ReDD-Observatory: Using the Web of Data for Evaluating the Research-Disease Disparity for Emerging Regions

Amrapali Zaveri*
Universität Leipzig

Ricardo Pietrobon+
Duke University

Sören Auer*
Universität Leipzig

Jens Lehmann*
Universität Leipzig

Michael Martin*
Universität Leipzig

Timofey Ermilov*
Universität Leipzig

* AKSW/BIS, Institut für Informatik
PF 100920, D-04009 Leipzig, Germany
{lastname}@informatik.uni-leipzig.de

+ Duke University Medical Center
Box 3094, Durham, NC 27710, USA
rpietro@duke.edu

ABSTRACT

It is widely accepted that there is a large disparity between the availability of treatment options and the prevalence of diseases in Emerging Regions of the World, thus placing individuals in danger. This disparity is partially caused by the restricted access to information that would allow healthcare and research policy makers to formulate more appropriate measures to mitigate this disparity. Specifically, this shortage of information is caused by the difficulty in reliably obtaining and integrating data regarding the disease burden for a given nation and the respective research investments. In response to these challenges, the Linked Data paradigm provides a simple mechanism for publishing and interlinking structured information on the Web. In conjunction with the ever increasing data on diseases and healthcare research available as Linked Data, an opportunity is created to reduce this information gap that would allow for better policy in response to these disparities. In this paper, we present the ReDD-Observatory, an approach for evaluating the Research-Disease Disparity based on the interlinking and integrating of various biomedical data sources. Specifically, we devise a method for representing statistical information as Linked Data and adopt interlinking algorithms for integrating relevant datasets.The assessment of the disparity is then performed with a number of parametrized SPARQL queries.We evaluate the results wrt. information quality and interlinking precision.As a consequence, we are for the first time able to provide reliable indicators for the extent of the research-disease disparity in emerging regions in an semi-automated fashion, thus enabling healthcare professionals and policy makers to make more informed decisions.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

1. INTRODUCTION

In this work, we explore how Web technologies and the Web itself can be employed to increase life expectancy and the quality of life in emerging regions. It is widely accepted that there is large disparity between the availability of treatment options and the prevalence of diseases in emerging regions of the World and the research investments to address these conditions. A classical example is *tuberculosis*, a widespread disease in most developing nations, but where antiquated treatment schemes have barely been updated for half a century [13]. Among the contributing factors to this disparity is the difficulty by policy makers in reliably obtaining and integrating information regarding the disease burden and associated research efforts for developing countries [16]. Without this information, the problem cannot be properly assessed and no corresponding policy measures can be taken. The consequences of this lack of appropriate policies are clearly damaging, costing lives and causing suffering in situations where appropriate information could lead to improved healthcare and research policy. To aggravate this situation, a feedback is created in that ineffectively treated diseases can become a starting point for epidemic or pandemic outbreaks for conditions such as HIV/AIDS, tuberculosis, malaria, and other global health challenges. Effective monitoring and evaluation is essential to: (a) Provide funders with evidence that their investments are improving programs and health. (b) Compare cost effectiveness of interventions. (c) Allocate resources most effectively[1].

While substantial efforts have already been undertaken to formally represent knowledge in the life sciences domain, a number of interesting questions such as the research-disease disparity investigated in this article can be only be answered by integrating information from multiple, dispersed, and heterogeneous sources. Efficient data and knowledge integration is therefore a crucial component for the attainment of new insights not only in the life science domain. Recently, the Linked Data paradigm emerged as a simple mechanism for employing the Web as a medium for data and knowledge integration, thus allowing for information publication and exchange in an interoperable and reusable fashion. Many different communities on the Internet already use Linked

[1] `http://www.path.org/measuring-impact.php`

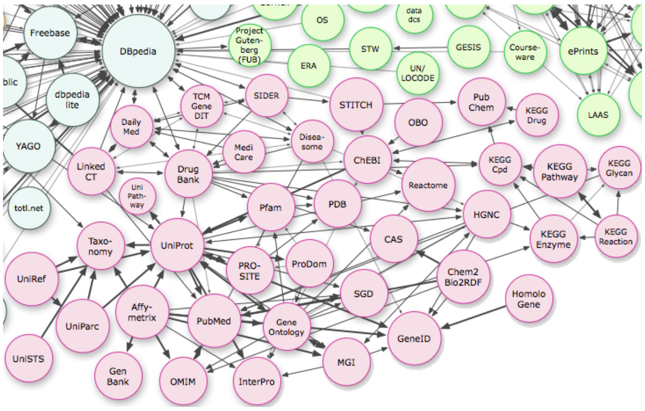


Figure 1: The life science Linked Data Web.

Data standards to provide and exchange information. This is confirmed by the dramatically growing Linked Data Web, where currently more than 25 billion facts are represented. In particular, the amount of information on diseases and healthcare research available as Linked Data is constantly increasing. Figure 1 demonstrates the part of the map of the Linked Data Web[2] covering the life science domain.

In this article, we present the ReDD-Observatory, an approach for evaluating the Research-Disease Disparity based on interlinking and integrating various biomedical data sources. Our approach is based on identifying and integrating datasets relevant for characterizing the disparity between biomedical research and disease burden, using data sources that are either already available as Linked Data or in proprietary formats. In order to deal with legacy formats, we devise a method for extracting statistical information from files in spreadsheet format and for representing the extracted information as Linked Data. We also adopt and evaluate interlinking algorithms for integrating the identified datasets - mainly the World Health Organization's *Global Health Observatory*, *ClinicalTrials.gov* and the US National Library of Medicine's *PubMed*. The assessment of the disparity is then performed with a number of parametrized SPARQL queries on the integrated data substrate. We evaluate the results with regard to information quality and interlinking precision. One particular observation originating from our results is that medical research being performed in developing regions might be biased towards diseases being prevalent in these regions but much less in the developed world (e.g. Malaria), even though a different allocation of resources might potentially be more effective.

The paper is structured as follows: We review related approaches both from the health-care and computer science perspective in Section 2. We give an overview on the general methodology in Section 3. The approach for dataset identification and conversion is described in Section 4. In Section 5, the interlinking and integration of the different datasets is presented. We describe the research-disease disparity assessment based on the integrated dataset in Section 6. Section 7 concludes with a review of the limitations of the ReDD-Observatory and an outlook on future work.

2. RELATED WORK

In this section we look at the current state-of-the-art regarding methods to evaluate the disparity between biomedical research and disease burden. Since these are mostly manual statistical methodologies we also review relevant related representation techniques and algorithms in the three categories: (a) biomedical data publishing and discovery; (b) knowledge interlinking and fusion and (c) assessment of data quality. Related work for each of these aspects is discussed in the following subsections.

Current methods to evaluate disparity. Measuring the Research-Disease Disparity allows healthcare policy makers determine whether research strategies developed by a given nation are aligned with their respective healthcare needs. For example, the Agency for Healthcare Research and Quality (AHRQ)[3] in the US publishes two annual National Healthcare Disparities reports [1]. However, these reports focus only on a limited number of socio-demographic groups and clinical conditions.

Although there is evidence suggesting that the investments in healthcare research are cost-effective, in a recent report the Commission on Health Research for Development highlighted the great imbalance between these investments and the global burden of disease [22]. The US government also reported a lack of a perfect association between the disease burden and amount of funding allocated for diseases in the United States [25]. Although the report only depicts the situation in the United States, there is a high likelihood that worse problems might exist in developing countries with a special focus on the poor, adolescents, and women [10]. Although people and countries in developing conditions are the ones at most risk, the lack of reliable, ongoing reports makes these disparities nearly impossible to monitor and therefore to appropriately address.

Among the many possibilities to measure this disparity is the generation of cross-sectional studies comparing estimates of disease-specific research productivity compared with different indices measuring burden of disease [7]. Other methods include the use of suitable statistical measures on samples of data to quantify the disparity [6]. These methods to calculate the disparity are not only cumbersome and time consuming but also are limited in that they use a limited sample of the data for analysis as opposed to using all data.

Although previous efforts have highlighted disparity issues between disease burden and research efforts in a given country that are beginning to be addressed, we see three major pending problems that we intend to address through the ReDD Observatory. First, since all information has to be manually collected by experts, current methods to generate reports that evaluate Research-Disease Disparity are burdensome and expensive. This problem is particularly pervasive in countries that need these evaluations the most, namely developing countries where the cost of such evaluations is prohibitive. As a direct consequence of this cost and expertise issue, current reports are not published as often, thus decreasing the ability of policy makers to obtain

[2] `http://lod-cloud.net/`

[3] `http://www.ahrq.gov`

a current perspective on the magnitude of these problems. And finally, since reports are scarce, comparison with other countries are not possible, thus making it difficult to search for successful policy models that could be used to level and decrease the disparity levels across nations.

Biomedical data publishing and discovery. There are numerous websites[4][5][6] and governmental efforts [21] containing information on healthcare disparities but this information is mainly published in textual or only partially machine-readable formats. Although these efforts bring together resources to either retrieve citations about healthcare disparities or display the statistics as charts or maps, their databases are not linked to other data sources. In addition, they do not directly address the disparity between research productivity and healthcare, thus missing the opportunity to guide policy makers. With efforts such as UMLS [20] and MeSH, a large number of taxonomies for the medical domain is available, although the reuse of common identifiers is still not at a level where queries crossing several datasets could be executed.

Knowledge Interlinking and Fusion. Interlinking occurs in the literature under a dozen of terms [24] such as Deduplication [11], Entity Identification [18], Record Linkage [15] and many more. Encountered problems are generally caused by data heterogeneity [8]. The processes of data cleaning [14] and data scrubbing [28] are common terms for resolving such identity resolution problems. As a new challenge, the Linked Data paradigm provides the means necessary to skip the data preparation step as they have already proliferated a shared structural representation of data. DBpedia [4] and other knowledge bases maintained by the Linked Data community are available as a crystallization points for new datasets. Combined with the proposed quality assessment and interlinking approaches, this allows lowering the access barrier for new open biomedical datasets to evolve into interlinked knowledge bases and join the network ecosystem. In a recent survey on Data Fusion [5], the semantic heterogeneity is considered as the greatest challenge for data integration and fusion, data fusion being the last step of a Data Integration process (preceded by schema mapping and duplicate detection) [19]. While ontologies or thesauri already play a major role when integrating several sources to overcome the semantic heterogeneity [29], the problem of creating a complete, concise and consistent integration has not been sufficiently addressed. In the context of the emerging Web of Data, new challenges include: 1. *on-the-fly integration* with a priori unknown data based on the discovered schema, 2. consideration of *provenance and trust* based on provided metadata, 3. creation of specific metrics to merge structured data based on the given vocabulary and *data quality*, 4. *handling of inconsistencies* in structured data (a key difference here is the defined semantics of ontology relations compared to e.g. relations in databases).

Assessment of data quality. Assessing the quality of data is essential due to the multiple and autonomous data sources that can be linked, which affects data accessibility and usability [3]. The dimensions of data quality commonly found in the literature include accuracy, consistency, timeliness, completeness, relevancy, inter-operability and trustworthiness. Ensuring high quality data requires two kinds of processing: (a) data validation and (b) consistency validation on knowledge fragments. However, values may often be missing in biomedical data due to several reasons such as lost or corrupted samples, patients not showing up for scheduled appointments or failing of measuring instruments. If there are too many missing values, ignoring them may invalidate the analysis that is performed [9]. One of the remedies is to fill in the gaps with suitable replacements such as either (a) fixed values or (b) existing values at random or (c) averaging neighboring values. There are numerous efforts employed for quality checking such as (a) data mining techniques, (b) comparing data on the web versus a gold standard, (c) using provenance information about the data on the web to assess the quality and trust-worthiness [17], (d) using a metrics for quality assessment [23] and others. Thus, assessing data quality is one of the important prerequisites for publishing linked data on the web.

3. METHODOLOGY

The overall strategy we followed in the development of the ReDD-Observatory is depicted in Figure 2. The methodology is based on the rationale of employing Linked Data datasets, which are published on the Web and interlinked using typed links. The use of Linked Data offers a number of significant benefits:

Uniformity. All datasets published as Linked Data share a uniform data model, the RDF statement data model. With this data model all information is represented in facts expressed as triples consisting of a subject, predicate and object. The components used in subject, predicate or object positions are mainly globally unique IRI/URI entity identifiers. At the object position also literals, i.e. typed data values can be used.

De-referencability. URIs are not just used for identifying entities, but since they can be used in the same way as URLs they also enable locating and retrieving resources describing and representing these entities on the Web.

Coherence. When an RDF triple contains URIs from different namespaces in subject and object position, this triple basically establishes a link between the entity identified by the subject (and described in the source dataset using namspace A) with the entity identified by the object (described in the target dataset using namespace B). Through the typed RDF links data items are effectively interlinked.

Integrability. Since all Linked Data source share the RDF data model, which is based on a single mechanism for representing information, it is very easy to attain a syntactic and simple semantic integration of different Linked Data sets. A higher level semantic integration can be achieved by employing schema and instance matching techniques and expressing found matches again as alignments of RDF vocabularies and ontologies in terms of additional triple facts.

[4] `http://www.ahrq.gov/data/dataresources.htm`
[5] `http://phpartners.org/health_stats.html`
[6] `http://apps.who.int/globalatlas/`

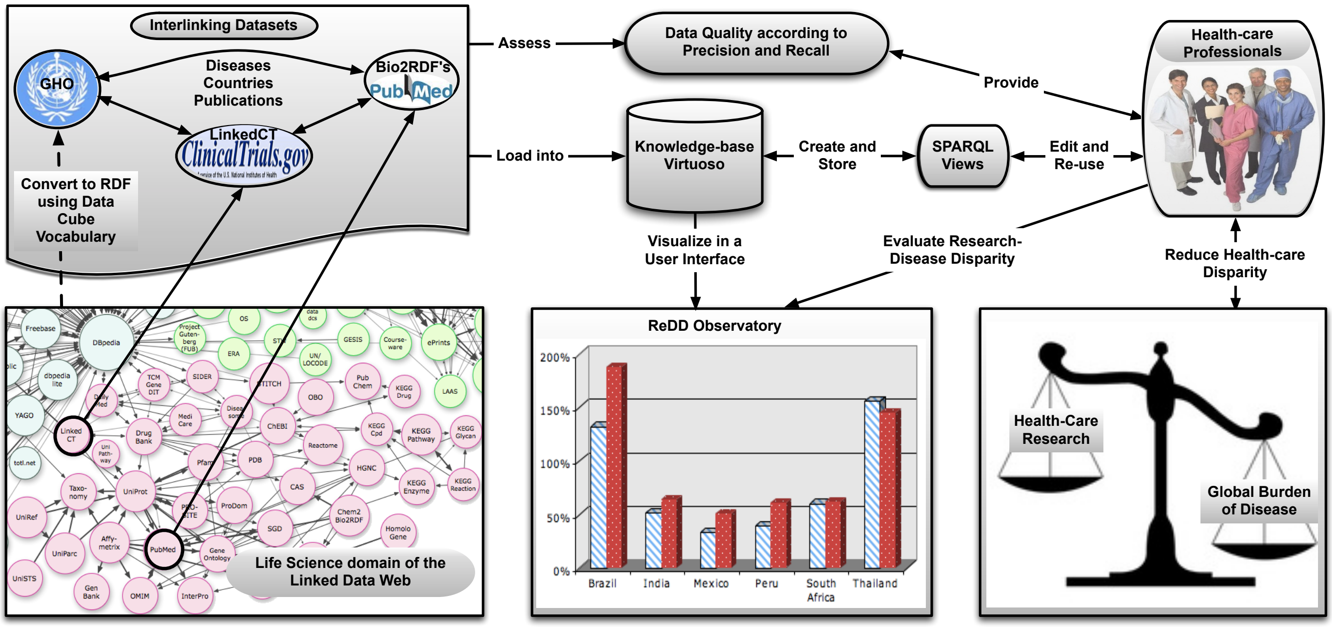


Figure 2: Bird's-eye view of the ReDD-Observatory.

Timeliness. Publishing and updating Linked Data is relatively simple thus facilitating a timely availability. In addition, once a Linked Data source is updated it is straightforward to access and use the updated data source, since time consuming and error prune extraction, transformation and loading is not required.

However, these advantages of pursuing a Linked Data integration approach can not be realized immediately, but will be attained by small iterative integration and refinement steps: 1. *Identification* of data sources. 2. *Conversion* and representation of legacy data. 3. *Data integration* by establishing RDF links between the datasets. 4. *Assessment* of the information quality. 5. *Storage and querying* of the integrated substrate using SPARQL views and a triple store. 6. *Evaluation* of the disparity according to various measures. In the next three sections we explain our proceeding in each of these steps and the respective results in more detail.

4. DATASET IDENTIFICATION AND CONVERSION

There are numerous biomedical datasets (cf. Figure 1) available that originate from different sources and span diverse biomedical domains. The integration of these datasets can enable the discovery of new hypotheses and powerful integrative analyses. However, after having performed an extensive analysis of relevant datasets we selected three particular ones, which are most relevant for answering the ReDD research question. We present these datasets in the first part of this section. Our analysis of relevant datasets also revealed that statistical data required for evaluating the disparity is available, but not yet published adhering to the RDF data model. Hence, in the second part of this section we devise a method for the semi-automatic conversion and representation of statistical data in RDF based on the DataCube vocabulary.

Dataset Identification.

In order to measure the research-disease disparity, we included indices for medical research intensity per disease and country as well as the metrics for disease burden within each country. In order to identify relevant datasets, we performed a comprehensive review of the life-science datasets available as Linked Data (as shown in Figure 1) and other available datasets. As a result of this analysis we identified the following three core datasets (cf. Table 1): (1) *ClinicalTrials.gov* (2) *PubMed* and (3) World Health Organization's *Global Health Observatory* (GHO).

ClinicalTrials.gov is an database of clinical trials, i.e. statistical studies aiming at providing evidence for the effectiveness of a treatment option (often a medication) for a particular disease. LinkedCT, is the Linked Data representation of ClinicalTrials.gov. It contains information about 61,920 governmentally and privately funded *clinical trials* conducted around the world, amounting to 9.8 million triples. LinkedCT comprises about 235,998 links to external datasets such as DBpedia [4] (in particular links to locations) and Bio2RDF.org's PubMed (in particular links to references). LinkedCT data is accessible via a SPARQL endpoint.

PubMed.gov is a service of the US National Library of Medicine that includes bibliographic information and abstracts of over 19 million publications from MEDLINE and other life science journals. It covers the fields of medicine, nursing, dentistry, veterinary medicine, the healthcare system and preclinical sciences. Bio2RDF is a mashup of about 42 different bio-medical knowledge bases, aiming to facilitate the creation of bioinformatics information systems. It also contains data from PubMed in RDF with about 800 million triples. PubMed contains about 30,000 links to other datasets such as GenInfo identifier, MeSH and DBpedia to name a few. PubMed (via Bio2RDF) is also available as SPARQL endpoint, which is constructed from 593 files of the 2009 MEDLINE release.

Table 1: Details of core datasets in the ReDD-Observatory.

Dataset	Linked Data	SPARQL Endpoint	Triples	Classes	Properties	RDF Links
ClinicalTrials.gov	LinkedCT.org	data.linkedct.org/snorql	7M	13	90	235,998
PubMed.gov	Bio2RDF.org	pubmed.bio2rdf.org/sparql	797M	120	362	30,000
WHO.int/gho	gho.aksw.org	gho.aksw.org/sparql	3M	8	5	6,000

WHO's GHO dataset is published in accordance to the WHO report ”Global burden of disease: 2004 update” [27]. The aim of GHO is to provide access to data and analyses for monitoring the global health situation. Specifically, it contains statistical information regarding the mortality and burden of disease classified according to the death and *DALY* (disability-adjusted life year) estimates grouped by countries and regions. It is further classified according to the age and gender as well as according to the WHO regions. For our purpose, we focus on the death and DALY (disability-adjusted life year) estimates that GHO reports according to countries per disease. DALY is the concept of reporting the the number of lost years of ”healthy” life. The burden of disease is thus an indirect measurement of the amount of suffering caused by a given condition to a well-established population in a specific period of time. The consistent and comparative description of the burden of diseases and the risk factors that cause them facilitates health decision-making and planning processes. However, the measures of death and DALY estimates are only available spreadsheet format thus hindering the linking and integration with other information. The details of the conversion of statistical data (such as GHO data) into RDF are described in the next section.

RDF Representation of Biomedical Statistical Data.

Biomedical statistical data is often represented in spreadsheets, i.e., by describing a single data item (e.g. disease prevalence) in several dimensions (e.g. country, year). As a consequence, a simple RDF transformation representing each spreadsheet row as an ontology instance of the said data is not meaningful. The Data Cube Vocabulary [12] is based on the popular SDMX standard[7] and designed particularly to represent multidimensional statistical data using RDF. The statistical dataset is considered a multi-dimensional cube which can be characterized by a set of dimensions that define what the values apply to (e.g. time, country, population), along with the metadata describing what was measured (e.g. death rate), how it is measured and how the observations are expressed (e.g. rate, status). Thus, a cube is organized according to a set of dimensions, attributes and measures collectively called components. A set of dimensions is sufficient to describe a single observation. The measure describes the phenomenon that is reported. The attribute, on the other hand, qualifies and interprets the observed value, such as the status of the observation. The dimensions, attributes and measures are represented as RDF properties. Each is an instance of the abstract `qb:ComponentProperty` class, which in turn has sub-classes `qb:DimensionProperty`, `qb:AttributeProperty` and `qb:MeasureProperty`. Another feature of the Data Cube vocabulary is that it allows defining the structure of the dataset, which enables verification that the dataset matches the expected structure. The `qb:DataStructureDefinition` allows a user to determine which dimensions are available for query. Thus, the data structure definition can be defined once and reused for similar structured files. The Data Cube vocabulary also uses the SDMX feature of content oriented guidelines (COG). COG's define a set of common statistical concepts and associated code lists that can be re-used across datasets.

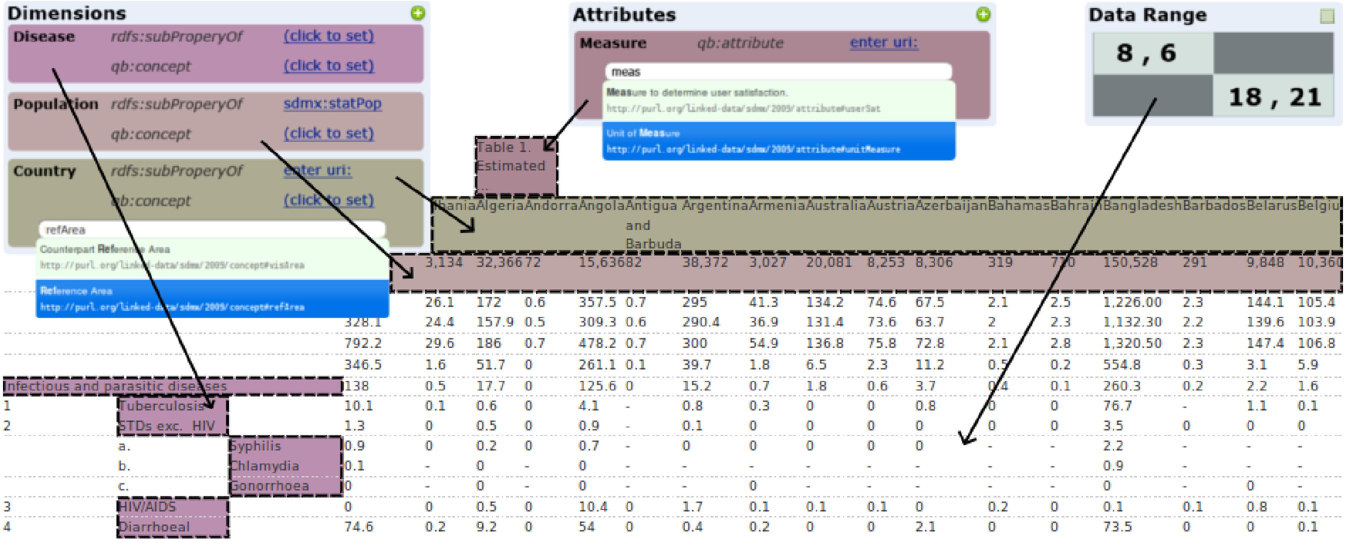
The GHO dataset contains statistical death and DALY rates per country for each disease. The data in GHO is published in yearly increments as spreadsheets. Generating a transformation from the GHO spreadsheets to RDF based on the Data Cube Vocabulary in a fully automated way is not feasible, since the spreadsheet publication format contains many implicit information (such as formatting and indentation). For example, dimensions are often encoded in the heading or label of a sheet or figures may be given as a fraction of 1000 so as to save space. To facilitate the transformation, we developed a semi-automatic approach by integrating the algorithm as a plug-in extension into OntoWiki [2]. OntoWiki is a tool which supports collaborative creation, maintenance and publication of RDF knowledge bases. In addition to ontology engineering tasks, OntoWiki provides ontology evolution functionality, which can be used to further transform the newly converted statistical data. Furthermore, OntoWiki provides various interfaces (in particular Linked Data and SPARQL interfaces) to publish and query RDF data.

As is illustrated in Figure 3, when a spreadsheet containing multi-dimensional statistical data is imported into OntoWiki, it is presented as tables. This presentation of the data gives the users the ability to configure (1) dimensions; (2) attributes by manually creating them and selecting all elements belonging to a certain dimension and; (3) the range of statistical items that are measured. The corresponding COG concepts are automatically suggested, using RDFa, when a user enters a word in the text box provided. It is also possible to save and reuse these configurations for other spreadsheets, which adhere to the same structure (e.g. for data in consecutive years). Once the transformation is configured by the user, the Data Cube importer plugin for OntoWiki takes care of automatically transforming the spreadsheets into RDF. After converting the data reported for the mortality and burden of disease in GHO, classified according to countries and region, we obtained an RDF dataset containing 3 million triples[8]. An example of the incidence value *1098* is illustrated in the following listing:

```
        qb:dataset        eg:dataset-in1 ;
        gho:Country       "Afghanistan" ;
        gho:Disease       "Tuberculosis" ;
        eg:incidence      "10.1" .
```

[7] `http://sdmx.org`

[8] Available at: `http://aksw.org/Projects/GHO2SCOVO`



Dimensions

Disease rdfs:subProperyOf (click to set)
qb:concept (click to set)

Population rdfs:subProperyOf sdmx:statPop
qb:concept (click to set)

Country rdfs:subProperyOf enter uri:
qb:concept (click to set)

refArea
Counterpart Reference Area
http://purl.org/linked-data/sdmx/2009/concept#visArea
Reference Area
http://purl.org/linked-data/sdmx/2009/concept#refArea

Attributes

Measure qb:attribute enter uri:

meas
Measure to determine user satisfaction.
http://purl.org/linked-data/sdmx/2009/attribute#userSat
Unit of Measure
http://purl.org/linked-data/sdmx/2009/attribute#unitMeasure

Data Range

8 , 6 | 18 , 21

Table 1. Estimated

				lbania	Algeria	Andorra	Angola	Antigua and Barbuda	Argentina	Armenia	Australia	Austria	Azerbaijan	Bahamas	Bahrain	Bangladesh	Barbados	Belarus	Belgiu
				3,134	32,366	72	15,636	82	38,372	3,027	20,081	8,253	8,306	319	710	150,528	291	9,848	10,360
				26.1	172	0.6	357.5	0.7	295	41.3	134.2	74.6	67.5	2.1	2.5	1,226.00	2.3	144.1	105.4
			328.1	24.4	157.9	0.5	309.3	0.6	290.4	36.9	131.4	73.6	63.7	2	2.3	1,132.30	2.2	139.6	103.9
			792.2	29.6	186	0.7	478.2	0.7	300	54.9	136.8	75.8	72.8	2.1	2.8	1,320.50	2.3	147.4	106.8
			346.5	1.6	51.7	0	261.1	0.1	39.7	1.8	6.5	2.3	11.2	0.5	0.2	554.8	0.3	3.1	5.9
Infectious and parasitic diseases			138	0.5	17.7	0	125.6	0	15.2	0.7	1.8	0.6	3.7	0.4	0.1	260.3	0.2	2.2	1.6
1	Tuberculosis		10.1	0.1	0.6	0	4.1	-	0.8	0.3	0	0	0.8	0	0	76.7	-	1.1	0.1
2	STDs exc. HIV		1.3	0	0.5	0	0.9	-	0.1	0	0	0	0	0	0	3.5	0	0	0
	a.	Syphilis	0.9	0	0.2	0	0.7	-	0	0	0	0	0	-	-	2.2	-	-	-
	b.	Chlamydia	0.1	-	0	-	0	-	-	-	-	-	-	-	-	0.9	-	-	-
	c.	Gonorrhoea	0	-	0	-	0	-	-	0	-	-	-	-	-	0	-	0	-
3	HIV/AIDS		0	0	0.5	0	10.4	0	1.7	0.1	0.1	0.1	0	0.2	0	0.1	0.1	0.8	0.1
4	Diarrhoeal		74.6	0.2	9.2	0	54	0	0.4	0.2	0	0	2.1	0	0	73.5	0	0	0.1

Figure 3: Screenshoot of the OntoWiki statistical data import wizard displaying a GHO table configured for conversion into RDF.

5. DATASET INTERLINKING AND FUSING

One of the major challenges when integrating heterogeneous information obtained from different sources is to establish links between different data items. In our case, we aim to establish links between the LinkedCT, PubMed and GHO datasets in particular for diseases and countries. This was carried out using two approaches: (a) using UMLS (Unified Medical Language System) and (b) using SILK, a tool for interlinking RDF resources[26].

Interlinking based on UMLS. In order to interlink the different sets of disease identifiers used in our core datasets, we used the *Unified Medical Language System* (UMLS) which integrates key terminology and associated resources in the medical domain. In particular, we used *MeSH terms* (which are part of UMLS) to generate the links. MeSH (Medical Subject Headings) is the National Library of Medicine's controlled vocabulary thesaurus. It consists of sets of terms (i.e. synonyms) naming descriptors (e.g. diseases) arranged in a hierarchical structure that permits searching at various levels of specificity. Both, the synonyms and the hierarchical structure are very important properties in our case: Firstly, synonyms allow us to establish `owl:sameAs` links (i.e. equivalence) between diseases labeled differently in the different datasets. Secondly, the hierarchical structure helps us to establish `rdfs:subClassOf` links (i.e. containedness) between diseases represented on different levels of granularity in the different datasets. To accomplish this interlinking we proceeded as follows:

Let s be a string, $terms(s)$ the terms occurring in this string, and $syn(\{s_1, \ldots, s_n\})$ the union of UMLS synonyms for each string s_i $(1 \leq i \leq n)$. Assuming we have a disease name d_{GHO} in GHO and a disease name d_{CT} from LinkedCT, we defined their UMLS-similarity sim as:

$$sim(d_{GHO}, d_{CT}) = \frac{syn(terms(d_{GHO})) \cap syn(terms(d_{CT}))}{syn(terms(d_{GHO})) \cup syn(terms(d_{CT}))}$$

We used a relational database system and the SQL MATCH AGAINST command to compute the similarity matrix and picked those above a user defined threshold. Upon manual evaluation of a sample of 500 links, the precision is approximately 96%.

Interlinking using SILK. Silk 2.0 [26] is a tool for discovering relationships between data items within different knowledge bases, usually available via SPARQL endpoints. Silk provides a declarative language for specifying (1) the types of RDF links that should be discovered between data sources and (2) the conditions which the data items must fulfill in order to be interlinked. The details of the interlinking results are displayed in Table 2. We used the *Jaro distance* as string metric where applicable. For the matching process, we used to confidence value thresholds: Links above 0.95 confidence were accepted and links between 0.90 and 0.95 were logged to a separate file for manual inspection. Figure 4 shows the links between the three datasets. The thick black lines with arrows on both sides represent the link between the different classes. They are labeled with the properties linking the instances of those classes. Some links were already present such as `owl:sameAs` links for publications between LinkedCT and PubMed with 42,219 links[9].

6. RESULTS AND EVALUATION

Description of Indices. The indices used to monitor the disparity between research performance and disease burden contain indicators for both measures, spanning across a wide range of countries and diseases. First, we selected *death rate* and *DALY* (disability-adjusted life year)[10] as an initial set of disease burden indicators following WHO guidelines. The death rate is defined as the number of people whose death

[9] `http://esw.w3.org/HCLSIG/LODD/Interlinking`

[10] `http://www.who.int/healthinfo/global_burden_disease/metrics_daly/en/index.html`

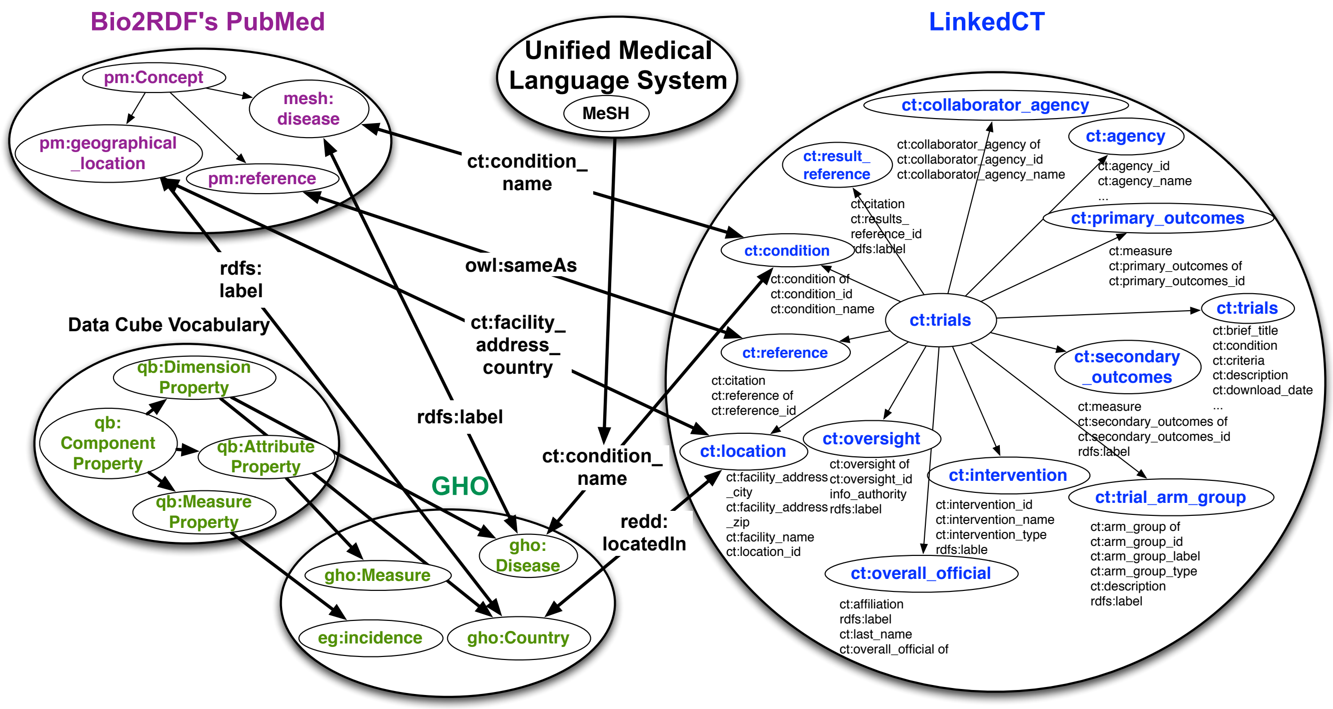


Figure 4: Depiction of the interlinking performed between the three core datasets (all used namespace prefixes are resolvable via `http://prefix.cc`).

Table 2: Number of links obtained between the three core datasets for disease and countries.

Link type	Source		Target		Links		Precision	
	Dataset	Instances	Dataset	Instances	Accepted	To verify	Accepted	To verify
Diseases	PubMed	23618	LinkedCT	5000	1910	1233	0.996	0.8418
Diseases	LinkedCT	5000	GHO	128	163	43	0.9625	-
Diseases	GHO	128	PubMed	23618	453	75	1	0.7083
Countries	PubMed	23618	LinkedCT	55000	4999	0	1	-
Locations	LinkedCT	757341	GHO	192	300000	0	1	-
Countries	GHO	192	PubMed	23618	201	12	1	0.9583

can be attributed to a given disease in a given country. DALY is defined as "a time-based measure that combines years of life lost due to premature mortality *and* years of life lost due to time lived in states of less than full health."[11] Disease burden indicators were all obtained from the GHO database. The disease burden indicator was normalized by representing it as a percent ratio between the disease burden for a given condition for a given country over the disease burden for all diseases for a given country. Indicators were placed in the denominator of the research-disease index so that 100 represented a perfect match between research effort and disease burden for a given country and for a given disease. Numbers over 100 represent an over investment in research for that area, whereas numbers under 100 represent underinvestment. Given that figures for the disease burden were only available till 2004, we followed a standard procedure of using it as the most recent disease burden information for that period while matching that information against research productivity data from more recent years. Second, to represent research productivity we made use of *number of articles* and *clinical trials* for a given country, disease and year. For some graphics we also used the total span of all available years. The number of articles was obtained from PubMed and that of clinical trials from LinkedCT. The research productivity index was normalized by creating a ratio between total productivity for a given disease in a given country over total research productivity for a given country. The research index was placed on the denominator to indicate that a surplus of research productivity in relation to disease burden would generate a number above 100. Given that certain articles and trials can focus on more than one disease at a time, this index can overestimate the total percentage of productivity for a given condition, but for the purposes of this articles this limitation was considered negligible.

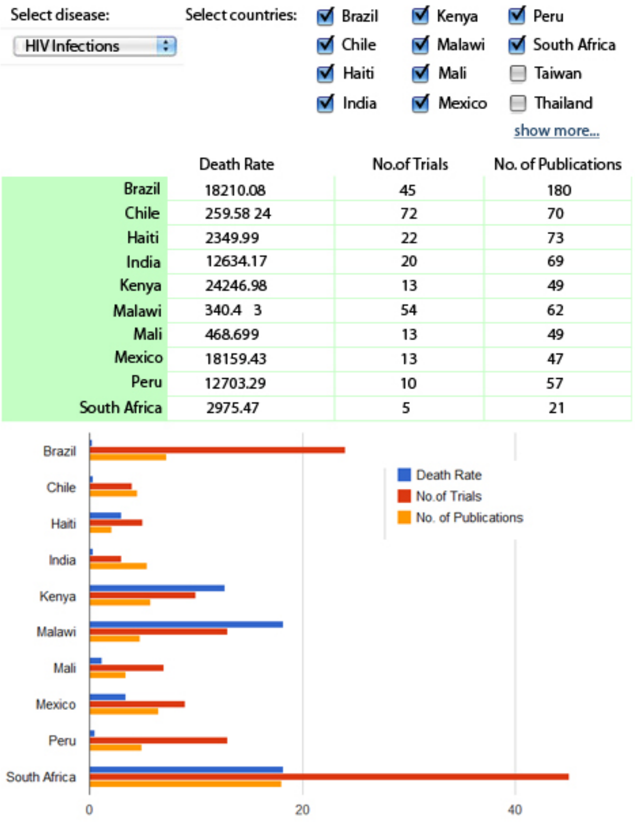
Results. Listing 1 shows the SPARQL query for retrieving the number of deaths and number of trials for Tuberculosis and AIDS. The queries for other indices and diseases look very similar. We selected the three diseases Tubercu-

[11] `http://who.int/healthinfo/global_burden_disease`

```
SELECT ?countryname ?diseasename ?value AS ?deaths
       count(?trial) AS ?number_of_trials
WHERE {
  ?item        a                 qb:Observation ;
               gho:Country       ?country ;
               gho:Disease       ?disease ;
               att:unitMeasure   gho:Measure ;
               eg:incidence      ?value .
  ?country     rdfs:label        ?countryname .
  ?disease     rdfs:label        ?diseasename .
  ?trial       a                 ct:trials ;
               ct:condition      ?condition ;
               ct:location       ?location .
  ?condition   owl:sameAs        ?disease .
  ?location    redd:locatedIn    ?country .
  FILTER (?diseasename IN
          ("Tuberculosis", "HIV/AIDS")) .
} GROUP BY ?countryname ?disedasename ?incidence
```

Listing 1: SPARQL query for retrieving the number of deaths and number of trials for Tuberculosis and HIV/AIDS in all countries.

losis, Malaria and Chronic Obstructive Pulmonary Disease (COPD), since they differ in many aspects and are known to be of high prevalence in emerging regions[12]. We calculated the four indices trials vs. death, trials vs. DALY, publications vs. death and publications vs. DALY for each of the diseases and selected six countries with sufficient available data and preferably with a Universal Health Care system (which ensures basic availability of treatment options). The results of the analysis are depicted in Figure 6. In general it can be observed, that there is a correlation between the indices comprising death and DALY as well as between the indices comprising trials and publications. However, there are exceptions from that general observation, such as in the case of Tuberculosis in India, which is over-resourced from the viewpoint of indices comprising publications, but under-resourced from the viewpoint of indices comprising clinical trials. A large gap between indicators comprising death rate and indicators comprising DALY (with one value below 100% and the other one above) would indicate, that it is difficult to balance between the two priorities longevity and quality-of-life. However, we still observed substantial discrepancies between both types of indices in some cases (e.g. Malaria in Colombia or COPD in South Africa), which might point towards a required re-allocation of research resources either towards life-saving or quality-of-life extending treatment options. An unexpected result of our analysis is, that the research resource allocation in developing countries is obviously biased towards certain diseases. In particular for Malaria, we can observe a high-prioritization of the research efforts in developing countries for this disease, although other similar prevalent diseases are less intensively tackled. A reason for this might be that medical research funders for developing countries expect treatment options for similar prevalent diseases which are also prevalent in developed countries to have already researched and developed by them. This confirms an impression commonly raised by life science researchers, but at the same time reveals that there might be a large potential for health-care improvements in emerging regions by diversifying health-care research portfolios. The results for the indices comprising publications (lower three diagrams) seem to be significantly affected by double-counting, i.e., articles that focus on more than one disease at the same time. Since our SPARQL queries did not weight for these factors (cf. Limitations Section below), these results should be evaluated with caution. In order to explore the results beyond the limited selection presented in this paper, we created a Web interface for the ReDD-Observatory, which is depicted in Figure 5.



Select disease: HIV Infections

Select countries: ☑ Brazil ☑ Kenya ☑ Peru ☑ Chile ☑ Malawi ☑ South Africa ☑ Haiti ☑ Mali □ Taiwan ☑ India ☑ Mexico □ Thailand

show more...

	Death Rate	No.of Trials	No. of Publications
Brazil	18210.08	45	180
Chile	259.58 24	72	70
Haiti	2349.99	22	73
India	12634.17	20	69
Kenya	24246.98	13	49
Malawi	340.4 3	54	62
Mali	468.699	13	49
Mexico	18159.43	13	47
Peru	12703.29	10	57
South Africa	2975.47	5	21

Figure 5: User interface of the ReDD-Observatory (available at `redd.aksw.org`) allowing users to visualize the disparity for particular diseases and regions as well as drill-down to the underlying data and publications.

7. CONCLUSIONS AND FUTURE WORK

Although, we are already able to obtain meaningful results, the work described in this article should be regarded as a first step towards a comprehensive, consistent, and real-time application for observation of the disparity between biomedical research efforts and disease disparity. During the work on the ReDD-Observatory, we encountered a number of obstacles, which were only partially solvable by ourselves. As major obstacles we identified the lack of *identifier reuse and interlinking*, *varying conceptualization* as well as deficiencies in *information quality and coverage*. By making these shortcomings explicit and measurable, we aim to trigger a development of incremental improvement, so that life science data integration efforts such as the ReDD-Observatory will deliver more precise and more extensive results over time. In the following we give account of the limitations of ReDD and outline planned future work.

[12] `http://who.int/mediacentre/factsheets/fs310/`

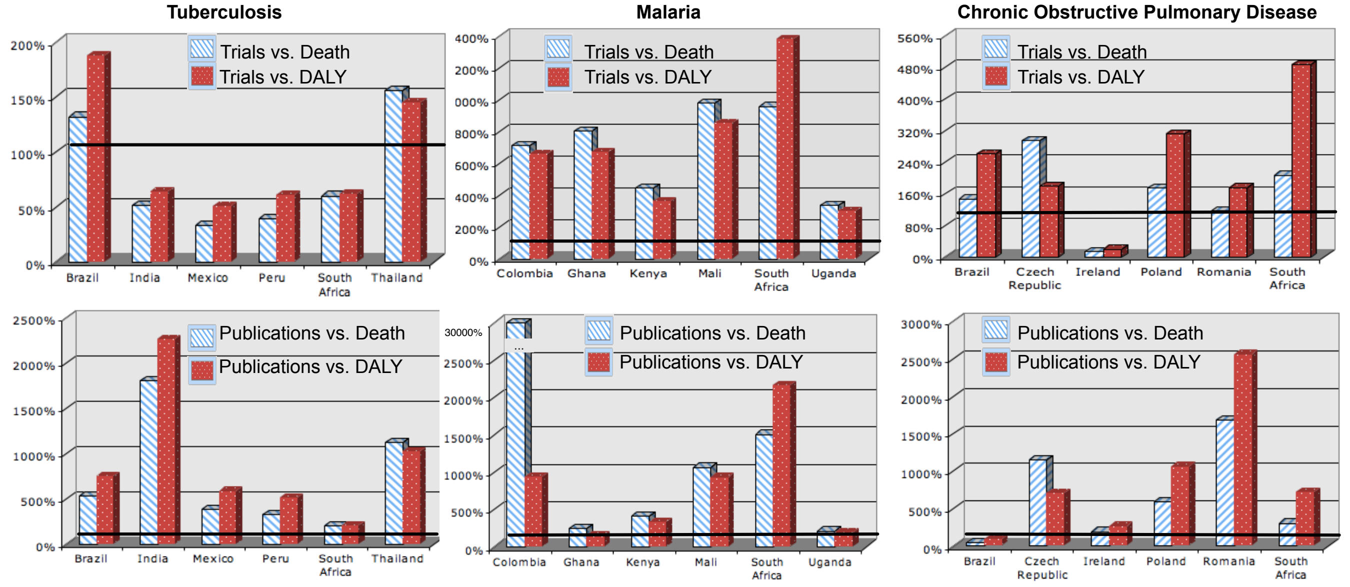


Figure 6: Depiction of the four ReDD indices for the diseases Tuberculosis, Malaria and COPD and selected countries. The black bar indicates the level of a balanced distribution of research resources.

Research indices completion and duplication. Although at this point ClinicalTrials.gov is the most complete clinical trial repository and, up to our knowledge, the only one directly available in RDF, it is not the only one. Other, less mature, trial repositories have been recently created in India, China, and other countries. Therefore, one of the future steps within this project is the incorporation of data from other repositories to increase the global coverage. Another limitation within the research indices is that since a single article or trial can cover more than one disease, but we currently assign credit to individual papers every time it describes a given disease, our indices are incurring in double counting in those cases. Although trials or manuscripts involving multiple diseases are not frequent, making this double counting negligible, in future version we intend to correct this index by creating queries that automatically tag individual trials or publications with the number of diseases addressed so that the index can be appropriately weighted.

Information Quality. Given the recent publication of most databases used in our study, their information presents a number of data quality issues. First, ClinicalTrials.gov reveals a number of issues as e.g. critical fields are not consistently reported during trial registration, including study contact, trial end date, and primary outcome. Second, PubMed only publishes and indices a subset of all trials conducted around the world. Even when published, the links to ClinicalTrials.gov might be missing. In addition, its RDF contains links to non-RDF URIs for most of its resources.

Coverage. Despite the exponential growth of data available on the Web coverage is still a major issue. When integrating data and performing analysis on the integrated dataset, the coverage of the base data is important and subsequent querying and analyzing the integrated data has to take limited coverage into account. In our case, the GHO data currently reports the statistics for death and DALY measures only till the year 2004, thus limiting the overall coverage of our integrated substrate. However, an updated release is scheduled for this year.

Interlinking Quality. The number of interlinks which could be automatically established was limited (see Table 2), as the datasets do not contain standardized identifiers for naming diseases, countries etc. For example, “AIDS” in LinkedCT could not be matched with “Acquired Immunodeficiency Syndrome” in PubMed using basic string similarity. Although we have attempted to address this limitation through the use of UMLS for interlinking as described in Section 5, there are limitations to this approach. Also, there was a smaller number of diseases present in the GHO dataset as compared to LinkedCT and PubMed (due to different levels of specificity), thus further complicating the automatic linking.

Error propagation. Given the above-mentioned issues with LinkedCT and PubMed, the interlinking of the overall dataset was compromised. This limitation was increased by the limited data coverage present of the GHO data while calculating the results. This amounts to the propagation of errors in the interlinked datasets and thus affects the final results.

These different limitations of a semi-automatic information integration approach as described in this paper cannot be addressed by one single party, but should be taken into account during the creation and publishing of data sets and when releasing new versions. In particular the use of shared identifiers (i.e. IRIs) would improve data integration.

Future Work. Future work will primarily tackle the mitigation of the above mentioned limitations. In particular, we will investigate ways to streamline the linking of biomedical data using the UMLS thesaurus. Specifically, given that UMLS is currently offered by the National Library of

Medicine under a service-oriented architecture, mechanisms should be designed to speed the conversion of large data sets such as article references in PubMed. By employing active machine learning techniques where the linking is improved based on user supplied examples, we aim at simplifying the linking process for end-users while at the same time improving precision and recall. In order to improve comprehensibility of aggregated and statistical information (e.g. represented using the DataCube vocabulary), we will create customizable views. To present massive statistical data to humans as comprehensible as possible, we will develop an ReDD dashboard as an OntoWiki extension that generates adequate charts for visualization. Since not all chart types are reasonable for special combinations of statistical attributes, the user will be given the facility to select the desired statistical attributes and the chart type such as a pie chart or a bar chart etc. If the statistical data sets also comprises a spatial dimension, aggregates of the information space will be presented on maps. Also, we will integrate further relevant biomedical datasets to the ReDD-Observatory so as to increase coverage as well as to provide additional views and measures to evaluate the disparity.

Acknowledgment

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 257943 for the project LOD2 (`http://lod2.eu`).

8. REFERENCES

[1] AHRQ. 2009 national healthcare quality and disparities reports. Technical report, Agency for Healthcare Research and Quality, 2009.

[2] S. Auer, S. Dietzold, and T. Riechert. Ontowiki - a tool for social, semantic collaboration. In *ISWC 2006*, volume 4273 of *LNCS*. Springer, 2006.

[3] E. Bertino, A. Maurino, and M. Scannapieco. *Data Quality in the Internet Era*, volume 14, chapter 4, pages 11 – 13. IEEE Computer Society, July 2010.

[4] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. DBpedia - a crystallization point for the web of data. *Journal of Web Semantics*, 7(3):154–165, 9 2009.

[5] J. Bleiholder and F. Naumann. Data fusion. *ACM Comput. Surv.*, 41(1):1–41, 2008.

[6] A. J. Bonito, C. R. Eicheldinger, and N. F. Lenfestey. Health disparities: Measuring health care use and access for racial/ethnic populations. Technical report, RTI International, 2005.

[7] G. F. A. Cary P. Gross and N. R. Powe. The relation between funding by the ntaional institute of health and the burden of diseases. *The New England Journal of Medicine*, 340(1881-1887), 1999.

[8] A. Chatterjee and A. Segev. Data manipulation in heterogeneous databases. *SIGMOD Rec.*, 20(4):64–68, 1991.

[9] D. Cook and D. F. Swayne. *Interactive and dynamic graphics for data analysis.* Springer, 2007.

[10] B. Crossette. Disparities in health: Inequities create great risks for poor, adolescents, and women in developing countries. Technical report, Disease Control Priorities Project, August 2006.

[11] A. Culotta and A. McCallum. Joint deduplication of multiple record types in relational data. In *CIKM '05*.

[12] R. Cyganiak, D. Reynolds, and J. Tennison. The rdf data cube vocabulary. Technical report, http://publishing-statistical-data.googlecode.com /svn/trunk/specs/src/main/html/cube.html, 8 2010.

[13] C. Dye and K. Floyd. *Disease Control Priorities in Developing Countries, 2nd edition*, chapter 16. World Bank, 2006.

[14] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19:1–16, 2007.

[15] I. P. Fellegi and A. B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, December 1969.

[16] D. T. Jamison, J. G. Breman, A. R. Measham, G. Alleyne, M. Claeson, D. B. Evans, P. Jha, A. Mills, and P. Musgrove, editors. *Disease Control Priorities in Developing Countries, 2nd edition*, chapter 4. World Bank, 2006.

[17] Knoesis. *Using Web Data Provenance for Quality Assessment*, 2009.

[18] P. S. Lim E P, Srivastava J and R. J. Entity identification in database integration. In *9th Int. Conf. on Data Engineering*, 1993.

[19] B. J. Naumann F., Bilke A. and W. M. Data fusion in three steps: Resolving schema, tuple and value inconsistencies. *IEEE Data Eng. Bull.*, 29(2):21–31, 2006.

[20] S. J. Nelson, T. Powell, Humphreys, and B. L. *The Unified Medical Language System (UMLS) Project*, pages 369–378. New York: Marcel Dekker, Inc., 2002.

[21] Public Health Information Development Unit. *The value of linked data for policy development, strategic planning, clinical practice and public health: An international perspective.* PHIDU, 2003.

[22] J. S. Ross, G. K. Mulvey, E. M. Hines, S. E. Nissen, and H. M. Krumholz. Trial publication after registration in clinicaltrials.gov: A cross-sectional analaysis. *PLoS Med*, 6(9), September 2009.

[23] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser. Assessing information quality of a community-based encyclopedia. In *Int. Conf. on Information Quality*, pages 442 – 454, Cambridge, MA, 2005.

[24] A. Thor. *Automatische Mapping-Verarbeitung auf Webdaten.* PhD thesis, Universität Leipzig, 2008.

[25] H. Varmus. Funding allocation for disease research. Statement, National Institutes of Health, May 1999.

[26] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov. Discovering and maintaining links on the web of data. In *ISWC*, 2009.

[27] WHO. The global burden of disease 2004 update. Technical report, World Health Organization, 2004.

[28] J. Wisdom. Research problems in data warehousing. In *CIKM '95*, pages 25 – 30. ACM.

[29] P. Ziegler and K. R. Dittrich. Three decades of data integration - all problems solved? In *In 18th IFIP World Computer Congress (WCC), Building the Information Society*, volume 12, pages 3 – 12, 2004.