01Comisión Sectorial de Investigación Científica

Universidad de la República

Solicitud de Fondos para Financiamiento
Proyectos de Investigación
Llamado 2004

1- Datos del investigador responsable

Nombre y Apellido	Adriana Marotta
Cédula de Identidad	2760527-1 Fecha Nasc.(AA-MM-DD) : 69/6/29
Nº. de funcionario	34070
Descripción del Cargo (*)	Grado: 3 Dedicación horaria: 40 Efec./Inter.: Efec D.T.: no
Servicio universitario y repartición	Instituto de Computación – Facultad de Ingeniería
Teléfono y fax (*)	711 42 44, fax 711 04 69
E-mail	amarotta@fing.edu.uy

(*) del Servicio y repartición donde se realizará el proyecto.

2- Datos del Proyecto

Título del proyecto	Análisis de factores de calidad en Sistemas de Información Multi-fuente.		
Duración (meses)	24		
Agraria ☐ Básica ☐ Salud ☐ Social ☐ Tecnológica ☒ (*)	Disciplina: Informática Subdisciplina: Sistemas de Información Palabras claves (hasta 3): Calidad, Sistemas de Información Multi-fuente, Sistemas de Información Heterogéneos		
Monto solicitado a C.S.I.C (en pesos uruguayos)	1er año 119963.6	2 do año 119971.8	Total 239935.4

(*) Marque una sola opción. Al marcar el área en la cuál se inscribe el Proyecto presentado, se debe señalar el área de conocimiento independientemente del servicio al que pertenece el investigador. Los Proyectos que no se ajusten al área señalada serán reasignados al área correspondiente.

3- Otras fuentes de financiamiento de este Proyecto (total o parcial)

Indique si este mismo proyecto ha solicitado otras fuentes de financiamiento

Institución	Monto ($)	En estudio	Aprobado (período de ejecución)	Nombre del Responsable	No aprobado

Adjuntar los comprobantes necesarios.

4- Otras fuentes de financiamiento relacionadas

Indique si el Proyecto se enmarca en algún Plan General, Línea de Investigación o Programa de algún grupo o Institución.

si

Especifique:
El proyecto se enmarca dentro de la línea de investigación “Calidad en Sistemas Multi-fuente”, del grupo Concepción de Sistemas de Información (CSI) del Instituto de Computación de la Facultad de Ingeniería.

Si contestó afirmativamente el punto 4- indique:

i) si el Plan General, Línea de Investigación o Programa de algún grupo o Institución aludido en el punto 4 tiene financiamiento. **No**

ii) si algún proyecto enmarcado en Plan General, Línea de Investigación o Programa de algún grupo o Institución con similitudes con el que se está presentando a CSIC, cuenta con financiación

Fuente	Monto ($)	Período	Título del Proyecto o Programa	Nombre del Responsable

5- Antecedentes

Proyectos financiados por C.S.I.C bajo su responsabilidad (Iniciación, I+D o Sector Productivo)

C.S.I.C	Nombre del proyecto	Programa
1991 si **no** 1992 si **no** 1993 si **no** 1994 si **no** 1996 si **no** 1999 si **no** 2001 si **no** (marque la opción que corresponda)		

6- Aspectos docentes

Indique si este Proyecto incorpora trabajos de Grado o Postgrado

Grado si Posgrado si

Especifique

Posgrado:

- Tesis de Doctorado en curso. Adriana Marotta. “Manejo de cambios en la calidad de datos de un Sistema de Información Multi-fuente”. Pedeciba – Universidad de la República. Supervisor: Raúl Ruggia.
- Tesis de Doctorado en curso. Verónika Peralta. “Calidad de Datos en Sistemas de Información Heterogéneos” Pedeciba – Universidad de la República (Uruguay) y Universidad de Versailles (Francia). Supervisores: Raúl Ruggia (Universidad de la República) y Mokrane Bouzeghoub (Universidad de Versailles)

- Tesis de Maestría a comenzar. Salvador Tercia. Pedeciba – Universidad de la República. Supervisor: Raúl Ruggia.

Proyectos de grado de la carrera Ingeniería en Computación:

- “Plataforma para Evaluación de la Calidad de Datos”. María José Rouiller, Javier Penengo. Supervisores: Verónika Peralta y Raúl Ruggia. En curso.
- “Implementación de un Herramienta de Evaluación de la Frescura de los Datos”. Proyecto correspondiente al curso Ingeniería de Software, a proponer en setiembre de 2004. Supervisor: Verónika Peralta.
- Proyecto de grado a proponerse en 2005.
- Proyecto de grado a proponerse en 2006.

7- Descripción del Proyecto

Detalles de la Investigación (no más de 20 carillas):

A.- Resumen de la investigación (no más de 250 palabras y en una carilla aparte.-.(*)

Los avances tecnológicos de los últimos años han permitido el desarrollo de sistemas de información de gran porte que brindan acceso a grandes volúmenes de información distribuida en múltiples fuentes de datos heterogéneas. A medida que la cantidad de datos potencialmente recuperados aumenta, los usuarios se preocupan más por la calidad de sus resultados. La calidad de los resultados depende principalmente de la calidad de los datos fuentes (su coherencia, su completitud, su frescura, etc.) y de las características de los procesos que combinan estos datos (cuáles son fuentes consultadas, cómo se integran los datos, qué transformaciones se realizan, etc).

Nuestra propuesta se desarrolla en el contexto de Sistemas de Información Multi-fuente (MSIS), y trabaja en base a la definición y manejo de propiedades de calidad. Para las propiedades de calidad distinguimos entre los valores de calidad provistos por el sistema y los valores de calidad requeridos por el usuario, siendo mejor la calidad del sistema cuanto más se acercan los primeros a los segundos.

Nuestro objetivo es brindar técnicas y mecanismos que permitan evaluar la calidad del MSIS, tomar decisiones de diseño del mismo considerando la calidad, y resolver los problemas que surgen a raíz de los cambios en la calidad de las fuentes de datos. Para esto proponemos construir un marco de trabajo donde se implementen dichas técnicas y mecanismos para un grupo, lo más general posible, de propiedades de calidad.

B.- Fundamentación y antecedentes.

Motivación

Los avances tecnológicos de los últimos años han permitido el desarrollo de sistemas de información de gran porte que brindan acceso a grandes volúmenes de información distribuida en múltiples fuentes de datos. Si bien dichos sistemas han sido propuestos y utilizados desde hace más de una década, han ido tomando más importancia en los últimos años, tanto a nivel académico y de investigación como a nivel de uso práctico en la industria. El interés creciente que se percibe en este tipo de sistemas se debe principalmente a la proliferación de información almacenada en diversas plataformas y formatos, la cual muchas veces necesita ser visualizada en forma integrada y presentada según los requerimientos del consumidor, y a los grandes avances que han habido en las comunicaciones, los cuales posibilitan la interconexión entre sistemas de información. Con el advenimiento y la generalización del uso de la Web, se ha comenzado a considerar a la misma como una fuente de información más, que puede ser utilizada por los sistemas de información como proveedora de datos.

A medida que la cantidad de datos potencialmente recuperados aumenta, los usuarios se preocupan más y más por la calidad de sus resultados. Varios estados del arte (*surveys*) y estudios empíricos han demostrado la importancia de la calidad en el diseño de sistemas de información, en particular, en sistemas de integración de datos [WS96][Shi03][MW04].

La calidad de los resultados depende principalmente de la calidad de los datos fuentes (su coherencia, su completitud, su frescura, etc.) y de las características de los procesos que combinan estos datos (cuáles son fuentes consultadas, cómo se integran los datos, qué transformaciones se realizan, etc). Por ejemplo, si un usuario desea realizar un viaje a España, puede consultar distintas fuentes de datos y así obtener gran cantidad de datos sobre pasajes, hoteles, estadías, paquetes turísticos, etc. Un primer problema que se presenta es el de decidir qué fuentes de datos consultar y luego cómo combinar los datos de dichas fuentes. La calidad de los resultados depende de las fuentes consultadas (si contienen información parcial de algunas compañías o presentan listas suficientemente completas, si los datos se actualización frecuentemente o no, etc.) y de la manera de combinar los datos (si se consideran los datos de la fuente más confiable o si se realiza la unión de datos de distintas fuentes, si el tiempo de cálculo de los resultados es razonable, etc).

Proveer a los sistemas cuya información proviene de diversas fuentes, la posibilidad de manejar información acerca de la calidad que reciben y que brindan a sus usuarios, es sin duda un importante aporte. Esto se cumple especialmente en los dominios de aplicación donde la calidad de la información cobra mayor importancia, como en bio-informática.

Características técnicas y problemas

Los Sistemas de Información Multi-Fuente (Multi-Source Information Systems, en adelante MSIS) son sistemas de información que integran datos provenientes de múltiples fuentes de datos y proveen a los usuarios un acceso uniforme a los mismos. Dichos sistemas pueden tener diferentes características, y atendiendo a éstas se han

propuesto distintas áreas de investigación que se concentran en distintos tipos de MSIS. Uno de estos tipos, por ejemplo, es el Data Warehouse, cuyo objetivo es satisfacer requerimientos de información para toma de decisiones, obteniéndose dicha información a partir de fuentes de datos, en general bases de datos operacionales, y transformándose la misma con procesos de limpieza, integración, cambios de formatos, etc. La arquitectura básica que consideramos para un MSIS, la cual se muestra en la Figura 1, cuenta con dos capas: (i) la capa de las fuentes de datos con sus traductores (conocidos como "wrappers"), y (ii) la capa del Mediador, que es el módulo que integra toda la información de las fuentes y presenta la visión de una única base de datos hacia los usuarios.

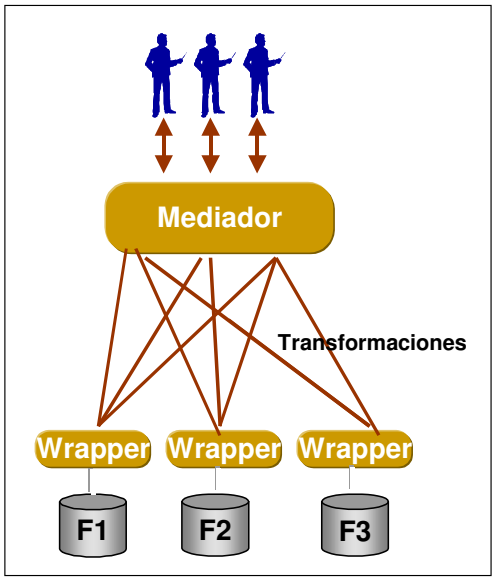

Figura 1 – Bosquejo de un MSIS

Dada la cantidad y variedad de fuentes que pueden participar en un MSIS y considerando además la autonomía de las mismas, es muy importante conocer y saber utilizar la información acerca de su *calidad*. En nuestro enfoque, la información de calidad se basa en un conjunto de propiedades sobre los datos de las fuentes (por ejemplo, *frescura*, *confiabilidad*, *completitud*) y sobre la forma de extraer, integrar y transformar esos datos (por ejemplo, *tiempo de respuesta, disponibilidad*). Dichas propiedades son evaluadas en las fuentes y en el sistema total (valores reales) y a la vez requeridas por los usuarios del MSIS (valores esperados). Consideramos que la calidad es mejor cuanto más se acercan los valores reales de las propiedades del sistema a los valores requeridos por el usuario.

Considerar la calidad brindada por las fuentes en un MSIS, es de gran utilidad para la selección de las fuentes de donde se va a extraer información y además puede tener gran incidencia en el diseño del propio sistema. Por ejemplo, si dos fuentes de datos proporcionan la misma información, se puede elegir alimentar el MSIS a partir de la fuente que tenga mejor calidad (datos más recientes, alta confiabilidad, etc). En otro caso, podría cambiarse el diseño del sistema de integración para mejorar ciertos valores de calidad, por ejemplo, si determinada fuente está accesible sólo durante ciertos períodos, pueden materializarse ciertos datos de manera de aumentar la disponibilidad

de la información. También podría ser necesario eliminar ciertos ítems de información que provee el MSIS como consecuencia de no contar con datos de calidad suficiente para poblarlos.

Es necesario entonces, estudiar cómo influyen las propiedades de calidad de las fuentes y la implementación del MSIS en la calidad del mismo. Aspiramos a proponer técnicas para: (i) calcular los valores de calidad alcanzados por el sistema a partir de los valores de calidad de las fuentes, es decir, *evaluar* la calidad del sistema, y (ii) mejorar la calidad del sistema estudiando la implementación del mismo (alternativas para diseñarlo y poblarlo).

La evaluación de la calidad de los datos en un MSIS implica: (i) la selección de las propiedades de calidad a evaluar, (ii) la selección de métricas apropiadas para esas propiedades, (iii) la determinación de las características del sistema sistema (costos, políticas, restricciones, etc.) que influyen en la medición de la calidad, (iv) la implementación de algoritmos de evaluación que tengan en cuenta las características del sistema y calculen los valores de calidad siguiendo las métricas, y (v) la ejecución de los algoritmos para medir la calidad de los datos producidos por el sistema.

Se deberá hacer un estudio en profundidad de un grupo de propiedades de calidad, ya que por la diversidad de su esencia, resulta difícil proponer técnicas generales aplicables para cualquier propiedad. Por lo tanto, creemos que lo mejor es construir un marco de trabajo (framework) donde se implementen técnicas y mecanismos para propiedades particulares, que puedan ser generalizadas para otras propiedades.

Otro problema importante a tener en cuenta en este contexto es el de cambios en las propiedades de calidad de las fuentes. Un cambio en un valor de calidad en una fuente puede tener como consecuencia que cambie un valor de calidad del sistema, dejando de satisfacer los requerimientos de calidad del mismo. Por lo tanto, un cambio de calidad en las fuentes puede significar que se deba re-diseñar una parte o todo el sistema para lograr satisfacer los requerimientos de calidad del mismo.

El fenómeno de cambios en la calidad de las fuentes tiene características particulares (diferentes, por ejemplo, a las de evolución de los esquemas de las fuentes), que se deben tener en cuenta a la hora de construir soluciones para manejarlo. Algunas de ellas las comentamos a continuación. Por su naturaleza, estos cambios pueden ser muy frecuentes y a la vez difíciles de predecir. Por ejemplo, si consideramos la propiedad *tiempo de respuesta*, vemos que su valor puede cambiar en cualquier momento, ya que depende de factores que son externos a la fuente, como puede ser el tráfico existente en la red de comunicaciones. Una alternativa posible para tener algún tipo de información acerca de cambios futuros de esta propiedad es utilizar datos estadísticos. Por otro lado, las consecuencias que pueden tener estos cambios en el MSIS son diversas; por ejemplo, puede suceder que una fuente que proveía información al sistema sea dejada fuera del mismo por haber cambiado cierto valor de calidad, o que un cambio de este tipo en una fuente cause que otra fuente deba alcanzar un mayor valor de calidad para que se mantengan los valores del MSIS dentro de los requeridos. Finalmente, este tipo de cambios en muchos casos no son originados por el administrador de la fuente, e incluso su causa no se genera en la fuente. En otros casos, podría suceder que un cambio en la calidad de una fuente sea provocado, cambiando algún procedimiento local, con la finalidad de mejorar la calidad del MSIS.

Sería de gran interés poder contar con técnicas de manejo de cambios en la calidad de las fuentes en MSIS, que permitan controlar y minimizar el impacto, así como automatizar dicho manejo. Estas técnicas deberán apoyarse en dos estrategias principales: (i) la modificación de valores de calidad (o valores relacionados con estos) en las fuentes y/o en las transformaciones de los datos, y (ii) la modificación de las transformaciones de los datos y/o del esquema de datos global provisto por el MSIS. Además, por las características propias del fenómeno es posible proponer "estrategias pro-activas", determinando a priori rangos de valores aceptados para las propiedades de calidad de las fuentes, e incluso utilizando modelos probabilísticos que nos provean información útil para lograr más estabilidad en el MSIS (mediante modificaciones en valores de calidad o en el diseño del sistema, en forma preventiva).

Este problema debería ser estudiado en el marco de trabajo mencionado anteriormente, comenzando también con algunas propiedades de calidad y teniendo como objetivo generalizar las soluciones a otras propiedades.

Trabajos relacionados

En un estudio del estado del arte en los temas de calidad y cambios en las fuentes en MSIS, hemos encontrado que todavía hay muchos problemas abiertos. Hay muchos enfoques diferentes; algunos se concentran en el análisis y definición de propiedades de calidad, mientras que otros además hacen propuestas para el manejo de calidad (cálculos, especificación de la metadata, etc.).

Algunos trabajos estudian la calidad de los datos desde la perspectiva del usuario. En [WS96] y [SLW97], Wang et al. proponen un marco de trabajo que captura las propiedades de calidad que son más importantes para los usuarios, basados en dos surveys, e identifican los problemas claves asociados a la calidad. Se basan en el concepto de que datos de alta calidad son datos que se adecuan para el uso de los consumidores de datos.

En [LSKW01] se propone una metodología para la evaluación y el mejoramiento de la calidad de información. También presentan dimensiones de calidad y clasificaciones en categorías, y proveen algunas tablas resumiendo la perspectiva de los académicos y la de los prácticos (especialistas de las organizaciones, consultores, etc.). En [BWP+98] modelan un sistema de información como un sistema de producción y presentan un conjunto de conceptos y procedimientos para determinar la calidad de la información. Presentan algunas propiedades relevantes de calidad tales como puntualidad (timeliness), exactitud (accuracy) y costo.

En [JV97] los autores presentan el problema de modelado y mediciones de la calidad en sistemas de Data Warehousing (DW), y expresan que debe existir un mapeo entre los componentes del DW y el modelo de calidad. Discuten varias relaciones entre parámetros de calidad y aspectos de diseño y operacionales en un DW. En [JQJ98] se presenta un meta-modelo formal para representar la formulación de objetivos de calidad y mediciones de calidad en un DW. En [HH02] se presenta también una propuesta para manejar calidad en DW a través de un sistema basado en metadata, mientras que en [MRV00] se presenta una propuesta para selección y ranking de fuentes de datos, basados en metadata sobre contenido y calidad de los datos de las fuentes.

Por otro lado, el trabajo presentado en [NLF99] propone tener en cuenta algunas propiedades de calidad en el diseño de sistemas de mediación. Estudian cómo propagar los valores de las propiedades de calidad de las fuentes al mediador. Los valores propagados para diversos factores de calidad se combinan en una suma ponderada. En [GTS+04] discuten diversos problemas relacionados con la evaluación y el aseguramiento de la calidad de los datos. En particular, prometen un álgebra para combinar valores reales de calidad de las fuentes y así calcular la calidad de los resultados.

El tema medición de la calidad es abordado con profundidad en [PLW02], donde la evaluación de la calidad se presenta como dependiente de "percepciones subjetivas" y "mediciones objetivas". Las evaluaciones subjetivas reflejan necesidades y experiencias de los consumidores y personas que trabajan con los datos, mientras que las evaluaciones objetivas involucran métricas para el conjunto de datos en cuestión. Los autores proponen, realizar los dos tipos de mediciones y comparar los resultados, identificando discrepancias y determinando acciones a tomar. Además, presentan un conjunto de dimensiones de calidad de datos y la métrica a utilizar en cada caso.

Existen numerosos trabajos que se centran en el estudio de una o dos propiedades de calidad y su impacto en el diseño del sistema, ya sea estudiando cómo mejorar los valores de una propiedad o cómo diseñar el sistema bajo restricciones de calidad. Como ejemplos podemos citar los siguientes trabajos: En [BR02] se balancea latencia y frescura en sistemas de caching, en [TB99] se balancea performance y frescura en un contexto del materialización de vistas, en [LPH+03] se combinan técnicas de materialización y caching para balancear tiempo de respuesta y frescura, en [HZ96] se construyen mediadores híbridos (virtuales / materializados) para asegurar la frescura y la consistencia de los datos. Si bien dichos trabajos presentan soluciones para escenarios o sistemas concretos, son propuestas interesantes para adaptar a otros contextos e intentar generalizar en un marco de trabajo más amplio.

No hemos encontrado trabajos que se centren en el problema de cambios de los valores de calidad en sistemas multi-fuente. Un problema relacionado es la evolución de los esquemas fuente en este tipo de sistemas. Existen varios artículos escritos por Rundensteiner y su equipo de investigación en donde abordan el problema de la adaptación de vistas materializadas ante cambios ocurridos en las fuentes. En [RLN97, NLR98, NR99] presentan taxonomías y clasificaciones de los problemas de adaptación de vistas, un ambiente para la resolución de estos problemas, y un lenguaje que permite especificar, en el momento de definirse la vista, criterios para la evolución. En [ZR99, ZR00, CCR02] se concentran en la resolución de problemas de concurrencia entre actualizaciones de datos y cambios de esquema. Finalmente, en [KR02] se focalizan en el mantenimiento de vistas materializadas, donde la evolución de los esquemas fuentes puede provocar en la vista tanto cambios de esquema como de datos. En [MP02] se presenta un enfoque de transformación de esquemas para propagar evolución en las fuentes hacia el esquema integrado. En [BK00] los autores estudian evolución de las fuentes con una granularidad a nivel de fuente, es decir los cambios considerados son sobre agregado o eliminación de fuentes completas en el MSIS. En un trabajo posterior [BFZS02] estudian en detalle la evolución en un MSIS a partir de las fuentes, considerando al esquema integrado como vistas SQL y modificando los grafos de las consultas de las vistas, al propagar evolución.

[Qui99] es el único trabajo que hemos encontrado que relaciona los problemas de evolución y calidad en MSIS. Proveen una taxonomía de operaciones de evolución y los factores de calidad a los que afectan cada una de ellas. El artículo se centra principalmente en la propuesta de un meta-modelo para arquitectura y procesos de DW, y una especialización de éste en un meta-modelo para evolución en DW. Junto con esto se presentan patrones de procesos de evolución.

Antecedentes del grupo CSI

Nuestro equipo de investigación, grupo CSI (Concepción de Sistemas de Información, www.fing.edu.uy/inco/grupos/csi), perteneciente al Instituto de Computación de la Facultad de Ingeniería, Universidad de la República, ha trabajado en el área de sistemas de información multi-fuente desde hace aproximadamente 7 años. Durante este período se han realizado dos proyectos, apoyados por CSIC, del tema Data Warehouse [CSI00] [CSI02], tres tesis de Maestría [Mar00][Car00][Per01] y numerosos proyectos de grado de la carrera Ingeniería en Computación, centrados en el mismo tema. También se han dictado cursos de actualización profesional, de posgrado y de posgrado a distancia, en esta área.

En el tema específico de Calidad en MSIS se han comenzado a desarrollar dos tesis de doctorado (Adriana Marotta y Verónika Peralta), y se mantiene un fuerte contacto de cooperación con el equipo del Profesor M. Bouzeghoub del Laboratorio PRISM de la Universidad de Versailles, Francia. Cabe agregar que la tesis de doctorado de la integrante de este proyecto Verónika Peralta se realiza en “modalidad sándwich”, en co-supervisión entre el Prof. Raul Ruggia (integrante de este proyecto) y el Prof. Mokrane Bouzeghoub (jefe del grupo de Versailles), y la tesis de Adriana Marotta (quien dirige el presente proyecto) se está realizando en coordinación con el grupo de Versailles, apoyándose en pasantías realizadas en el mismo (una fue realizada en el pasado año y otra está planificada para enero del año próximo).

En base al trabajo que se viene realizando en este tema, hemos realizado una publicación presentando el problema general de Calidad en MSIS [MR03] y otra que se focaliza en el análisis de un factor de calidad en particular [BP04].

También en el contexto de este trabajo, se realizaron el pasado año dos proyectos de grado de la carrera Ingeniería en Computación. Uno implementó una herramienta para la generación automática de vistas para MSIS [OO04], basándose en la propuesta de [KB99], la cual se prevé como de mucha utilidad para combinar con las técnicas que se propongan para diseño del MSIS considerando calidad y para manejo de cambios. El otro proyecto realizó una primer experiencia de prototipación de plataforma para evaluación de calidad en MSIS [FC04].

El grupo CSI, además ha trabajado en el tema evolución de esquemas fuentes en MSIS, primeramente en una parte de la tesis de maestría [Mar00] se analiza el problema de evolución de esquemas fuentes en DW y se proponen algunas soluciones para esto, y por otro lado en el trabajo [MMR01] se propone una arquitectura para Web DW y se plantean taxonomías de cambios en las fuentes y en los distintos componentes del sistema, junto con soluciones para casos particulares planteados.

C.- Objetivos generales y específicos.

Objetivo general:

Proponer un marco de trabajo (framework) para el manejo de propiedades de calidad en un MSIS[1], que nos permita evaluar la calidad del sistema, tomar decisiones de diseño del mismo, y resolver los principales problemas relacionados con los cambios en la calidad de las fuentes de datos.

Objetivos específicos:

- Identificar un conjunto minimal de propiedades de calidad relevantes para un MSIS, que serán tomadas como base para el proyecto. No se trata de identificar un conjunto completo de propiedades, sino un conjunto mínimo que pueda usarse como base representativa para la investigación.
- Analizar técnicas de evaluación para las propiedades de calidad seleccionadas y determinar las características de las fuentes o del sistema que influyen en la calidad del MSIS.
- Analizar el impacto de las propiedades de calidad de las fuentes y requerimientos de calidad del sistema, en el diseño del MSIS, considerando solamente las propiedades de calidad seleccionadas.
- Estudiar el problema de gestión de cambios en los valores de calidad de las fuentes, para las propiedades de calidad seleccionadas, y proponer estrategias para el manejo.
- Especificar un mecanismo general para evaluación de la calidad en un MSIS.
- Especificar técnicas generales para el manejo de cambios de calidad en las fuentes en un MSIS.
- Implementar un prototipo de un marco de trabajo para el manejo de propiedades de calidad en un MSIS.

D.- Especificación de las preguntas que busca responder el proyecto.

- ¿Cómo influyen las propiedades de calidad que poseen las fuentes en la calidad global de un MSIS?
- ¿Cómo influyen las propiedades de calidad que poseen las fuentes en el diseño de un MSIS?
- ¿Qué decisiones de diseño se deberían tomar frente a los valores de calidad ofrecidos por las fuentes y teniendo en cuenta los valores de calidad requeridos por el usuario del sistema?
- ¿Cómo y en qué casos influyen en un MSIS los cambios en los valores de calidad de las fuentes?

[1] MSIS – Sistema de Información Multi-fuente

- ¿Cuál sería un mecanismo adecuado para adaptar el MSIS cuando ocurren cambios en los valores de calidad de las fuentes?

E.- Estrategia de investigación.

Para alcanzar los objetivos que nos fijamos, proponemos trabajar en base a dos centros de interés, (i) el problema de evaluación de calidad y diseño en base a la calidad en un MSIS, y (ii) el problema de cambios en la calidad de las fuentes y su repercusión en el MSIS. La estrategia a seguir es la de atacar estos dos problemas en paralelo, compartiendo resultados que puedan ser de utilidad a ambos. Además, dado que las propiedades de calidad pueden ser muy diversas en su naturaleza es muy difícil encontrar técnicas generales para los problemas relacionados con ellas. Por lo tanto, en ambos casos ((i) y (ii)) se propone comenzar por buscar soluciones para un conjunto pequeño de propiedades de calidad (podrían ser 3 propiedades), y luego generalizar los resultados para otras propiedades de calidad. Gran parte del trabajo, por ejemplo el relativo al estudio de las características generales de las propiedades de calidad elegidas, será realizado en común por las personas que se encuentren trabajando en los problemas (i) y (ii).

Desde el punto de vista de organización del trabajo, y esto es una estrategia común en el grupo de investigación, se trata de escalonar el trabajo de doctorado, maestría y estudiantes de grado. En este sentido, desde 1999 se trabaja con proyectos de grado de Ing. en Computación que corresponden, ya sea a implementaciones de prototipos o partes de ellos bajo la dirección de un estudiante de Maestría, o también a la resolución de un caso real usando las técnicas y herramientas propuestas.

Otra estrategia relevante para este proyecto es la cooperación con grupos internacionales. En este sentido, este proyecto tiene una fuerte cooperación con el Laboratorio PRISM (Universidad de Versailles – Francia). El grupo de Concepción de Sistemas de Información del Laboratorio PRISM se especializa (desde fines de los años ’80) en el área de técnicas y herramientas para Sistemas de Información, y en los últimos años se ha orientado al diseño y manejo de sistemas de información multi-fuente, y en particular al manejo de propiedades de calidad en estos sistemas. Este ultimo es el tema principal de cooperación actualmente.

F.- Actividades específicas.

A1. Definición y caracterización de un conjunto minimal de propiedades de calidad (posiblemente 3).

A2. Estudio de una de las propiedades definidas, en diferentes escenarios, determinando métricas, características del sistema relacionadas a su medición y algoritmos de evaluación para los diferentes escenarios y métricas.

A3. Estudio del impacto de la propiedad estudiada en A2 en el diseño del MSIS.

A4. Propuesta de un marco de trabajo para el manejo de la calidad basándonos en el estudio realizado en A2 y en A3.

A5. Estudio de las otras propiedades caracterizadas en A1, aplicando los conocimientos adquiridos con el estudio de la primer propiedad. Definición de métricas, características y algoritmos de evaluación para dichas propiedades.

A6. Estudio del impacto de las propiedades estudiadas en A5 en el diseño del MSIS.

A7. Generalización del marco de trabajo definido en A4 incorporando los conocimientos adquiridos durante el estudio de las restantes propiedades.

A8. Caracterización del problema de cambios en la calidad de las fuentes en un MSIS.

A9. Estudio de posibles estrategias para actuar en forma preventiva (respecto a los cambios de calidad en las fuentes), basándose en predicciones de cambios futuros y en la definición de restricciones para los cambios ("estrategias pro-activas").

A10. Estudio de la aplicabilidad de técnicas existentes en el área de Investigación Operativa (técnicas de optimización, modelos probabilísticos, etc.) para las estrategias de A9.

A11. Propuesta de una estrategia concreta para actuar en forma preventiva (de las estudiadas en A9 y A10) aplicada a la propiedad de calidad estudiada en A2 y A3, y en el marco propuesto en A4.

A12. Estudio de posibles estrategias de propagación de los cambios de calidad de las fuentes hacia el MSIS.

A13. Propuesta de técnicas de propagación de los cambios de calidad de las fuentes hacia el MSIS, para la propiedad de calidad estudiada en A2 y A3, y en el marco propuesto en A4.

A14. Adaptación de la propuesta de A11 para las restantes propiedades de calidad definidas en A1.

A15. Adaptación de la propuesta de A13 para las restantes propiedades de calidad definidas en A1.

A16. Generalización de las técnicas para manejo de cambios de calidad en las fuentes, definidas para las propiedades particulares, como parte del marco de trabajo definido en A7.

A17. Implementación de prototipo del marco de trabajo para manejo de calidad propuesto.

G.- Materiales y métodos. (Especificar las facilidades con que cuenta: incluir área física y equipos disponibles y los equipos a adquirir en este proyecto debidamente fundamentados).

Se cuenta con un espacio físico de aproximadamente 40m2, divididos en 2 salas.
En la actualidad se cuenta con 1 PC disponible para cada investigador involucrado en el proyecto, y una máquina SUN común al grupo de investigación que funciona como servidor de disco. Este equipamiento es compartido con estudiantes que realizan sus proyectos de grado asociados al grupo de investigación.

En este proyecto se prioriza la inversión económica en recursos humanos.

H.- Cronogramas de ejecución.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	X																							
2		X	X	X	X	X																		
3							X	X	X	X	X													
4												X	X											
5														X	X	X	X							
6																		X	X	X	X			
7																						X	X	
8	X	X																						
9			X	X	X																			
10						X	X	X																
11									X	X														
12											X													
13												X	X	X	X									
14																X	X	X						
15																			X	X	X			
16																						X	X	
17				X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Nota: Las columnas de esta tabla corresponden al número de mes del proyecto. Las filas corresponden al número de actividad del proyecto.

Aclaración: La actividad 17, correspondiente a la implementación del prototipo, comienza en el mes nro. 4, porque desde ese momento ya se necesita, y ya se puede, implementar la plataforma de diseño y manejo de propiedades del MSIS, y a partir de esto, el prototipo se irá desarrollando en forma incremental, integrando de a poco las distintas propuestas que se vayan desarrollando.

I.- Describa el personal asignado al proyecto así como el personal a contratar (tanto docente como no docente). Detalle las tareas a realizar por cada uno de los integrantes del equipo de investigación.

Persona	Cargo	Tarea a realizar
Raul Ruggia	Grado 4	Dirección académica, a través de la supervisión de las tesis de doctorado y maestría.
Adriana Marotta	Grado 3	Dirección general del proyecto. A8 – A16, dirección de parte de A17.
Veronika Peralta	Grado 2	A1 - A7, dirección de parte de A17.
Salvador Tercia (a contratar)	Grado 1	Parte de A5 y A6, parte de A14 y A15.
A contratar	Grado 1	A17

J.- Resultados esperados.

Se espera obtener la especificación de estrategias y técnicas, y la prototipación de un marco de trabajo, para el manejo de propiedades de calidad en un MSIS. Este manejo incluye la evaluación de la calidad del sistema, decisiones de diseño del mismo en base a factores de calidad, y la resolución de los principales problemas relacionados con los cambios en la calidad de las fuentes de datos.

Los resultados técnicos se especificarán en forma de reportes técnicos, los cuales se generarán a medida que se van realizando las actividades.

Reporte 1: A1, A2.
Reporte 2: A3, A4.
Reporte 3: A5, A6, A7.
Reporte 4: A8.
Reporte 5: A9, A10, A11.
Reporte 6: A12, A13.
Reporte 7: A14, A15, A16.
Reporte 8: A17.

K.- Estrategias de difusión.

Se difundirán los resultados a través de la página web del grupo de investigación, así como a través de la publicación y presentación de artículos en conferencias locales e internacionales.
Se presentarán resultados en cursos de grado y postgrado, así como en cursos de actualización profesional, a los efectos de estudio y eventual uso práctico.

L.- Impacto y/o beneficios de los resultados.

Tal como se describió anteriormente, una de las áreas de problema más importantes que se presentan en los sistemas de información multi-fuente y distribuidos concierne a la evaluación de la calidad de los resultados en las consultas que se realizan sobre estos sistemas. Estos problemas, que afectan directamente a los usuarios finales, inhiben el uso efectivo de sistemas de este tipo e impiden que se aprovechen en todo su potencial.

Este proyecto tendrá impacto directo en la generación de técnicas y herramientas para resolver consultas sobre sistemas de información teniendo en cuenta las propiedades de calidad de las fuentes de información.

Los beneficios esperados consisten en el desarrollo de técnicas y prototipos para avanzar en la resolución de los problemas planteados y formar recursos humanos en el área que hagan posible la continuidad en la investigación.

M.- Referencias bibliográficas.

[BFZS02] M. Bouzeghoub, B. Farias Loscio, Z. Kedad, A.C. Salgado. Managing the Evolution of Mediation Queries. Reporte interno, Universidad de Versailles, Francia. 2002.

[BK00] Bouzeghoub, Kedad. A Logical Model for Data Warehouse Design and Evolution. DaWaK'00

[BP04] Bouzeghoub, M.; Peralta, V.: "A Framework for Analysis of Data Freshness. To appear in proc. of the 1st. International Workshop on Information Quality in Information Systems. Paris, June 2004.

[BR02] Bright, L.; Raschid, L.: "Using Latency-Recency Profiles for Data Delivery on the Web". In Proc. of the 28th Int. Conf. on Very Large Databases (VLDB'02), China, 2002.

[BWP+98] Ballow, D.; Wang, R.; Pazer, H.; Tayi, G.: "Modelling Information Manufacturing Systems to Determine Information Product Quality". Management Science, Vol. 44 (4), April 1998.

[Car00] F. Carpani. CMDM: A conceptual multidimensional model for Data Warehouse.
Master Thesis. InCo - Pedeciba, Facultad de Ingeniería, Universidad de la República. Montevideo, Uruguay. October 2000.

[CCR02] J. Chen, S.Chen, Rundensteiner. TxnWrap: A Transactional Approach to Data Warehouse Maintenance. ER'02.

[CSI00] Grupo CSI. "Técnicas y herramientas para diseño lógico y mantenimiento de Data Warehouses Relacionales". Proyecto CSIC, 2000.

[CSI02] Grupo CSI. “Diseño Lógico de Data Warehouses: Técnicas y Desarrollo de Herramientas CASE”, Proyecto CSIC, 2002. http://www.fing.edu.uy/inco/grupos/csi/esp/Proyectos/dwd_csic2002/

[FC04] F. Fajardo, I. Crispino. Una herramienta para evaluar la calidad de los datos en un sistema de información multi-fuente. Proyecto de Grado de la carrera Ingeniería en Computación. 2004.

[GTS+04] Gertz, M.; Tamer Ozsu, M.; Saake, G.; Sattler, K.: “Report on the Dagstuhl Seminar: Data Quality on the Web”. SIGMOD Record Vol. 33(1), March 2004.

[HH02] M. Helfert, C. Herrmann. Proactive Data Quality Management for Data Warehouse Systems. DMDW 2002: 97-106

[HZ96] Hull, R.; Zhou, G.: "A Framework for Supporting Data Integration Using the Materialized and Virtual Approaches". In Proc. of the 1996 ACM Int. Conf. on Management of Data (SIGMOD'96), Canada, 1996.

[JQJ98] M. A. Jeusfeld, C. Quix, M. Jarke. Design and Analysis of Quality Information for Data Warehouses. ER 1998: 349-362

[JV97] M. Jarke, Y. Vassiliou. Data Warehouse Quality: A Review of the DWQ Project. Invited Paper, Proc. 2nd Conference on Information Quality. MIT, Cambridge, 1997.

[KB99] Z. Kedad, M. Bouzeghoub. Discovering View Expressions From a Multi-Source Information System. Proceedings of 4th. Int. Conf. In Cooperative Infromation Systems (CoopIS). Scotland, 1999.

[KR02] Koeller, Rundensteiner. Incremental Maintenance of Schema-Restructuring Views. EDBT’02

[LPH+03] Li, W.S.; Po, O.; Hsiung, W.P.; Selçuk Candan, K.; Agrawal, D.: “Freshness-driven adaptive caching for dynamic content Web sites”. Data & Knowledge Engineering (DKE), Vol.47(2), 2003.

[LSKW01] Y. W. Lee, D. M. Strong, B. K. Kahn, R. Y. Wang. AIMQ: A Methodology for Information Quality Assessment. Forthcoming in Information & Management, published by Elsevier Science (North Holland). (Accepted in November 2001)

[Mar00] A. Marotta. Data Warehouse Design and Maintenance through Schema Transformations. Master Thesis - 2000. Universidad de la República. Montevideo, Uruguay. Technical Report INCO TR-01-10. ISSN 0797-6410.

[MMR01] A. Marotta, R. Motz, R. Ruggia. Managing Source Schema Evolution in Web Warehouses. Proceedings, WIIW '2001. Artículo seleccionado para el Journal of the Brazilian Computer Society. Special Issue on Information Integration on the Web. Volume 8, number 2, November 2002. ISSN 0104-6500

[MP02] Mc.Brien, Poulovassilis. Schema Evolution in Heterogeneous Database Architectures, A Schema Transformation Approach. CAISE’02

[MR03] A. Marotta, R. Ruggia. Quality Management in Multi-Source Information Systems. II Workshop de Bases de Datos. Jornadas Chilenas de Computación 2003. Chillán, Chile. Nov. 2003.

[MRV00] Using Quality of Data Metadata for Source Selection and Ranking. G. Mihaila, L. Raschid, M. Vidal. WebDB ’ 2000

[MW04] Mannino, M.; Walter, Z.: “A Framework for Data Warehouse Refresh Policies”. Technical report CSIS-2004-001, University of Colorado at Denver, 2004.

[NLF99] Felix Naumann, Ulf Leser, Johann Christoph Freytag. Quality-driven Integration of Heterogenous Information Systems. VLDB 1999: 447-458

[NLR98] Nica, Lee, Rundensteiner. The CVS Algorithm for view synchronization in evolvable large-scale information systems. EDBT’98.

[NR99] Nica, Rundensteiner. View Maintenance after View Synchronization. IDEAS’99

[OO04] A. Odriozola, D. Oliveros. Herramienta para la generación automática de expresiones de vista para Sistemas de Información Multi-fuente. Proyecto de grado de la carrera Ingeniería en Computación. 2004.

[Per01] Diseño logico de Data Warehouses a partir de esquemas conceptuales multidimensionales. Master Thesis. InCo - Pedeciba, Facultad de Ingeniería, Universidad de la República. Montevideo, Uruguay. November 2001. Technical Report INCO TR-01-17.

[PLW02] L. L. Pipino, Y. W. Lee, R. Y. Wang. Data Quality Assessment. Communications of the ACM. April 2002 / Vol. 45, No. 4ve.

[Qui99] C. Quix. Repository Support for Data Warehouse Evolution. DMDW’99

[RLN97] Rundensteiner, Lee, Nica. On Preserving views in evolving environments. KRDB’97.

[Shi03] Shin, B.: “An exploratory Investigation of System Success Factors in Data Warehousing”. Journal of the Association for Information Systems, Vol. 4(2003), 141-170, 2003.

[SLW97] D. M. Strong, Y. W. Lee, R. Y. Wang. Data Quality in Context. Communications of the ACM. May 1997/Vol. 40, No. 5.

[TB99] Theodoratos, D.; Bouzeghoub, M.: "Data Currency Quality Factors in Data Warehouse Design". In Proc. of the Int. Workshop on Design and Management of Data Warehouses (DMDW'99), Germany, 1999.

[WS96] Wang, R.; Strong, D.: "Beyond accuracy: What data quality means to data consumers". Journal on Management of Information Systems, Vol. 12, 4:5-34, 1996.

[ZR00] Zhang, Rundensteiner. DyDa: Dynamic Data Warehouse Maintenance in a Fully Concurrent Environment. DAWAK’00

[ZR99] Zhang, Rundensteiner. The SDCC Framework for Integrating Existing Algorithms for Diverse Data Warehouse Maintenance Tasks. IDEAS’99

N.- Nombre dos referentes académicos en el tema de este proyecto.

Mokrane Bouzeghoub. Laboratorio PRISM de la Universidad de Versailles, Francia. Temas: Calidad, Sistemas de Información Multi-fuente.

Elke Rundensteiner. Computer Science Department of Worcester Polytechnic Institute (WPI). U.S.A. Tema: Evolución en Sistemas de Información Multi-fuente.

Detalle de los Recursos Solicitados

8- Rubro Sueldos

8.1 Creación de becas

Escf/ Gr.	Dedicación horaria 1er. Año	Monto 1er. Año ($) (**)	Dedicación horaria 2do. Año	Monto 2do. Año ($) (**)	Monto total ($)

Sub-total $ _____ _____ _____

(**) Ver forma de cálculo en planilla excel y/o instructivo

8.2- Extensiones de cargos docentes y no docente

Tipo (*)	Grado	Dedicación actual	Dedicación a la que aspira 1er. Año	Monto 1er. Año ($) (**)	Dedicación a la que aspira 2do. Año	Monto 2do. Año ($) (**)	Monto total ($)

(*) Tipo: 1 Docente 2 No docente

Sub-total $ _____ _____ _____

(**) Ver forma de cálculo en planilla excel y/o instructivo

8. 3 Dedicaciones Compensadas Docentes

(hasta 30% del total del Proyecto)

Escf/ Gr	Horas sobre las que se calcula la Compensación	Duración meses 1er.Año	Monto de la compensación 1er. Año ($) (**)	Duración meses 2do.Año	Monto de la compensación 2do. Año ($) (**)	Monto total ($)
3	40	6	44715.5	3	22357.8	67073.3
		Sub-total $	**44715.5**		**22357.76**	**67073.28**

(**) Ver forma de cálculo en planilla excel y/o instructivo

8.4 Creación de Cargos Docentes

Escf/ Gr.	Dedicación horaria 1er. Año	Monto 1er. Año ($) (**)	Dedicación horaria 2do. Año	Monto 2do. Año ($) (**)	Monto total ($)
1	20	32949.1 (9 meses)	20	43932	76881.1
1	20	32949.1 (9 meses)	20	43932	76881.1
Sub-total $	**65898.1**		**87864.1**		**153762.3**

(**) Ver forma de cálculo en planilla excel y/o instructivo

9- Rubro Gastos (*)

*(Leer con atención las Bases del Llamado "Modalidad del Gasto" numeral **2**, inciso **e**)*

9.1- Materiales

Cantidad	Descripción	Monto anual ($)	Monto ($) total
2	Tonner para impresora	3300	6600
10000	Hojas	1300	2600
50	cds	250	500
	Sub-total $	**4850**	**9700**

9.2- Otros gastos

ATENCIÓN: No se financiarán actividades que estén contempladas en los Programas implementados por la Unidad de Recursos Humanos de la C.S.I.C.

Concepto	Monto anual ($)	Monto Total ($)
Sub-total $	_____	_____

(*) Justificación detallada de los gastos (adjuntar otra hoja si es necesario):

10- Rubro Inversiones (*)

*(Leer con atención las Bases del Llamado "Modalidad del Gasto" numeral **2**, inciso **e**)*

10.1- Equipos: describir y cuantificar los equipos que solicita a la C.S.I.C. para realizar el presente proyecto

Cantidad	Descripción	Monto ($)

Sub-total $ _____

(*) Justificación detallada de las inversiones (adjuntar otra hoja si es necesario):

10.2- Bibliografía solicitada

Tipo de publicación	Monto ($)
Publicaciones periódicas y "proceedings" de conferencias especializadas en el área.	9400

Sub-total $ 9400

Resumen de Montos

Período	Sueldos (Total del ítem 8)	Gastos (Total del ítem 9)	Inversiones (Total del ítem 10)	Total
1er año	110613.6	4850	4500	$ 119963.6
2do año	110221.8	4850	4900	$ 119971.8
Total	220835.4	9700	9400	**$ 239935.4**

CONSTANCIA DE PRESENTACIÓN EN EL SERVICIO

Fecha de presentación en el Servicio: _____

Sello del Servicio:

Firma del receptor del Servicio del Proyecto: _____

Firma del contador:_____

En caso de que corresponda,
Aprobación del Cómite de Ética: **Firma C. de Ética: _____**

Los proyectos que necesitan la aprobación del Comité de Ética (tanto para experimentación animal como la realizada en seres humanos) deberán consultar a las Comisiones responsables en el Servicio que corresponda.

Firma del solicitante: _____

Este formulario tiene valor de Declaración Jurada

Fecha de presentación ante la C.S.I.C.: _____

IMPORTANTE:

Anexar

A.- El Curriculum Vitae del solicitante y de los demás participantes en el proyecto.

B.- En caso de que el proyecto cuente con otras fuentes de financiamiento, adjuntar la documentación que establezca los montos aprobados y los plazos de ejecución de estos fondos.

C.- Entregar una copia del formulario y los C.V. en diskette 3.5 o CD, formato Word (Indispensable para el envío a evaluadores externos).