universe and their sizes are proportional to the number of attributes in the corresponding biobanks. The connecting lines represent the similarities (defined as the number of matches or the biobank similarities) between biobanks, the more similar they are and the closer they are next to each other in the universe. The online version is dynamic so you can see the numbers more clearly

$$\overrightarrow{\text{Biobank } B} = (\text{Vegetables}:\ 0.8,\ \text{Beans}:\ 1, \text{Tomatoes}:\ 1)$$

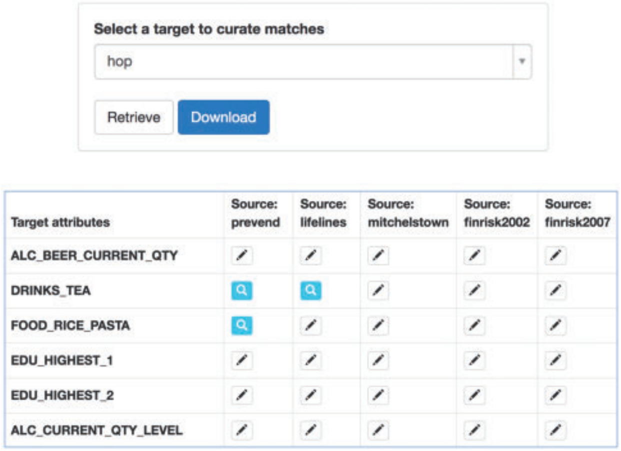
The cosine similarity between them is 0.978. Based on this measure, we can generate a matrix containing all pairwise similarities between all biobank collections available. We then visualize the matrix in a network using the Vis 3D JavaScript library to provide users with a visual representation of which biobank collections are closest to each other (see Section 4).

3 Implementation

We have made the biobank matchmaker algorithm available in a user-friendly web application (http://www.biobankuniverse.org). It can be also downloaded as part of MOLGENIS (http://www.molgenis.org). It uses a domain model (see the file data_model.pdf in Supplementary material) that extends the MIABIS standard model for 'Biobank' and 'SampleCollection' description (Norlin *et al.*, 2012). The system works as follows.

3.1 Biobankers upload collection metadata and match their attributes

Biobankers can upload data collection descriptions, i.e. the list of data items of an existing biobank or study for which data items can be shared via CSV. An example file can be found in Supplementary material prevend_biobank.csv. At upload, each attribute is automatically tagged with ontology terms. The tag groups and their quality measures (cosine similarity and matched words) are stored in the database for fast retrieval. The software then generates a list of candidate matches for each of the previously loaded biobanks. For example, the attribute 'Have you ever had high blood pressure' is matched with the tag group (Hypertension), a record of explanation is as follows, query string = 'high blood pressure'; matched words = 'high blood pressure'; ontology terms = 'Hypertension'; cosine similarity = 50%. All of the information on the matched source attributes, cosine similarities and matched words are stored in the AttributeMappingCandidate table. The tag groups cannot be edited at the moment but will be in the future.



Select a target to curate matches: hop

Target attributes	Source: prevend	Source: lifelines	Source: mitchelstown	Source: finrisk2002	Source: finrisk2007
ALC_BEER_CURRENT_QTY	✎	✎	✎	✎	✎
DRINKS_TEA	🔍	🔍	✎	✎	✎
FOOD_RICE_PASTA	🔍	✎	✎	✎	✎
EDU_HIGHEST_1	✎	✎	✎	✎	✎
EDU_HIGHEST_2	✎	✎	✎	✎	✎
ALC_CURRENT_QTY_LEVEL	✎	✎	✎	✎	✎

Fig. 3. Curating candidate matches by data owners. Users can curate all generated matches available in the universe. Users first choose a leading 'target', based on which a match table is generated. (Any biobanks can be a target because of the pairwise match). Users then need to go through each of the cells in the table to make decisions about the generated matches

3.2 Finding matching biobanks

Researchers and other prospective biobank users can use the system to find biobanks with relevant data and can explore the matching relationships between those attributes using a data discovery user interface (shown in Fig. 2).

When the page is first loaded, a biobank 'universe' is shown in the center of the page beneath the search box. The circles represent biobank members of the universe. The size of the circle indicates the number of attributes the biobanks contains. The connecting lines between circles represent the number of matching attributes between biobank members. Users can define their own queries in the search box at the top of the page. In order to retrieve attributes with high precision, the search box is equipped with an auto-complete function that provides suggestions from the UMLS ontology. Depending on the filter, the biobank universe will be reduced in size and the circles and number of matches will change dynamically. Users can also display the universe showing only human curated matches or using the semantic similarities between biobanks, as described above.

3.3 Exploring and curating attribute matches

Users can drill down to view and compare the attribute matches for a subset of biobanks. To start a comparison session, users first choose one of the biobanks as the 'target'. For each of its attributes, matches available in the other biobanks are then shown (see Fig. 3). Users can manually curate these matches using an editing interface in which they can select or reject matches. To more efficiently curate the large number of matches, we have introduced a batch acceptance feature that enables users to accept/reject all matches at once based on a quality criterion.

Table 1. Recall and precision performance for the HOP project (0–100)

	Lifelines		Mitchelstown		Prevend		Total		Biobank connect	
Rank	R	P	R	P	R	P	R	P	R	P
1	23	64	23	87	39	41	25	66	24	58
2	39	55	33	66	61	38	38	55	37	45
3	45	45	42	58	70	34	46	47	45	39
4	52	41	48	52	71	32	52	44	50	35
5	56	38	56	50	73	30	58	42	54	32
6	59	35	58	46	74	30	60	39	57	30
7	64	34	62	44	74	29	64	37	60	29
8	66	32	66	43	74	28	67	36	63	27
9	68	30	69	42	77	29	69	35	65	26
10	70	29	72	41	77	29	71	34	67	25
20	85	25	81	36	77	28	82	30	76	19
50	88	20	85	34	77	28	85	26	77	16

Note: P, precision; R, recall.

3.4 Searching for research variables

One of the main challenges in biobank research is finding datasets suitable for a particular analysis or for testing a particular hypothesis. To speed up this discovery process, users can also upload a complete list of desired research attributes and then start a data discovery job. This list is then shown as an additional circle within the universe. This search interface then works in the same way as the matching curation interface, enabling curation of the matches between desired research variables and biobank data items. The results can be downloaded for use as the basis for a data request.

4 Results

The main goal of BiobankUniverse is automatic generation of high quality lists of matching attributes between biobanks. To evaluate precision and recall, we re-ran our evaluation procedure from BiobankConnect (Pang *et al.*, 2015), which compares automatically found matches against human curated (relevant or 'correct') matches as follows:

$$Recall = \frac{Found\_relevant\_matches}{All\_relevant\_matches}$$

$$Precision = \frac{Relevant\_found\_matches}{All\_found\_matches}$$

We applied this to a new version of the validation data we used in Molgenis/Connect (Pang *et al.*, 2016): a human-curated matching set from the BioSHaRE Healthy Obese Project (HOP) consisting of 92 target attributes in three different biobanks (Wolffenbuttel, 2013). In addition, we also used a curation set between two large biobank collections from the FINRISK project.

4.1 BioSHaRE healthy object project performance

We evaluated BiobankUniverse‘s performance using the complete set of HOP, which consists of 92 target attributes, and three sets of biobank attributes (from the LifeLines, Mitchelstown and Prevend biobanks). There are 66 884 possible matches, out of which 633 were classified as relevant. We observed new average precisions and recalls over ranks ranging from 1st, to 50th (see Table 1) that are better than those of BiobankConnect (see Table 1) while providing major user time- and cost-savings because substantial manual tagging is no longer required. In addition, the new matching algorithm is more efficient than that of BiobankConnect. It took 2 min on average for BiobankUniverse to generate candidate matches between HOP and any of the biobanks, while 1 and half hour approximately for BiobankConnect to generate the candidate matches for the same pair.

Table 2. Recall and precision performance for the FINRISK project (including 550 manual matches)

Rank	Recall	Precision	Retrieved
1	0.813	0.592	755
2	0.878	0.325	1486
3	0.891	0.223	2197
4	0.898	0.171	2889
5	0.904	0.139	3563
6	0.911	0.119	4214
7	0.913	0.104	4834
8	0.915	0.092	5438
9	0.918	0.084	6032
10	0.922	0.077	6614
20	0.929	0.044	11605
50	0.938	0.027	19088

4.2 FINRISK large collection matching performance

We also evaluated the performance of BiobankUniverse using the National FINRISK Study, survey years 2002 and 2007, which involved matching two large biobank collections against each other with potentially 581 742 possible matches (798*729), of which 550 of were classified as 'correct' by human curators. Although the two surveys were conducted by the same research group, they were created in different time periods and the questions asked changed over time, thus requiring this integration effort. The motivation for matching these two collections is that they are often used together in analyses.

For example, the attribute ‘Siblings diagnosed with asthma’ collected in FINRISK 2002 changed to ‘sisters diagnosed with asthma’ and ‘brothers diagnosed with asthma’ in FINRISK 2007. Researchers who want to use data from both of the collections usually need to match the two sets of attributes with each other manually. In order to manually match all attributes in these two collections, the FINRISK researchers performed the following process: they organized and tabulated all attributes into topics one study at a time, and then compared the attributes against the items in the other collection, first inside each topic and then across the full collection if no match was found inside a topic. The quality of the matches was scored using SKOS mapping system (Miles and Pérez-Agüera, 2007). The full tabulation and comparison of the two collections was labor-intensive, taking approximately 2 working days. It is important to note that this work was done by a person highly familiar with these collections—the work would have taken longer for someone not familiar with them. We applied BiobankUniverse to FINRISK 2002 and FINRISK 2007 tabulated attributes and generated a set of matches between them. These matches were compared to the manually created list of matches (see Supplementary material FINRISK2002-FINRISK2007-relevant-matches.xlsx). We computed

Table 3. The overall performance comparison while enabling and disabling the matching criteria from the HOP experiment (including 633 manual matches)

	Matching criteria enabled			Matching criteria disabled		
Rank	R	P	RE	R	P	RE
1	0.25	0.66	240	0.24	0.56	268
2	0.38	0.55	443	0.36	0.44	516
3	0.46	0.47	613	0.43	0.37	735
4	0.52	0.44	753	0.50	0.34	931
5	0.58	0.42	877	0.54	0.31	1089
6	0.60	0.39	987	0.58	0.30	1235
7	0.64	0.37	1085	0.61	0.28	1373
8	0.67	0.36	1173	0.63	0.26	1506
9	0.69	0.35	1250	0.65	0.25	1630
10	0.71	0.34	1320	0.68	0.25	1751
20	0.82	0.30	1724	0.76	0.18	2723
50	0.85	0.26	2054	0.80	0.13	3848

Note: P, precision; R, recall; RE, number of retrieved matches.

precision and recall using the procedure described above, and found a recall of 0.81 precision of 0.59 at rank 1st and recalls of 0.92, 0.93 and 0.94 at rank 10th, rank 20th and rank 50th respectively, the complete set can be found in Table 2. According to the FINRISK researchers, approximately identifying a correct match within the top 10 candidate matches takes 10–20 s (ignore candidates outside the top 10). The complete curation process for 800 pairs of matches would take about 2–4.5 h and identify 92% of the true matches.

5 Discussion

Below we discuss improvements over BiobankConnect, how to reduce false positives, potential improvements of the matching procedure beyond lexical and semantic matching and other future work.

5.1 Improvements over BiobankConnect

BiobankUniverse is the successor to BiobankConnect, which was developed to find matches between a small target schema describing variables for a research project and large biobank schemas that (hopefully) provide these variables. BiobankConnect, however, required an unacceptable level of user interaction to achieve matching results with high precision. In BiobankUniverse, we therefore worked to reduce manual effort as much as possible. First, we enhanced automatic tagging to capture as many tag groups as possible. Second, we used UMLS semantic types to automatically remove false positives. Third, we introduced an objective measure to calculate the cosine similarity score and to discover matched words in order to provide users with a fairly good idea how the matches were generated. All together, these improvements enabled us to match large biobank collections against each other, and it is very encouraging to see that BiobankUniverse performs similarly to the more human-labor-intensive BiobankConnect.

5.2 Use of strict matching criteria to reduce false positives

Users questioned the added value of filtering using key-concepts. In response, we compared recall, precision and the number of matches retrieved with and without this filter using the HOP project data (see Table 3 for results). Applying the key-concept filters resulted in many fewer candidate matches while systematically increasing recall and precision. This is exactly as desired because the main purpose of these criteria is to improve precision by removing false positives so that users need to review fewer invalid candidate matches before finding all relevant matches. As shown in the examples in Table 3, users had to check 431 (1751–1320), 999 (2723–1724) and 1794 (3848–2054) fewer matches when applying the strict matching criteria at rank 10th, 20th and 50th. Suppose that rejecting a false positive would take a minimum of 10 s (in reality it could be more), users would have to spend at least 1, 3 and 5 h more to curate candidate matches at rank 10th, 20th and 50th respectively.

5.3 Improving ontology coverage of the domain

We could account for some of the poorer attribute matches because they were based on attribute labels from HOP that don't exist in the UMLS ontology, for which the system consequently couldn't use semantic matching. For example, the target attribute 'Current Consumption Frequency of Bakery Products' is manually matched to eight source attributes (e.g. Pancakes, Fruit Pies) in Mitchelstown, but the system failed to retrieve any of the relevant attributes. We know, retrospectively, that if the concept 'Bakery Products' had been annotated with the ontology term 'Starchy food' then all of the relevant matches would have been found by the system because all eight matches have been annotated with the ontology terms that are the subclasses of 'Starchy food' (e.g. Pancake is a descendant of Starchy Food).

5.4 Limiting the query expansion in the parent direction

During the development of BiobankUniverse, we realized that expanding queries towards the parent direction might result in unexpected matches as these include very broad concepts such as Disease or Food. We therefore experimented with various heuristics to remove these matches. The most promising results were achieved by limiting the distance from the root of the ontology at which the query expansion would stop. We therefore calculated recall and precision using the HOP data for 1-6 levels from the root (results shown in Supplementary Table S4). What we found was that precision increased with level up to level 5 from the root. This is because concepts are less general at higher levels and thus fewer false positives are produced. However, precision started to decline beyond the level 6. We also found that recall was relatively steady from the root up to level 5, then started to drop at the level 6. Apparently level 6 contains some informative ontology terms that help in the semantic matching. More importantly, the level 5 cut-off produces the best f-measure compared to other levels, we therefore chose level 5 as the final cut-off.

5.5 The limitation of the lexical and semantic based matching algorithms

The use of ontologies in matching algorithms has been effective in matching attributes, especially in resolving the differences between datasets in case of synonyms, hypernyms and hyponyms (Pang *et al.*, 2015). However, we still often encounter difficult cases where the attribute is described in a non-standard way and ambiguously. For example, the LifeLines attribute FOOD7A1 'How many cups did you on average use on such a day?' should be matched to the target attribute 'Current Consumption Quantity Of Coffee'. In this case the source attribute doesn't have any mention of 'Coffee' in the description and it‘s not clear that the question is referred to coffee, tea or something else. Thus only humans having inside knowledge are able to find such attribute matches.

We have piloted technical solutions for such ambiguities. For instance, we can use the language model GloVe, which is an unsupervised learning algorithm for obtaining the vector representations for words (Pennington *et al.*, 2014). The trained GloVe model outputs the probability for the word pair that indicates the likelihood of its co-occurrence. In the previous example of matching the key word 'tea' to 'coffee', we could use the GloVe model to find a list of the most frequently co-occurred words for 'coffee'. Because 'cup' and 'coffee' tend to appear quite often, we should see the word 'cup' ended up in the list and hence be able to succeed in matching 'Current Consumption Quantity Of Coffee' to 'How many cups did you on average use on such a day?'. We envision use of such technologies to further improve the matching algorithm.

5.6 Future perspectives for BiobankUniverse

Currently BiobankUniverse is used as a mapping tool where users can generate, curate and download the attribute matches. Our ultimate goal is to have a community powered service where everybody can submit their data dictionary to the existing 'universe'. The use case doesn't need to be restricted to the biobank domain only. We envision that other universes can be created using the same toolset. Currently we ask collaborators to send us data collections for uploading but plan to provide comprehensive documentation and video trainings for data contributors to enable self-service. We also want to start collaborations with registries such as EU directory (containing 500+ collections) to incorporate more data collection metadata (Holub *et al.*, 2016). Additionally we encourage not only data owners but also researchers to identify matches between datasets to improve the quality of the universe. BiobankUniverse will be particularly useful for discovering relevant datasets by searching certain combinations of selection criteria (certain ontology concepts) and determine harmonization potentials by quickly uploading their own data schema to find data sources in the universe. We realize we need to develop more advanced user interface components to accommodate these advanced use cases. For example, we plan to add more details about attribute matches in the universe for users to interact with. Finally we must invest in performance. In the current system it takes approximately 20 minutes for a laptop with a 4 core CPU and 8 GB RAM to generate matches between one pair of biobanks each containing 1000 attributes. In a biobank universe with 10 members, we would need to calculate 45 pairs. If all these biobanks also contain 1000 attributes, it would take 15 hours to construct the universe. As the universe grows, the computation time will grow near exponentially {$\text{time} = N^*(N - 1)/2$}. To address this problem, we plan to implement a more scalable pipeline to generate matches that can farm the matching across a parallel computer cluster.

6 Conclusion

We have created the BiobankUniverse system for quickly matching data attributes between biobanks by fully automating the matching procedure and by providing new user interfaces for data discovery and matchmaking. While saving much time and eliminating handwork, the performance of the system is also improved compared to the previous system BiobankConnect. In conclusion, we not only increased the speed of the system but also in the mean time we managed to maintain and improve the quality of the candidate matches.

Acknowledgements

We thank the communities of MOLGENIS, CORBEL, BBMRI-NL, BBMRI-ERIC, ELIXIR and RD-connect for assistance in the software development; THL Biobank for the FINRISK test case (to KS, NE); and BioSHARE for the HOP test case; and Kate Mc Intyre for editing the manuscript.

Funding

This work has been supported by CORBEL and BBMRI-NL (to FK, DvE, CP, and MS). CORBEL is an ESFRI cluster project in the European Union's Horizon 2020 research and innovation programme under grant agreement No 654248. BBMRI-NL is a research infrastructure financed by the Netherlands Organization for Scientific Research (NWO), grant number 184.033.111.

Conflict of Interest: none declared.

References

Fortier,I. *et al.* (2010) Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. *Int. J. Epidemiol.*, **39**, 1383–1393.

Merino-Martinez,R. *et al.* (2016) Toward Global Biobank Integration by Implementation of the Minimum Information About BIobank Data Sharing (MIABIS 2.0 Core). *Biopreserv. Biobank.*, **14**, 298–306.

Holub,P. *et al.* (2016) BBMRI-ERIC Directory: 515 Biobanks with Over 60 Million Biological Samples. *Biopreserv. Biobank.*, **14**, 559–562.

Maelstrom Research (2015) Maelstrom Research, https://www.maelstrom-research.org/ (9 March 2017, date last accessed).

Miles,A. and Pérez-Agüera,J.R. (2007) SKOS: Simple Knowledge Organisation for the Web. *Catalog. Classif. Q.*, **43**, 69–83.

Norlin,L. *et al.* (2012) A minimum data set for sharing biobank samples, information, and data: MIABIS. *Biopreserv. Biobank.*, **10**, 343–348.

Pang,C. *et al.* (2015) BiobankConnect: software to rapidly connect data elements for pooled analysis across biobanks using ontological and lexical indexing. *J. Am. Med. Inf. Assoc.*, **22**, 65–75.

Pang,C. *et al.* (2016) MOLGENIS/connect: a system for semi-automatic integration of heterogeneous phenotype data with applications in biobanks. *Bioinformatics*, **32**, btw155.

Pennington,J. *et al.* (2014) GloVe: Global Vectors for Word Representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543.

Scholtens,S. *et al.* (2015) Cohort Profile: LifeLines, a three-generation cohort study and biobank. *Int. J. Epidemiol.*, **44**, 1172–1180.

Shima,H. (2011) WordNet similarity for Java, https://code.google.com/p/ws4j/ (9 March 2017, date last accessed).

Swertz,M.A. *et al.* (2010) The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button. *BMC Bioinformatics*, **11**, S12.

The Apache Software Foundation (2006) Apache Lucene. *Agenda*, 2009.

Wolffenbuttel,B. (2013) Healthy Obese Project. 1.

Wu,Z. and Palmer,M. (1994) Verb Semantics and Lexical Selection. In: *32nd annual meeting on Association for Computational Linguistics*, p. 6.