Journal of Web Engineering, Vol.12, No.3&4 (2013) 232-248

A MODEL FOR ANALYSING DATA PORTAL PERFORMANCE: THE BIODIVERSITY CASE

PEDRO LUIZ PIZZIGATTI CORRÊA PABLO SALVANHA ANTONIO MAURO SARAIVA

University of São Paulo, São Paulo
{pedro.correa, salvanha, saraiva}@.usp.br

PAULO SCARPELINI NETO CARLOS ROBERTO VALÊNCIO
ROGÉRIA CRISTIANE GRATÃO DE SOUZA

São Paulo State University, São José do Rio Preto
pauloscarpelini@gmail.com, {valencio, rogeria}@ibilce.unesp.br

Received December 24, 2012
Revised February 20, 2013

Currently, many museums, botanic gardens and herbariums keep data of biological collections and using computational tools researchers digitalize and provide access to their data using data portals. The replication of databases in portals can be accomplished through the use of protocols and data schema. However, the implementation of this solution demands a large amount of time, concerning both the transfer of fragments of data and processing data within the portal. With the growth of data digitalization in institutions, this scenario tends to be increasingly exacerbated, making it hard to maintain the records updated on the portals. As an original contribution, this research proposes analysing the data replication process to evaluate the performance of portals. The Inter-American Biodiversity Information Network (IABIN) biodiversity data portal of pollinators was used as a study case, which supports both situations: conventional data replication of records of specimen occurrences and interactions between them. With the results of this research, it is possible to simulate a situation before its implementation, thus predicting the performance of replication operations. Additionally, these results may contribute to future improvements to this process, in order to decrease the time required to make the data available in portals.

Key words: Data Portal Performance, Distributed Database Systems, Data Replication Processes
Communicated by: D. Schwabe & S. Auer

1 Introduction

Portals integrate web applications and need to handle requests in large-scale simultaneously, to access information from different providers in an unified way. Considering the performance of portals, some issues can be caused by a large number of requests and too much waiting time for processing requests [1]. The large number of approaches for data replication also contributes to the requirement of efficient analytical performance evaluation [2].

In the biodiversity application domain, access to ecological data is important to help to understand how much the ecosystems have been changed by humans more intensively since the last century, once

the human population increased from 2.5 billion in 1950 to more than 6 billion in 2000, and it resulted in loss of several genes and species [3, 4]. Also, biodiversity can be impacted by climate changes and other elements; therefore, life on Earth is studied in this science, which involves habitats, genetics, ecology, palaeontology, and others [5]. In this situation, biodiversity scientists have developed approaches to understand the causes and consequences of biodiversity changes, which requires information extraction from data sets and results must be published and made available to the community [6].

The Global Biodiversity Information Facility (GBIF) provides an infrastructure to support the open access to biodiversity data over the Internet – over 377 million of indexed records from several localities [6, 7]. Another one that can be considered and uses the same format is the portal of pollinator data from the Inter-American Biodiversity Information Network – IABIN [7]. In order to enrich this information, there are a lot of relevant biodiversity informatics websites and a list of them was introduced by Agrawal, Archak and Tyagi [9].

The communication of data providers and biodiversity portals is necessary for sharing the information with the research community and requires data replication. During this process, the data schema Darwin Core (DwC) is used with protocols such as Distributed Generic Information Retrieval – DiGIR and TDWG Access Protocol for Information Retrieval – TAPIR, which were developed for data transferring between providers and biodiversity portals [10, 11, 12]. Darwin Core is a standard supported by GBIF and aims to support the integration of biodiversity data collections of the world [13]. However, such operations can be complex considering the information: there are several ecosystems around the world and a large number of species, which means biodiversity data replication in large-scale [14, 15].

Regarding the biodiversity case, it is important to consider the interaction of computer scientists, biologists and natural resource managers in order to deal with complex challenges found in the biodiversity domain [15], besides the need of sharing information between providers and biodiversity data portals and the high complexity of data replication process caused by often the heterogeneity and the large-scale of data.

This work presents a model for analysing performance of the data replication process applied to data portals, which can be a contribution towards an improvement of information transfer between portals and providers. The model was validated by a case study conducted using the IABIN biodiversity portal of pollinators, which supports the replication of records of specimen occurrences and the interactions between them. The results showed that the difference between the real application and the simulation varied by 6.04%, acceptable considering the complexity the process involves.

This paper is organized as follows: in section 2 some concepts and information related to the developed work are presented; in section 3, the data replication stages are detailed; in section 4, the model to analyse the proposed performance is presented; in section 5, the case study is described; in section 6, results are discussed; in section 7, the conclusions of this work are presented.

2 Basic concepts and related works

In this chapter are presented some concepts and information related to distributed databases, data replication management protocol, biodiversity data portal and performance analysis method, which

provided fundamental basis to this work. Finally, a comparison between the proposed model and its related works is described, in order to highlight its relevance.

2.1 Distributed databases

Distributed database systems use database technology allied to computer network technology to enable the sharing of geographically distributed information [16]. Many digital libraries were developed to share research results. However, the format, focus and the locality of this information are quite diverse, making it necessary to standardize them for sharing [10].

An example of a data schema which allows information standardizing is Darwin Core (DwC). This schema aims to provide the exchange of information about geographical occurrences of specimens in collections [10]. Its use requires that databases have a conceptual model that supports DwC and enables interaction between providers, as well as integration between different provider schemas and the portal [17].

2.2 Protocol for data replication management

Data schemas are logical references that can be properly identified and mapped when being transported between providers and portal [11]. The process is performed by transport protocols which enable data replication between providers and portals.

One of them is named DiGIR Protocol, based on interpreted languages to facilitate their use and communication between different databases in several platforms. Distributed Generic Information Retrieval (DiGIR) has a data schema to carry out the mapping between a local database and remote requester. The communication language is based on eXtensible Markup Language (XML) to facilitate requests and the return of data; this protocol is applied to both provider and portal, but with different focuses [11].

Another protocol is TAPIR, TDWG Access Protocol for Information Retrieval, that was developed by the Biodiversity Information Standards (TDWG). Like the DiGIR, it supplies resources for data transport between heterogeneous databases. However, the TAPIR uses extensions and resources from both DiGIR and the Biological Collection Access Service for Europe (BioCASE).With this, TAPIR optimises the communication process [11, 12].

Since it is generic, TAPIR may be used with different portals. This allows for the creation of multiple data resources, which makes it possible to find different information and even different data schema. Such a characteristic supports the work of providers and is widely used with data interaction. Other advantages of TAPIR protocol are: it is open source and has no link to the operational system, data schema or data exit model [12].

2.3 A Biodiversity Data Portal

GBIF data portal aims to make the access to biodiversity data available from any part in the world [7] and is used as a model of architecture to the others biodiversity data portals. The main resources of this portal are:

- *Centralized storage of specimens homogeneous data*: stored in specific register fields, from different providers, made available in the portal;

- *Search for occurrences, datasets and countries*: by means of specific filters, the user can select biodiversity registers based on a scientific name, family, country and provider;
- *Exploring taxonomical tree*: locating registers based on the taxonomical tree in a rapid and standardized way;
- *Geo-referenced positioning*: for registers that supply geographic information about the collections, the portal has a map where the geographic locality of the register is exhibited;
- *Multiple language support*: portal information is available in various languages.

Information is made available on the GBIF data portal by means of data replication. The portal contacts several providers and uses their data schema and portal, even if they are different, to import the data to the local database [4].

2.4 Performance analysis method

The diversity of transaction models makes it necessary for them to be adapted and optimised so that long processes may be executed. Stages of the process can be segmented to optimise them and generate useful information [18].

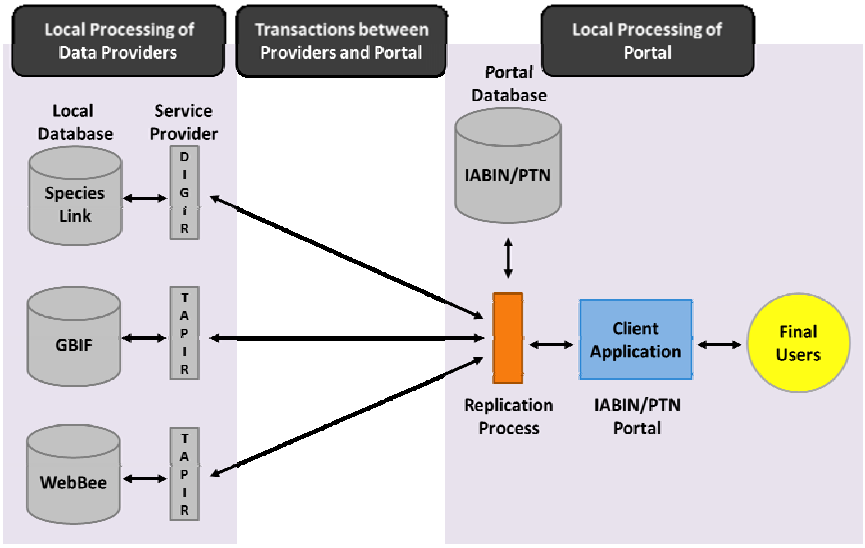


Figure 1. Architecture for the pollinators' portal TAPIR/DiGIR system.

The metrics and formula proposed for the simulation and performance analysing are based on an allocation project defined in [18], which considers the Open System Interconnection (OSI) application layer. Figure 1 shows the architecture for the pollinators' portal TAPIR/DiGIR system.

Details of each stage of the architecture are as follows:

- *Local processing of data providers* - maps the local database fields to make them available on the portal by means of a protocol;

- *Transactions between providers and portals* - executes network processes and encapsulates data;
- *Local processing of portal* - the portal collects data. All providers are interlinked; data is indexed and allocated so as to replicate source data.

Several models have been proposed and some of them are organized in surveys [19] [20]. Performance in data portals is not only a concern related to database technology, but it is also addressed by software engineering in topics like web quality [21][22].

Models can be analytical or simulation and some important elements have to be considered: simulations have been applied to validate the results of analytical analysis, although this approach has a higher cost related to time of execution and more complex than the other [19]. In addition, some proposed model was validated by theoretical model or simulation strategy [19] [20], differently of the work described in this paper, which adopted an approach of validation based on measurement of experiments considering three conditions: the communication with a local, a national and an international node.

The result of this work considers the response time and the data transferring rate and can be used as a simulation resource to enable analysis of situations before their implementation, once it was applied to a real world case and demonstrated a suitable variation between simulation and real case.

3 Data replication stages

This work aims to simulate and to evaluate the performance of the data replication process and the replication stages are detailed in this section. Figure 2 shows the data replication stages between portal and provider.

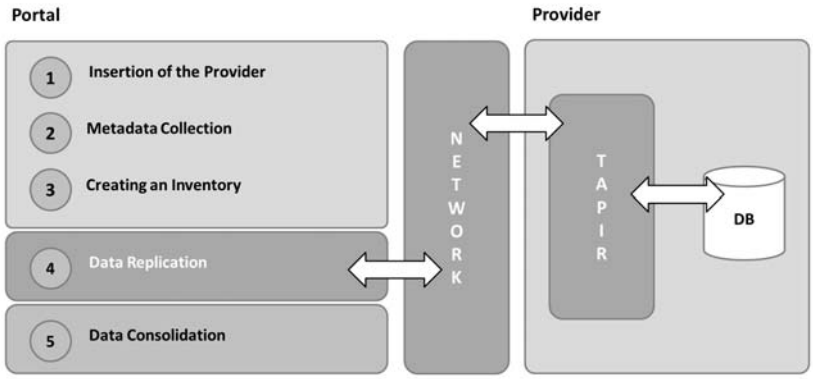


Figure 2. Data replication stages between portal and provider.

Details of each provider replication stages are following:

- *Insertion of the provider in the portal* - data providers are inserted in the database, such as name and URL, without the need for them to communicate with each other;
- *Metadata Collection* - an initial communication is established and metadata providers are supplied, technical details, providers' representatives and resources which will be available;

- *Creating an inventory* - the provider requests the inventory based on scientific names that will be used in the next stage to divide requests;
- *Data replication from providers to portal* - this is the most important stage of the system, because the replication is actually carried out. The aims is to analyse the performance of this stage;
- *Data consolidation in the portal* - data is adapted and stored in the portal's database. At this stage, the portal no longer needs to remotely access the provider.

4 The model for analysing portal performance

The proposed model for a simulation of a data replication process is presented in this section.

4.1 Experimental structure

The scenarios of the providers for the experiments are:

- *Local* node - provider in the same server as portal, in order to eliminate network impact;
- *National remote node* - provider from a server in Brazil to consider the impact of the network;
- *International remote node* - provider from a server outside Brazil which significantly increases package latency.

Three provider configurations were defined for each experimental scenario to evaluate the different situation in each process, in relation to the number of registers that are sent per package. In this work, the quantities were defined as being: 50; 200 and 400. A database was extracted from a GBIF biodiversity system in order to perform the experiments and the number of selected registers was 88,491 with 26 attributes, including a geo-referential one.

So as to evaluate the costs of each element of a scenario, as well as the performance of the network which can change in different experimental tests in scenarios 2 and 3, notification points were inserted in the codes, as shown in Figure 3.

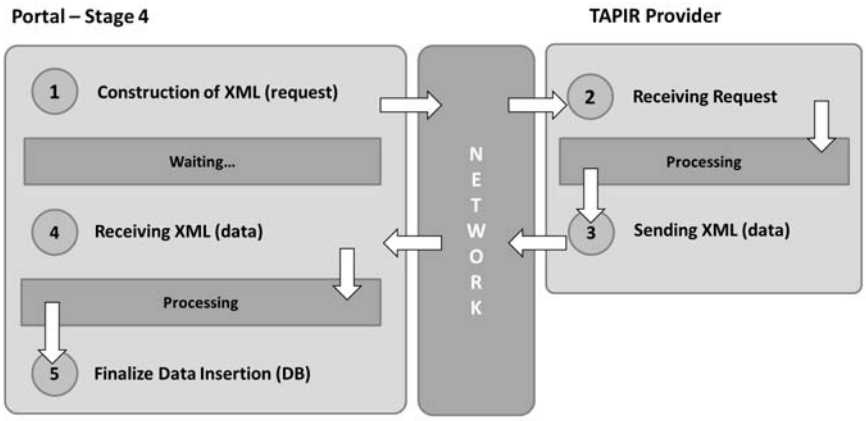


Figure 3. Notification points.

Details of the notification points:

1) *Construction of XML (request)* - the portal constructs XML for the request of the package for the provider;

2) *Receiving request* - the provider receives the package with the portal's request;

3) *Sending XML (data)* - the provider sends the requested package to the portal;

4) *Receiving XML (data)* - the portal receives the package from the provider;

5) *Finalize data insertion (BD)* - the portal processes the received package and inserts the data in the local database.

4.2 Verifying notification points

Timetables were synchronised by means of a single timetable server so that admeasurements be coherent. Collections were made for each provider according to the programming language used by each stage of the replication.

In the case of the GBIF portal, the collection of notification points was made as follows:

- *TAPIR provider* - was developed in PHP programming language, so a temporary text file was used to store the notification points;
- *Portal* - based on JAVA programming language, a class with a list was developed that stored all the information in the memory, later transferred to a simple file.

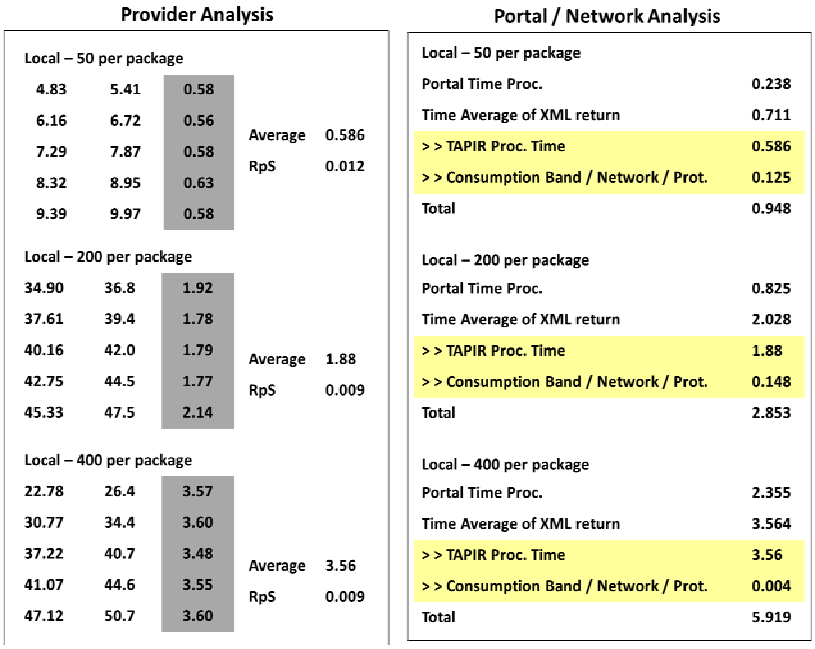
4.2.1 Verification for local scenario

Several reports were generated from the survey of notification points. Figure 4 shows local scenario results.

In the Providers Analysis quadrant, some time data samples for Notification Point Construction of XML, Receiving Request and Sending XML are analysed. The mean difference of the samples results in time, in seconds, spent to process the XML package in the provider – highlighted column. The time to process a register is described by the variable RpS (Registers per Second) which shows the processing performance.

In the Portal/Network Analysis quadrant, results for each number of package registers are presented, considering a mixed analysis of portal and network. The Portal Processing Time refers to how long the portal took to process a package, considering the time for processing of package receipt and finalizing the insertion of data in a database; The Time Average of XML Return covers both network and TAPIR processing (from XML construction to the receipt of the request); TAPIR Processing Time refers to the time spent for processing with TAPIR, from receiving the request to sending the return to the portal; and Consumption Band/Network/Protocol is the time lacking to complete the return of XML added with the network time.

In the local scenario, network times were given by the minimum processing times of Apache (PHP server) and Tomcat (Java server), therefore not checked.



Provider Analysis

Local – 50 per package

4.83	5.41	0.58		
6.16	6.72	0.56	Average	0.586
7.29	7.87	0.58	RpS	0.012
8.32	8.95	0.63		
9.39	9.97	0.58		

Local – 200 per package

34.90	36.8	1.92		
37.61	39.4	1.78		
40.16	42.0	1.79	Average	1.88
42.75	44.5	1.77	RpS	0.009
45.33	47.5	2.14		

Local – 400 per package

22.78	26.4	3.57		
30.77	34.4	3.60		
37.22	40.7	3.48	Average	3.56
41.07	44.6	3.55	RpS	0.009
47.12	50.7	3.60		

Portal / Network Analysis

Local – 50 per package	
Portal Time Proc.	0.238
Time Average of XML return	0.711
> > TAPIR Proc. Time	0.586
> > Consumption Band / Network / Prot.	0.125
Total	0.948
Local – 200 per package	
Portal Time Proc.	0.825
Time Average of XML return	2.028
> > TAPIR Proc. Time	1.88
> > Consumption Band / Network / Prot.	0.148
Total	2.853
Local – 400 per package	
Portal Time Proc.	2.355
Time Average of XML return	3.564
> > TAPIR Proc. Time	3.56
> > Consumption Band / Network / Prot.	0.004
Total	5.919

Figure 4. Local scenario reports.

4.2.2 Verification for remote scenario (National provider)

In this experimental test, data were checked in the situation in which the TAPIR provider is located in another server, yet in the same country - in this case, Brazil. Figure 5 shows the description of results.

In this situation, results presented in Provider Analysis and Portal Analysis/Network follow the same pattern, with the addition of considerable network consumption values. In Network Summary, a survey was made to find the average size of the package, in bytes; the average checked latency in seconds between provider and portal; the network time with and without latency; and lastly, the speed of traffic in the network.

4.2.3 Verification for remote scenario (International provider)

In this scenario, the same data were checked, but now in the case of the provider being based outside the country - in this case, the United States.

It was noted that the results, shown in Figure 6, differ basically due to network reasons and also the configuration of the server used by the provider.

Provider Analysis

National – 50 per package				
17.76	19.2	1.45		
20.49	21.9	1.45	Average	1.458
22.98	24.5	1.49	RpS	0.029
25.40	26.8	1.44		
27.88	29.3	1.46		
National – 200 per package				
7.47	12.3	4.86		
15.88	20.8	4.87		
23.76	28.7	4.91	Average	4.902
31.14	36.1	4.98	RpS	0.025
38.55	43.4	4.89		
National – 400 per package				
21.00	30.6	9.63		
37.68	47.2	9.56		
52.89	62.4	9.50	Average	9.446
0.22	9.39	9.17	RpS	0.024
14.69	24.1	9.37		

Portal / Network Analysis

National – 50 per package	
Portal Time Proc.	0.238
Time Average of XML return	2.143
> > TAPIR Proc. Time	1.458
> > Consumption Band / Network / Prot.	0.685
Total	2.381
National – 200 per package	
Portal Time Proc.	0.825
Time Average of XML return	6.488
> > TAPIR Proc. Time	4.902
> > Consumption Band / Network / Prot.	1.586
Total	7.313
National – 400 per package	
Portal Time Proc.	2.355
Time Average of XML return	12.65
> > TAPIR Proc. Time	9.446
> > Consumption Band / Network / Prot.	3.204
Total	15.01

NETWORK SUMMARY

National – 50 per package	
Average Package Size (bytes)	67463
Average Checked Latency (sec)	0.023
Network Time	0.6846
Network Time (without latency)	0.6616
Network Speed (bytes per second)	101969.47
Network Speed (kbps)	815.7557
National – 200 per package	
Average Package Size (bytes)	268191
Average Checked Latency (sec)	0.023
Network Time	1.586
Network Time (without latency)	1.563
Network Speed (bytes per second)	171587.33
Network Speed (kbps)	1372.699
National – 400 per package	
Average Package Size (bytes)	507410
Average Checked Latency (sec)	0.023
Network Time	3.204
Network Time (without latency)	3.181
Network Speed (bytes per second)	159512.73
Network Speed (kbps)	1276.102

Figure 5. Remote scenario reports (national provider).

Provider Analysis

International – 50 per package

27.98	28.5	0.56		
30.83	31.4	0.56	Average	0.562
33.45	34.0	0.56	RpS	0.011
36.04	36.6	0.57		
38.64	39.2	0.56		

International – 200 per package

23.41	25.2	1.8		
28.58	30.4	1.8		
34.10	35.9	1.8	Average	1.8
39.29	41.1	1.8	RpS	0.009
44.57	46.4	1.8		

International – 400 per package

0.71	4.35	3.64		
11.62	15.3	3.66		
21.27	23.2	1.93	Average	3.3
26.75	30.4	3.64	RpS	0.008
35.86	39.5	3.63		

Portal / Network Analysis

International – 50 per package	
Portal Time Proc.	0.238
Time Average of XML return	2.197
> > TAPIR Proc. Time	0.562
> > Consumption Band / Network / Prot.	1.635
Total	2.435
International – 200 per package	
Portal Time Proc.	0.825
Time Average of XML return	4.444
> > TAPIR Proc. Time	1.8
> > Consumption Band / Network / Prot.	2.644
Total	5.269
International – 400 per package	
Portal Time Proc.	2.355
Time Average of XML return	6.704
> > TAPIR Proc. Time	3.3
> > Consumption Band / Network / Prot.	3.404
Total	9.059

NETWORK SUMMARY

International – 50 per package	
Average Package Size (bytes)	67463
Average Checked Latency (sec)	0.169
Network Time	1.635
Network Time (without latency)	1.466
Network Speed (bytes per second)	46018.417
Network Speed (kbps)	368.1473
International – 200 per package	
Average Package Size (bytes)	268191
Average Checked Latency (sec)	0.169
Network Time	2.644
Network Time (without latency)	2.475
Network Speed (bytes per second)	108360
Network Speed (kbps)	866.88
International – 400 per package	
Average Package Size (bytes)	507410
Average Checked Latency (sec)	0.169
Network Time	3.404
Network Time (without latency)	3.235
Network Speed (bytes per second)	156850.08
Network Speed (kbps)	1254.801

Figure 6. Remote scenario reports (international provider).

4.3 Making of the proposed model

From the results, constants and variables were determined, and it was possible to produce the proposed model.

4.3.1 Constants used in the formulas and their origins

The constants used, as well as their value, description and origins are:

- PRC – Portal Register Cost, in seconds
  - Value: 0.00492;
  - Obtained from the average time spent in the portal, considering an individual cost per individual.
- CPRP – Cost per performance Point per Register in the Provider
  - Value: 0.003;
  - TAPIR provider was implemented with different configurations and its performance was exhaustively verified, making it possible to produce a point report to determine processing time.
- AS – Average Size per mapped field, in bytes
  - Value: 52.2;
  - As each TAPIR provider can be configured in a different way, it was verified that the best way to obtain the average size of a data package is for it to be based on the number of mapped fields. An average size for each XML field was obtained.

4.3.2 Variable fields involved in the formulas

The variable fields are:

- nreg – Number of registers per package - determines how many registers will traffic in each information package – Valid entries: between 50 and 400;
- pf – Performance factor - determines the TAPIR provider processing level. Valid entries: 1 – 3, Xeon 3-4Gb RAM; 4-6, Intel Core 2 Duo 2Gb RAM; 6-8, Athlon/Dual Core 1Gb RAM; above 9, inferior computers;
- qfields – Number of mapped fields in the TAPIR provider. More than twenty – which covers most cases. Lower numbers may distort results;
- lat – Average latency between parties (provider and portal). There are no limits;
- totreg – Number of registers to be replicated. Also, no limits;
- band – Average speed in bytes per second.

Support formulas for the mathematical model are shown in Table I.

Finally, the formula which determines the total foreseen time for the whole data replication process is defined by the equation:

$$FC = \sum_{\square}^{AQP} \left(PC_{pa} + NC_{pa} + TPC_{pa}\right)$$

being:

- FC - Final cost of the replication (seconds);
- AQP - Average quantity per package (units);
- PC - Portal cost per package (seconds);
- NC - Network cost per package (seconds);
- TPC - TAPIR provider cost per package (seconds).

Aiming to prove the viability of the model, the same experimental test scenarios were applied to it. The maximum results distortion was 6% in relation to those previously checked, which can be considered acceptable, given the total time for its execution.

TABLE I. SUPPORT FORMULAS FOR MATHEMATICAL MODEL

Formula	Description
$PC_{pa} = nreg * PRC$	Portal cost per package
$TPC_{pa} = \sum_{pf}^{nreg} pf * CPRP$	TAPIR provider cost per package (seconds)
$SP_{pa} = qfields * AS * nreg$	Average size of a data package (bytes)
$ANS = \frac{band}{8}$	Average network speed (bytes per second)
$NC_{pa} = lat + \frac{SP}{ANS}$	Network cost per data package (seconds)
$AQP = \frac{totreg}{nreg}$	Average quantity of packages to be in the traffic

5 Applying the model: Case study

In order to demonstrate the proposed model based on the previously described formula, a replication of a real database in Brazil was used - CEPANN (*Coleção Entomológica Paulo Nogueira Neto*) of *Laboratório de Abelhas do Departamento de Ecologia do IBUSP* (*Instituto de Biociências da USP*). CEPANN database stores data in DwC format and experimental tests were performed in the register replication stage of the provider to the portal, which is based on GBIF technology.

The first experimental test was performed on a single data provider, having a standard configuration adopted by most TAPIR providers, in which 200 registers per package were defined. In this execution, 35,067 registers having 61 fields were transferred; the band speed was 725 kbps, the communication latency was 32 ms and the provider factor was six; the total data transfer process time was 2,159 seconds.

The second experimental test was conducted in an interaction data provider. With this provider, the schema that was used for data interaction was restricted to 6 fields. With this, should the number of packages be maintained, very small packages would be generated, causing an impact on the performance. Therefore, to approximate package sizes to the first test, registers per package were set at 800. In this experiment, 17,534 registers having six fields were transferred; the band speed was 807 kbps; the communication latency was 31 ms and the provider factor was six; the total data transfer process time was 1,161 seconds; the formula was applied to each of the tests and results are shown in Table II.

From the experimental tests it was verified that the total data process transfer time was greater for the singular data provider than for the interaction data provider. Also, formula time was greater for the first provider than the second, 2,041 and 456 seconds respectively. The formula variation for the first test was 5.47% and 60.72% for the second.

TABLE II. RESULTS OBTAINED FROM APPLYING THE FORMULAS

Formula	Singular data provider	Interaction data provider
PC_{pa}	0.984	3.936
TPC_{pa}	3.6	14.4
SP_{pa}	636,840	250,560
ANS	90,625	100,875
NC_{pa}	7.0592	2.5148
AQP	175.335	21.91
FC	2041.46	456.84

6 Analysis of the results

In this section, results are minutely analysed.

6.1 Performance

Several experimental tests with providers were performed to achieve the proposed model and some characteristics related to performance were observed, among them:

- The transfer of small packages (less than 50 registers each) between the provider and the portal or by the browser did not reach the expected performance. On the other hand, packages with close to 200 to 400 registers had a significant performance gain;
- Real performance values and formula simulated values had a variation of 6.04%, which represents the reliability of results obtained with the formula. Considering the complexity of the process and the number of variables, this value is satisfactory for the purposes of this work;
- Improvements made to the portal do not always contribute towards a better process on a whole, once the portal accounts for only a small part of the process.

The individual cost of Provider, Portal and Network were also measured during the process, using the proposed model. A comparison of those costs is shown in Table III where 100,000 registers, 61 fields and a 0.030 mean latency per second were defined. Figure 7 shows the results.

By analysing Figure 7, it can be verified that the variation between the number of registers did not significantly alter the costs between provider, portal or network. Performance alteration generates an oscillation of around 10% in the cost of the provider; however, a large part of the processing time depends on the band at which it is available. Network conditions may vary during the process and the use of small packages may cause an unsatisfactory performance, however, an increase in performance; generated a significant improvement to the network process cost.

TABLE III. EXPERIMENTAL TEST SCENARIOS FOR A COMPARISON OF COSTS

Scenario	Register per package	Provider's performance	Average band (kbit/seg)
T1	50	5	1000
T2	200	5	1000
T3	400	5	1000
T4	200	7	1000
T5	200	3	1000
T6	200	5	500
T7	200	5	2000

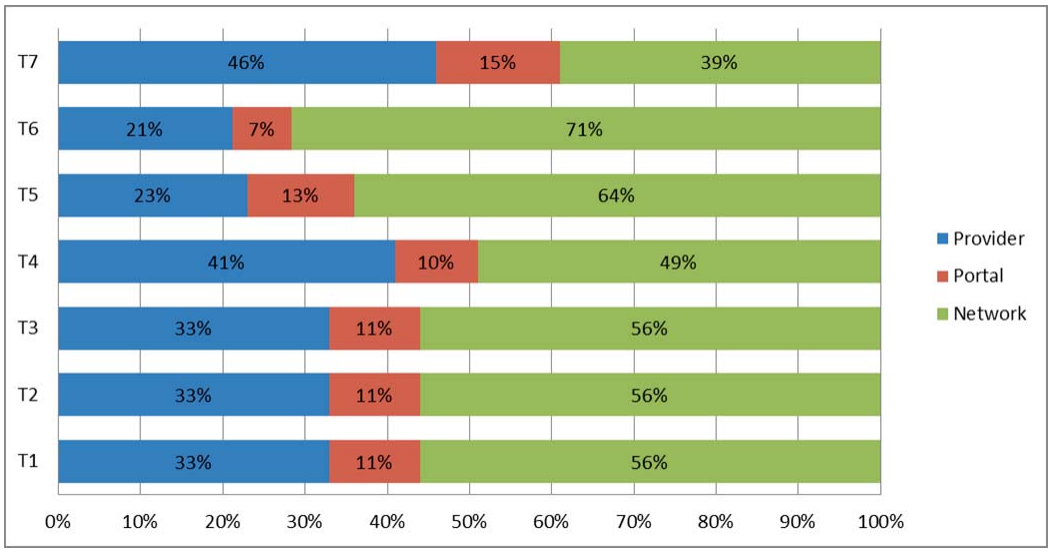


Figure 7. Results of costs of each process in each experimental test scenario.

6.2 Case study

In the case study, the replication process was conducted in the same way as in a real situation. When applying the model formula to the provider of single data, results had a variation of 5.47% within the bracket limit of the initial experimental tests; thus, the performance of the model can be considered acceptable for this case. In the second experimental test, the formula was applied to an interaction data provider where although a variation of results was expected the variation was 60.72%, which is impracticable to be considered adequate for a simulation.

The elements that made up the second experimental test were assessed to identify the reasons for the variation:

- The mapping of few fields and differentiated contents distorted the size of packages and, consequently, the application of the formula;
- The cost of processing registers is high for a schema having only 6 fields and 800 registers per package, whereas with packages having 61 fields and 200 registers, the cost is included in the average within the cost per register;
- The exit model of experimental test 2 is not a native to the TAPIR provider, which requires an external consultation;
- The portal software was not as efficient as expected with a reduced number of fields.

7 Conclusions

Portals permit the sharing research results related to a certain area of knowledge, coming from several different sources and available to the scientific community. The heterogeneity of data, as well as the computational cost of processing and transferring them, are factors that hinder data integration in an infrastructure that provides researchers access to information.

Data portals can execute data replication from local databases in order to make data available for remote access. Due to the importance of the data replication stage, this work proposed a model for the simulation and analysis of the performance of this task in data portals. With this model, it was possible to identify some characteristics that influence this process performance. Among them, it was seen that the real performance and the simulated values of the formula varied by 6.04%. Considering the complexity of the process and the number of variables treated, this percentage is acceptable and represents the reliability of the results with the proposed model.

With the case study, it was possible to assess the behaviour proposed model with both single and multiple interaction data providers. The proposed model showed it can be applied to real cases – mainly in the biodiversity area, which requires large-scale information sharing.

As an original contribution, this work offers data and tools capable of aiding the analysis of the replication process performance between portals and providers, which has adopted an approach of validation based on measurement of experiments considering three conditions: the communication with a local, a national and an international node; differently of some related works, which were validated by theoretical model or simulation strategy. Therefore, it is possible to simulate situations

that can help with the timely decisions making regarding the publication of new data providers and the distribution of information in data portals.

References

1. Shanping, L., Jiaqi, T.,A collaborative performance tuning approach for Portal-based web sites. In: Sixth International Conference on Networked Computing and Advanced Information Management (NCM), 2010, pp.113-117.

2. Nicola, M. and Jarke, M. 2000. Performance Modeling of Distributed and Replicated Databases. In: IEEE Transactions on Knowledge and Data Engineering. 12, 4 (2000), 645-672.

3. Schneiders, A., Van Daele, T., Van Landuyt, W., Van Reeth, W. Biodiversity and ecosystem services: Complementary approaches for ecosystem management?. In: Ecological Indicators, 21, pp. 123-133, 2012. doi: 10.1016/j.ecolind.2011.06.021

4. Halkos, G.E. and Tzeremes, N.G., Measuring biodiversity performance: A conditional efficiency measurement approach. In: Environmental Modelling & Software. 25, 12 (2010), 1866-1873.

5. Enkea, N. et al. 2012. The user's view on biodiversity data sharing — Investigating facts of acceptance and requirements to realize a sustainable use of research data. In: Ecological Informatics. 11, (2012), 25-33.

6. Flemons, P., Guralnick, R., Krieger, J., Ranipeta, A., Neufeld, D. A web-based GIS tool for exploring the world's biodiversity: The Global Biodiversity Information Facility Mapping and Analysis Portal Application (GBIF-MAPA). In: Ecological Informatics, 2, 1 (2007), pp. 49-60. doi: 10.1016/j.ecoinf.2007.03.004.

7. GBIF. Global Biodiversity Information Network. Available at: <www.gbif.org>.

8. IABIN. Inter-American Biodiversity Information Network. Available at: <www.oas.org/en/sedi/dsd/iabin>.

9. Agrawal, R.C. et al., An overview of biodiversity informatics with special reference to plant genetic resources. In: Computers and Electronics in Agriculture. 84, 92-99, (2012).

10. DwC. DarwinCore Schema. Available at: <www.tdwg.org/activities/darwincore>.

11. DiGIR. Distributed Generic Information Retrieval. Available at: <www.digir.network>.

12. TAPIR. Tapir TDWG Task Group. Available at: <www.tdwg.org/activities/tapir>.

13. Bafnaa, S. et al., Schema driven assignment and implementation of life science identifiers (LSIDs). In: Journal of Biomedical Informatics. 41, 5 (2008), 730-738.

14. Salvanha, P., Najm, L. H., Corrêa, P. L. P. and Saraiva, A. M., Model of management and sharing distributed interaction pollinators information for centralized biodiversity portals, In: Proc. 5th

Contecsi International Conference on Information Systems and Technology Management. São Paulo, Brazil, 2009.

15. Schnasea, J.L. et al., Information technology challenges of biodiversity and ecosystems informatics. In: Information Systems. 28, 4 (2003), 339-345.

16. Ozsu, T. M., Valduriez, P., Principles of Distributed Database Systems, 2 [S.I.]: Prentice Hall, 1999.

17. Batini, C., Lenzerini, M. and Navathe, S. B., A comparative analysis of methodologies for database schema integration, In: ACM Comput. Surv., New York, USA, vol. 18, (1986), pp. 323-364.

18. Côrrea, P. L. P., Guidelines and procedures for the project database. (Thesis) Department of Computer Engineering and Digital Systems of the Polytechnic University of São Paulo, 2002 [in Portuguese].

19. Nicola, M. and Jarke, M., Performance Modeling of Distributed and Replicated Databases. In: IEEE Transactions on Knowledge and Data Engineering. 12, 4 (2000), 645-672.

20. Osman, R. and J., W.K., Database system performance evaluation models: A survey. In: Performance Evaluation. 69, (2012), 471-493.

21. Calero, C., Ruiz, J., Piattini, M., Classifying web metrics using the web quality model. In: Online Information Review, vol. 29, 3(2005), pp. 227-248.

22. Olsina, L. et al., Using web quality models and a strategy for purpose-oriented evaluations. In: Journal of Web Engineering. 10, 4 (2011), 316-352.