<table>
<thead>
<tr>
<th>N° PROYECTO</th>
<th>11130648</th>
<th>DURACIÓN</th>
<th>3 años</th>
<th>AÑO ETAPA</th>
<th>2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>TÍTULO PROYECTO</td>
<td>QUERY LANGUAGES AND VIEWS FOR GRAPHS WITH DATA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DISCIPLINA PRINCIPAL</td>
<td>INGENIERIA EN COMPUTACION</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRUPO DE ESTUDIO</td>
<td>INGENIERIA 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INVESTIGADOR(A) RESPONSABLE</td>
<td>JUAN LORENZO REUTTER DE LA MAZA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIRECCIÓN</td>
<td>COMUNA:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CIUDAD:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>REGIÓN</td>
<td>METROPOLITANA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**FONDO NACIONAL DE DESARROLLO CIENTIFICO Y TECNOLOGICO (FONDECYT)**
Moneda 1375, Santiago de Chile - casilla 297-V, Santiago 21
Teléfono: 2435 4350 FAX 2365 4435
Email: informes.fondecyt@conicyt.cl
### OBJETIVOS

Cumplimiento de los Objetivos planteados en la etapa final, o pendientes de cumplir. Recuerde que en esta sección debe referirse a objetivos desarrollados, NO listar actividades desarrolladas.

<table>
<thead>
<tr>
<th>Nº</th>
<th>OBJETIVOS</th>
<th>CUMPLIMIENTO</th>
<th>FUNDAMENTO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Study data graph query languages in terms of their expressive power, and find a graph query language that can be considered the core of each graph query language that aims to express relevant graph properties.</td>
<td>TOTAL</td>
<td>We studied two families of data graph languages. On one hand we specialized in theoretical data fragments such as graph XPath or regular queries with memories, and from a more practical side we studied SPARQL. The results about data graph languages were published in the ICDT2014 conference, and we have prepared a journal version of this article which is currently under revision. The results of the expressivity of SPARQL graph to graph queries was published in the ICDT2015 conference, and we also published a specific study about property paths, the navigational fragment of SPARQL, in the ISWC2015 conference.</td>
</tr>
<tr>
<td>2</td>
<td>Study the problem of query containment, when the classes of queries involved correspond to different choices among current data graph query languages and their various fragments.</td>
<td>TOTAL</td>
<td>In our ICDT2014 submission we also studied the containment problem of various data graph query languages. Furthermore, in ICDT2015 we also published a paper about the containment problem for what is - in our view - the main navigational core of any graph language, that we call regular queries. We have also published a full version of this conference paper in a ISI journal.</td>
</tr>
<tr>
<td>3</td>
<td>Study the problem of query answering using materialized views when both queries and views are defined with the languages studied in the previous two goals. Devise algorithms to solve this problem, pinpoint its computational complexity, and analyze its expressive power from a practical point of view.</td>
<td>TOTAL</td>
<td>While there was no time to study views for the data graph languages, we did a thorough study of SPARQL CONSTRUCT queries, a graph-to-graph query language that is currently the main alternative for defining SPARQL views. Within this framework, one can show that query answering reduces to studying the composition of SPARQL queries. We studied the feasibility, complexity and other formal aspects of this problem in our ICDT2015 submission about SPARQL CONSTRUCT queries, and we are currently preparing a full version that will be sent to a journal.</td>
</tr>
<tr>
<td>4</td>
<td>Continue studying our model of views for graph data, but now in the context of data integration. One should be able to show that the good properties of this model carry over when used as a tool to specify data integration systems.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>In our third year we understood that the focus on graph data integration must change from a classical perspective to a more dynamic one, where one now integrates classical database queries with other volatile sources of information, such as web APIs. This change of focus led us to shift to researching how to integrate SPARQL queries with web APIs; we developed a Demo that does this, that was presented at ISWC2016. Our demo system makes it possible to integrate SPARQL graph queries and API answers. However, further research is still needed in order to fully develop this solutions, and it is part of our future work.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Otro(s) aspecto(s) que Ud. considere importante(s) en la evaluación del cumplimiento de objetivos planteados en la propuesta original o en las modificaciones autorizadas por los Consejos.
RESULTS OBTAINED:

First Goal: Expressiveness of Query languages. As mentioned in the proposal, there has not been many studies regarding the expressiveness of different graph query languages. As such, the first goal of our project was:

Study data graph query languages in terms of their expressive power, and find a graph query language that can be considered the core of each graph query language that aims to express relevant graph properties.

To the fulfilment of this goal we advanced in two different angles.

1. First, we studied and compared all the classic query languages that were available in the literature. This study has led to the preparation of a survey of graph query languages, that is currently under revision. Furthermore, this first study gave us tremendous insight on the inner workings of graph database languages, and was vital for the dissemination activities of our project.

2. On a different angle, we set to study the expressive power of SPARQL, one of the most important query languages used in practice nowadays [12], devised specifically to query RDF graphs [9]. Here we realized that almost nothing was again known for those SPARQL queries of the CONSTRUCT form (see again [12]), which are effectively graph-to-graph query languages in the sense that they take graphs as inputs and deliver graphs as outputs. This is obviously a good thing to have in a graph query language, and would lend itself to a very elegant definition of the concept of views.

Together with the international collaborator and a student from my group, we started analysing the expressive power of those queries. In particular, we managed to show that this language is equivalent in expressive power to first order logic. Moreover, we also studied a fragment of SPARQL queries of the CONSTRUCT form denoted as well-designed patterns ([10]), discovering that this language is essentially positive first order logic. Finally, we studied the expressiveness of this language if a certain feature called blank nodes were allowed (see again [12] for more details). Here it was impossible to obtain a direct logic-based characterization. However, we managed to establish the equivalence of this language to a certain data exchange setting.

It is worth mentioning that these expressivity results lead to important conclusions on composability of these languages. In particular, the fragments without
blank nodes are composable, that is, the composition of two queries can be always expressed by another query, but if blank nodes are allowed, then this important property is lost. These results have been submitted to one of the top conferences in databases [6], and we are currently preparing a full version.

Second Goal: The problem of query containment. Our second goal was to study the problem of containment of various forms of query languages. That is, the problem of, given queries $Q_1$ and $Q_2$, decide whether $Q_1$ is contained in $Q_2$, i.e., if the result of the evaluation of $Q_1$ over a graph $G$ is contained in the results of evaluating $Q_2$ over $G$, over all possible graph databases. We also expected to pinpoint their exact computational complexity. Our advancements were as follows.

1. To begin we set up to study the containment problem for a very expressive form of queries that we call Regular Queries. In a nutshell, these queries are obtained by closing the language of CRPQs with a transitive closure operator that is to be computed solely on binary relations. Results from the first goal told us that this was, essentially, the most expressive query language that one could hope to have a decidable containment problem, as it more powerful than any of the graph query languages outlined in our proposal.

We showed that the containment problem for these languages is not only decidable, but EXPSPACE-complete. This means that the containment for these queries is essentially as complex as CRPQs, and not more. This is an outstanding result that tells us that we can focus on Regular Queries for the definition of our views without sacrificing computability, while gaining much more expressive power at the same time. Furthermore, there was no need to study other graph query languages, as this language already subsumes all of our other candidates. These results have been submitted to a conference [13], and during the beginning of the third year was extended to a full paper, already published in ToCS, a good journal in theoretical computer science [14].

2. However, it remains to see whether Regular Queries will be simple enough to be the main language to define views for graphs, as it is too soon to tell. In the meanwhile, we also studied the navigational features of SPARQL. This involved the study of query containment of SPARQL queries when these queries are allowed to use property paths, a navigational construct added in the last revision of the SPARQL standard [12]. Property paths are very similar to other well studied graph query languages such as regular path queries (see e.g. [4] for work on the containment of those queries). So, the question that motivated our study was: Can we leverage our knowledge of graph query languages and apply
them to study the containment of SPARQL queries with property paths? It
turned out that this was not an easy task, and at the end we had to reinvent all
the known framework that was developed in [8, 11] to tackle the containment
problem for SPARQL queries in the absence of property paths. Nevertheless,
we manage to show that, in most cases, the computational complexity of the
containment problem remains the same when we move from graph databases
to SPARQL queries. These results were published in ISWC [7].

**Third goal: Query answering using views.** As mentioned in our last year
report, since RDF is the most widely use graph database in industry, we settled for
SPARQL CONSTRUCT queries [12], one of the result forms of SPARQL, the query
language for RDF databases [9], instead of regular queries. The third goal of our
project can thus be rephrased as follows:

Study the problem of query answering using materialised views when
both queries and views are defined with CONSTRUCT SPARQL queries.
Devise algorithms to solve this problem, pinpoint its computational com-
plicity, and analyse its expressive power from a practical point of view.

Our advancements are as follows.

1. The study of query answering using views can be understood as the problem of
composability for SPARQL queries: given CONSTRUCT queries $Q_1$ and $Q_2$,
is there a query $Q_{1,2}$ such that it represents correctly the result of applying first
$Q_1$ and then $Q_2$? We have studied this problem, and shown that this is possible
as long as CONSTRUCT queries do not use blank in templates (intuitively, if
we do not allow our views to invent new values), but not possible if we allow
them. Preliminary results have been published in a conference last year [6],
but this investigation opens up the path for a more detailed study, with several
additional open questions: What should we add to the language to maintain
composability in the most general case? Can any sequence $Q_1, \ldots, Q_n$ of queries
be represented as a composition of just two CONSTRUCT queries? During our
third year we continued researching these questions. We have some additional
results, and a journal submission is under preparation.

2. As mentioned on previous year’s reports, we identified an opportunity to apply
our results in a somewhat different setting: to devise a recursive functionality
for SPARQL that is defined as the consecutive application of multiple CON-
STRUCT queries (views). On our second year we finished the study and sent
the results to a conference [15]. This functionality can be easily implemented
on top of existing SPARQL processors, and forms a language that is much more powerful than other previous proposals based on iteration. We also showed that our proposal can be effectively deployed in practice: our experiments show very adequate computing times when considering the sort of queries we were computing. It is also worth mentioning that this paper was nominated for the best paper award at ISWC 2015, the conference where it was presented.

**Fourth goal: Data Integration.** Since we stood with SPARQL queries for our second year, it made sense to continue studying SPARQL. In this respect, the most interesting data integration scenarios for SPARQL demand leveraging web data, integrating it into SPARQL. Thus, the third goal of our project can be stated in more detail as follows.

Continue studying our model of views for graph data, but now in the context of integrating RDF data with web data.

Our advancements are as follows.

1. The most ubiquitous web data present is that of Web APIs. Thus, it makes sense to understand how to combine RDF answers and web APIs into a single system. As a format for Web APIs we chose JSON, the default API language used nowadays. During our third year we had numerous meetings with Elena Botoeva, a researcher from Bolzano that was itself studying data integration between ontologies and JSON data (see e.g. [3]). These interactions were vital to design a way of querying both SPARQL and web API data, into a single querying formalism. We have submitted a demo to this year’s ISWC [5], and we are currently working on a conference version. Some of these results were also incorporate into the thesis of Fernando Surez.

2. The other possibility is to make use of the web infrastructure present in RDF itself, as part of the Web of Linked Data and the linked data principles (see [2] for a more detailed discussion). We have started a research plan to understand what is the best way to implement navigational SPARQL queries over the web; we already have a workshop paper with some preliminary results [1].

**References**

[1] J. Baier, D. Daroch, J. L. Reutter, and D. Vrgoc. Property paths over linked data: Can it be done and how to start?


ACHIEVEMENTS OF THE PROJECT:
- Research visit(s) to other institution(s).
- Outreach activities related to the project’s main topic.
- Any other contribution, not addressed elsewhere, that you consider important.

The maximum length for this section is 1 page. (Arial or Verdana, font size 10).

Technical seminars and talks.
- I was invited to give a 3-hour tutorial on RW School 2016 (http://www.abdn.ac.uk/events/rr-2016/rw-summer-school-2016/), where I mostly talked about the results on expressivity and containment of graph query languages, developed on the first two years of this project.

- A Keynote at the 7th COLD workshop, co-located with ISWC (https://www.dcc.uchile.cl/cold2016/). Here I talked about the problem of data integration and graph queries in a web context, which we studied in the third year of this project.

Training of human capital. During the third year we saw the completion of Fernando Suarez’s Master thesis, informed as part of the project. We also worked with Adrin Soto, a first year phD student, and with Jaime Castro, a student just starting his masters.

Non-technical talks and other products
- On July 2016 I went to Universidad Técnica Federico Santa María, in Valparaíso, Chile, to present to local students the implications of my project and the future problems uncovered by my results.

- On June 2016 I participated on the 2016 version of the AMW School in Ciudad de Panamá, Panamá, and presented the basics of graph databases to a mix audience of students and people linked with databases.

- On January 2015 I organised a Summer School in my university, which was open to undergraduate and early postgraduate students of Chilean and latin-american universities. One of the recurrent topics of this school was querying graph databases, and in fact the invited researcher Elena Botoeva presented a tutorial in this school. Needles to say, this tutorial had a introductory note, aimed as I said at undergraduate students.
COOPERACIÓN INTERNACIONAL

N° Proyecto: 11130648
Nombre Colaborador (a) Extranjero (a): ELENA BOTOEVA
Afiliación Institucional Actual: FREE UNIVERSITY OF BOZEN-BOLZANO
Fechas de estadía Desde: 28/11/2015 Hasta: 31/01/2016

Describa las actividades realizadas y resultados obtenidos. Destaque su contribución al logro de los objetivos del proyecto. Si es pertinente, indique las publicaciones conjuntas generadas, haciendo referencia a lo informado en la etapa Productos. Agregue en la etapa anexos la información necesaria.

With Elena we researched the possibility of using graph data languages in the context of description logics, as well as a means to integrating other forms of semistructured data (such as JSON) into an ontology-based setting. The discussions were important at the time of selecting the best way of integrating graph queries with other forms of data, but unfortunately the discussions have not yet materialized into a publication.

It is also worth mentioning that Elena gave a few talks during her stays, including a tutorial for students at PUC Chile in description logics, her area of expertise.

We do expect to be able to materialize our discussions into a paper soon, but probably after a second visit.

PRODUCTOS

ARTÍCULOS

Para trabajos en Prensa/Aceptados/Enviados adjunte copia de carta de aceptación o de recepción.

<table>
<thead>
<tr>
<th>Nº</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autor (a)(es/as)</td>
<td>Juan L. Reutter; Miguel Romero; Moshe Y. Vardi</td>
</tr>
<tr>
<td>Nombre Completo de la Revista</td>
<td>Theory of Computing Systems</td>
</tr>
<tr>
<td>Título (Idioma original)</td>
<td>Regular Queries on Graph Databases</td>
</tr>
<tr>
<td>Indexación</td>
<td>ISI</td>
</tr>
<tr>
<td>ISSN</td>
<td>1432-4350</td>
</tr>
<tr>
<td>Año</td>
<td>2016</td>
</tr>
<tr>
<td>Vol.</td>
<td>1</td>
</tr>
<tr>
<td>Nº</td>
<td>1</td>
</tr>
<tr>
<td>Páginas</td>
<td>1-53</td>
</tr>
<tr>
<td>Estado de la publicación a la fecha</td>
<td>Publicada</td>
</tr>
</tbody>
</table>

Otras Fuentes de financiamiento, si las hay:
Iniciativa Científica Milenio, NC120004

Enviar documento en papel: no
Archivo(s) Asociado(s) al artículo: Reut_RQ.pdf
<table>
<thead>
<tr>
<th>N°</th>
<th>Título (Idioma original)</th>
<th>Autor (a)(es/as)</th>
<th>País</th>
<th>Ciudad</th>
<th>Fecha Inicio</th>
<th>Fecha Término</th>
<th>Nombre Publicación</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Recursion in SPARQL</td>
<td>Reutter; Soto; Vrgoc</td>
<td>Estados Unidos de América</td>
<td>Bethlehem</td>
<td>12/10/2015</td>
<td>15/10/2015</td>
<td>International Semantic Web Conference</td>
</tr>
<tr>
<td>2</td>
<td>SPARQL with Property Paths</td>
<td>Kostylev; Reutter; Romero; Vrgoc</td>
<td>Estados Unidos de América</td>
<td>Bethlehem</td>
<td>12/10/2015</td>
<td>15/10/2015</td>
<td>International Semantic Web Conference</td>
</tr>
<tr>
<td>3</td>
<td>Regular Queries on Graph Databases</td>
<td>Reutter, J.; Romero, M.; Vardi, M.</td>
<td>Estados Unidos de América</td>
<td>Bethlehem</td>
<td>12/10/2015</td>
<td>15/10/2015</td>
<td>International Semantic Web Conference</td>
</tr>
</tbody>
</table>
Nombre del Congreso: ICDT 2015
País: BELGICA
Ciudad: Bruselas
Fecha Inicio: 23/03/2015
Fecha Término: 27/03/2015
Nombre Publicación:
Año:
Vol.:
Nº:
Páginas:
Envía documento en papel: no
Archivo Asociado:
Gmail_-_resubmission_of_ICDT_2015_paper_43.pdf

Nº: 4
Autor (a)(es/as): Kostylev, E.; Reutter, J.; Ugarte, M.
Título (Idioma original): CONSTRUCT Queries in SPARQL
Nombre del Congreso: ICDT 2015
País: BELGICA
Ciudad: Bruselas
Fecha Inicio: 23/03/2015
Fecha Término: 27/03/2015
Nombre Publicación:
Año:
Vol.:
Nº:
Páginas:
Envía documento en papel: no
Archivo Asociado:

TESIS/MEMORIAS

Nº: 1
Nombre y Apellidos del(de la) Alumno(a): Fernando Suarez Barría
Nombre y Apellidos del(de la) Tutor(a): Juan Lorenzo Reutter de la Maza
Título Grado: Magister
Institución: pontificia universidad católica de chile
País: CHILE
Ciudad: SANTIAGO
Estado de Tesis: Terminada
Fecha Inicio: 31/03/2015
Fecha Término: 14/09/2016
Envía documento en papel: no
Archivo Asociado:
tesisFdo.pdf


ANEXOS

Nº: 1
Archivo Asociado: Informe_Difusion_11130648.pdf


A continuación se detallan los anexos físicos/papel que no se incluyen en el informe en formato PDF.