ParlBench: a SPARQL-benchmark for electronic publishing applications

Tatiana Tarasova and Maarten Marx

ISLA, University of Amsterdam, Science Park 904, 1098 XH Amsterdam
`T.Tarasova@uva.nl, maartenmarx@uva.nl` *

Abstract. ParlBench is a scalable RDF benchmark modelling a large scale electronic publishing scenario. The benchmark offers large collections of the Dutch parliamentary proceedings together with information about members of the parliament and political parties. The data is real, but free of intellectual property rights issues. On top of the benchmark data sets, several realistic application benchmarks as well as targeted micro benchmarks can be developed. This paper describes the benchmark data sets and 28 analytical queries covering a wide range of SPARQL constructs. The potential use of ParlBench is demonstrated by executing the query set for 8 different scalings of the benchmark data sets on Virtuoso RDF store. Measured on a standard laptop, data loading times varied from 43 seconds (for 1% of the data set) to 48 minutes (for the complete data set), and execution of the complete set of queries (1520 queries in total) varied from 9 minutes to 13 hours.

Keywords: SPARQL, RDF benchmark, parliamentary proceedings

1 Introduction

RDF stores are the backbones of RDF data driven applications. There is a wide range of RDF stores systems available[1] together with various benchmark systems[2] to assess performances of the systems.

As discussed in the Benchmark Handbook [1], different applications impose different requirements to a system, and the performance of the system may vary from one application domain to another. This creates the need for domain specific benchmarks. The existing application benchmarks for RDF store systems often employ techniques developed by the Transaction Processing Performance Council [4] (TPC) for relational databases and use synthetically generated data sets for their workloads. However, performance characteristics for loading and querying such data may differ from those that were measured on real life data sets, as it was shown by the DBpedia benchmark [2] on DBpedia [8]. To the best of our knowledge, among the existing benchmarks for RDF store systems, only the DBpedia benchmark provides a real data set.

With this work we propose the ParlBench application benchmark that closely mimics a real-life scenario: large scale electronic publishing with OLAP-type queries. ParlBench consists of (1) real life data and (2) a set of analytical queries developed on top of these data.

* This work was supported by the EU- FP7 (FP7/2007-2013) project ENVRI (grant number 283465).

[1] `http://www.w3.org/wiki/LargeTripleStores`

[2] `http://www.w3.org/wiki/RdfStoreBenchmarking`

The benchmark data sets include the Dutch parliamentary proceedings, political parties and politicians. The ParlBench data fit very well the desiderata of Gerhard Weikum's recent Sigmod blog[3]: it is open, big, real, useful, linked to other data sources, mixing data-values and free text, and comes with a number of real-life workloads.

The queries in the benchmark can be viewed as coming from one of two use cases: create a report or perform a scientific research. As an example of the latter, consider the question whether the performance of males and females differs in parliament, and how that has changed over the years. To enable more comprehensive analysis of the RDF stores' performances, we grouped the benchmark queries into four micro benchmarks [5] with respect to the their analytical aims *Average*, *Count*, *Factual* and *Top 10*.

The paper is organized as follows. Section 2 gives an overview of the related work. Section 3 describes the benchmark data sets. In Section 4 we define the benchmark queries and present the micro benchmarks. The evaluation of the ParlBench benchmark on the Virtuoso RDF store is discussed in Section 5.

2 Related Work

There is a number of RDF store benchmarks available. The most relevant benchmarks to our work are discussed further. The Berlin SPARQL Benchmark (BSBM) [6] implements an e-commerce application scenario. Similarly to ParlBench, BSBM employed the TPC [4] techniques, such as query permutations (for the Business Intelligence use case) and system ramp-up.

The SPARQL Performance Benchmark (SP^{2}Bench) [7] is settled in the DBLP scenario. SP^{2}Bench queries are carefully designed to test the behavior of RDF stores in relation to common SPARQL constructs, different operator constellations and RDF access patterns. SP^{2}Bench measures query response time in cold runs settings, i.e., when query execution time is measured immediately after the server was started.

Both the Berlin and SP^{2}Bench use synthetically generated data sets, whereas, the DBpedia SPARQL Benchmark (DBPSB) [2] uses a real data set, DBpedia [8]. In addition to using a real data set, the DBPSB benchmark uses real queries that were issued by humans and applications against DBpedia. These queries cover most of the SPARQL features and enable comprehensive analysis of RDF stores' performance on a single feature as well as combinations of features. The main difference between the ParlBench and DBPSB benchmarks is that the latter is not developed with a particular application in mind. Thus, it is more useful for a general assessment of the performance of different RDF stores' implementations, while ParlBench is particularly targeted on developers of e-publishing applications and can support them in choosing systems that are more suitable for analytical query processing.

3 Benchmark Data Sets

The benchmark consists of five conceptually separate data sets summarized in Table 1:

Members : describes political players of the Dutch parliament.
Parties : describes Dutch political parties.

[3] `http://wp.sigmod.org/?p=786`

Proceedings : describes the structure of the Dutch parliamentary proceedings.
Paragraphs : contains triples linking paragraphs to their content.
Tagged entities : contains triples linking paragraphs to DBpedia entities indicating that these entities were discussed in the paragraphs.

Fig. 1. Statistics of the benchmark datasets

dataset	**# of triples**	**size**	**# of files**
members	33,885	14M	3,583
parties	510	612K	151
proceedings	36,503,688	4.15G	51,233
paragraphs	11,250,295	5.77G	51,233
tagged entities	34,449,033	2.57G	34,755
TOTAL:	82,237,411	∼13G	140,955

The data model of the benchmark data sets is described in Appendix A.

3.1 Scaling of the Benchmark Data Sets

The size of the ParlBench data sets can be changed in different ways. The data set can be scaled by the number of included proceedings. All proceeding files are ordered chronologically. The scaled data set of size $1/n$ consists of every n-th file in this list, plus the complete *Parties* and *Members* sets. Optionally, one can include *Tagged entities* and/or *Paragraphs* data sets to the test collection. In this case *Paragraphs* and *Tagged Entities* are scaled accordingly to the included proceedings, i.e., only paragraphs and/or tags pointing to id's in the chosen proceedings are included.

4 Benchmark Queries

ParlBench provides 19 SPARQL queries. The queries were grouped into four micro benchmarks:

Average: 3 queries, numbered from `A0` to `A2`, retrieve aggregated information.
Count: 5 queries, numbered from `C0` to `C4`, count entities that satisfy certain conditions.
Factual: 6 queries, numbered from `F0` to `F5`, retrieve instances of a particular class that satisfy certain conditions.
Top 10: 5 queries, numbered from `T0` to `T4`, retrieve the top 10 instances of a particular class that satisfy certain filtering conditions.

All the queries are listed in Appendix B. Their SPARQL representations can be seen in Appendix C. The benchmark queries cover a wide range of the SPARQL language constructs. Table 1 shows the usage of SPARQL features by individual query and distribution of the features across micro benchmarks.

5 Experimental Run of the Benchmark

In this section we demonstrate the application of our benchmark on the OpenLink Virtuoso RDF native store (OSE)[4]. Tested on the Berlin benchmark, Virtuoso showed one of the best performance results among other systems [6].

[4] `http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/`

Table 1. SPARQL characteristics of the benchmark queries.

micro benchmark	**Average**			**Count**					**Factual**						**Top 10**				
	A0	**A1**	**A2**	**C0**	**C1**	**C2**	**C3**	**C4**	**F0**	**F1**	**F2**	**F3**	**F4**	**F5**	**T0**	**T1**	**T2**	**T3**	**T4**
FILTER						+	+	+	+	+	+	+		+					
UNION	+	+	+					+		+	+	+					+		+
LIMIT										+			+		+	+	+	+	+
ORDER BY										+			+		+	+	+	+	+
GROUP BY	+	+	+				+	+		+			+		+	+	+	+	+
COUNT	+	+	+	+	+	+	+	+		+	+	+	+		+	+	+	+	+
DISTINCT									+			+	+	+					
AVG	+	+	+																
negation														+					
OPTIONAL									+					+					
subquery	+	+	+					+			+	+	+						
blank node scoping				+	+	+	+	+	+	+	+	+							
# of triple patterns	10	9	12	5	5	5	6	13	8	16	6	6	2	4	2	4	9	3	11

5.1 Experimental Setup

Test Environment For the benchmark experiment we used a personal laptop Apple MacBook Pro with Intel i7 CPU (2x2 cores) running at 2.8 GHz and 8GB memory. See Appendix D for a more detailed specification of the test environment.

Evaluation Metrics

Loading Time The loading time is the time for loading RDF data into an RDF store. The benchmark data sets are in RDF/XML format. The time is measured in seconds. Loading of data into Virtuoso was done one data set at a time. For the loading of *Parties* and *Members* we used the Virtuoso RDF bulk load procedure. For *Proceedings* we used the Virtuoso function `DB.DBA.RDF_LOAD_RDFXML_MT` to load large RDF/XML text.

Query Response Time The query response time is the time it takes to execute a SPARQL query. To run the queries programmatically, we used `isql`, the Virtuoso interactive SQL utility. The execution time of a single query was taken as the real time returned by the bash `/usr/bin/time` command. 10 permutations of the benchmark queries were created, each containing 19 SPARQL queries.

Before starting measuring the query response time, we *warmed-up* Virtuoso by running 5 times 10 different permutations of all 19 queries of the benchmark. In total, 950 queries were executed in the *warm-up* phase, and each query was run 50 times. After that, we run the same permutations 3 more times and measured the execution time of each query. The query response time was computed as the mean response time of executing each query 30 times.

Test Collections Experiments are run on 8 test collections. Each collection includes the *Parties* and *Members* data sets and a scaled *Proceedings* data set ranging from 1 to 100% . Table 2 gives an overview of the sizes of each test collection.

Table 2. Sizes of the test collections for different scalings of *Proceedings*

Scaling Factor	1%	2%	4%	8%	16%	32%	64%	100%
# of triples	494,875	1,027,395	1.906,880	3,851,642	7,554,304	15,129,621	23,341,602	36,542,431

5.2 Results

We report on three experiments, relating database size to execution time: (1) time needed to load the test collections (fig. 2), (2) total time needed to execute all the queries in micro benchmarks[5] (fig. 3), and (3) query execution time of all the queries on the largest collection (fig. 4).

The y-axes on fig. 2 and fig. 3 are presented in a log scale, and the numbers represent the loading and query response time in seconds. Appendix E contains larger versions of these plots.

To make the results reproducible, we publish the benchmark data sets, queries and scripts at `http://data.politicalmashup.nl/RDF/data/`.

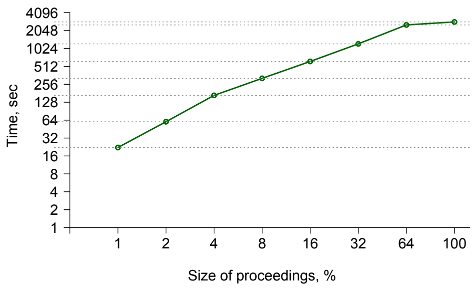


Fig. 2. Loading Time in sec of the benchmark collections

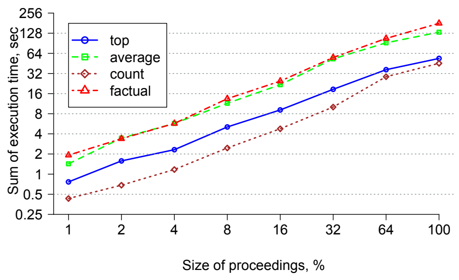


Fig. 3. Query Execution Time in sec of micro benchmarks on the test collections

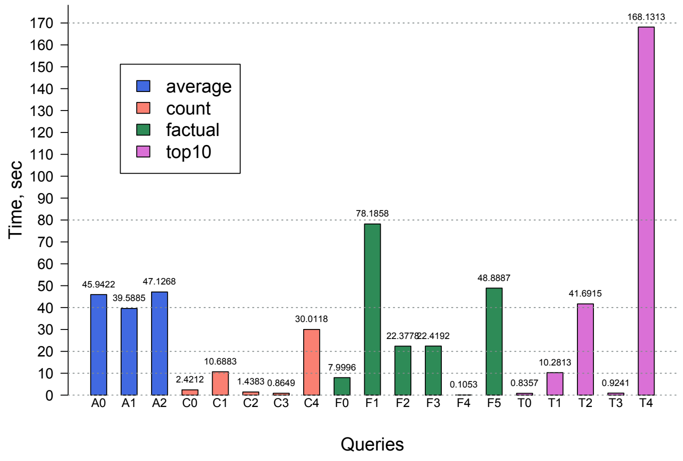


Fig. 4. Query Execution Time in sec of the benchmark queries on the largest test collection

[5] For each group we summed the execution time of each query in the group.

6 Conclusion

ParlBench has the proper characteristics of an RDF benchmark: it can be scaled easily and it has a set of intuitive queries which measure different aspects of the SPARQL engine.

We believe that ParlBench is a good proxy for a realistic large scale digital publishing application. ParlBench provides real data that encompass major characteristics shared by most of the e-publishing use cases including rich metadata and hierarichal organization of the content into text chunks.

The data set is large enough to perform non-trivial experiments. In addition to the analytical scenario presented, one can think of several other application scenarios that can be developed on the same data sets. Due to the many and strong connections of the benchmark to the Linked Open Data Cloud through the DBpedia links, natural Linked Data integration scenarios can be developed from ParlBench. The ParlBench data is also freely available in XML format [9], enabling cross-platform comparisons of the same workload.

As a future work we will consider the execution of the benchmark on multiple RDF stores and comparison of the results with the ones achieved on Virtuoso.

Another interesting direction for future work could be to extend the set of queries. Currently, there are only two queries with the `OPTIONAL` operator, which was proved to be the reason of the high complexity of the SPARQL language [10]. Queries that use features of SPARQL 1.1 could be a good addition to the benchmark. ParlBench has many queries that extensively use the `UNION` operator to explore the transitive `hasPart` relation. We could re-write these queries through the SPARQL 1.1. path expressions. Such queries would be a good ground to test the reasoning capabilities of RDF store systems.

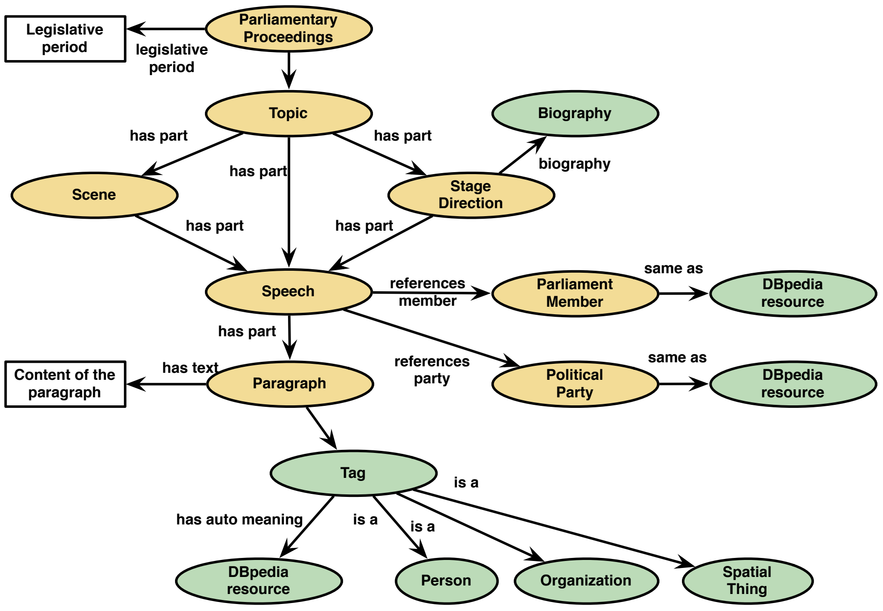
References

1. Gray, J.: The Benchmark Handbook for Database and Transaction Systems, (2nd Edition), Morgan Kaufmann, ISBN 1-55860-292-5.
2. Morsey, M., Lehmann, J., Auer, S., Ngomo, A.-C. N.: DBpedia SPARQL benchmark: performance assessment with real queries on real data. In *ISWC*, (2011). LNCS, vol. 7031, pp. 454-469. Springer, Heidelberg (2011).
3. Transaction Processing Performance Council (2008): TPC Benchmark H, Standard Specification Revision 2.7.0. Retrieved March 2, (2009), `http://www.tpc.org/tpch/spec/tpch2.7.0.pdf`
4. Transaction Processing Performance Council. `http://www.tpc.org/`
5. Afanasiev L., Manolescu I., Michiels P.: MemBeR: a micro-benchmark repository for XQuery. In *XSym* (2005). LNCS, vol. 3671, pp. 144–161. Springer, Heidelberg (2005).
6. Bizer, C., Schultz, A.: The Berlin SPARQL benchmark. *Int. J. On Semantic Web and Information Systems.* 5(2), 1-24 (2009).
7. Schmidt, M., Hornung, T., Lausen, G., Pinkel, C.: SP^{2}Bench: A SPARQL Performance Benchmark. In *ICDE*, 222-233 (2009).
8. Bizer C., Lehmann J., Kobilarov G., Auer S., Becker C., Cyganiak R., Hellmann S.: DBpedia - a crystallization point for the web of data. *J. of Web Semantics* 7(3), 154–165 (2009).
9. Maarten M.: Advanced Information Access to Parliamentary Debates. *J. of Dig. Inf.*. 10(6), (2009).
10. Pérez, J., Arenas, M., Gutierrez, C.: Semantics and complexity of SPARQL *ACM Trans. Database Syst.*. 34(3), (2009).

A Data Model

Figure 5 gives an overview of the data model.

Fig. 5. Data model of the benchmark datasets.



The data model consists of the following classes:

- `PoliticalParty` represents political parties. Parties are linked to their equivalent resources in DBpedia.
- `ParliamentMember` represents politicians, i.e., members of the parliament. Members are linked to the DBpedia resources that correspond to the same politicians. Additionally, members have biographical information such as gender, birthday and the place of birth, and the death date and place if applicable.
- `ParliamentaryProceedings` are written records of parliamentary meetings.
- `Topic` represents a different point on the agenda in the proceedings.
- `Scene` and `StageDirection` are important structural elements of the Dutch parliamentary proceedings. They contain information about current speakers and their interrupters. We didn't include this information into the benchmark data sets, but we kept the elements to have a complete structure of the proceedings.
- `Speech` is a constituent of a *Topic*, a *Scene* or *Stage Direction*. *Speech* represents a beginning of a speech of a new speaker.
- `Paragraph` is a container for all spoken text; can be part of any structural element of the proceedings described above.

The structural elements of the proceedings are connected to `Proceedings` through the `dcterms:hasPart` property. The speaker of the speech and the affiliated party of the speaker are attached to `Speech` via the `refMember` and `refParty` properties correspondingly.

Vocabularies Relevant existing vocabularies and ontologies to model documents of the parliamentary proceedings fall into two categories. In the first category there are vocabularies that are too generic, such as the SALT Document Ontology[6] or the DoCo, the Document Components Ontology[7]. They do not provide means to represent such specific concepts as *stage direction* or *scene*. Vocabularies in the second category are too specific, like the Semantic Web Conference Ontology[8] or the Semantic Web Portal Ontology[9] which model proceedings of conferences.

We defined our own RDF vocabulary to model parliamentary proceedings, the Parliamentary Proceedings vocabulary[10], and integrated it with other existing vocabularies. To represent biographical information of politicians, we used: BIO: A vocabulary for biographical information[11] together with the Friend of a Friend Vocabulary (FOAF) [12] and the DBpedia Ontology[13]. The Modular Unified Tagging Ontology[14] (MUTO) was used to represent information about tagged entities of paragraphs. The Dublin Core Metadata Terms[15] was used to encode metadata information.

[6] http://salt.semanticauthoring.org/ontologies/sdo#
[7] `http://purl.org/spar/doco/Paragraph`
[8] `http://data.semanticweb.org/ns/swc/swc_2009-05-09.html#`
[9] `http://sw-portal.deri.org/ontologies/swportal#`
[10] `http://purl.org/vocab/parlipro#`
[11] `http://vocab.org/bio`
[12] http://xmlns.com/foaf/0.1/
[13] `http://dbpedia.org/ontology/`
[14] `http://muto.socialtagging.org/`
[15] `http://purl.org/dc/terms/` and `http://purl.org/dc/elements/1.1/`

B ParlBench Queries

Table 3. List of the benchmark queries

		average
11.	A0	Retrieve average number of people spoke per topic.
12.	A1	Retrieve average number of speeches per topic.
13.	A2	Retrieve average number of speeches per day.
		count
1.	C0	Count speeches of females.
2.	C1	Count speeches of males.
3.	C2	Count speeches of speakers who were born after 1960.
4.	C3	Count speeches of male speakers who were born after 1960.
5.	C4	Count speeches of a female speaker from the topic where only one female spoke.
		factual
14.	F0	What members were born after 1950, their parties and dates of death of exist?
15.	F1	What gender of politicians who spoke most within a certain timeframe?
16.	F2	What is the percentage of male speakers?
17.	F3	What is the percentage of female speakers?
18.	F4	What politician has most number of Wikipedia pages in different languages?
19.	F5	What speeches are made by politicians without Wikipedia pages?
		top 10
6.	T0	Retrieve top 10 members with the most speeches.
7.	T1	Retrieve top 10 topics when most of the people spoke.
8.	T2	Retrieve top 10 topics with the most speeches.
9.	T3	Retrieve top 10 days with the most topics.
10.	T4	Retrieve top 10 longest topics (i.e., number of paragraphs).

C ParlBench SPARQL Queries

Table 4. Prefixes used in the SPARQL queries.

rdf:	`<http://www.w3.org/1999/02/22-rdf-syntax-ns#>`
parlipro:	`<http://purl.org/vocab/parlipro#>`
dcterms:	`<http://purl.org/dc/terms/>`
dc:	`<http://purl.org/dc/elements/1.1/>`
bio:	`<http://purl.org/vocab/bio/0.1/>`
foaf:	`<http://xmlns.com/foaf/0.1/>`
dbpedia:	`<http://dbpedia.org/resource/>`
owl:	`<http://www.w3.org/2002/07/owl#>`

Table 5: SPARQL representation of the benchmark queries.

A0: Retrieve average number of people spoke per topic.

```
SELECT AVG(?numOfMembers) as ?avgNumOfMembersPerTopic
WHERE {{
SELECT COUNT(?member) AS ?numOfMembers
WHERE {
  ?topic rdf:type parlipro:Topic .
  ?speech rdf:type parlipro:Speech .
  ?speech parlipro:refMember ?member .
  {?topic dcterms:hasPart ?speech .}
UNION{
  ?topic dcterms:hasPart ?sd .
  ?sd rdf:type parlipro:StageDirection .
  ?sd dcterms:hasPart ?speech .}
UNION{
  ?topic dcterms:hasPart ?scene .
  ?scene rdf:type parlipro:Scene .
  ?scene dcterms:hasPart ?speech .}}
GROUP BY ?topic}}
```

A1: Retrieve average number of speeches per topic.

```
SELECT AVG(?numOfSpeeches) as ?avgNumOfSpeechesPerTopic
WHERE {{
SELECT COUNT(?speech) AS ?numOfSpeeches
WHERE {
  ?topic rdf:type parlipro:Topic .
  ?speech rdf:type parlipro:Speech .
  {?topic dcterms:hasPart ?speech .}
UNION{
  ?topic dcterms:hasPart ?sd .
  ?sd rdf:type parlipro:StageDirection .
  ?sd dcterms:hasPart ?speech .}
UNION{
  ?topic dcterms:hasPart ?scene .
  ?scene rdf:type parlipro:Scene .
  ?scene dcterms:hasPart ?speech .}}
GROUP BY ?topic}}
```

A2: Retrieve average number of speeches per day.

```
SELECT AVG(?numOfSpeeches) as ?avgNumOfSpeechesPerDay
WHERE {{
SELECT ?date COUNT(?speech) AS ?numOfSpeeches
WHERE {
  ?proc dcterms:hasPart ?topic .
  ?proc rdf:type parlipro:ParliamentaryProceedings .
  ?proc dc:date ?date .
  ?speech rdf:type parlipro:Speech .
  ?topic rdf:type parlipro:Topic .
  {?topic dcterms:hasPart ?speech .}
UNION{
  ?topic dcterms:hasPart ?sd .
  ?sd rdf:type parlipro:StageDirection .
  ?sd dcterms:hasPart ?speech .}
UNION{
  ?topic dcterms:hasPart ?scene .
  ?scene rdf:type parlipro:Scene .
  ?scene dcterms:hasPart ?speech .}}
GROUP BY ?date}}
```

C0: Count speeches of females.

```
SELECT COUNT(?speech)
WHERE {
   ?speech rdf:type parlipro:Speech .
   ?speech parlipro:refMember ?member .
   ?member bio:biography _:bio .
   _:bio rdf:type bio:Biography .
   _:bio foaf:gender dbpedia:Female .}
```

C1: Count speeches of males.

```
SELECT COUNT(?speech)
WHERE {
   ?speech rdf:type parlipro:Speech .
   ?speech parlipro:refMember ?member .
   ?member bio:biography _:bio .
   _:bio rdf:type bio:Biography .
   _:bio foaf:gender dbpedia:Male .}
```

C2: Count speeches of speakers who were born after 1960.

```
SELECT COUNT(?speech)
WHERE {
  ?speech rdf:type parlipro:Speech .
  ?speech parlipro:refMember ?member .
  ?member bio:biography _:bio .
  _:bio rdf:type bio:Biography .
  _:bio foaf:birthday ?birthday .
FILTER (year(?birthday) > 1960)}
```

C3: Count speeches of male speakers who were born after 1960.

```
SELECT COUNT(?speech)
WHERE {
  ?speech rdf:type parlipro:Speech .
  ?speech parlipro:refMember ?member .
  ?member bio:biography _:bio .
  _:bio rdf:type bio:Biography .
  _:bio foaf:gender dbpedia:Male .
  _:bio foaf:birthday ?birthday .
FILTER (year(?birthday) > 1960)}
```

C4: Count speeches of a female speaker from the topic where only one female spoke.

```
SELECT ?topic ?member COUNT(?speech) as ?numOfSpeeches
WHERE {{
SELECT ?topic ?member COUNT(?member) AS ?numOfFemales ?speech
WHERE {
  ?topic rdf:type parlipro:Topic .
  ?speech rdf:type parlipro:Speech .
  ?speech parlipro:refMember ?member .
  ?member bio:biography _:bio .
  _:bio rdf:type bio:Biography .
  _:bio foaf:gender dbpedia:Female .
  {?topic dcterms:hasPart ?speech .}
UNION{
  ?topic dcterms:hasPart ?sd .
  ?sd rdf:type parlipro:StageDirection .
  ?sd dcterms:hasPart ?speech .}
UNION{
  ?topic dcterms:hasPart ?scene .
  ?scene rdf:type parlipro:Scene .
  ?scene dcterms:hasPart ?speech .}}
GROUP BY ?topic ?member ?speech}
FILTER (?numOfFemales = 1 )}
GROUP BY ?topic ?member
```

F0: What members were born after 1950, their parties and dates of death of exist?

```
SELECT DISTINCT ?member ?party ?birthday
WHERE {
  ?speech rdf:type parlipro:Speech .
  ?speech parlipro:refMember ?member.
  ?speech parlipro:refParty ?party .
  ?member rdf:type parlipro:ParliamentMember .
  ?member bio:biography _:bio .
  _:bio rdf:type bio:Biography .
  _:bio foaf:birthday ?birthday .
FILTER (year(?birthday) > 1960)}
FILTER (year(?birthday) > 1950)
OPTIONAL{_:bio dbpedia-ont:deathDate ?deathDate .}}
```

F1: What gender of politicians who spoke most within a certain timeframe?

```
SELECT ?gender COUNT(?member) AS ?numOfMembers
WHERE {
  ?proc rdf:type parlipro:ParliamentaryProceedings .
  ?proc dc:date ?date .
  ?proc dcterms:hasPart ?topic .
  ?speech rdf:type parlipro:Speech .
  ?speech parlipro:refMember ?member .
  ?member rdf:type parlipro:ParliamentMember .
  ?member bio:biography _:bio .
  _:bio rdf:type bio:Biography .
  _:bio foaf:gender ?gender .
  {?topic dcterms:hasPart ?speech .}
UNION{
  ?topic dcterms:hasPart ?sd .
  ?sd rdf:type parlipro:StageDirection .
  ?sd dcterms:hasPart ?speech .}
UNION{
  ?topic dcterms:hasPart ?scene .
  ?scene rdf:type parlipro:Scene .
  ?scene dcterms:hasPart ?speech .}
FILTER (year(?date) > 1995 AND year(?date) < 2005)}
GROUP BY ?gender
ORDER BY DESC(?numOfMembers)
LIMIT 1
```

F2: What is the percentage of male speakers?

```
SELECT (?numOfMaleMembers*100)/?numOfMembers
WHERE{{
SELECT COUNT(DISTINCT ?memberMale) as ?numOfMaleMembers
       COUNT(DISTINCT ?member) as ?numOfMembers
WHERE {{
   ?speech rdf:type parlipro:Speech .
   ?speech parlipro:refMember ?member .
   ?member rdf:type parlipro:ParliamentMember .}
UNION{
   ?memberMale bio:biography _:bio .
   _:bio rdf:type bio:Biography .
   _:bio foaf:gender dbpedia:Male .
FILTER (sameTerm(?member,?memberMale))}}}}
```

F3: What is the percentage of female speakers?

```
SELECT (?numOfFemaleMembers*100)/?numOfMembers
WHERE{{
SELECT COUNT(DISTINCT ?memberFemale) as ?numOfFemaleMembers
       COUNT(DISTINCT ?member) as ?numOfMembers
WHERE {{
  ?speech rdf:type parlipro:Speech .
  ?speech parlipro:refMember ?member .
  ?member rdf:type parlipro:ParliamentMember .}
UNION{
  ?memberFemale bio:biography _:bio .
  _:bio rdf:type bio:Biography .
  _:bio foaf:gender dbpedia:Female .
FILTER (sameTerm(?member,?memberFemale))}}}}
```

F4: What politician has most number of Wikipedia pages in different languages?

```
SELECT ?member ?numOfPages
WHERE {{
SELECT DISTINCT ?member COUNT(?dbpediaMember) AS ?numOfPages
WHERE {
  ?member rdf:type parlipro:ParliamentMember .
  ?member owl:sameAs ?dbpediaMember .}
GROUP BY ?member}}
ORDER BY DESC(?numOfPages)
LIMIT 1
```

F5: What speeches are made by politicians without Wikipedia pages?

```
SELECT DISTINCT ?speech ?member
WHERE {
  ?speech rdf:type parlipro:Speech .
  ?speech parlipro:refMember ?member .
  ?member rdf:type parlipro:ParliamentMember .
OPTIONAL {?member owl:sameAs ?dbpediaMember .}
FILTER (!bound(?dbpediaMember))}
```

T0: Retrieve top 10 members with the most speeches.

```
SELECT ?member COUNT(?speech) as ?numOfSpeeches
WHERE {
  ?member rdf:type parlipro:ParliamentMember .
  ?speech parlipro:refMember ?member .}
GROUP BY ?member
ORDER BY DESC(?numOfSpeeches)
LIMIT 10
```

T1: Retrieve top 10 topics when most of the people spoke.

```
  SELECT ?topic COUNT(?member) as ?numOfMembersSpokeInTopic
  WHERE {
    ?topic rdf:type parlipro:Topic .
?topic dcterms:hasPart ?speech .
?speech rdf:type parlipro:Speech .
?speech parlipro:refMember ?member .}
  GROUP BY ?topic
  ORDER BY DESC(?numOfMembersSpokeInTopic)
  LIMIT 10
```

T2: Retrieve top 10 topics with the most speeches.

```
  SELECT ?topic
  COUNT(?speech) as ?numOfSpeeches
  WHERE {
    ?topic rdf:type parlipro:Topic .
    ?speech rdf:type parlipro:Speech .
    {?topic dcterms:hasPart ?speech .}
  UNION{
    ?topic dcterms:hasPart ?sd .
    ?sd rdf:type parlipro:StageDirection .
    ?sd dcterms:hasPart ?speech .}
  UNION{
    ?topic dcterms:hasPart ?scene .
    ?scene rdf:type parlipro:Scene .
    ?scene dcterms:hasPart ?speech .}}
  GROUP BY ?topic
  ORDER BY DESC(?numOfSpeeches)
  LIMIT 10
```

T3: Retrieve top 10 days with the most topics.

```
  SELECT ?date COUNT(?topic) as ?numOfTopics
  WHERE {
    ?proc rdf:type parlipro:ParliamentaryProceedings .
    ?proc dcterms:hasPart ?topic .
    ?proc dc:date ?date .}
  GROUP BY ?date
  ORDER BY DESC(?numOfTopics)
  LIMIT 10
```

T4: Retrieve top 10 longest topics (i.e., number of paragraphs).

```
SELECT ?topic COUNT(?par) as ?numOfPars
WHERE {
  ?topic rdf:type parlipro:Topic .
  ?speech rdf:type parlipro:Speech .
  ?speech dcterms:hasPart ?par .
  ?par rdf:type parlipro:Paragraph .
  {?topic dcterms:hasPart ?speech .}
UNION{
  ?topic dcterms:hasPart ?sd .
  ?sd rdf:type parlipro:StageDirection .
  ?sd dcterms:hasPart ?speech .}
UNION{
  ?topic dcterms:hasPart ?scene .
  ?scene rdf:type parlipro:Scene .
  ?scene dcterms:hasPart ?speech .}}
GROUP BY ?topic
ORDER BY DESC(?numOfPars)
LIMIT 10
```

D Test Machine Specification

For the benchmark evaluation we used a personal laptop Apple MacBook Pro. The operating system running is Mac OS X Lion 10.7.5 x64. The specification of the machine is the following:

Hardware

– CPUs: 2.8 GHz Intel Core i7 (2x2 cores)
– Memory: 8 GB 1333 MHz DDR3
– Hard Disk: 750GB

Software

– OpenLink Virtuoso: Open Source Edition v.06.01.3127 compiled from source for OS X
– MySQL Community Server (GPL) v. 5.5.15
– Scripts (bash 3.2, Python 2.7.3) to scale and upload RDF datasets, to create permutations of queries and run them on Virtuoso. The scripts are available for downloading at `http://data.politicalmashup.nl/RDF/scripts/`.

Virtuoso Configuration We configured the Virtuoso Server to handle load of large data sets[16].

NumberOfBuffers = 680000
MaxDirtyBuffers = 500000

[16] `http://www.openlinksw.com/dataspace/doc/dav/wiki/Main/VirtRDFPerformanceTuning`

The RDF Index Scheme remained as it was supplied with the default Virtuoso installation. Namely, the scheme consists of the following indices:

- `PSOG` - primary key.
- `POGS` - bitmap index for lookups on object value.
- `SP` - partial index for cases where only S is specified.
- `OP` - partial index for cases where only O is specified.
- `GS` - partial index for cases where only G is specified.

E Large Plots

Fig. 6. Log_2 of Loading Time in sec of the benchmark collections.

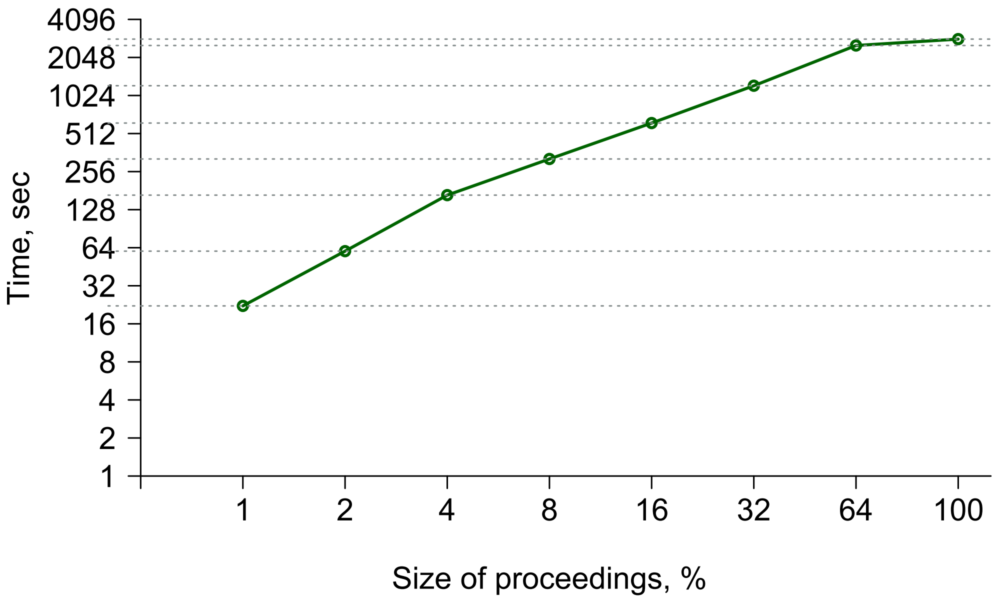
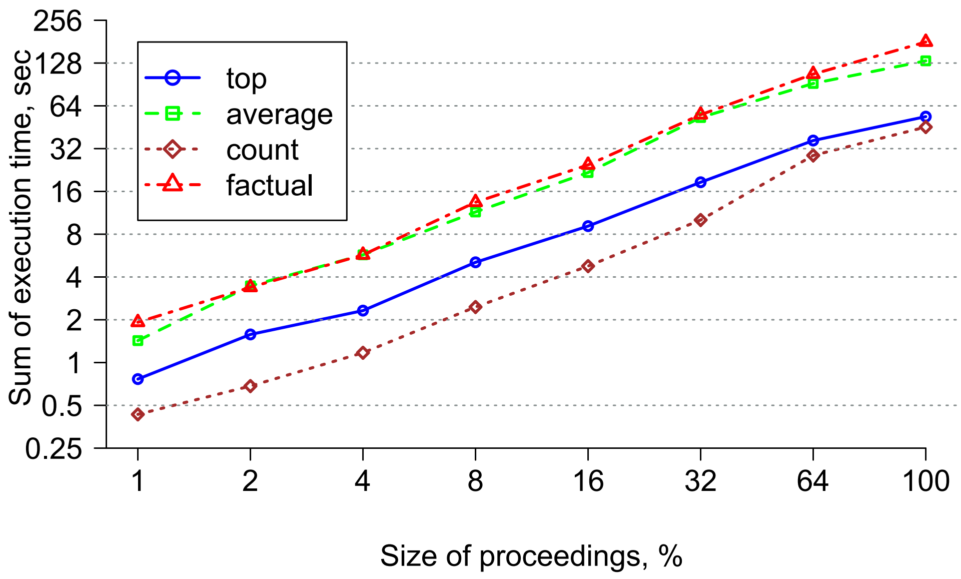


Fig. 7. Log_2 of Query Execution Time in sec of micro benchmarks on the test collections.


Sum of execution time, sec
top
average
count
factual
Size of proceedings, %