Eurostars Project

3DFed – Dynamic Data Distribution and Query Federation

Project Number: E!114681 **Start Date of Project:** 2021/04/01 **Duration:** 36 months

Deliverable 2.2
Report on Monitoring the Data Storages

Dissemination Level	Public
Due Date of Deliverable	March 31, 2023
Actual Submission Date	March 31, 2023
Work Package	WP2, Data Storage Monitoring and Profiling
Deliverable	D2.2
Type	Report
Approval Status	Final
Version	1.0
Number of Pages	15

Abstract:
Deliverable D2.2 (Report on Monitoring the Data Storages) aims to provide an overview and describe the results from the development and deployment of a SPARQL endpoint monitoring platform by the consortium. As part of T2.2, the team from OpenLink has successfully deployed a cloned version of the open-source solution SPARQLES, which has been automatically monitoring the uptime, average response time for different types of queries, etc. of SPARQL endpoints registered on Datahub. Additionally, the monitoring data is made available in RDF format, as part of a generated RDF dataset.

The information in this document reflects only the author's views and Eurostars is not liable for any use that may be made of the information contained therein. The information in this document is provided "as is" without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/ her sole risk and liability.

3DFed Project by Eurostars.

History

Version	Date	Activity	Author
0.1	21/02/2023	Initial Draft	Milos Jovanovik
0.2	01/03/2023	Extended Draft	Milos Jovanovik
0.3	28/03/2023	Input on Draft	Mirko Spasić
0.4	28/03/2023	Issued for review	Milos Jovanovik
0.5	30/03/2023	Review	Muhammad Saleem
1.0	31/03/2023	Final approval and submission	Milos Jovanovik

Author List

Organization	Name	Contact Information
OpenLink Software	Milos Jovanovik	mjovanovik@openlinksw.com
OpenLink Software	Mirko Spasić	mspasic@openlinksw.com
OpenLink Software	Hugh Williams	hwilliams@openlinksw.com
University of Paderborn	Muhammad Saleem	saleem@informatik.uni-leipzig.de

Contents

1 Introduction

The aim of this deliverable, D2.2 "Report on Monitoring the Data Storages", is to provide a report on the work done on T2.2 of the project, titled "Monitoring the data storage solutions". In the reporting period, OpenLink deployed a web application - 3DFed SPARQL Endpoint Monitoring Service - which serves as an automated tool for monitoring and profiling of public SPARQL endpoints, i.e. RDF data storage solutions. As per T2.2, these SPARQL endpoints were collected from Datahub (`https://datahub.io/`).

The application is publicly available online, at `https://sparqles.demo.openlinksw.com`. It monitors the availability, uptime, average response time, performance, interoperability and discoverability of the SPARQL endpoints of interest. The collected data and the calculated statistics are stored in the application, but are also available as RDF for the creation of the monitoring and profiling dataset defined in T2.1.

The application currently works with SPARQL endpoints only, but this can be extended and generalized in the future to other NoSQL databases which have public interfaces.

2 Monitoring of SPARQL Endpoints

The 3DFed SPARQL Endpoint Monitoring Service has been operating for 11 months, monitoring a total of 581 SPARQL endpoints (Figure 1). It is a clone of the famous SPARQLES application, which has been modified for the purposes of the project and has been upgraded to use Docker and Docker Compose for easier deployment. The application is open-source, and the code is available on GitHub (`https://github.com/OpenLinkSoftware/sparqles`).

2.1 Monitoring Statistics

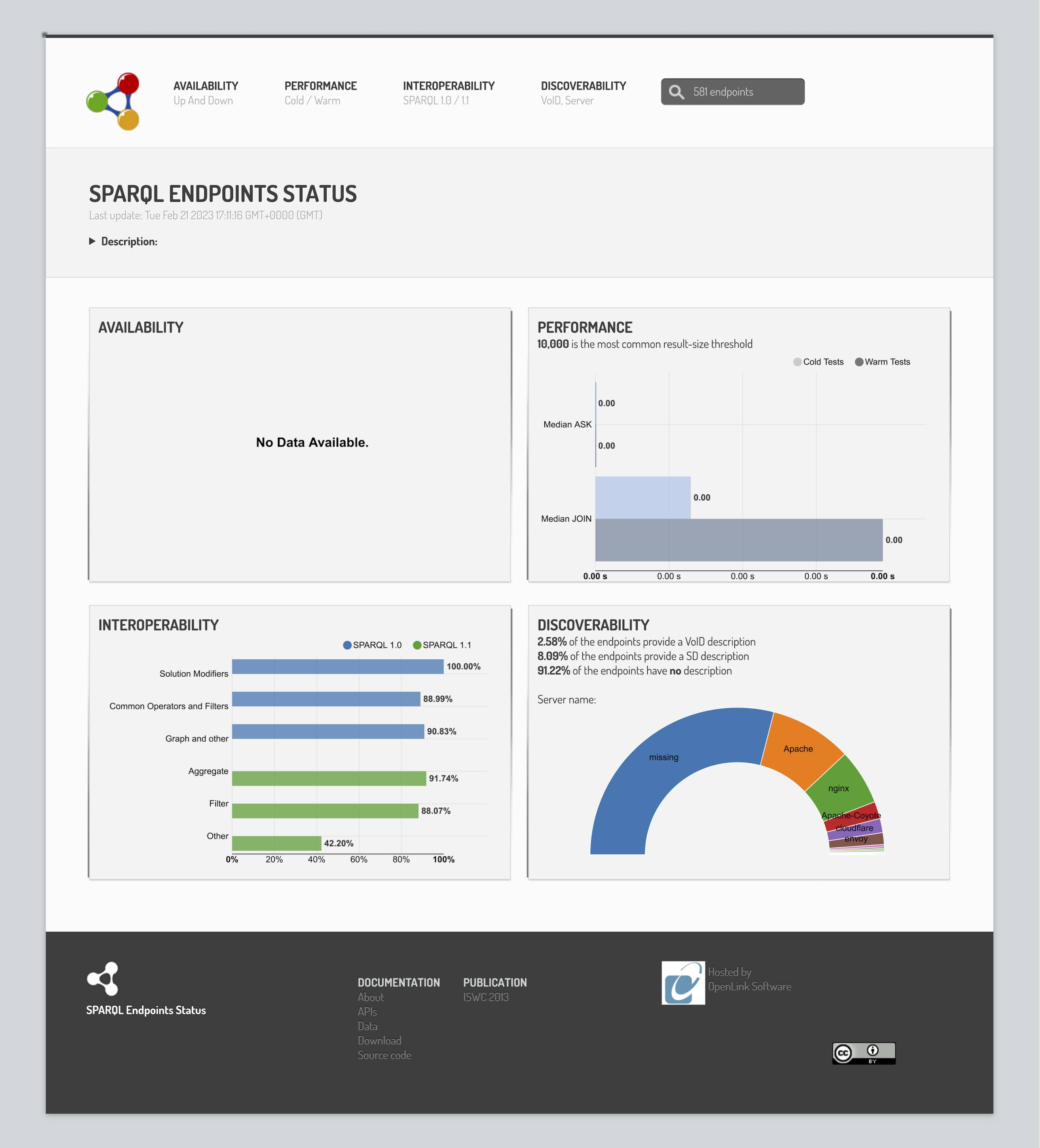
In this section we present some important statistics collected in the past 11 months about the datasets monitored by our application. This is achieved using the following two methods:

REST API: SPARQLES portal provides 7 different publicly available APIs[1], where 3 of them are related to the endpoints (listing all of them, listing only endpoints whose URL, label or dataset label is partly specified and providing information about the specific endpoint) and 4 of them are the analytic ones (availability, discoverability, interoperability and performance) based on the last test executed, or (in the case of availability) based on all tests performed in the last 24 hours or 7 days.

MongoDB queries: In cases where different statistics are needed (different time windows, analytics covering more features, etc.), it is possible to run MongoDB queries directly against the database managed by the SPARQLES portal within a Docker container, and retrieve useful information accordingly.

The total number of monitored endpoints by our platform is 581, which are public SPARQL endpoints registered on the Datahub portal. In the continuation of the project, within Task 2.1, we plan to add the ability to add individual endpoints that we want to monitor, regardless of whether they are registered on Datahub or not. The vast majority of the 581 monitored endpoints (94.84%) contain only one dataset each, while there are a small number that contain more than one dataset. Table 1 shows how many endpoints contain how many datasets.

[1] `https://sparqles.demo.openlinksw.com/api`



SPARQL ENDPOINTS STATUS

Last update: Tue Feb 21 2023 17:11:16 GMT+0000 (GMT)

▶ Description:

AVAILABILITY

No Data Available.

PERFORMANCE

10,000 is the most common result-size threshold

INTEROPERABILITY

DISCOVERABILITY

2.58% of the endpoints provide a VoID description
8.09% of the endpoints provide a SD description
91.22% of the endpoints have **no** description

Server name:

Figure 1: The 3DFed SPARQL Endpoint Monitoring Service.

Table 1: The number of endpoints and their percentage that contain the corresponding number of datasets.

Number of datasets	Number of different endpoints	Percentage
1	551	94.84%
2	18	3.10%
3	1	0.17%
4	2	0.34%
5	2	0.34%
6	2	0.34%
7	1	0.17%
8	1	0.17%
9	1	0.17%
14	1	0.17%
61	1	0.17%

Availability

The SPARQLES portal monitors the availability of endpoints, i.e., the ratio of time that a given endpoint is responsive via the SPARQL protocol. This is done by a generic ASK or SELECT query opting for any triple stored in the endpoint. This type of test is executed once per hour, so we can aggregate the results for each endpoint in a specific time interval, such as the last 24 hours or the last 7 days. At the time of writing the deliverable, only 101 endpoints were available, while 102 different endpoints were available at least once in the last 24 hours. In the last 7 days, this number increases to 109. In the last 24 hours, only 90 endpoints were available without interruption, meaning that all 24 tests were successful, while the number of endpoints available constantly in the last week is only 62. If we take into account the maximal time frame, i.e., last 11 months, from the availability chart (Figure 2) it can be concluded that there is a slight drop in the endpoint availability. Portions of the screenshot showing all endpoints and their availability over the last 24 hours and 7 days, sorted in descending order, are shown in Figure 3.

The MongoDB query shown in Figure 4 can be used to calculate the average percentage of endpoint availability. It groups all results of the availability tasks by endpoint URI, counts how many of them are considered as successful, and calculates its percentage. Then, it groups these endpoints again based on their average availability into the following groups: [0-5%], [5-55%], [5-95%], [95-99%], and [99-100%]. Its result shows that the majority of endpoints (463, or 79,69% of all endpoints) were almost always unavailable, i.e., their availability were recorded only in less than 5% of tests. There are 16 endpoints (2.75%) mostly unavailable, i.e., their average availability were in the range between 5% and 75% and 18 endpoints (3.10%) mostly available where the average availability varies between 75% and 95%. There were 33 (5.68%) reliable endpoints with availability measured in the range of 95% to 99%, and 51 (8.78%) very reliable endpoints whose availability is greater than 99%.

Comparing these results to the similar results from the original SPARQLES paper from 2013 [1], in which they analyzed 427 monitored endpoints, we notice that the availability of endpoints in general is much lower, both in absolute and relative numbers. For example, there were 137 very reliable endpoints (32.2% of total) with the average availability greater than 99%, which is more than twice as many as today. In other groups the situation is similar, except for the group with very low availability, i.e., the group where the average availability is less than 5%.

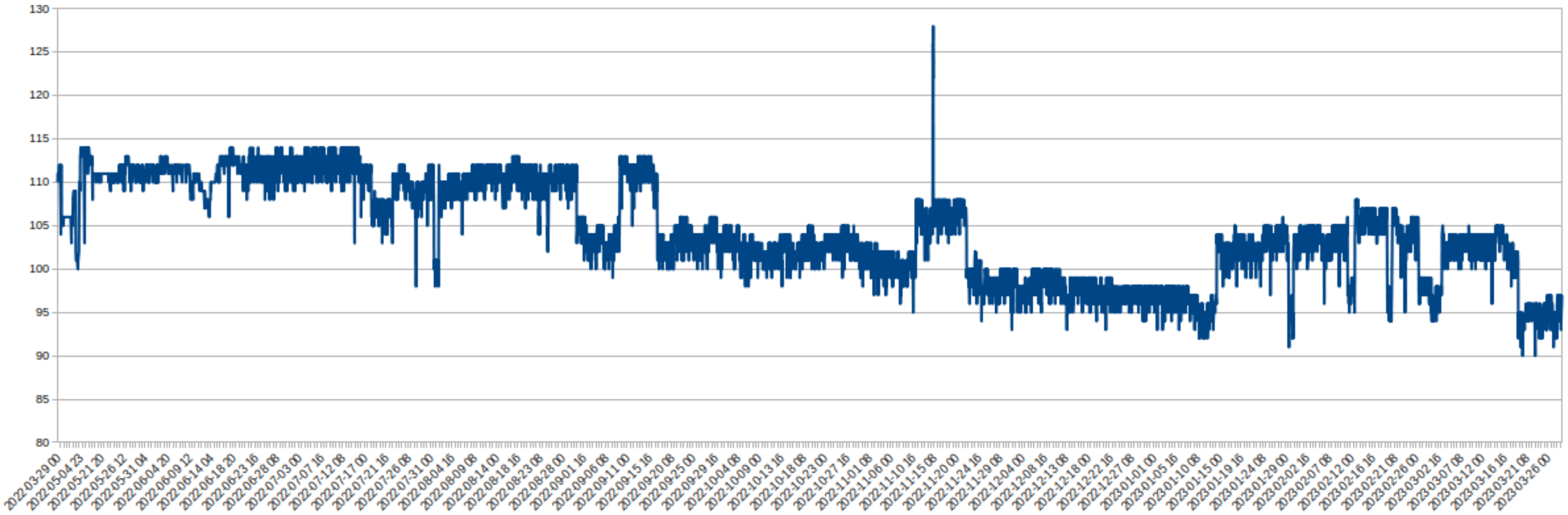
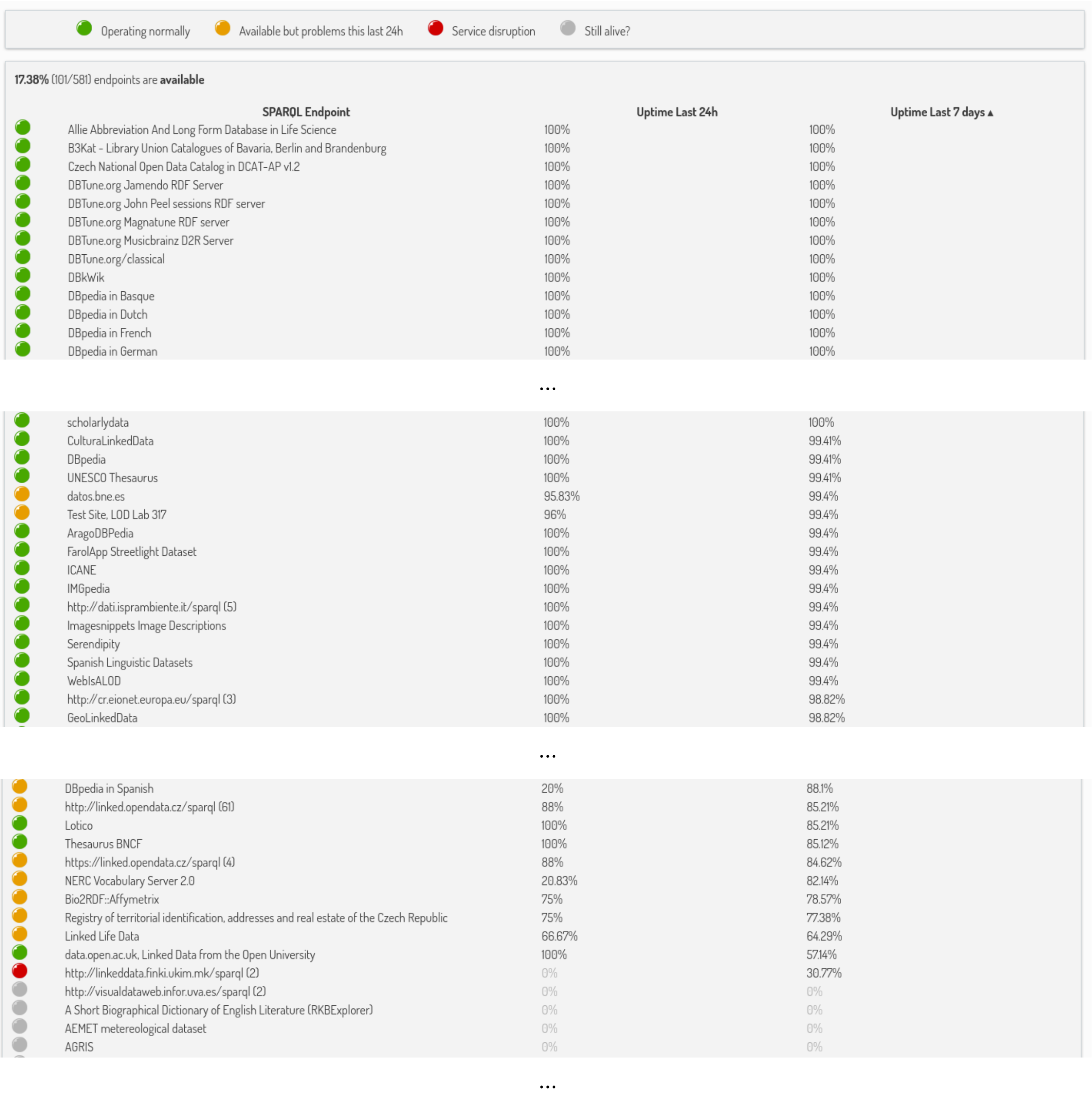


Figure 2: Number of available endpoints in the last 11 months, on an hourly basis.

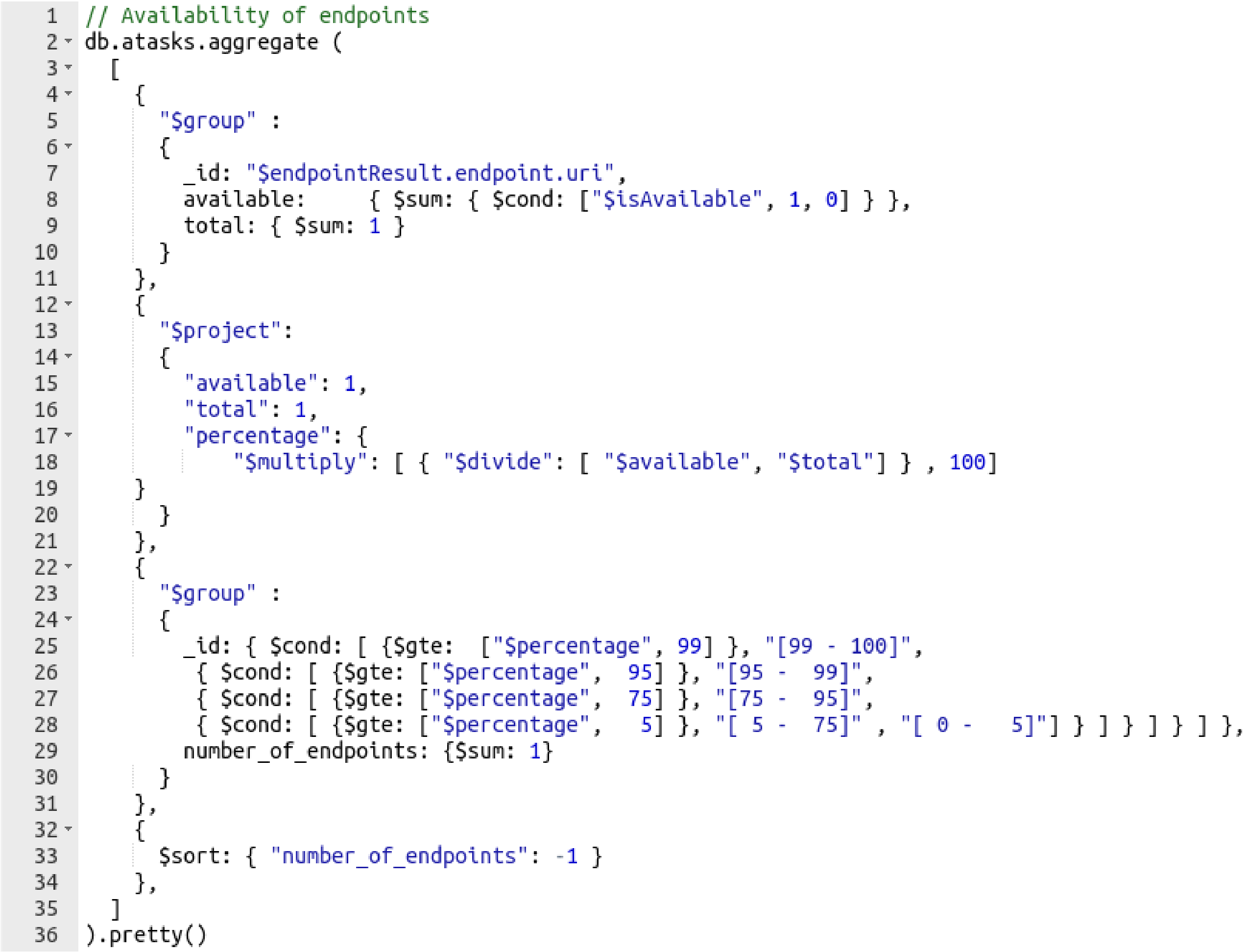


🟢 Operating normally 🟠 Available but problems this last 24h 🔴 Service disruption ⚪ Still alive?

17.38% (101/581) endpoints are **available**

	SPARQL Endpoint	**Uptime Last 24h**	**Uptime Last 7 days ▲**
🟢	Allie Abbreviation And Long Form Database in Life Science	100%	100%
🟢	B3Kat - Library Union Catalogues of Bavaria, Berlin and Brandenburg	100%	100%
🟢	Czech National Open Data Catalog in DCAT-AP v1.2	100%	100%
🟢	DBTune.org Jamendo RDF Server	100%	100%
🟢	DBTune.org John Peel sessions RDF server	100%	100%
🟢	DBTune.org Magnatune RDF server	100%	100%
🟢	DBTune.org Musicbrainz D2R Server	100%	100%
🟢	DBTune.org/classical	100%	100%
🟢	DBkWik	100%	100%
🟢	DBpedia in Basque	100%	100%
🟢	DBpedia in Dutch	100%	100%
🟢	DBpedia in French	100%	100%
🟢	DBpedia in German	100%	100%
	…		
🟢	scholarlydata	100%	100%
🟢	CulturaLinkedData	100%	99.41%
🟢	DBpedia	100%	99.41%
🟢	UNESCO Thesaurus	100%	99.41%
🟠	datos.bne.es	95.83%	99.4%
🟠	Test Site, LOD Lab 317	96%	99.4%
🟢	AragoDBPedia	100%	99.4%
🟢	FarolApp Streetlight Dataset	100%	99.4%
🟢	ICANE	100%	99.4%
🟢	IMGpedia	100%	99.4%
🟢	http://dati.isprambiente.it/sparql (5)	100%	99.4%
🟢	Imagesnippets Image Descriptions	100%	99.4%
🟢	Serendipity	100%	99.4%
🟢	Spanish Linguistic Datasets	100%	99.4%
🟢	WebIsALOD	100%	99.4%
🟢	http://cr.eionet.europa.eu/sparql (3)	100%	98.82%
🟢	GeoLinkedData	100%	98.82%
	…		
🟠	DBpedia in Spanish	20%	88.1%
🟠	http://linked.opendata.cz/sparql (61)	88%	85.21%
🟢	Lotico	100%	85.21%
🟢	Thesaurus BNCF	100%	85.12%
🟠	https://linked.opendata.cz/sparql (4)	88%	84.62%
🟠	NERC Vocabulary Server 2.0	20.83%	82.14%
🟠	Bio2RDF::Affymetrix	75%	78.57%
🟠	Registry of territorial identification, addresses and real estate of the Czech Republic	75%	77.38%
🟠	Linked Life Data	66.67%	64.29%
🟢	data.open.ac.uk, Linked Data from the Open University	100%	57.14%
🔴	http://linkeddata.finki.ukim.mk/sparql (2)	0%	30.77%
⚪	http://visualdataweb.infor.uva.es/sparql (2)	0%	0%
⚪	A Short Biographical Dictionary of English Literature (RKBExplorer)	0%	0%
⚪	AEMET metereological dataset	0%	0%
⚪	AGRIS	0%	0%
	…		

Figure 3: The 3DFed SPARQL Endpoint Monitoring Service: Availability page.



```
// Availability of endpoints
db.atasks.aggregate (
  [
    {
      "$group" :
      {
        _id: "$endpointResult.endpoint.uri",
        available:     { $sum: { $cond: ["$isAvailable", 1, 0] } },
        total: { $sum: 1 }
      }
    },
    {
      "$project":
      {
        "available": 1,
        "total": 1,
        "percentage": {
            "$multiply": [ { "$divide": [ "$available", "$total"] } , 100]
    }
      }
    },
    {
      "$group" :
      {
        _id: { $cond: [ {$gte:  ["$percentage", 99] }, "[99 - 100]",
         { $cond: [ {$gte: ["$percentage",  95] }, "[95 -  99]",
         { $cond: [ {$gte: ["$percentage",  75] }, "[75 -  95]",
         { $cond: [ {$gte: ["$percentage",   5] }, "[ 5 -  75]" , "[ 0 -   5]"] } ] } ] } ] },
        number_of_endpoints: {$sum: 1}
      }
    },
    {
      $sort: { "number_of_endpoints": -1 }
    },
  ]
).pretty()
```

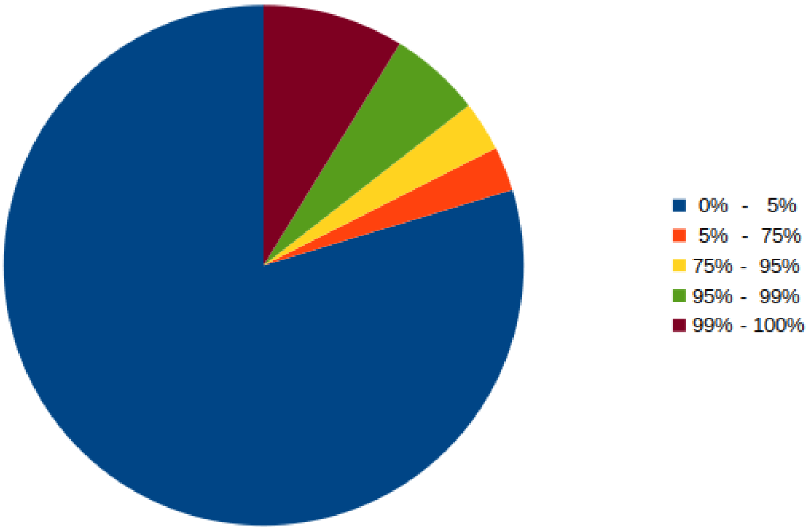


Figure 4: MongoDB query for summarizing the average availability of endpoints and a visual representation of the results.

Table 2: The number of endpoints and their percentage supporting the corresponding number of SPARQL features.

Number of SPARQL 1.0 features	Number of different endpoints	Percentage	Number of SPARQL 1.1 features	Number of different endpoints	Percentage
0	480	82.62%	0	481	82.79%
3	2	0.34%	1	1	0.17%
4	1	0.17%	5	3	0.52%
20	2	0.34%	7	3	0.52%
21	5	0.86%	9	1	0.17%
22	5	0.86%	10	1	0.17%
23	2	0.34%	11	1	0.17%
24	84	14.46%	12	1	0.17%
			13	3	0.52%
			16	1	0.17%
			17	40	6.88%
			18	45	7.75%

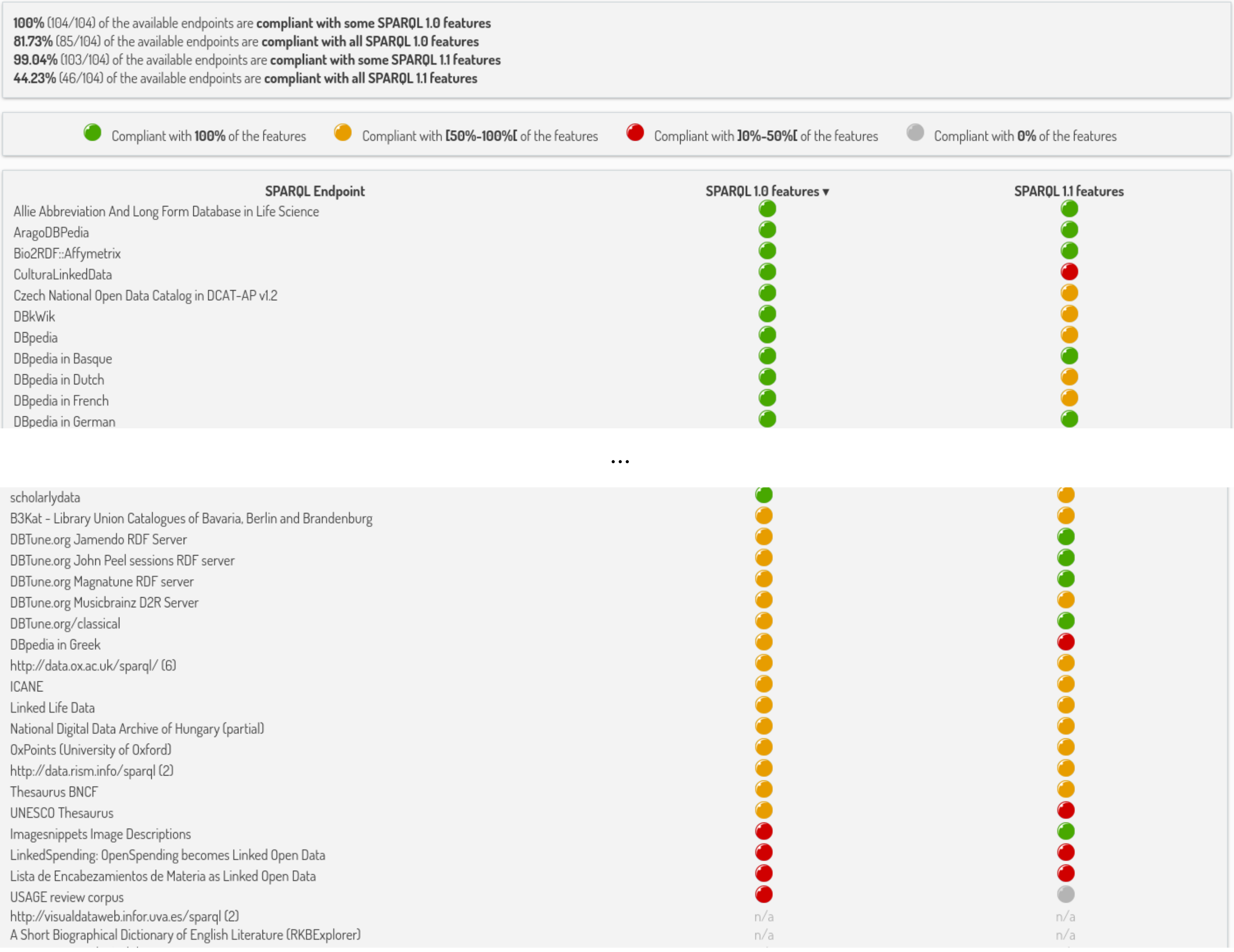
Interoperability

The SPARQLES portal analyses the interoperability of the monitored endpoints, i.e., it checks which SPARQL features (SPARQL 1.0 and SPARQL 1.1) are supported. For each endpoint, a series of queries of different types (SELECT, CONSTRUCT and ASK) are executed, containing specific operators (joins, unions, optionals, filters, negations, property-paths, binding, etc.), functions (regex, datatype, string functions, etc.) and solution modifiers (limit, order by, offset, distinct, etc). If an endpoint returns a valid SPARQL response (even though it may not be correct, which cannot be tested since the contents of the database are not known), the test is considered successful. If the query engine raises an exception, the test is unsuccessful and the corresponding feature is considered unsupported.

Since the SPARQL feature support of an endpoint is not a dynamic feature, i.e., it cannot be changed very often, this type of test is run once a week. In the last run while writing the deliverable, we found that most endpoints do not support neither of SPARQL 1.0 not SPARQL 1.1 features (these endpoints were not available at the time of this test, therefore the number of endpoints supporting 0 features corresponds to the number of unavailable endpoints in the last availability test). Table 2 summarizes the number of endpoints that support a different number of SPARQL features from both standards. Portions of the screenshot showing all endpoints and their compliance level with the standards, sorted in descending order, are shown in Figure 5. The numbers are lower than reported in [1] due to the same reason as in the case of availability.

Discoverability

The purpose of the SPARQLES portal discoverability analysis is to determine the extent to which endpoints offer descriptions of themselves and their content. It is tested whether SPARQL 1.1 Service Descriptions (SD) (describing the endpoint's capabilities in terms of query features, I/O formats or supported entailments, configuration of default and named graphs, etc.) and VoID metadata (describing an RDF dataset, including statistics about size, schema terms used, frequency of terms, etc.) are present. In addition, this type of test also



100% (104/104) of the available endpoints are **compliant with some SPARQL 1.0 features**
81.73% (85/104) of the available endpoints are **compliant with all SPARQL 1.0 features**
99.04% (103/104) of the available endpoints are **compliant with some SPARQL 1.1 features**
44.23% (46/104) of the available endpoints are **compliant with all SPARQL 1.1 features**

🟢 Compliant with **100%** of the features 🟠 Compliant with **[50%-100%[** of the features 🔴 Compliant with **]0%-50%[** of the features ⚪ Compliant with **0%** of the features

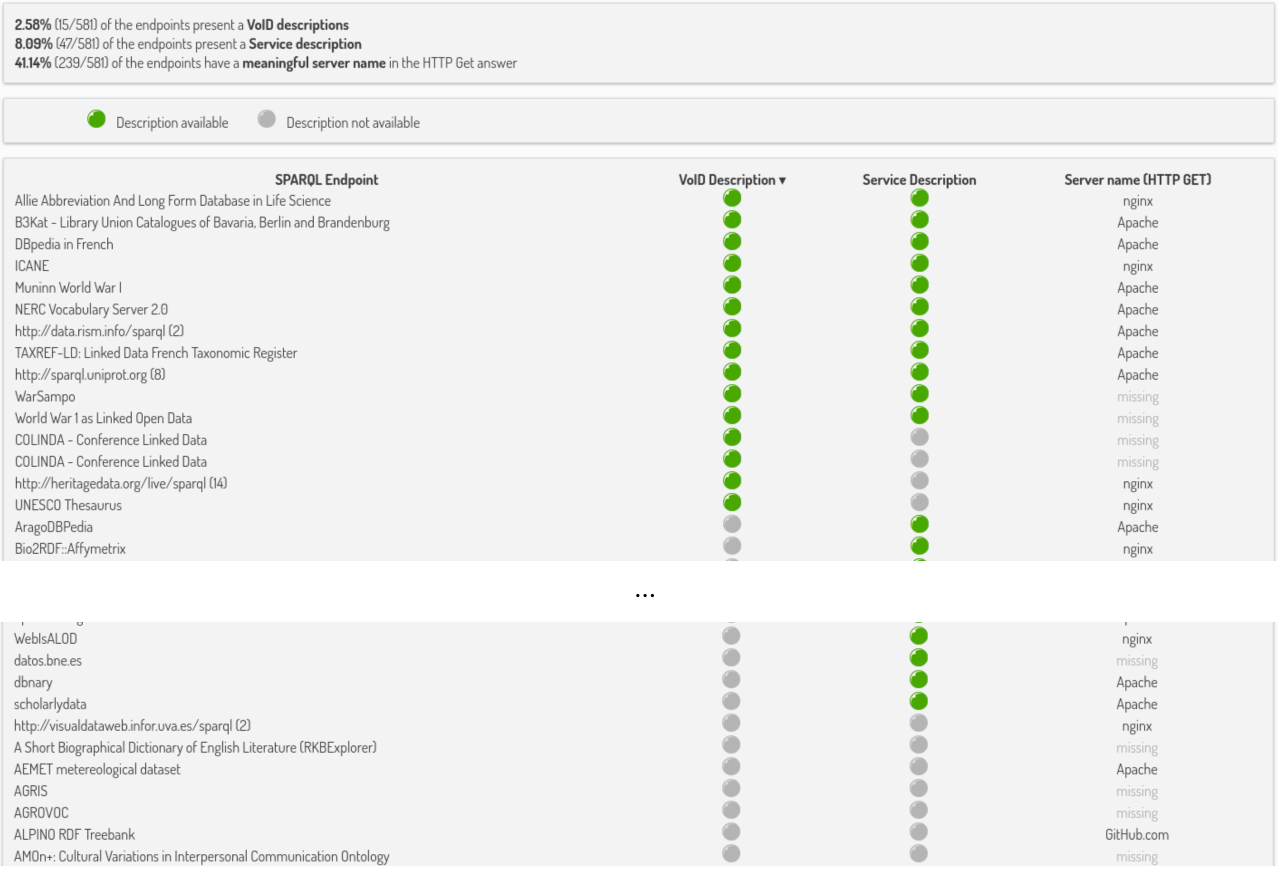
SPARQL Endpoint	SPARQL 1.0 features ▾	SPARQL 1.1 features
Allie Abbreviation And Long Form Database in Life Science	🟢	🟢
AragoDBPedia	🟢	🟢
Bio2RDF::Affymetrix	🟢	🟢
CulturaLinkedData	🟢	🔴
Czech National Open Data Catalog in DCAT-AP v1.2	🟢	🟠
DBkWik	🟢	🟠
DBpedia	🟢	🟠
DBpedia in Basque	🟢	🟢
DBpedia in Dutch	🟢	🟠
DBpedia in French	🟢	🟠
DBpedia in German	🟢	🟢

...

scholarlydata	🟢	🟠
B3Kat - Library Union Catalogues of Bavaria, Berlin and Brandenburg	🟠	🟠
DBTune.org Jamendo RDF Server	🟠	🟢
DBTune.org John Peel sessions RDF server	🟠	🟢
DBTune.org Magnatune RDF server	🟠	🟢
DBTune.org Musicbrainz D2R Server	🟠	🟠
DBTune.org/classical	🟠	🟢
DBpedia in Greek	🟠	🔴
http://data.ox.ac.uk/sparql/ (6)	🟠	🟠
ICANE	🟠	🟠
Linked Life Data	🟠	🟠
National Digital Data Archive of Hungary (partial)	🟠	🟠
OxPoints (University of Oxford)	🟠	🟠
http://data.rism.info/sparql (2)	🟠	🟠
Thesaurus BNCF	🟠	🟠
UNESCO Thesaurus	🟠	🔴
Imagesnippets Image Descriptions	🔴	🟢
LinkedSpending: OpenSpending becomes Linked Open Data	🔴	🔴
Lista de Encabezamientos de Materia as Linked Open Data	🔴	🔴
USAGE review corpus	🔴	⚪
http://visualdataweb.infor.uva.es/sparql (2)	n/a	n/a
A Short Biographical Dictionary of English Literature (RKBExplorer)	n/a	n/a

...

Figure 5: The 3DFed SPARQL Endpoint Monitoring Service: Interoperability page.



2.58% (15/581) of the endpoints present a **VoID descriptions**
8.09% (47/581) of the endpoints present a **Service description**
41.14% (239/581) of the endpoints have a **meaningful server name** in the HTTP Get answer

● Description available ○ Description not available

SPARQL Endpoint	VoID Description ▾	Service Description	Server name (HTTP GET)
Allie Abbreviation And Long Form Database in Life Science	●	●	nginx
B3Kat - Library Union Catalogues of Bavaria, Berlin and Brandenburg	●	●	Apache
DBpedia in French	●	●	Apache
ICANE	●	●	nginx
Muninn World War I	●	●	Apache
NERC Vocabulary Server 2.0	●	●	Apache
http://data.rism.info/sparql (2)	●	●	Apache
TAXREF-LD: Linked Data French Taxonomic Register	●	●	Apache
http://sparql.uniprot.org (8)	●	●	Apache
WarSampo	●	●	missing
World War 1 as Linked Open Data	●	●	missing
COLINDA - Conference Linked Data	●	○	missing
COLINDA - Conference Linked Data	●	○	missing
http://heritagedata.org/live/sparql (14)	●	○	nginx
UNESCO Thesaurus	●	○	nginx
AragoDBPedia	○	●	Apache
Bio2RDF::Affymetrix	○	●	nginx

...

WebIsALOD	○	●	nginx
datos.bne.es	○	●	missing
dbnary	○	●	Apache
scholarlydata	○	●	Apache
http://visualdataweb.infor.uva.es/sparql (2)	○	○	nginx
A Short Biographical Dictionary of English Literature (RKBExplorer)	○	○	missing
AEMET metereological dataset	○	○	Apache
AGRIS	○	○	missing
AGROVOC	○	○	missing
ALPINO RDF Treebank	○	○	GitHub.com
AMOn+: Cultural Variations in Interpersonal Communication Ontology	○	○	missing

...

Figure 6: The 3DFed SPARQL Endpoint Monitoring Service: Discoverability page.

examines the type of query engine running a SPARQL endpoint. This information may be relevant to a user, as different engines support some non-standard query features that may be important to a particular user's needs.

The frequency of this type of test is once a week. In the last while writing the deliverable, only 15 endpoints had VoID metadata available, representing 2.58% of the total number of all endpoints. This problem identified here (small number of endpoints with available VoID) will be addressed in Task 2.1, and for all endpoints for which there is no VoID, it will be automatically generated. In the case of SD, the situation is slightly better, and this type of description is present in 47 endpoints, representing 8.09% of all endpoints. These two features of the endpoints are less frequent than in the previous report [1]. Portions of the screenshot showing all endpoints, the relevant engines used to power them, and indicators whether their VoID and SD metadata are available, are shown in Figure 6. Table 3 gives an overview of the distribution of query engines corresponding to the endpoints.

Perfomance

The SPARQLES portal performs a series of performance-based tests on all monitored endpoints. It analyses three main aspects of a query engine (streaming, lookups, and joins) in a generic manner, regardless of the content stored in the endpoint. The streaming analysis is used to estimate the maximum throughput of the service, but also to determine the maximum number of items in the result sets. From the latest performance test, it appears that there is a single endpoint with the largest result set of less than 1,000 items, 28 endpoints with a maximum size of at most 10,000, 16 endpoints with a maximum size of at most 100,000, and 64 endpoints with a maximum size of more than 100,000. The second goal of these tests is to measure the time required to perform

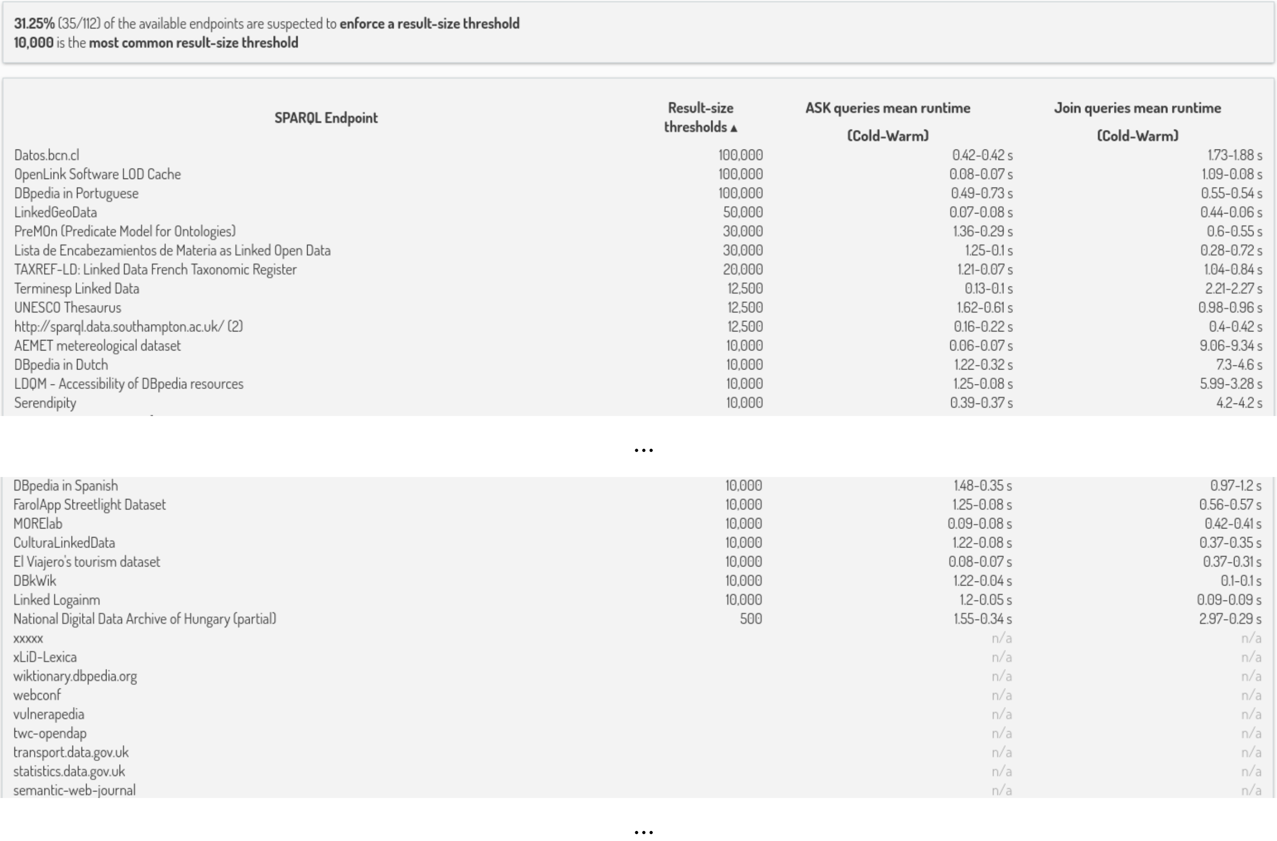
Table 3: Distribution of query engines across the endpoints.

Server name	Number of endpoints	Percentage
missing	342	58.86%
Apache	103	17.73%
nginx	72	12.39%
cloudflare	20	3.44%
envoy	15	2.58%
Apache-Coyote	14	2.41%
Virtuoso	3	0.52%
GitHub.com	2	0.34%
CloudFront	2	0.34%
Jetty	2	0.34%
openresty	1	0.17%
PasteWSGIServer	1	0.17%
Microsoft-IIS	1	0.17%
redir-httpd	1	0.17%
GlassFish Server Open Source Edition 3.1.2.2	1	0.17%
AmazonS3	1	0.17%

an atomic lookup. The platform runs 17 queries for each endpoint twice, the first time with a cold index and another time with a warm index. The left part of Table 4 shows the run times in seconds for these queries in percentiles. The third objective of these tests belongs to join analytics. It measures the generic join performance, namely s-s joins, s-o joins, and o-o joins. The right part of Table 4 shows the join performance results. For the available endpoints, the results are similar to the results presented in the paper [1]. Portions of the screenshot showing all endpoints and their results on the last performance test are shown in Figure 7.

Table 4: Runtimes for ASK and JOIN queries (%-iles).

ASK queries			**JOIN queries**		
Percentile	Time (cold)	Time (warm)	Percentile	Time (cold)	Time (warm)
0	0.03	0.03	0	0.07	0.05
25	0.15	0.07	25	0.28	0.18
50	1.18	0.14	50	1.31	1.16
75	1.25	0.28	75	2.53	1.99
90	1.51	0.67	90	5.74	4.77
100	4.98	6.24	100	20.87	14.45



31.25% (35/112) of the available endpoints are suspected to **enforce a result-size threshold**
10,000 is the **most common result-size threshold**

SPARQL Endpoint	Result-size thresholds ▲	ASK queries mean runtime	Join queries mean runtime
		(Cold-Warm)	(Cold-Warm)
Datos.bcn.cl	100,000	0.42-0.42 s	1.73-1.88 s
OpenLink Software LOD Cache	100,000	0.08-0.07 s	1.09-0.08 s
DBpedia in Portuguese	100,000	0.49-0.73 s	0.55-0.54 s
LinkedGeoData	50,000	0.07-0.08 s	0.44-0.06 s
PreMOn (Predicate Model for Ontologies)	30,000	1.36-0.29 s	0.6-0.55 s
Lista de Encabezamientos de Materia as Linked Open Data	30,000	1.25-0.1 s	0.28-0.72 s
TAXREF-LD: Linked Data French Taxonomic Register	20,000	1.21-0.07 s	1.04-0.84 s
Terminesp Linked Data	12,500	0.13-0.1 s	2.21-2.27 s
UNESCO Thesaurus	12,500	1.62-0.61 s	0.98-0.96 s
http://sparql.data.southampton.ac.uk/ (2)	12,500	0.16-0.22 s	0.4-0.42 s
AEMET metereological dataset	10,000	0.06-0.07 s	9.06-9.34 s
DBpedia in Dutch	10,000	1.22-0.32 s	7.3-4.6 s
LDQM - Accessibility of DBpedia resources	10,000	1.25-0.08 s	5.99-3.28 s
Serendipity	10,000	0.39-0.37 s	4.2-4.2 s
…			
DBpedia in Spanish	10,000	1.48-0.35 s	0.97-1.2 s
FarolApp Streetlight Dataset	10,000	1.25-0.08 s	0.56-0.57 s
MORElab	10,000	0.09-0.08 s	0.42-0.41 s
CulturaLinkedData	10,000	1.22-0.08 s	0.37-0.35 s
El Viajero's tourism dataset	10,000	0.08-0.07 s	0.37-0.31 s
DBkWik	10,000	1.22-0.04 s	0.1-0.1 s
Linked Logainm	10,000	1.2-0.05 s	0.09-0.09 s
National Digital Data Archive of Hungary (partial)	500	1.55-0.34 s	2.97-0.29 s
xxxxx		n/a	n/a
xLiD-Lexica		n/a	n/a
wiktionary.dbpedia.org		n/a	n/a
webconf		n/a	n/a
vulnerapedia		n/a	n/a
twc-opendap		n/a	n/a
transport.data.gov.uk		n/a	n/a
statistics.data.gov.uk		n/a	n/a
semantic-web-journal		n/a	n/a

…

Figure 7: The 3DFed SPARQL Endpoint Monitoring Service: Performance page.

2.2 Monitoring Data as an RDF Dataset

As part of the T2.2 activities, we developed a procedure for transforming the collected data by the monitoring service into an RDF dataset, i.e. an RDF Knowledge Graph. The SPARQLES monitoring service uses mongo as a database, which stores the collected data as a JSON object. Aside from the explicitly collected data, it also generates aggregate data, on a scheduled basis, which expresses condensed information regarding the characteristics which have been monitored in the past.

The data stored in mongo is selected as a collection of JSON objects, which are then properly mapped into an appropriate RDF schema. An example JSON entry in the database, which contains aggregated information about the availability of a given endpoint, is shown in Listing 1. As we can see, the entry contains information about the endpoint in question, its default dataset's URI and label, information about it being currently available, it being available over the past 24 hours, and it being available over the past 7 days. Additionally, it has a timestamp (in UNIX time) about when this last monitoring took place.

The RDF entity generated from the example JSON entry in the database, is shown in Listing 2. As we can see, the RDF entity contains the same information, now formatted as RDF, with proper classes and properties which ensure interoperability with tools and systems which use Semantic Web technologies, and work with RDF Knowledge Graphs. The main purpose of the graph data is to be used by the dynamic data exchange service and the 3DFed federation engine in order to have sufficient data about the available endpoints and make an informed algorithm-based decision about where to store data and where to read the needed data from.

Listing 1: Source JSON Object

```
{
   "_id": {"$oid":"641320f8e4b0ab570805617c"},
   "endpoint": {
      "uri":"http://pt.dbpedia.org/sparql",
      "datasets": [{
         "uri":"http://pt.dbpedia.org/sparql",
         "label":"DBpedia in Portuguese"
      }]
   },
   "upNow":true,
   "uptimeLast24h":1.0,
   "uptimeLast7d":1.0,
   "lastUpdate": {
      "$numberLong":"1679644809450"
   }
}
```

Listing 2: Output RDF Entity

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix sd: <http://www.w3.org/ns/sparql-service-description#> .
@prefix sp: <http://sparqles.demo.openlinksw.com/> .

sp:641320f8e4b0ab570805617c sp:lastUpdate "1679644809450" ;
    sp:upNow true ;
    sp:uptimeLast24h 1.0 ;
    sp:uptimeLast7d 1.0 ;
    sd:defaultDataset <http://pt.dbpedia.org/sparql> ;
    sd:endpoint <http://pt.dbpedia.org/sparql> .

<http://pt.dbpedia.org/sparql> rdfs:label "DBpedia in Portuguese" .
```

3 Conclusion

The main goal of T2.2 was to deploy an automatic tool for monitoring SPARQL endpoints, with URLs collected from Datahub, which is what our team has done with our modified SPARQLES instance. The purpose of this monitoring is to measure and report the uptime and availability, average response time for different types of queries, support for different version of the SPARQL standard, etc., which is exactly what our tool does.

Additionally, the collected data is transformed into an RDF dataset, for improved interoperability of the tool with other semantic tools, which work with RDF, Linked Data and Knowledge Graphs.

The public availability of the modified platform, the collected data and the source code, significantly increase the impact of the project efforts.

References

[1] Carlos Buil-Aranda, Aidan Hogan, Jürgen Umbrich, and Pierre-Yves Vandenbussche. SPARQL Web-Querying Infrastructure: Ready for Action? In *The Semantic Web – ISWC 2013*, pages 277–293, Berlin, Heidelberg, 2013. Springer.