ELSEVIER

A case study: semantic integration of gene–disease associations for type 2 diabetes mellitus from literature and biomedical data resources

Dietrich Rebholz-Schuhmann[1,2], Christoph Grabmüller[1], Silvestras Kavaliauskas[1], Samuel Croset[1], Peter Woollard[3], Rolf Backofen[4], Wendy Filsell[5] and Dominic Clark[1]

[1] European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK
[2] Computerlinguistik, Universität Zürich, Binzmühlestrasse 14, 8050 Zürich, Switzerland
[3] GlaxoSmithKline, GlaxoSmithKline Medicines Research Centre, Gunnels Wood Road, Stevenage SG1 2NY, UK
[4] Albert-Ludwigs-University Freiburg, Fahnenbergplatz, D-79085 Freiburg, Germany
[5] Unilever R&D, Colworth Science Park, Sharnbrook MK44 1LQ, UK

In the Semantic Enrichment of the Scientific Literature (SESL) project, researchers from academia and from life science and publishing companies collaborated in a pre-competitive way to integrate and share information for type 2 diabetes mellitus (T2DM) in adults. This case study exposes benefits from semantic interoperability after integrating the scientific literature with biomedical data resources, such as UniProt Knowledgebase (UniProtKB) and the Gene Expression Atlas (GXA). We annotated scientific documents in a standardized way, by applying public terminological resources for diseases and proteins, and other text-mining approaches. Eventually, we compared the genetic causes of T2DM across the data resources to demonstrate the benefits from the SESL triple store. Our solution enables publishers to distribute their content with little overhead into remote data infrastructures, such as into any Virtual Knowledge Broker.

Type 2 diabetes mellitus (T2DM) is a disease with unresolved questions

The genetic causes of diabetes are still not fully understood, although different types of diabetes can be distinguished, including neonatal diabetes (transient and permanent), noninsulin-dependent, maturity-onset diabetes of the young (MODY) and T2DM in adults [1]. Several genes are under investigation for their involvement in the development of this disease [2–5].

In the case of neonatal diabetes, the causes can be found in modifications of the insulin gene [6,7] as well as in other molecular defects (e.g. involving transcriptional and translational factors). For MODY, only six genes account for 80% of the disease development; using selected clinical traits, it is possible to distinguish eight genetic subgroups of MODY [1,8]. By contrast, all loci associated with the risk of diabetes explain no more than 1% of the risk variance and, for most loci, there is a lack of clues about their function in diabetes pathogenesis [9,10]. In addition, only specific phenotypes in diabetes, such as insulin resistance and β cell dysfunctions, indicate heritability [11]. Finally, the genetic parameters only marginally improve the prediction of the disease risk and only if they have been added to the phenotypic factors in the analysis [9].

Furthermore, diabetes is linked to other diseases, such as obesity, which has its own genetic preconditions [12,13]. In addition, the risk for T2DM increases under the genetic predisposition for obesity. In recent years, even patients with type 1 diabetes mellitus (T1DM) have developed obesity, leading to an obscured border between T1DM and T2DM [14]. Therefore, clinical symptoms and genetic criteria have to be reassessed for improved diagnostics and treatments possibly leading to novel drugs [3,14–16].

Trying to determine the causes of T2DM (e.g. insulin resistance in comparison to β cell dysfunction) leads to novel hypotheses for improved disease treatment. For this goal, pharmaceutical companies have to bring together their experts from different disciplines, such as molecular biologists, medicinal chemists and

Corresponding author:. Rebholz-Schuhmann, D. (rebholz@ebi.ac.uk), (d.rebholz.schuhmann@gmail.com)

1359-6446/06/$ - see front matter © 2013 Published by Elsevier Ltd. http://dx.doi.org/10.1016/j.drudis.2013.10.024

Please cite this article in press as: Rebholz-Schuhmann, D. et al., A case study: semantic integration of gene–disease associations for type 2 diabetes mellitus from literature and biomedical data resources, Drug Discov Today (2013), http://dx.doi.org/10.1016/j.drudis.2013.10.024

toxicologists, to make use of existing data from the public domain and from local repositories and, thus, to improve their productivity. Indeed, the information from the gene to the phenotype must be readily accessible in an interoperable way to explore any complex disease fully [17–19].

However, data repositories are often focused on one type of entity only (e.g. proteins in UniProt Knowledgebase (UniProtKB)) or possibly two (e.g. drugs and their targets in Drugbank) [20,21]. A comprehensive information system for complex biomedical problems would require the integration of facts while, at the same time, considering large numbers of entities, as well as relevant data resources, data providers, or heterogeneous data from the scientific literature, to serve the research community efficiently.

Exploring the genetic causes and the pathogenesis of T2DM requires combining different data resources, such as functional annotations of proteins in UniProtKB as well as gene–disease associations (GDAs) from Online Mendelian in Men (OMIM), which are provided from specific tables or from the scientific literature, respectively [21]. This integration generates unnecessary extra work, since the data resources do not comply with standardized, transparent, or interoperable data formats [22,23]. Above all, facts from scientific manuscripts are still kept in monolithic electronic documents. A few attempts have been made to standardize and enrich such documents, but the facts are not yet delivered as structured data or as linked data into any public repository [24–27]. In particular, the use of data standards for scalable data resources (e.g. semantic web technologies, nanopublications and the Virtual Knowledge Broker) would improve the accessibility of information from the scientific literature significantly [28,29].

Sharing biomedical data with semantic web technologies

Semantic web technologies form a framework for public data exchange and data sharing, and serve as an alternative to proprietary relational databases; strong semantic support is part of the infrastructure. It enables the deliver of evidence from the literature as information pieces into publicly available biomedical data resources [30,31]. Following the 'Linked Data Principles' for the semantic web according to Berners-Lee, the first requirement is the use of universal resource identifiers (URIs) to label things or entities (e.g. for a protein or chemical entity, and also when they appear in the scientific manuscript) [21,32,33].

The next requirement is to combine names with web addresses (i.e. 'http://URIs'), which should lead to useful information in readily available representation standards. For scientific manuscripts, only the metadata of the documents has been exploited up to now, in contrast to the use of the scientific content itself [34–36]. Finally, links to further URIs should be provided to enable discovery (i.e. the entities, or 'things' in the document should be linked through URIs to publicly available biomedical data resources), so-called 'Semantic Enrichment'. Ideally, all entities from the document can be linked to a public data resource and the facts can be verified against the content from biomedical databases.

The Resource Description Framework (RDF) is a semantic web standard for access to public data. Each fact or statement is represented as a triple comprising a subject (e.g. 'P53'), a predicate (e.g. 'is-a') and an object (e.g. 'protein'), and, at best, all three parts are specified uniquely using web-based resources. RDF data triples enable semantic interoperability between resources and provide considerable advantages [37–39] including: (i) consistent reuse of content across distributed resources with well-specified concepts and relations [40,41]; (ii) better error handling through standardized and transparent data representations [28] and (iii) large-scale and seamless exploitation of data through the simplicity and generality of the data representation. Eventually, open standards also enable publishing companies to take part in the data integration and data distribution activities [24,42,43].

Sharing biomedical data in the semantic web

Integrating the literature with public data repositories, such as UniProtKB, requires that its content is structured in a formalized and standardized way: entities (e.g. *p53* gene) and concepts (e.g. transcription regulation activity) from the text have to be referenced through terminologies, ontologies and fact repositories to achieve interoperability [44–47]. For this goal, the Foundry for Open Biomedical Ontologies (OBO) determined principles enabling scientific data resources to communicate with minimum uncertainty (e.g. without ambiguity): for example, the Human Phenotype Ontology (HPO) uses cross-references to available ontologies, such as Gene Ontology (GO), Chemical Entities of Biological Interest (ChEBI), Foundational Model of Anatomy (FMA) and so on, to define entities logically [33,48–50].

The data repositories should link their data entries in a readable form to relevant information for interactive use [51,52]. In the biomedical domain, this has been achieved by assigning metadata information to experimental data, thus improving information retrieval: for example, transforming table data into a representation using triples integration of health-related data (e.g. in Chinese medicine or for the modelling of neurological receptors) [41,53–57]. For translational medicine in Alzheimer's disease, a fact repository of 350 million triples have been built using ontologies and their domain knowledge (in OWL) in a structured way using well-defined concepts [58,59]. Similarly, selected pathway repositories have been integrated, including Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome and BioCyc, together with EntrezGene [60–65].

Current solutions integrating the literature with semantic web technologies make use of metadata only, once it has been provided by the author or extracted from the textual content [36,66,67]. The existing solutions demonstrate the potential of semantic web technologies for the integration of data and services. By contrast, the proposed semantic web solutions do not sufficiently integrate facts from the scientific literature or demonstrate the requirements necessary for this goal.

Decomposing the biomedical scientific literature

Members from academia, life science and publishing companies have worked together in the Semantic Enrichment of the Scientific Literature (SESL) project to integrate public and proprietary data using semantic web technologies. The project has produced technical achievements in biomedical semantics in addition to explaining the wider perspective [68].

In total 638,088 scientific publications were contributed by the publishing companies involved (Elsevier, Nature Publishing Group, Oxford University Press and Royal Society of Chemistry). A further

Please cite this article in press as: Rebholz-Schuhmann, D. et al., A case study: semantic integration of gene–disease associations for type 2 diabetes mellitus from literature and biomedical data resources, Drug Discov Today (2013), http://dx.doi.org/10.1016/j.drudis.2013.10.024

232,665 documents were accessible from Biomed Central and from the Europe PMC (EUPMC) distribution [69]. In total, 20,168 publications (from 870,753) contained information relevant to T2DM (i.e. those documents provided a reference to a GDA for T2DM).

All documents were processed with the same routines: sentencization, tokenization and entity recognition [70]. The identification of the genes and protein names was based on a large terminological resource (LexEBI) in combination with basic disambiguation [71,72]. The identification of diseases again relied on standard terminological resources (Unified Medical Language System® [UMLS]). The text and the entities were annotated using the IeXML format [73]. The annotation services are available from service infrastructure of EBI [74].

All sentences that contained a pair comprising a gene and a disease were identified and integrated into the SESL triple store. The triple store keeps the sentence, the provenance of the data (i.e. the reference to the paragraph and the document that contains the statement) as well as the DOI and the reference data to the publication.

The UniProtKB data repository is already available as a triple store, but has been reduced to the content covering human proteins only. The subselection makes reference to 20,272 proteins and contains 100,723 triples linking a gene to a functional annotation based on GO terms. Furthermore, 13,897 protein–protein interactions are represented as 111,176 triples, and 120,224 DOIs are covered by this portion of the triple store.

The integration of the Gene Expression Atlas (GXA) required the export of the experiments through the provided Java application programming interface (API). Different serializations in XML and JSON exported 138 experiments. The annotations of the experiments are based on the Experimental Factor Ontology (EFO), which contains concepts from a range of semantic resources, such as the ontology Foundational Model of Anatomy [50,75,76]. The disease annotations from EFO used in the GXA data repository had to be normalized to the Disease Ontology (DO) to use the existing mappings to UMLS [77]. Several entries in the GXA contained explicit mentions of the disease name instead of the EFO identifier, which again required normalization to the standard terminological resources. As alternative, we tested all proteins that have a location in an organ involved in diabetes mellitus, or the selection of genes to a predefined species. Both approaches only slightly increased the retrieval results.

The metafile for distributing the GDA was imported from OMIM in contrast to processing the narratives of OMIM, and other data resources for the same data (e.g. the Bio2Rdf distribution) were not available. The disease representation was again normalized to UMLS to be semantically compliant with the other data.

Open data standards lead the way to interoperability

The SESL prototype relies on the RDF (rdf:type, rdfs:label) in N3 notation, on the web Ontology Language (OWL; owl:sameAs) for data representation, and on SPARQL as the query language [78].[a] Further data standards are provided from Dublin Core and from the Simple Knowledge Organization System. Only the data from the scientific manuscripts required defining novel types of relation.

For the textual annotation, the IeXML format was used that enables annotations of different kinds of entities and concepts, and enables the referencing any type of data resource. IeXML has been successfully integrated by a large number of text-mining groups to distribute and share annotations [79].

The scientific literature was processed using standard terminological resources distributed through LexEBI [72]. The LexEBI terminological resource makes reference to 1,178,659 clusters or unique concept ids from public resources, 3,848,775 terms, and 2,665,753 unique terms. The terminological resource can serve two different purposes: (i) mining the entities from the scientific literature and (ii) linking the entities to reference data resources.

Additional conceptual or semantic data resources contribute to the standardization of literature content by providing the identifier and the label for the type or concept: for example, the DO for disease concepts and the NCBI taxonomy for species mentions. In particular, DO was necessary to align, cross-link and reference diseases across resources (i.e. between OMIM and GXA). UniProtKB supplies annotations for proteins (and genes), whereas GXA provides metadata information for genes and the experimental conditions leading to the modified expression levels of those genes. UniProtKB is publicly distributed as a triple store in addition to other data formats and was reduced to the human-relevant content only.

All content was grouped according to the original data resources and contained in four separate triple store repositories: (i) literature; (ii) GXA; (iii) UniProtKB and (iv) OMIM. Combining the data in a single triple store delivered the same retrieval results, but showed lower retrieval performance. The triple stores were implemented based on Jena TDB using the N triple notation.

The normalization of the data content from the different resources enabled retrieval of the data from one repository and established a Virtual Knowledge Broker [68]. Alternative approaches would be a federated database or a data mart instance, but would be less suitable for the distributed content of the SESL prototype [80,81]. Access to the triple store was achieved using SOAP and REST web services, as well as SPARQL queries and a graphical user interface (GUI) for browsing of results (http://www.pistoia-sesl.org).

Retrieval of distributed biomedical data

The triple store has been tested under the following conditions: (i) distribution of the content to separate compute engines; (ii) distribution of redundant content to separate compute engines and (iii) data integration of content from the triple store with external semantic data resources, such as Wikipedia. None of the approaches led to impairments of the functionality of the triple store, and the last task demonstrated the interoperability of the triple store technology for data integration with external data resources. Thus, we suggest that virtual knowledge brokering is ready to deliver content from public and proprietary sources and from disparate locations through a shared infrastructure: the Virtual Knowledge Broker [68].

The retrieval speed was correlated to the amount of content in the data resource and the amount of retrieved results. The response time was considerably longer (i.e. performance was reduced) if a query combined data from different resources (i.e. for complex and

[a] http://www.openrdf.org/; http://www.w3.org/designissues/notation3.html; http://www.w3.org/tr/rdf-sparql-protocol; http://www.w3.org/rdf/ and http://www.w3.org/tr/rdf-sparql-query/.

Please cite this article in press as: Rebholz-Schuhmann, D. et al., A case study: semantic integration of gene–disease associations for type 2 diabetes mellitus from literature and biomedical data resources, Drug Discov Today (2013), http://dx.doi.org/10.1016/j.drudis.2013.10.024

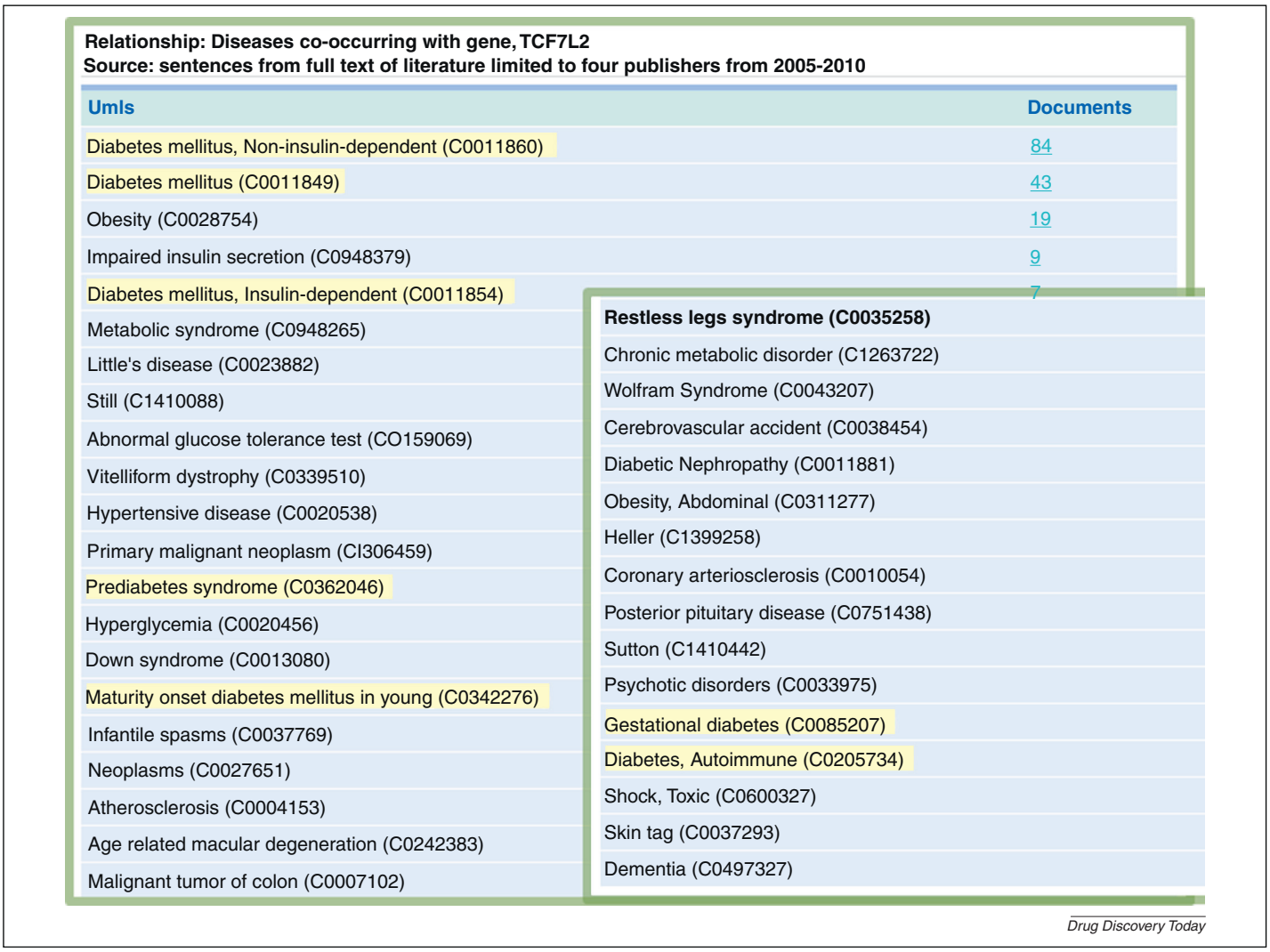

FIGURE 1

According to the literature analysis, several genes can be found in the context of type 2 diabetes mellitus (T2DM) in addition to Transcription factor 7-like 2 (*TCF7L2*). Peroxisome proliferator-activated receptor gamma (*PPARG*) and Hepatocyte nuclear factor 1-alpha (*HNF1A*) have already been confirmed by genome-wide association studies. Other genes are also endorsed by the literature analysis.

extensive queries). Thus, queries for interactive tasks through the GUI had to be optimized (as 'materialized views') to serve best the most common tasks taking the underlying data into consideration. In particular, the annotation of gene names from UniProtKB was enriched with content from OMIM and GXA.

The SESL GUI accepts gene and disease keywords, and produces concepts appropriate to the query term ('auto-completion'). The querying for T2DM led to the retrieval of genes that have been prioritized for the number of documents that make reference to them (Fig. 1), for example Transcription factor 7-like 2 (*TCF7L2*), Peroxisome proliferator-activated receptor gamma (*PPARG*), Hepatocyte nuclear factor 4-alpha (*HNF4A*) and Hepatocyte nuclear factor 1-alpha (*HNF1A*). For Transcription factor 7-like 2 (TCF7L) (a T2DM candidate gene), the retrieval delivers a list of diseases, which is not restricted to T2DM, because a variety of diseases is encountered in the context of the queried gene. For example, 'Malignant tumour of colon' is known to be associated with TCF7L2. The list of GO annotations for the gene includes 'activation of insulin secretion – GO 0032024', 'β catenin binding – GO 0008013' and others.

Towards hypothesis generation for T2DM

The integration of data was focused on T2DM only. The selection of relevant data was linked to the mention of T2DM in the primary bioinformatics data resources and in the scientific literature. This approach required the mapping of disease mentions to standardized terminological resources, but also restricted the disease-relevant data to the pre-selection from the primary resources: that is, literature as well as the database content.

Querying facts as triples

The SESL prototype enables cross-evaluation of data resources in terms of the amount of data and their compatibility. In total, 490,228 manuscripts made reference to at least one gene or protein, 938,081 sentences contained at least one disease reference and 118,868 sentences had both entities. In total, 2075 genes co-occurred with diabetes in text, and were mentioned in at least one experiment in the GXA (Table 1).

According to OMIM, 2707 GDAs in total were confirmed and 2032 were monogenetic diseases (i.e. one gene has been reported to be relevant for the given disease). In addition, 306 GDAs in

Please cite this article in press as: Rebholz-Schuhmann, D. et al., A case study: semantic integration of gene–disease associations for type 2 diabetes mellitus from literature and biomedical data resources, Drug Discov Today (2013), http://dx.doi.org/10.1016/j.drudis.2013.10.024

TABLE 1

The total number of triples from different resources contained in the SESL triple store[a]

	Public and proprietary data	(%)	Public data only	(%)
ArrayExpress	182,840	0.5%	182,840	0.7%
EFO	49,026	0.1%	49,026	0.2%
UMLS, homebrew	6,906,735	18.8%	6,906,735	26.5%
DO	1,863,664	5.1%	1,863,664	7.2%
GO	495,595	1.3%	495,595	1.9%
UniProt filtered for Human	12,552,239	34.1%	12,552,239	48.2%
Triples on metadata from full-text literature	3,485,212	9.5%	1,949,293	7.5%
Triples with gene annotation from full-text literature	2,373,584	6.5%	300,773	1.2%
Triples with disease annotation from full-text literature	4,983,788	13.6%	662,824	2.5%
Triples with GO annotation full-text literature	3,870,834	10.5%	1,099,410	4.2%
Total number of triples	36,763,517		26,062,399	
Total number of public triples	14,713,418	40.0%	4,012,300	15.4%

[a] The largest portion has been contributed from UniProt, UMLS and the scientific full-text literature. 'Public' refers to those resources that are accessible without any license or access restrictions.

MorbidMap (OMIM) were confirmed through facts (triples) extracted from the scientific literature, and each fact was on average supported by 6.36 statements from the text. Most GDAs were not related to T2DM, because other diseases had also been detected in the manuscripts.

UniProtKB makes reference to 20,272 unique human gene entries. It also provides 7598 distinct GO concepts leading to 100,599 GO annotations of human genes in the Triple Store, and 13,897 interaction annotations. The data have now been integrated into the SESL triple store.

From the literature, 137,216 GO annotations were identifed in the context of 6410 genes: 21.4 GO annotations for a given gene on average (Table 1). For the same 6410 genes, 99,260 GO annotations were identified in the GOA database, which resulted in an average of 15.5 GO annotations per gene. From both sets of GO annotations, 6788 GO annotations of genes could be filtered out that were shared between both resources for the same gene (average 1.1). These analyses demonstrate that the scientific literature contributes information that could be relevant for the interpretation of the causes of the disease linked to the genes under scrutiny.

The processing of the GXA data repository led to the retrieval of 138 experiments that made reference to 36,568 genes and provided expression levels for 15,135 distinct genes. All experiments were annotated with a total of 183 unique EFO annotations. Comparing the overexpressed genes from the 138 experiments in GXA with the gene–disease pairs in OMIM led to the retrieval of zero associations, which is not necessarily surprising, given that the experiments in GXA are not primarily meant to confirm known GDAs from OMIM (Fig. 2).

By retrieving the GDAs that are linked to diabetes (C0011849), a list was generated of 561 associated genes according to GXA and of 2121 genes from the scientific literature, with an overlap of 12 (Fig. 2 and Table 3). None of the genes has direct links to either insulin signalling or to glucose metabolism and, thus, form a core set of highly relevant candidate genes.

We also analyzed whether a gene had already been mentioned in any of the previously mentioned review articles [3–5]. Several genes have been extracted from the literature and confirmed from the review articles, but have not yet been included in the public data resources, showing that the scientific literature provides relevant underexposed data (Table 2).

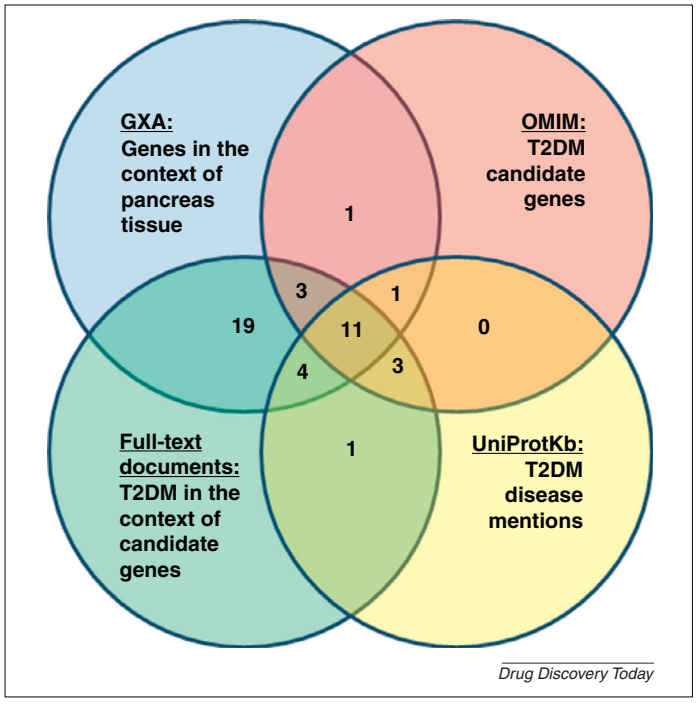

FIGURE 2

An overview of the gene–disease associations (GDAs) that were shared between the different resources. In total, ten GDAs were known to all four resources [Online Mendelian in Men (OMIM), Gene Expression Atlas (GXA), UniProt Knowledgebase and the scientific literature]. The most candidate genes were shared between GXA and the scientific literature, demonstrating the potential to explore the integration of data resources. All candidate genes that were shared between OMIM and UniProtKB could also be confirmed from a third resource (i.e. from the literature or GXA), although each resource also contributed a single candidate gene that was confirmed from a third resource.

Please cite this article in press as: Rebholz-Schuhmann, D. et al., A case study: semantic integration of gene–disease associations for type 2 diabetes mellitus from literature and biomedical data resources, Drug Discov Today (2013), http://dx.doi.org/10.1016/j.drudis.2013.10.024

TABLE 2

The table gives an overview on the genes found in the different resources. For all genes at least one reporting is provided from the SESL literature content in the context of T2DM. The resources Omim and UniProtKb from the columns and the rows indicate whether a gene has been found in GXA or in one of the review articles (see text)

		OIMIM (+) UniPro (+)	OIMIM (+) UniProt (−)	OMIM (−) UniPro (+)	OMIM (−) UniProt (−)
Review (+)	GXA (+)	ABCC8, CAPN10, HNF1A, HNF1B (TCF2), HNF4A, INSR, NeuroD1, PPARG, TCF7L2	WFS1	IRS1, PDX1	HHEX, JAZF1
Review (+)	GXA (−)	GCK, KCNJ11	IGF2BP2		
Review (−)	GXA (+)	MAPK8IP1, PAX4	LIPC, PTPN1	GBP28 (ADIPOQ), PPP1R3A	ACVR2A, ADCP2 (DPP4), ARCN1, FFAR1, GCG, GLP1R, IAPP, IDE, IL1B, MAP4K2, NEFA (NUCB2), NIF3 (CTDSP1), NOS3, PGC1A (PPARGC1A), PPARA, RBP4, UCP2
Review (−)	GXA (−)	SLC2A4	IL6, RETN	INS	

TABLE 3

An overview of the most T2DM relevant genes according to the SESL prototype

Gene symbol	Protein encoded by gene	Number of documents
PPARA	Nuclear receptor subfamily 1 group C member 1	325
GBP28	Adipocyte complement-related 30-kDa protein	227
GLP1R	Glucagon-like peptide 1 receptor	146
OB	Obese protein	127
GCG	GLP-1(7–37)	96
TCF7L2	T cell factor 4	84
PPARG	Peroxisome proliferator-activated receptor gamma	72
ADCP2	Dipeptidyl peptidase 4	58
IAPP	Amylin	57
INSR	Insulin receptor subunit β	51
KCNJ11	Potassium channel, inwardly rectifying subfamily J member 11	48
FIZZ3	Adipose tissue-specific secretory factor	47
PTP1B	PTP-1B	44
PLANH1	Serpin E1	43
NR2A1	Transcription factor 14	42
HNF1A	Hepatocyte nuclear factor 1-alpha	39
PGC1A	PGC1α	37
PRKACG	cAMP-dependent protein kinase catalytic subunit γ	35
NOS3	Endothelial nitric oxide synthase	34
DPP9	DPLP9	33
ACVR2A	Activin receptor type IIA	32
KIAA1845	Calcium-activated neutral proteinase 10	32
CTRP1	GIP	31
GLUT4	Glucose transporter type 4, insulin-responsive	31
IDE	Insulinase	31
TNF	Tumor necrosis factor	30
HNF1B	Variant hepatic nuclear factor 1	27
IL6	CTL differentiation factor	27
RBP4	PRBP	27
IRS1	IRS-1	26
PRH	Homeobox protein PRH	26
UNQ524/PRO1066	Ghrelin-28	26
ARCN1	Archain	25
RENBP	RnBP	25
IGF2BP2	IGF2 mRNA-binding protein 2	24
APOE	Apolipoprotein E	22
IL1B	Catabolin	22
SELENBP1	SBP56	22
ALT2	Glutamic–alanine transaminase 2	21
GFR	Guanine nucleotide exchange factor for Rap1	21
LEPR	Leptin receptor	20
NAMPT	Visfatin	19

Unified data sharing and integration

The objective of the SESL project was the integration of data resources for T2DM (i.e. proprietary and public resources) into a public web-based infrastructure for biomedical researchers. In particular, the integration of the scientific literature with biomedical data resources was a primary achievement. Semantic web technologies support data sharing and data integration based on semantic resources [28,37]. In particular, the use of openly available terminological resources that make reference to public data repositories should improve the integration of literature and data repositories, leading to innovative data infrastructures, such as the Virtual Knowledge Broker [68,72].

The first objective was the harmonization of data resources across different repositories. For the scientific literature, the content is processed by openly available text-mining solutions. After the extraction step, the data were transformed into the RDF representation [35,36,46,67,75,82–84]. The next objective was

Please cite this article in press as: Rebholz-Schuhmann, D. et al., A case study: semantic integration of gene–disease associations for type 2 diabetes mellitus from literature and biomedical data resources, Drug Discov Today (2013), http://dx.doi.org/10.1016/j.drudis.2013.10.024

not only the sharing of data using semantic web technologies as standards (i.e. RDF and SPARQL), but also integrating facts from public content, such as UniProtKB and OMIM.

The provision of services for data integration and data sharing to the public formed another objective [85–87]. Ideally, the available data standards enable an infrastructure where the literature and data resources can be delivered through different distribution channels. A particular open source broker solution can be instantiated as a public broker, either as a registry to web services or as a data-sharing point [86].

Altogether, the new public infrastructure demonstrates: (i) its relevance for an important topic (i.e. T2DM); (ii) the means for data access (i.e. SPARQL queries, REST or SOAP web services, and even a user interface); (iii) the integration of public and proprietary resources and (iv) the exploitation of literature content. This infrastructure instantiates a public resource for hypothesis generation and testing, including of the scientific literature, and for cross-validation of disparate public and proprietary data resources against a focal topic (i.e. the GDAs for T2DM). The overlapping IMI OpenPhacts project is now demonstrating similar capability in drug discovery fields, especially chemistry [42,68]. Although hypothesis validation is not fully supported, the SESL prototype gathers and exposes the evidence from primary protein domain data resources (i.e. GXA, UniProtKB and OMIM) and the scientific literature. As a benefit, the researcher receives timely access to the relevant genes, their functional annotations from the different resources, and possibly the relevance of selected genes for alternative diseases.

Acknowledgements

Special thanks to Samuel Croset for support in the transformation of literature content into a triple store representation. Misha Kapushesky is acknowledged for support in the integration of the gene expression atlas. Ian Harrow, Ian Stott, Nigel Wilkinson and Catherine Marshalls provided valuable project management and quality assurance support. Mike Westaway gave input on the optimization of the SESL triple store. The SESL project was funded by the Pistoia Alliance (http://www.pistoiaalliance.org). Particular thanks to the publishing houses, Reed Elsevier, Nature Publishing Group, Oxford University Press and the Royal Society for Chemistry, for providing literature content for the development of the SESL prototype.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at http://dx.doi.org/10.1016/j.drudis.2013.10.024.

References

1 McCarthy, M.I. and Hattersley, A.T. (2008) Learning from molecular genetics: novel insights arising from the definition of genes for monogenic and type 2 diabetes. *Diabetes* 57, 2889–2898
2 Kota, S.K. *et al.* (2012) Genetics of type 2 diabetes mellitus and other specific types of diabetes; its role in treatment modalities. *Diabetes Metab. Syndr.* 6, 54–58
3 Malandrino, N. and Smith, R.J. (2011) Personalized medicine in diabetes. *Clin. Chem.* 57, 231–240
4 Herder, C. and Roden, M. (2011) Genetics of type 2 diabetes: pathophysiologic and clinical relevance. *Eur. J. Clin. Invest.* 41, 679–692
5 McCarthy, M.I. (2004) Progress in defining the molecular basis of type 2 diabetes mellitus through susceptibility-gene identification. *Hum. Mol. Genet.* 13 (Spec. no. 1), 33–41
6 Stoy, J. *et al.* (2010) Clinical and molecular genetics of neonatal diabetes due to mutations in the insulin gene. *Rev. Endocr. Metab. Disord.* 11, 205–215
7 Mlinar, B. *et al.* (2007) Molecular mechanisms of insulin resistance and associated diseases. *Clin. Chim. Acta* 375, 20–35
8 Peltonen, L. *et al.* (2006) Lessons from studying monogenic disease for common disease. *Hum. Mol. Genet.* 15 (Spec. no. 1), 67–74
9 Meigs, J.B. (2009) Prediction of type 2 diabetes: the dawn of polygenetic testing for complex disease. *Diabetologia* 52, 568–570
10 McCarthy, M.I. (2009) Exploring the unknown: assumptions about allelic architecture and strategies for susceptibility variant discovery. *Genome Med.* 1, 66
11 Clee, S.M. and Attie, A.D. (2007) The genetic landscape of type 2 diabetes in mice. *Endocr. Rev.* 28, 48–83
12 Li, S. *et al.* (2011) Genetic predisposition to obesity leads to increased risk of type 2 diabetes. *Diabetologia* 54, 776–782
13 O'Rahilly, S. (2009) Human genetics illuminates the paths to metabolic disease. *Nature* 462, 307–314
14 Smith, R.J. *et al.* (2010) Individualizing therapies in type 2 diabetes mellitus based on patient characteristics: what we know and what we need to know. *J. Clin. Endocrinol. Metab.* 95, 1566–1574
15 Altshuler, D. *et al.* (2008) Genetic mapping in human disease. *Science* 322, 881–888
16 Bell, J. (2010) Redefining disease. *Clin. Med.* 10, 584–594
17 Hoehndorf, R. *et al.* (2011) PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic Acids Res.* 39, e119
18 Robinson, P.N. *et al.* (2008) The Human Phenotype Ontology: a tool for annotating and analysing human hereditary disease. *Am. J. Hum. Genet.* 83, 610–615
19 Kohler, S. *et al.* (2009) Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am. J. Hum. Genet.* 85, 457–464
20 Wishart, D.S. *et al.* (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 36, D901–D906
21 Apweiler, R. *et al.* (2011) On-going and future developments at the Universal Protein Resource. *Nucleic Acids Res.* 39, D214–D219
22 Amberger, J. *et al.* (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.* 37, D793–D796
23 Parkinson, H. *et al.* (2011) ArrayExpress update-an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res.* 39, D1002–D1004
24 Rebholz-Schuhmann, D. *et al.* (2012) Text-mining solutions for biomedical research: enabling integrative biology. *Nat. Rev. Genet.* 13, 829–839
25 Shotton, D. *et al.* (2009) Adventures in semantic publishing: exemplar semantic enhancements of a research article. *PLoS Comput. Biol.* 5, e1000361
26 Sansone, S.A. *et al.* (2012) Toward interoperable bioscience data. *Nat. Genet.* 44, 121–126
27 Mons, B. *et al.* (2011) The value of data. *Nat. Genet.* 43, 281–283
28 Samwald, M. and Stenzhorn, H. (2010) Establishing a distributed system for the simple representation and integration of diverse scientific assertions. *J. Biomed. Semantics* 1 (Suppl. 1), S5
29 Hassanzadeh, O. *et al.* (2012) Data management issues on the semantic web. *2012 IEEE 28th Int. Conf. Data Eng.* pp. 1204–1206
30 Antezana, E. *et al.* (2009) Biological knowledge management: the emerging role of the semantic web technologies. *Brief. Bioinform.* 10, 392–407
31 Reformat, M.Z. and Hossein Zadeh, P.D. (2012) Assimilation of information in RDF-based knowledge base. *Adv. Comput. Intell.* LCNS 7311, 191–200
32 Bizer, C. *et al.* (2009) Linked data – the story so far. *Int. J. Semantic Web Inf. Syst.* 5, 1–22
33 Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29
34 Belleau, F. *et al.* (2008) Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J. Biomed. Inform.* 41, 706–716
35 Witte, R. *et al.* (2007) Enhanced semantic access to the protein engineering literature using ontologies populated by text mining. *Int. J. Bioinform. Res. Appl.* 3, 389–413
36 Roos, M. *et al.* (2009) Structuring and extracting knowledge for the support of hypothesis generation in molecular biology. *Brief. Bioinform.* 10 (Suppl. 10), S9
37 Cheung, K. (2009) Semantic web for health care and life sciences: a review of the state of the art. *Brief. Bioinform.* 10, 111–113
38 Neumann, E. and Prusak, L. (2007) Knowledge networks in the age of the Semantic Web. *Brief. Bioinform.* 8, 141–149

Please cite this article in press as: Rebholz-Schuhmann, D. et al., A case study: semantic integration of gene–disease associations for type 2 diabetes mellitus from literature and biomedical data resources, Drug Discov Today (2013), http://dx.doi.org/10.1016/j.drudis.2013.10.024

39 Splendiani, A. *et al.* (2011) Knowledge sharing and collaboration in translational research, and the DC-THERA Directory. *Brief. Bioinform.* 12, 562–575
40 Machado, C. *et al.* (2013) The semantic web in translational medicine: current applications and future directions. *Brief. Bioinform.* bbt079
41 Hoehndorf, R. *et al.* (2011) A common layer of interoperability for biomedical ontologies based on OWL EL. *Bioinformatics* 27, 1001–1008
42 Cheung, K.H. *et al.* (2009) A journey to semantic web query federation in the life sciences. *BMC Bioinformatics* 10 (Suppl. 10), S10
43 Williams, *et al.* (2012) Open PHACTS: semantic interoperability for drug discovery. *Drug Discov. Today* 17, 1188–1198
44 Groth, *et al.* (2010) The anatomy of a nanopublications. *Inf. Services Use* 30, 51–56
45 Cannata, N. *et al.* (2008) A semantic web for bioinformatics: goals, tools, systems, applications. *BMC Bioinformatics* 9, S1
46 Courtot, M. *et al.* (2011) Controlled vocabularies and semantics in systems biology. *Mol. Syst. Biol.* 7, 543
47 Thompson, P. *et al.* (2011) The biolexicon: a large-scale terminological resource for biomedical text mining. *BMC Bioinformatics* 12, 397
48 Hettne, K.M. *et al.* (2010) Automatic vs. manual curation of a multi-source chemical dictionary: the impact on text mining. *J. Cheminform.* 2, 4
49 Degtyarenko, K. *et al.* (2009) ChEBI: an open bioinformatics and cheminformatics resource. *Curr. Protoc. Bioinformatics* http://dx.doi.org/10.1002/0471250953.bi1409s26
50 Smith, B. *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* 25, 1251–1255
51 Rosse, C. and Mejino, J.L. (2003) A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *J. Biomed. Inform.* 36, 478–500
52 Burgun, A. and Bodenreider, O. (2008) Accessing and integrating data and knowledge for biomedical research. *Yearb. Med. Inform.* 91–101
53 Rebholz-Schuhmann, D. and Nenadic, G. (2010) Biomedical semantics: the hub for biomedical research 2.0. *J. Biomed. Semantics* 1, 1
54 Cheung, K.H. and Chen, H. (2010) Semantic web for data harmonization in Chinese medicine. *Chin. Med.* 5, 2
55 Chen, H. *et al.* (2009) Semantic web for integrated network analysis in biomedicine. *Brief. Bioinform.* 10, 177–192
56 Cheung, K.H. *et al.* (2010) Structured digital tables on the semantic web: toward a structured digital literature. *Mol. Syst. Biol.* 6, 403
57 Mukherjea, S. (2005) Information retrieval and knowledge discovery utilising a biomedical Semantic Web. *Brief. Bioinform.* 6, 252–262
58 Chen, H. *et al.* (2013) Semantic Web meets Integrative Biology: a survey. *Brief. Bioinform.* 14, 109–125
59 Ruttenberg, A. *et al.* (2009) Life sciences on the Semantic Web: the Neurocommons and beyond. *Brief. Bioinform.* 10, 193–204
60 Ruttenberg, A. *et al.* (2007) Advancing translational research with the Semantic Web. *BMC Bioinformatics* 8 (Suppl. 3), S2
61 Sahoo, S.S. *et al.* (2008) An ontology-driven semantic mashup of gene and biological pathway information: application to the domain of nicotine dependence. *J. Biomed. Inform.* 41, 752–765
62 Joshi-Tope, G. *et al.* (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* 33, D428–D432
63 Ogata, H. *et al.* (1999) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 27, 29–34
64 Karp, P.D. *et al.* (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.* 33, 6083–6089
65 Maglott, D. *et al.* (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* 39, D52–D57
66 Casher, O. and Rzepa, H.S. (2006) SemanticEye: a semantic web application to rationalize and enhance chemical electronic publishing. *J. Chem. Inf. Model.* 46, 2396–2411
67 Smith, A. *et al.* (2007) Leveraging the structure of the Semantic Web to enhance information retrieval for proteomics. *Bioinformatics* 23, 3073–3079
68 Harrow, I. *et al.* (2013) Towards virtual knowledge broker services for semantic integration of life science literature and data sources. *Drug Discov. Today* 18, 428–434
69 McEntyre, J.R. *et al.* (2011) UKPMC: a full text article resource for the life sciences. *Nucleic Acids Res.* 39, D58–D65
70 Kirsch, H. *et al.* (2006) Distributed modules for text annotation and IE applied to the biomedical domain. *Int. J. Med. Inform.* 75, 496–500
71 Rebholz-Schuhmann, D. *et al.* (2007) EBIMed-text crunching to gather facts for proteins from Medline. *J. Bioinformatics* 23, e237–e244
72 Rebholz-Schuhmann, D. *et al.* (2013) Evaluation and cross-comparison of lexical entities of biological interest (LexEBI). *PLoS ONE* 8, e75185
73 Rebholz-Schuhmann, D. *et al.* (2008) IeXML: towards a framework for interoperability of text processing modules to improve annotation of semantic types in biomedical text. *ISMB SIG BioLink*, Fortaleza, Brazil, 2006
74 Rebholz-Schuhmann, D. *et al.* (2008) Text processing through web services: calling Whatizit. *J. Bioinformatics* 24, 296–298
75 Malone, J. *et al.* (2008) Developing an application focused experimental factor ontology: embracing the OBO community. In *Proceedings of ISMB 2008 SIG Meeting on BioOntologies*
76 Malone, J. *et al.* (2010) Modeling sample variables with an Experimental Factor Ontology. *J. Bioinformatics* 26, 1112–1118
77 Du, P. *et al.* (2009) From disease ontology to disease-ontology lite: statistical methods to adapt a general-purpose ontology for the test of gene-ontology associations. *J. Bioinformatics* 25, i63–i68
78 Makris, K. *et al.* (2012) SPARQL-RW: transparent query access over mapped RDF data sources. In *Proceedings of the 15th International Conference on Extending Database Technology* (Rundensteiner, E. *et al.* eds), In pp. 610–613, ACM
79 Rebholz-Schuhmann, D. *et al.* (2013) Evaluating gold standard corpora against gene/protein tagging solutions and lexical resources. *J. Biomed. Semantics* 4
80 Verona, G. *et al.* (2007) Innovation and virtual environments: towards virtual knowledge brokers. *Organ. Stud.* 27, 765–788
81 Farley, T. *et al.* (2013) The BioIntelligence Framework: a new computational platform for biomedical knowledge computing. *J. Am. Med. Inform. Assoc.* 20, 128–133
82 Cruz-Toledo, J. *et al.* (2010) RKB: a Semantic Web knowledge base for RNA. *J. Biomed. Semantics* 1 (Suppl. 1), S2
83 Pezik, P. *et al.* (2008) Static dictionary features for term polysemy identification. *Building & evaluating resources for biomedical text mining* European Language Resources Association (ELRA) http://www.lrec-conf.org/proceedings/lrec2008/
84 Rebholz-Schuhmann, D. *et al.* (2010) CALBC silver standard corpus. *J. Bioinformatics Comput. Biol.* 8, 163–179
85 Gessler, D.D. *et al.* (2009) SSWAP: a Simple Semantic Web Architecture and Protocol for semantic web services. *BMC Bioinformatics* 10, 309
86 Wilkinson, M.D. *et al.* (2011) The semantic automated discovery and integration (SADI) web service design-pattern, API and reference implementation. *J. Biomed. Semantics* 2
87 Wilkinson, M.D. *et al.* (2010) SADI, SHARE, and the in silico scientific method. *BMC Bioinformatics* 11 (Suppl. 12), S7

Please cite this article in press as: Rebholz-Schuhmann, D. et al., A case study: semantic integration of gene–disease associations for type 2 diabetes mellitus from literature and biomedical data resources, Drug Discov Today (2013), http://dx.doi.org/10.1016/j.drudis.2013.10.024