Semantic, Technical and Legal Interoperability of European Company Open Data in Practice: The STIRData Approach

Jakub Klímek[1][a], Alexandros Chortaras[2][b], Jakub Míšek[3][c], Jim J. Yang[4], Steinar Skagemo[5] and Vassilis Tzouvaras[2]

[1] *Charles University, Faculty of Mathematics and Physics, Department of Software Engineering, Czech Republic*
[2] *School of Electrical and Computer Engineering, National Technical University of Athens, Greece*
[3] *Institute of Law and Technology, Masaryk University, Czech Republic*
[4] *Norwegian Digitalisation Agency, Norway*
[5] *Brønnøysund Register Centre, Norway*

Keywords: Interoperability, Linked Data, Open Data, Data Spaces, High-Value Datasets (HVDs), Company Data.

Abstract: As part of the Open Data Directive, the European Commission has published a list of high-value datasets (HVDs) that public sector bodies must make available as open data. The list also contains specific data items that must be included in these datasets. However, it does not prescribe any technical means of how the data should be published, severely hindering the interoperability of the datasets once they are published. One of the HVD topics is company data. In this practice report paper, we present results of STIRData, a project co-financed by the Connecting Europe Facility Programme of the European Union, focusing on technical, semantic, and legal interoperability of open data from business registries, covering the company data HVDs topic. The results include a data architecture and a data specification to make the published data technically and semantically interoperable, and legal interoperability guidelines to ensure legal interoperability of the published data. Moreover, proof-of-concept transformations of data from selected European business registries are shown using open source tools and according to the specification. Finally, a user-orientated platform for browsing and analysing the data is presented as an example of the possibilities of using the data published in an interoperable way.

1 INTRODUCTION

As part of the Open Data Directive, the European Commission has published a list of high-value datasets (HVDs) that public sector bodies must make available as open data. For each of the HVDs, the list also contains specific data items that must be included in these datasets. At the same time, the European Commission is in the process of creating Common European Data Spaces, domain-specific ecosystems where data producers and consumers can exchange data, ideally in an interoperable way. In the case of HVDs, no technical guidance is given on how to publish these datasets. As a result, each publisher will publish their HVDs in their own way, making the result noninteroperable. The same problem will arise in the Common European Data Spaces, unless technical specifications are provided for each of the exchanged datasets.

STIRData,[1] a project co-financed by the Connecting Europe Facility Programme of the European Union, investigates this issue and sees what exactly needs to be done to ensure technical, semantic, and legal interoperability of the datasets and what the pitfalls are, by using open data from business registries, which is one of the HVD topics.

Other approaches to integration of company data typically keep the source data as it is, i.e., noninteroperable for others, and try to build value for their project by ingesting and cleaning the data for profit. STIRData, on the other hand, aims to improve the datasets at their source, making the datasets interop-

[a] https://orcid.org/0000-0001-7234-3051
[b] https://orcid.org/0000-0001-8591-2540
[c] https://orcid.org/0000-0002-8465-6087

[1] https://stirdata.eu

KlÃ■mek, J., Chortaras, A., MÃ■Åąek, J., Yang, J., Skagemo, S. and Tzouvaras, V.
Semantic, Technical and Legal Interoperability of European Company Open Data in Practice: The STIRData Approach.
DOI: 10.5220/0012132100003541
In *Proceedings of the 12th International Conference on Data Science, Technology and Applications (DATA 2023)*, pages 183-194
ISBN: 978-989-758-664-4; ISSN: 2184-285X

erable for everyone. The STIRData approach to technical interoperability is based on linked data, and the approach to semantic interoperability is based on a common data specification that reuses the European Core Vocabularies.[2] Finally, our approach to legal interoperability of the datasets is based on guidelines to establish the appropriate terms of use and an overview of the current state of terms of use of business registry datasets.

The main contributions of this practice report paper are as follows:

1. We demonstrate how to tackle technical and semantic interoperability in a given domain using an example of company data.
2. We summarise our legal interoperability framework for open data and apply it to the domain of data from business registers.
3. We verify the approach on datasets from business registers of 13 countries.
4. By creating a user-orientated platform, we show an example of how anyone can build an application on top of this interoperable data.

The remainder of the paper is structured as follows. In Section 2 we introduce the data architecture that supports technically interoperable open data. In Section 3 we introduce the STIRData specification that supports the semantic interoperability of company data. In Section 4 we introduce the legal framework to address existing legal interoperability issues. In Section 5 we introduce the datasets on which we verified our approach. Section 6 shows the STIRData platform and describes the necessary data preprocessing steps. In Section 7 we evaluate our approach, and in Section 8 we discuss related work. In Section 9 we conclude.

2 ARCHITECTURE

To achieve semantic data interoperability, there must be a common data format with clearly defined semantics. This is defined by the STIRData specification introduced later in Section 3. To achieve technical data interoperability not only for the STIRData platform (Section 6), but also for all potential data consumers, there must be a uniform data interface at each of the data providers.

This is in contrast to other approaches to company data integration, where technical interoperability was achieved by integrating data from various data sources into a centralised database using ad-hoc transformation scripts, as is the case of the euBuisnessGraph project (see Section 8). The obvious disadvantage of the latter approach is the lack of sustainability and scalability. When another data source is to be added to the system, there needs to be someone who creates the transformation script and adds the data source to the centralised system manually. When that person is no longer available, e.g. due to the end of project funding, the original data sources are left noninteroperable and the invested effort goes to waste. A similar approach can be intentionally followed to create business value from being in control of the integrated data, which is the case of OpenCorporates (see Section 8), however, that is not the goal of high-value datasets, which should be open to everyone, preferably in an interoperable way.

Given the motivation mentioned above, the choice of RDF (Lanthaler et al., 2014) and the principles of linked data[3] as the way to publish company data on the Web was natural. Regarding the technical interface, since users will typically need to query the data, the SPARQL endpoint (Harris and Seaborne, 2013) interface was chosen. Even though from a research point of view, this architecture is not novel, it is still rarely applied to publishing interoperable public sector information, if at all. Since it is a perfect fit for the HVDs, we demonstrate in this paper how it can be applied on an example of company data, and what are the hurdles on the way.

Ideally, each data provider would publish the company data according to the proposed STIRData specification in their own SPARQL endpoint. We denote this way of publishing the data as `LD-STIR`. However, this will happen gradually, one provider at a time. Until then, there are multiple options of where the company data can be hosted, in which data format, and how it gets to the desired data format and the SPARQL endpoint.

The options are shown in Figure 1, where the compliant data is indicated by the green colour. In all cases, our aim is to simulate the ideal state. We prepare the necessary data transformation and load each resulting dataset into a separate SPARQL endpoint so that the solution can be taken over by the business registry data providers if and when they wish to do so. This should be the case of at least the Czech Business Registry, since the Ministry of Justice hosting the business registry dataset is part of the STIRData project's Experts and Collaborators Group, and the Norwegian registry, since the Brønnøysund Register Centre is a partner in the STIRData project. For the transformations themselves, we use open-source tools

[2] https://joinup.ec.europa.eu/node/145983

[3] https://www.w3.org/DesignIssues/LinkedData

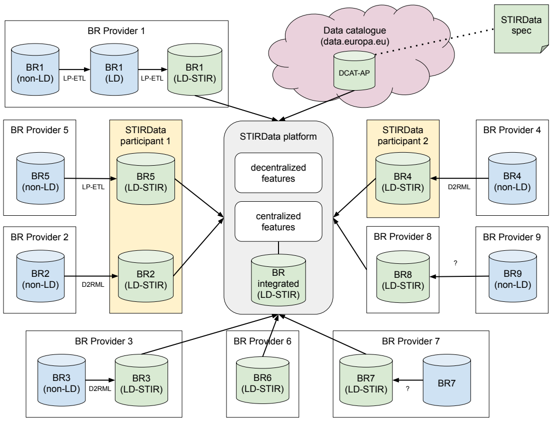


Figure 1: STIRData data architecture.

D2RML (Chortaras and Stamou, 2018) and Linked-Pipes ETL (Klímek and Škoda, 2017).

The ideal case is represented in Figure 1 by the BR provider 6, who publishes its data primarily in the `LD-STIR` form. However, we anticipate that most data providers will have their data primarily in a non-linked data form and transform it to the `LD-STIR` form directly using one of our tools (BR providers 2, 3, 4, and 5), other tools (BR providers 7 and 9) or through another linked data form such as a national one using national RDF vocabularies (BR provider 1).

The next question is where the compliant data is hosted. This would ideally be at the data provider's premises (BR providers 1, 3, 6, 7 and 8), other provider's premises (BR provider 9) or, in the worst case, at the premises of a third-party with a time-limited responsibility (BR providers 2, 4 and 5), in which case the data will eventually become unavailable. However, at least the transformations themselves will remain available to be adopted in the future.

To access the published company data, the STIRData platform, as well as any other application searching for the data, only needs to know the URLs of their respective SPARQL endpoints. These can be found in a data catalogue listing the STIRData compliant datasets, e.g., in the Official Portal for European Data,[4] via a SPARQL query searching for catalogue records linking to the STIRData specification. Part of the specification is a DCAT-AP record template that shows how to do that. Then, the platform, and any other potential data consumer, can implement both decentralized features, distributing queries to the individual SPARQL endpoints, and centralized features, such as precomputing statistics, which would normally take a longer-than-acceptable time, for real-time user interaction, to compute.

3 DATA SPECIFICATION

To ensure semantic interoperability of company data, a data specification[5] was created based on the EU Core Vocabularies by extending them and covering the requirements on the topic of company data from High-Value Datasets. As an additional input to the specification, we also used the results of the analysis of the contents of the datasets of European business registries discussed in Section 5.

The conceptual view of the specification is shown in Figure 2. The central concept is a company. Apart from typical fields such as legal name and founding date, a company may have economic activities (primary, secondary, auxiliary), it may be split in units, and it may be located in sites that may correspond to different establishments, each having an address. Apart from standard information, addresses should include the administrative units to which they belong. The specification also supports several types of company identifiers (e.g. registry identifier, tax identifier). The LEI code field offers interoperability with company data published by GLEIF.[6]

For encoding administrative units, the specification prescribes the use of the controlled vocabularies of the Territorial Units for Statistics (NUTS)[7] and the Local Administrative Units (LAU).[8] NUTS is a four-level hierarchy (NUTS-0 to NUTS-3), where the top level is the country level, and LAU further divides the lowest NUTS level to the most fine-grained local administrative units particular to each country. For company activities, the usage of the Statistical Classification of Economic Activities in the European Community Rev. 2 (NACE Rev2)[9] controlled vocabulary is prescribed, and in particular the national extensions thereof, as each country typically extends the base NACE Rev2 classification with more fine-grained, country-specific economic activity categories. NACE Rev2 is a four-level hierarchy, and national extensions typically add a fifth layer, but may extend it up to seven levels.

It should be noted that while for NUTS and NACE Rev2 there are SKOS-based linked data versions published at the EU Vocabularies site, this is not the case for LAU and NACE national extensions. Therefore, their SKOS versions were created within the STIRData project. The LAU SKOS version was created from the relevant data provided by Eurostat, while the

[4] https://data.europa.eu

[5] https://stirdata.github.io/data-specification/

[6] https://www.gleif.org/

[7] https://op.europa.eu/s/yzPW

[8] https://data.europa.eu/data/datasets/https-lkod-mff-cuni-cz-zdroj-datove-sady-stirdata-lau-stirdata

[9] https://op.europa.eu/s/yzPY

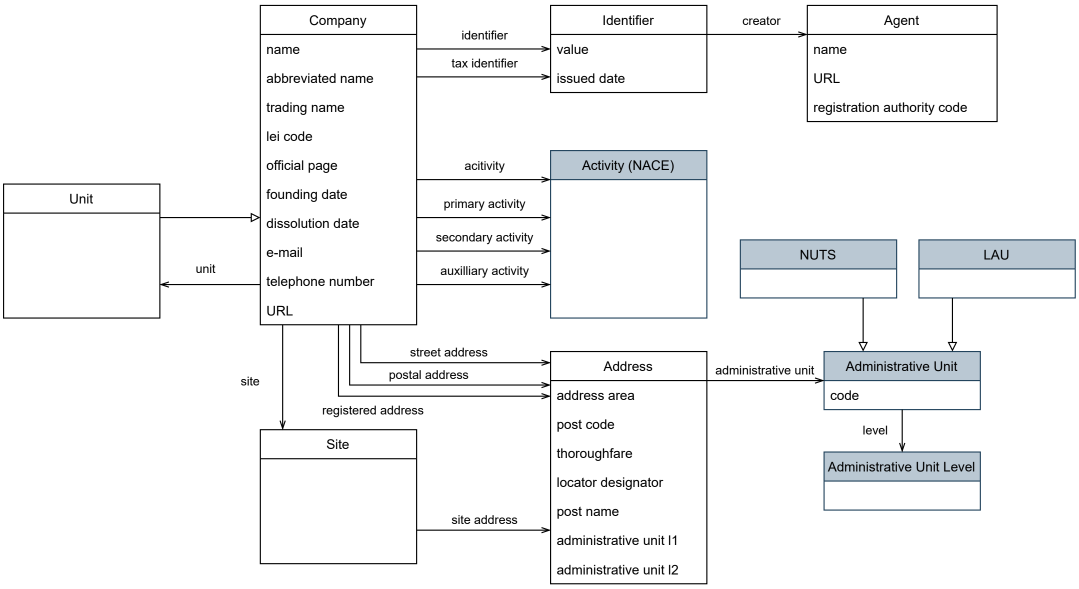


Figure 2: STIRData conceptual model. Gray classes represent controlled vocabularies.

NACE national extensions SKOS versions were created from the data provided by each country's relevant national authority. The transformations from the raw data were made using techniques similar to those discussed in Section 5.

The STIRData specification prescribes that data items should include only the lowest relevant level values in the hierarchies of those vocabularies (e.g. only the NUTS-3 and LAU level), since the upper levels can be inferred from them. Through the use of these common vocabularies, company location, encoded by administrative units, and economic activity, can be seen as two core, hierarchically structured, categorisation dimensions for companies.

The specification also contains a sample DCAT-AP[10] metadata record to be used to register the published dataset in an open data portal, so that it can be found later in `https://data.europa.eu` and from there by the STIRData platform or other consumers. The most crucial part of such a record is the inclusion of a SPARQL endpoint distribution of the data and a statement about the conformance of the data with the STIRData specification. Finally, the specification allows for licensing, data provenance and data freshness information.

4 LEGAL INTEROPERABILITY

Technical interoperability is not the only condition for successful integration and further use of open data, in our case, data from European business registers. Achieving legal interoperability is also essential; the data provider must ensure that there are no legal obstacles that would limit the further use of the provided data and must properly formulate the terms of use of the data. On the other hand, the data consumer must ensure that they know and understand the terms of use before using the data.

4.1 Data Provider

The first question the data provider must answer when publishing their dataset, e.g. a business registry, is *to what extent* the data should be made public. There is no legal instrument at the level of international or EU law that is widely applicable and comprehensively regulates the obligation of public administration to provide data and information. The issue is largely governed by national law.

Regarding our case of company data, the implementing regulation of EU Open Data Directive 2019/1024, the Implementing act on a list of High-Value Datasets, sets out the obligation to provide business register data in the required scope. However, harmonisation is not complete here either, as this obligation only applies to data already held by Member States. The regulation does not impose an obligation to create data to the required extent.This means

[10] DCAT Application Profile for data portals in Europe (DCAT-AP), https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe

that the richness of the published business registry datasets differs greatly among EU member states.

The second issue is rights to data. A potential obstacle is the *ownership of plain data*, a legislation based on national civil law and not regulated at the European or international level. In practice, however, we are not aware of any legislation in any European jurisdiction that protects plain data by property rights.

The next level of data rights is *intellectual property rights*, an area harmonised at the EU level.[11] Plain data is generally not protected by copyright because copyright law imposes relatively high standards on the originality of the subject matter and on the fact that it is "the author's own intellectual creation" (Hugenholtz and Quintais, 2021). In the context of public sector information, the mere fact that "mere intellectual effort and skill" were required for the creation of the intangible result are not relevant for its copyrightability.

The situation is different in the case of *legal protection of databases*. The latter is based at the European level by Directive 96/9/EC and consists of *copyright protection for the structure of the database* and *sui generis protection for the maker of the database* (Hugenholtz, 2016). Where database protection is present, as may arise in the publication of business register data, it is essential that the provider of the register licences the further use of its contents. The obligation to do so then also follows directly from Article 1(6) of Directive 2019/1024 (Open Data Directive). The appropriate way to deal with any obstacles arising from copyright and database law is to use a CC BY 4.0 licence.

A final area to consider is *personal data protection*. Business registers contain records of natural persons involved in registered companies and are therefore subject to EU Regulation 2016/679, the General Data Protection Regulation (GDPR). Although this is a harmonised European regulation, the final decision on whether selected personal data can be published in open data quality is up to the national legislator.

In summary, to ensure legal interoperability, it is essential that the data provider provides clear terms of use for the published data in the dataset. Theoretically, the terms of use may contain the custom conditions that the data provider imposes on the recipient, but in practice, the possibilities of setting new restrictions are severely limited by the Open Data Directive.

From a technical point of view, given the relative complexity of the national legislation governing publishing data, simply stating "this dataset is licensed as CC BY 4.0", which is a common practise, is actually not always sufficient, as this statement does not cover all potentially applicable rights summarised above with necessary legal certainty. This insufficient practice is often caused by a divide between people knowledgeable in copyright law and people implementing technical standards such as DCAT-AP in their data catalogues. For instance, the Czech terms of use are expressed in the necessary detail covering the main 4 areas based on the relevant Czech legislation as the following data structure in its metadata record in RDF Turtle:

```
terms: a pu:Specifikace ;
  pu:autorské-dílo terms:neobsahuje-autorská-díla ;
  pu:databáze-jako-autorské-dílo
  terms:není-autorskoprávně-chráněnou-databází ;
  pu:databáze-chráněná-zvláštními-právy
  terms:není-chráněna-zvláštním-právem-pořizovatele-databáze ;
  pu:osobní-údaje terms:neobsahuje-osobní-údaje .
```

The obvious downside of this legally proper approach is that it may be difficult for a foreign user of the data to properly understand the terms as they will not be experts in Czech legislation, even though this particular specification says that the data is free of any legal obstacles, but still not easily comparable to, e.g., a CC0 license, in the pure legal sense.

4.2 Data Consumer

The legal situation of a consumer of data from European business registries is not an easy one. The data consumer must ensure that they know and understand the terms of use before using data they find in a data catalog. However, dealing with data from multiple countries, where each dataset is described by a different terms of use document, based on a different national legislation, is hard even for people with legal background, not to mention determining a correct license to be assigned to the combined dataset. Technical people without such background therefore often simply skip the legal analysis, risking legal issues in the process. The types of current terms of use of business registry datasets used in STIRData can be seen in Table 2. The HVD regulation attempts to improve the situation by mandating that each HVD dataset should have terms of use comparable or less restrictive than "CC BY 4.0".

5 DATASETS

To verify our approach, we used open data datasets from several European countries' business registries, with the purpose of transforming them to the LD-STIR form, i.e. to linked data according to the STIRData specification, making them thus part of the data architecture of Figure 1. These source datasets were obtained from either the European or the respective

[11] See especially Directive 96/9/EC on the legal protection of databases and Directive 2001/29/EC on the harmonisation of certain aspects of copyright and related rights in the information society.

country's open data portal, or directly from the respective business registry. Each dataset should be available directly for bulk download or accessible through an API that allows obtaining its full contents, so that transformation of the entire dataset into `LD-STIR` form and its subsequent publication in an RDF store could be possible. The datasets that satisfied these conditions and are currently incorporated into the STIRData platform are for the business registers of Belgium, Cyprus, Czechia, Estonia, Finland, France, Greece (Athens area), Latvia, Moldova, Netherlands, Norway, Romania, and the UK.

These datasets, all of which are in `non-LD` form, are considerably heterogeneous in format and content. Most are available as single or multiple CSV files, but there are also spreadsheets, XML and JSON files. The size of the datasets also varies considerably, depending on the country and the encoded information. For example, the French dataset contains information for about 23M main legal entities and 33M establishments, the UK dataset more than 5M main legal entities, while the Moldovan dataset contains about 235K.

To prepare the datasets for transformation, the data fields and contents of each dataset were analysed, and, eventually, each dataset was characterised along a set of dimensions relevant to the STIRData specification. These dimensions and brief descriptions are provided in Table 1. Following the analysis, to convert the source datasets from their original format to STIRData compliant linked data datasets, appropriate transformation workflows were developed and applied using two transformation tools: LinkedPipes ETL[12] and D2RML.[13] Both are powerful generic transformation tools capable of handling multiple data formats and performing complex data mappings to linked data representations.

Table 1: Dataset dimensions.

Dimension	Description/Values
Entity types	The types of provided entities, e.g. main entities, establishments.
Person types	All datasets include registered legal persons, but some include data also about natural persons (sole traders).
Names	The types of provided names for legal entities. Apart from the legal name some registers provide trading names, or abbreviated names.
Identifiers	The types of provided identifiers. Usually the registry identifier is provided, but some datasets also provide tax and other identifiers.
Reg. date	It indicates whether a registration date is provided.
Dissolution date	It indicates whether a dissolution date is provided. Most registries provide information only about currently registered entries. Dissolved and deregistered companies are in most cases not included.
Economic activities	The types and number of provided economic activity codes. Some registries provide a single or fixed number of uncharacterized business activity codes, while others distinguish between main, secondary and auxiliary activities.
Address	The types of provided addresses (e.g. business, postal address).
Legal form	It indicates whether the legal form of each legal entity is provided. The possible legal forms for each country are usually a closed list.
Legal status	It indicates whether information about the current status of the company is provided (e.g. active, in liquidation, etc.)
Foreign entities	It indicates whether the dataset includes entities whose base registration is in a foreign country.

The defined transformations had a varying level of complexity, depending on the structure of the original data representation. In this respect, we should note that, given the heterogeneity of the source data, not only with respect to format but also to structure due to the different underlying legislations, several modelling and transformation issues had to be addressed. Such an issue is, e.g., the structure of large companies. In some countries, a company is registered as a whole in a business registry, where it has its identifier and a registered address, with no information about actual points of service, while in other countries, each company's point of service has an individual registration in the business registry, possibly with a different identifier and set of economic activities. As mentioned in Section 3, such issues were taken into account when designing the STIRData data specification, so as to make it generic enough to cover different registration practises regarding the structure of companies.

An important part of the company data transformation process was also the mapping of relevant source data entry values to linked data resources of

[12] https://etl.linkedpipes.com/

[13] https://apps.islab.ntua.gr/d2rml/. As an example, the D2RML document for transforming data from the Norwegian main business unit is https://stirdata-semantic.ails.ece.ntua.gr/api/content/no/mapping/2635b100-7f4b-44ca-bcd6-93854ceb644c.

the appropriate vocabularies, as in the case of economic activity codes, that were mapped to the respective NACE national extensions resources, and the addresses from which the underlying administrative units, as NUTS and LAU resources, had to be inferred. Since NACE codes are closed code lists and were included in the data, their transformations were straightforward. For addresses, the mapping was achieved by exploiting company address postcode information in combination with data provided by GISCO.[14] To enhance data interoperability, the mappings also produced LEI code properties using GLEIF open data.[15] Finally, in order to reduce the size of the resulting datasets, when the original datasets contained long historical data of dissolved companies, only the more recent data were kept.

The resulting `LD-STIR` datasets were published and made available through different SPARQL endpoints, one for each country (using Virtuoso Open Source Edition as underlying triple stores). Their size and information on which of five basic dimensions they include is provided in Table 2. The SPARQL endpoints are https://stirdata-semantic.ails.ece.ntua.gr/api/content/*xy*/sparql, where *xy* is `be`, `cy`, `ee`, `fi`, `fr`, `el`, `lv`, `md`, `nl`, `no`, `ro`, `uk`, for each country in the order it appears in Table 2. For Czechia, the endpoint is https://obchodní-rejstřík.stirdata.opendata.cz/vsparql. These endpoints are currently managed by the project partners. However, they are ready to be taken over by the individual business registries, including the data transformations creating their content.

As prescribed by the STIRData data specification, all published datasets include information about their provenance (the source dataset), licence, and date of last update. Given that the source datasets are updated periodically, the published STIRData-compliant datasets are also updated periodically after reapplying the transformations on the newer versions of the source datasets.

6 THE STIRData PLATFORM

To demonstrate in a concrete way the value of the proposed approach, we developed the STIRData platform,[16] which, on top of the data architecture described in Section 2, and the published datasets described in Section 5, provides a user-friendly interface to explore in a uniform manner all business registry data.

The platform in principle adopts a fully decentralised architecture. It assumes that each dataset resides in a separate remote SPARQL endpoint. Apart from some basic information about each dataset, it also centrally stores copies of the shared NUTS, LAU and NACE vocabularies. In addition, to improve performance of the user facing platform, centrally stored precomputed statistics data and indexes have been added as extensions to the basic platform architecture, making it less dependent on the performance characteristics of the source SPARQL endpoints. We discuss this addition in greater detail later in the paper.

The STIRData compliant business registry datasets offered by the platform are discovered automatically by scheduled tasks that periodically check for new datasets in the Official portal for European data, as well as for updates of already included datasets. Datasets not yet available in the official portal for European data can be registered manually; in either case, the only required information is a link to the respective SPARQL endpoint.

For the end-user, the platform offers access to the underlying data through three main types of queries: retrieval queries, search queries, and statistics queries.

6.1 Retrieval Queries

Retrieval queries are the simplest queries that retrieve information about specific legal entities. A legal entity is identified by its STIRData IRI; based on this, the platform identifies the corresponding business registry and issues an appropriate query to the respective SPARQL endpoint to get the legal entity details.

Apart from the details provided by the STIRData specification, the implementation of retrieval queries allows also for obtaining additional information about legal entities from other sources that have relevant data published as linked data. These sources can be added to the platform through a generic add-on mechanism. An example is the data of the Czech Trade Inspection Authority[17] which is used by the platform to show the inspections that Czech legal entities have undergone. An example of a retrieval result page, showcasing also that feature, is shown in Figure 3.

[14] https://ec.europa.eu/eurostat/web/gisco/

[15] https://www.gleif.org/en/lei-data/gleif-concatenated-file

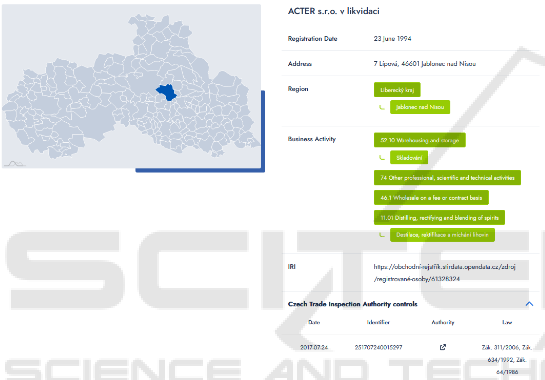
[16] https://portal.stirdata.eu

[17] https://data.europa.eu/data/datasets/https-lkod-mff-cuni-cz-zdroj-datove-sady-stirdata-c-oi-kontroly-zame-r-eni-sankce

Table 2: Overview of published STIRData-compliant linked data datasets.

Country	# Main legal entities*	Legal name	Registration date	Dissolution date	Economic activity	Location	Terms of use
Belgium	1,885,610	✓	✓		✓	✓	Custom (French)
Cyprus	509,648	✓	✓	✓		✓	CC BY 4.0
Czechia	556,854	✓	✓	✓	✓	✓	Custom (Czech)
Estonia	342,997	✓	✓			✓	CC BY-SA 3.0
Finland	143,354	✓	✓		✓	✓	CC BY 4.0
France	19,949,924	✓	✓	✓	✓	✓	Custom (French)
Greece†	36,247	✓	✓		✓	✓	Not specified
Latvia	449,031	✓	✓	✓		✓	CC0 1.0
Moldova	235,563	✓	✓	✓	✓	✓	Not specified
Netherlands‡	3,521,128		✓		✓		Not specified
Norway	1,097,878	✓	✓		✓	✓	Custom (Norwegian)
Romania	1,683,823	✓				✓	CC BY 4.0
United Kingdom	5,253,635	✓	✓		✓	✓	Not specified

*At the time of writing. †Athens region only. ‡Netherlands provides anonymised data.



ACTER s.r.o. v likvidaci

Registration Date: 23 June 1994

Address: 7 Lípová, 46601 Jablonec nad Nisou

Region: Liberecký kraj
- Jablonec nad Nisou

Business Activity: 52.10 Warehousing and storage
- Skladování

74 Other professional, scientific and technical activities

46.1 Wholesale on a fee or contract basis

11.01 Distilling, rectifying and blending of spirits
- Destilace, rektifikace a míchání lihovin

IRI: https://obchodní-rejstřík.stirdata.opendata.cz/zdroj/registrované-osoby/61328324

Czech Trade Inspection Authority controls

Date	Identifier	Authority	Law
2017-07-24	251707240015297		Zák. 311/2006, Zák. 634/1992, Zák. 64/1986

Figure 3: Example company details page. The bottom levels of the relevant NUTS/LAU and NACE hierarchy are displayed, as well as Czech Trade Inspection Authority data.

6.2 Search Queries

Search queries retrieve lists of companies that satisfy conditions based on location, economic activity, and registration date. For example, a user may request all companies registered in the Oslo area in Norway and in the Prague area in Czechia after a certain date that perform one of a specific set of economic activities. The conditions regarding location and economic activities are expressed using NUTS, LAU and NACE Rev2 vocabulary concepts, respectively.

Search queries are, in principle, federated queries, since they involve data residing in different SPARQL endpoints: those hosting the vocabularies and a different endpoint for the data of each involved country. Moreover, answering such queries involves SKOS hierarchy based reasoning since, e.g., asking for companies in a certain region is actually asking for companies in any subregion thereof, and similarly for economic activities. These features pose significant challenges to query efficiency. One option is to directly use the SPARQL constructs by which such queries can be realised, i.e., federated SPARQL queries and path expressions. Figure 4(a) shows a direct formulation of an example SPARQL query using these two constructs. However, the evaluation of such complex queries on public endpoints may turn out problematic. Table 3 (first query column) shows the results of an experimental evaluation of the above query on three different triple stores. As we can see, one did not

```
SELECT ?entity WHERE {
  ?entity a legal:LegalEntity .
  ?entity legal:companyActivity ?nace .
  ?entity legal:registeredAddress/m8g:adminUnit [
     m8g:code ?nuts3 ;
     m8g:level sd-adminUnitLevel:NUTS-3 ] .
  SERVICE endpoint-nace: {
    ?nace skos:exactMatch/skos:broader* nace:46 }
  SERVICE endpoint-nuts: {
    ?nuts3 skos:broader* nuts:NO0  }
}
```

(a) Federated query using property path expressions.

```
SELECT ?entity WHERE {
  ?entity a legal:LegalEntity .
  ?entity legal:companyActivity ?nace .
  ?entity legal:registeredAddress/m8g:adminUnit [
     m8g:code ?nuts3 ;
     m8g:level sd-adminUnitLevel:NUTS-3 ] .
  VALUES ?nace { nace-no:46.110 ... nace-no:46.900 }
  VALUES ?nuts3 { nuts:NO020 ... nuts:NO0B2 }
}
```

(b) Non-federated query using exhaustive values listing. The `VALUES ?nace` statement includes 71 effective values and the `VALUES ?nuts` 13 effective values.

Figure 4: Example queries asking for all companies in the NO0 NUTS-1 region (Norge) performing a subactivity of the nace:46 class (Wholesale trade, except of motor vehicles and motorcycles).

Table 3: Evaluation of the queries of Figure 4 on three triple stores.

Triple store	Query	
	Fig.4(a)	Fig.4(b)
Virtuoso Open Source Edition	Internal error: Unsupported combination of subqueries and service invocations	< 1 sec
Apache Jena Fuseki	~ 7 sec	~ 5 sec
GraphDB Free	No response after 2 hours	< 1 sec

support that query, one failed to efficiently evaluate it, and only one was able to answer in an acceptable time.

To avoid such problems, our platform leverages domain knowledge and the closed form of such queries, by expanding and splitting each query to a set of simpler queries addressed to the appropriate endpoints so that they can be answered more efficiently. In particular, a query of the form of Figure 4(a) is executed in three steps, corresponding to the three parts of the federated query. The first two steps, which we will call effective values computation, is to obtain the list of subactivities of the activity specified in the query by issuing a simple query to the respective endpoint, and do the same thing for the subregions of the region of interest; as the last step, a non-federated query is issued directly to the company data endpoint explicitly listing the effective activity and region values using the `VALUES` SPARQL construct. That query is shown in Figure 4(b) and the results of its evaluation in the second query column of Table 3. The effective values computation is more complex in case of multiple conditions.

As a further example of data interoperability, in addition to the above functionality, the implementation of search (and retrieval) queries by the STIRData platform also allows one to use conditions based on relevant Eurostat statistics.[18] So, in addition to the conditions described above, a user may ask, e.g., for companies located in the more urbanised areas of a country, or in areas with high availability of touristic accommodation. The implementation of this feature relies on the availability of such statistics as linked data using the RDF Data Cube Vocabulary (Cyganiak and Reynolds, 2014). Because Eurostat data are not currently available in this format (they are available as CSV data only), transformation and publication of selected Eurostat statistics has been done in a similar way to the business registry datasets using D2RML transformations. At query time, constraints expressed using Eurostat statistics are translated by issuing an appropriate SPARQL query to the platform's endpoint holding the statistics to an effective value list of the regions satisfying the constraints, and that list is then used as described above in the `VALUES` construct of the final SPARQL query to the registries' endpoints holding the actual business registry data.

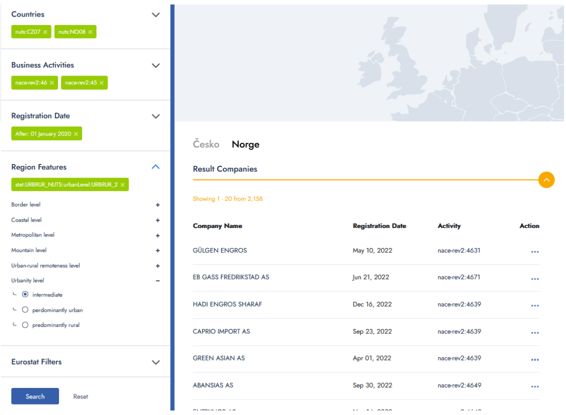
A sample search query page, which also demonstrates this feature, is shown in Figure 5.

6.3 Statistics Queries

Statistics queries are similar to search queries, but instead of lists of companies, they return aggregated statistical information, namely the number of companies satisfying the desired criteria, along with an analysis of the distribution of companies in the subregions and subactivities specified in the query. Statistics queries have been implemented similarly to search queries.

Because statistics queries provide useful, compact overviews of the underlying data, an important feature of the platform is that it allows users to browse through the location and/or the economic activity hierarchies, displaying the corresponding statistical information. However, because such browsing again requires the execution of multiple complex SPARQL queries against public triple stores containing potentially millions of RDF triples, their real-time computation would result in poor aggregate response times.

For this reason, the STIRData platform adopts the approach of pre-computing offline several of those statistics (for the location, economic activity, and registration date dimensions, and pairs thereof) and caches them in a database, so that they can be served instantly. The statistics precomputation process is activated each time a new business registry dataset is registered or already published data are updated. Precomputation of the statistics for a country can take from a few minutes to several days, depending on the size and dimensions of the dataset. The results are stored in a PostgreSQL database. A sample page



Countries
nuts:CZ07 ✕ nuts:NO08 ✕

Business Activities
nace-rev2:46 ✕ nace-rev2:45 ✕

Registration Date
After: 01 January 2020 ✕

Region Features
stat:URBRUR_NUTS:urbanLevel:URBRUR_2 ✕

Border level +
Coastal level +
Metropolitan level +
Mountain level +
Urban-rural remoteness level +
Urbanity level −
● intermediate
○ perdominantly urban
○ predominantly rural

Eurostat Filters

Česko Norge

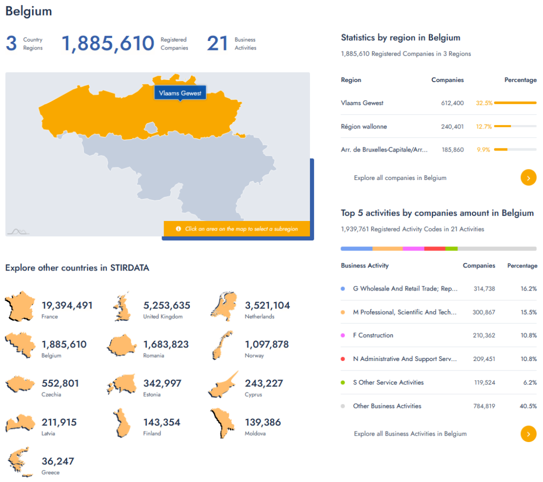
Result Companies

Showing 1 - 20 from 2,158

Company Name	Registration Date	Activity	Action
GÜLGEN ENGROS	May 10, 2022	nace-rev2:4631	...
EB GASS FREDRIKSTAD AS	Jun 21, 2022	nace-rev2:4671	...
HADI ENGROS SHARAF	Dec 16, 2022	nace-rev2:4639	...
CAPRIO IMPORT AS	Sep 23, 2022	nace-rev2:4639	...
GREEN ASIAN AS	Apr 01, 2022	nace-rev2:4639	...
ABANSIAS AS	Sep 30, 2022	nace-rev2:4649	...

Figure 5: Example search page, requesting Czech and Norwegian wholesale companies founded after 2020, located in intermediate urbanity level areas.

[18] https://ec.europa.eu/eurostat/data/database



Belgium

3 Country Regions | 1,885,610 Registered Companies | 21 Business Activities

Explore other countries in STIRDATA

19,394,491 France	5,253,635 United Kingdom	3,521,104 Netherlands
1,885,610 Belgium	1,683,823 Romania	1,097,878 Norway
552,801 Czechia	342,997 Estonia	243,227 Cyprus
211,915 Latvia	143,354 Finland	139,386 Moldova
36,247 Greece		

Statistics by region in Belgium

1,885,610 Registered Companies in 3 Regions

Region	Companies	Percentage
Vlaams Gewest	612,400	32.5%
Région wallonne	240,401	12.7%
Arr. de Bruxelles-Capitale/Arr...	185,860	9.9%

Explore all companies in Belgium ›

Top 5 activities by companies amount in Belgium

1,939,761 Registered Activity Codes in 21 Activities

Business Activity	Companies	Percentage
● G Wholesale And Retail Trade; Rep...	314,738	16.2%
● M Professional, Scientific And Tech...	300,867	15.5%
● F Construction	210,362	10.8%
● N Administrative And Support Serv...	209,451	10.8%
● S Other Service Activities	119,524	6.2%
● Other Business Activities	784,819	40.5%

Explore all Business Activities in Belgium ›

Figure 6: Interface for browsing country and economic activity statistics.

showing administrative units and economic activity statistics for Belgium is shown in Figure 6.

7 EVALUATION

The platform described in Section 6 is fully operational and allows users to explore multiple business registries, provides interoperability with other datasets (e.g. Eurostat data), and is periodically updated with new versions of the business registry datasets. In this section, we discuss the performance optimizations needed to be implemented to overcome the inherent performance issues of the public SPARQL endpoint-based decentralized architecture.

As mentioned above, our architecture relies on possibly third-party managed SPARQL endpoints to get access to company data. This achieves the desired decentralised data interoperability, and the SPARQL language is expressive enough to meet complex data querying needs. However, as the complexity of the queries posed by a data consumer increases, and also depending on the size of the underlying dataset and the computational resources available to an endpoint, performance problems may arise that can result in poor response times.

Given the decentralised nature of our approach, where ideally each business registry provides its data through an own-managed infrastructure, much in this respect depends on the software and computational resources that each business registry disposes, and the continuous availability of the endpoints, which is beyond the control of the platform. We experienced such problems, which were overcome by increasing the resources (allocated memory) available to the SPARQL endpoints and by rewriting queries using more efficient execution plans, as discussed in Section 6. As explained there, we also experimented with several software platforms that implement SPARQL endpoints, and their performance varied considerably, with some types of queries answered more efficiently by some of them than by others.

It is important to note that much of the complexity of the queries that leads to the above problems arises from the fact that, as discussed in Section 3, the STIRData specification requires keeping for each company only the lowest-level information about the administrative unit and economic activity hierarchies it belongs to. Although this is a sound data modelling assumption that reduces redundancy and is compliant with the linked data principles, it leads to the need of inferring at query time the higher levels of the hierarchies a company belongs to, which may result in computationally demanding queries. In Section 6 we showed how we alleviated the problem by performing the effective values computation, without compromising on the fully decentralised approach, and insisting on an on-the-fly execution of SPARQL queries. We also discussed how we precomputated certain statistics, whose on-the-fly computation is problematic.

However, because not all statistics (e.g. for multiple conditions) can be precomputed, and complex queries cannot be avoided without severely limiting the data exploration capabilities of the platform, another architectural compromise had to be implemented. As an alternative, a periodically updated, platform-managed materialised cache of the published data was introduced into the platform. It contains all the required inferred and pre-computed information, so that effective values computation on SKOS vocabulary hierarchies is not needed at query time, and the actual SPARQL queries evaluated on the remote SPARQL endpoints are simplified. The cache can be stored in a local triple store or in an index of the data, for even faster query times.

We are currently in the process of incorporating in the platform such an index using ElasticSearch, which stores for each company the fields on which search is expected, namely administrative units, economic activities, and registration dates, including materialised (i.e. SKOS hierarchy-based inferred) values. Some indicative results are shown in Table 4, which compares the response time for several statistics queries, using direct SPARQL evaluation, precomputed statistics, and the index. The size of the datasets on which the queries were executed is that shown in Table 2. As expected, precomputed data, when available, are served faster, with performance comparable to the index, which is significantly faster than direct SPARQL

Table 4: Response time for several statistics queries, using direct SPARQL evaluation, precomputed statistics, and materialized indexed data.

Query conditions		Response time (in secs)		
admin unit	activity	SPARQL	Prec.	Index
nuts:NO0A	-	206.0	0.2	0.2
nuts:NO0A	nace:B	9.1	0.1	0.1
nuts:NO0A, stat*	-	138.6	-	4.5
nuts:FI1	-	82.1	0.1	0.2
nuts:FI1	nace:B	2.5	0.1	0.1
nuts:FI1, stat*	-	14.1	-	6.5
nuts:UKL	-	192.9	0.1	0.1
nuts:UKL	nace:B	27.4	0.1	0.1
nuts:UKL, stat*	-	126.5	-	82.9

*Eurostat statistics-based condition involving NUTS-3 urbanity level classification (level 3, predominantly rural) and the number of establishments by NUTS-2 region statistic (> 100 establishments).

query evaluation. Direct SPARQL query response times depend clearly on the size of the underlying data and the complexity of the queries (i.e. on the effective values computation time). It is important to note that the statistic queries return both the number of companies satisfying the specified conditions, as well as a distribution thereof in the relevant subregions and subactivities; this means that not a single but multiple queries have to be executed in each case, which explains the relatively long response times. It is also interesting to note that, for the query involving Eurostat conditions, the index also appears to be relatively slow. This can be explained by the fact that in this case the first part of the query evaluation (the effective values computation) cannot be delegated fully to the index (which contains only SKOS hierarchy-based inferred values), since it requires issuing SPARQL queries to the endpoint holding the Eurostat statistics in order to get the effective values for the desired conditions to be used in the actual query to the index. This shows that in more general queries involving the combination of data from different sources, the index cannot always guarantee immediate response times. In conclusion, as expected, the performance improvement using an index is significant, although queries relying (even partially) on SPARQL query evaluation may still suffer slower response times.

8 RELATED WORK

STIRData is, of course, not the first project dealing with the integration of data on companies from various data sources. OpenCorporates[19] makes business out of the integration and cleaning of company data, and the euBusinessGraph[20] project built a marketplace for such datasets. However, both of these approaches have one thing in common, which is that they keep the source data as it is, i.e., noninteroperable for others, and they build value for their project by ingesting and cleaning the data for profit. In contrast, STIRData aims at improving the datasets at their source, making the datasets interoperable for everyone, and showing how such interoperable datasets can be reused.

There is also the Business Registers Interconnection System (BRIS),[21] which allows human users to manually search for companies in the integrated business registers using a web page. However, this is all that BRIS offers. It is a specialised information system connecting directly to the individual business registries, in a completely closed manner, and has nothing to do with open data or the Common European Data Spaces.

9 CONCLUSIONS

In this paper, we present the results of STIRData, a project that implements a linked data-based approach to the publication of open data from European business registries in an interoperable fashion. Our interoperability approach addresses the technical, semantic, and legal dimensions of interoperability of data. The semantic interoperability approach is based on the European Core Vocabularies, the technical interoperability approach is based on the linked data technologies and we suggest a legal interoperability framework for open data in general and emphasize the non-ideal situation of today's data consumers as to the legal certainty when using open data. The main difference of STIRData compared to other company data integration approaches is that we make the data interoperable on the publisher's side, i.e. for everyone, not centrally, and not for profit. Finally, we demonstrate a way to build applications on top of interoperable data, including difficulties coming from the linked data-based architecture, by presenting the STIRData platform for data browsing and analysis. In addition, we see the need for a similar approach also in the Common European Data Spaces, which are currently being established and which go beyond the scope of open data.

During the course of the project, we faced some performance issues inherent to the usage of SPARQL endpoints for publishing larger datasets and using

[19] https://opencorporates.com/

[20] https://www.eubusinessgraph.eu/

[21] https://e-justice.europa.eu/content_business_registers_at_european_level-105-en.do

aggregation queries on top of them. We worked around the problem by pre-computing the necessary statistics, only querying the endpoints using simpler queries, and creating materialised data indexes. As part of our future work, we will therefore investigate the possibilities of application of alternative linked data interfaces such as the Linked Data Fragments (Verborgh et al., 2014) and the Linked Data Event Streams (Lancker et al., 2021) to see whether they can help with the issue.

ACKNOWLEDGEMENTS

This work was co-financed by the Connecting Europe Facility Programme of the European Union, under GA n. INEA/CEF/ICT/A2019/2063078, action number 2019-EU-IA-0068: STIRData: Specification and Tools for Interoperable and Reusable data.

REFERENCES

Chortaras, A. and Stamou, G. (2018). D2RML: Integrating Heterogeneous Data and Web Services into Custom RDF Graphs. In Berners-Lee, T., Capadisli, S., Dietze, S., Hogan, A., Janowicz, K., and Lehmann, J., editors, *Workshop on Linked Data on the Web co-located with The Web Conference 2018, LDOW@WWW 2018, Lyon, France April 23rd, 2018*, volume 2073 of *CEUR Workshop Proceedings*. CEUR-WS.org. http://ceur-ws.org/Vol-2073/article-07.pdf.

Cyganiak, R. and Reynolds, D. (2014). The RDF Data Cube Vocabulary. W3C Recommendation, W3C. https://www.w3.org/TR/2014/REC-vocab-data-cube-20140116/.

Harris, S. and Seaborne, A. (2013). SPARQL 1.1 Query Language. W3C Recommendation, W3C. https://www.w3.org/TR/2013/REC-sparql11-query-20130321/.

Hugenholtz, P. B. (2016). Directive 96/9/EC. In Dreier, T. and Hugenholtz, P. B., editors, *Concise European copyright law*, pages 379–420. Kluwer Law International, Alphen aan den Rijn, The Netherlands, second edition edition.

Hugenholtz, P. B. and Quintais, J. P. (2021). Copyright and Artificial Creation: Does EU Copyright Law Protect AI-Assisted Output? *International Review of Intellectual Property and Competition Law*, 52(9):1190–1216. https://doi.org/10.1007/s40319-021-01115-0.

Klímek, J. and Škoda, P. (2017). LinkedPipes ETL in use: practical publication and consumption of linked data. In Indrawan-Santiago, M., Steinbauer, M., Salvadori, I. L., Khalil, I., and Anderst-Kotsis, G., editors, *Proceedings of the 19th International Conference on Information Integration and Web-based Applications & Services, iiWAS 2017, Salzburg, Austria, December 4-6, 2017*, pages 441–445. ACM. https://doi.org/10.1145/3151759.3151809.

Lancker, D. V., Colpaert, P., Delva, H., de Vyvere, B. V., Meléndez, J. A. R., Dedecker, R., Michiels, P., Buyle, R., Craene, A. D., and Verborgh, R. (2021). Publishing Base Registries as Linked Data Event Streams. In Brambilla, M., Chbeir, R., Frasincar, F., and Manolescu, I., editors, *Web Engineering - 21st International Conference, ICWE 2021, Biarritz, France, May 18-21, 2021, Proceedings*, volume 12706 of *Lecture Notes in Computer Science*, pages 28–36. Springer. https://doi.org/10.1007/978-3-030-74296-6_3.

Lanthaler, M., Wood, D., and Cyganiak, R. (2014). RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation, W3C. https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/.

Verborgh, R., Hartig, O., Meester, B. D., Haesendonck, G., Vocht, L. D., Sande, M. V., Cyganiak, R., Colpaert, P., Mannens, E., and de Walle, R. V. (2014). Low-Cost Queryable Linked Data through Triple Pattern Fragments. In Horridge, M., Rospocher, M., and van Ossenbruggen, J., editors, *Proceedings of the ISWC 2014 Posters & Demonstrations Track a track within the 13th International Semantic Web Conference, ISWC 2014, Riva del Garda, Italy, October 21, 2014*, volume 1272 of *CEUR Workshop Proceedings*, pages 13–16. CEUR-WS.org. http://ceur-ws.org/Vol-1272/paper_10.pdf.