Leveraging Domain Expertise to Support Complex, Personalized and Semantically Meaningful Queries Across Separate Data Sources

Cormac Hampson, Owen Conlan
Knowledge and Data Engineering Group (KDEG)
Trinity College Dublin
Ireland
{cormac.hampson, owen.conlan}@cs.tcd.ie

***Abstract*—Almost all information domains have witnessed an exponential increase in the amount of structured data available. However, there is still a lack of support for ordinary users to create complex queries spanning multiple information sources. Until this occurs the real benefits of having such a proliferation of metadata will not be realized by the general public. This paper describes SARA (Semantic Attribute Reconciliation Architecture), which is a framework that helps users leverage expert knowledge to discover relevant information, and to draw correlations across separate information sources. These sources can be in various data formats, and are accessed by users in a consolidated fashion. Users are supported in their information exploration with the knowledge of experts, which they can further tailor to better suit their needs. SARA offers tools and support for domain experts with no computing experience to encode their expertise, thus opening up SARA's use to almost any domain where rich metadata is available. This paper discusses the SARA framework in detail, as well as describing the applications to which it has been successfully applied in a number of different domains.**

Keywords-Semantic Attributes; Domain Expert; Personalization; Complex Queries; Information Exploration

I. INTRODUCTION

Structured and semi-structured data has become increasingly prevalent and more freely accessible in recent years. With the advent of the Linking Open Data community project [1] and the proliferation of web services, this trend is set to continue. Even within single enterprises it is not unusual to have many separate databases storing similar information in different formats and schemas. This situation has necessitated better techniques to allow users to correlate information, and elicit useful knowledge from separate data sources. However, while one may intuitively expect the additional structure in the data to have been exploited to provide sophisticated query capabilities, this has largely not proved to be the case [2]. Applications such as Freebase [3] do provide access to underlying data stores via query languages, however this is suitable primarily for application developers with a knowledge of the language rather than regular users wishing to ask very specific questions through a usable human interface.

As the prevalence of and dependence on digital data escalates, and the increasing use of such repositories by casual users, the need for user-friendly systems capable of supporting meaningful exploration will be greatly increased. SARA (Semantic Attribute Reconciliation Architecture) [4] is a framework that addresses this area by supporting ordinary users to make complex queries over heterogeneous sources from a domain. This paper describes the design of SARA and its component parts. Section 2 discusses some work related to this research and section 3 details the underlying design of the system. Section 4 describes the applications of SARA in various domains and section 5 summarizes the paper

II. RELATED WORK

The main aim of this research is to support user exploration of information stored over separate data sources from a domain. This section briefly discusses some pertinent research related to this field with more related work discussed here [4]. In terms of user exploration, faceted search [5] is one approach that has been successfully implemented within domain specific search. However, while well suited for bespoke applications working over individual data sources from a domain, there are many challenges to be overcome before faceted search is successfully used over multiple heterogeneous sources [6]. Similar to faceted search, expert systems are not suitable for general Web searches, though they can help users find reliable information in a narrow area such as medicine or accounting [7]. Hence, a generic platform for non-technical experts to encode their experience and knowledge of a domain would be very useful. Such a platform would allow end users to benefit by leveraging this expertise while exploring their domain of interest.

Because many interesting correlations can only be drawn by referencing multiple data sources (even within a single domain), the benefits offered by structured and semi-structured data will not be fully realized until this problem is overcome. This has seen a development in *pay as you go* integration inspired by dataspaces, because traditional methods of data integration such as data warehousing and mediated schemas are not adequate for web scale heterogeneity [8, 9]. With the rapid increase of Web Services and Linked Data [1] there has been an increasing interest in lightweight integration of data sources that can rapidly link data sources together. In terms of Web Services this has led to the proliferation of mashups and aggregated feeds [10, 11]. However, of much more potential is the use of dereferenceable URIs in Linked Data which pushes integration down to the instance level. This type of data integration can be more agile and potentially much more powerful than traditional methods.

III. Design

SARA is a framework that supports ordinary users (using an application connecting to its API) leverage domain expertise in order to query different information sources in a consolidated fashion. This section discusses how SARA integrates different sources, details the semantic attributes which are central to its operation, and outlines its mechanism for consolidating data from multiple sources.

A. *Integrating Sources*

Central components of SARA are its domain superclasses that must be chosen for each domain installation. These can be seen as any key entities from a domain that a user would typically like to get information about. For instance in the music domain you could select superclasses such as *Artist, Song, Album, Venue* etc. and in the Astronomy domain you could choose *Planet, Star, Astronaut, Satellite* etc. There is no limit to the number of superclasses you select, and there is no need to define any relationships or properties for them which can be an arduous task when creating domain ontologys. Furthermore, new superclasses can be added to SARA at any time so you are not limited to your initial selection.

Any source that is to be accessed by a SARA installation must register with SARA's source registry. The source registry is simply an XML file that contains details on the source's name, location (database address, SPARQL endpoint etc.), and any relevant metadata from its schema. The type of details needed by the registry varies depending on the kind of source that is being added, and new sources can be added seamlessly at any time. Sources that can be accessed directly via a query language such as SPARQL, XQuery or SQL (whether hosted locally or remotely) do not require any more manual effort to integrate with SARA. On the other hand web services do require a wrapper to act as an intermediary, however they can be reused in other SARA installations that want to connect to that source. Any metadata from an information source that has been added to the source registry can then be used to create semantic attributes which are a key mechanism within SARA and are explained in the next section.

Each source connected to SARA will contain many instances of the domain superclasses, and in their schema they will have metadata relating to them. Any metadata from the schema that is of interest (e.g. *song_duration* and *year_released* in a music database source) gets associated with one or more superclasses from the domain (e.g. *song_duration* with *Song* and *year_released* with *Album, Song* and *Artist*) in the source registry. This allows different data sources with different schemas to co-exist in SARA without having to go through the time consuming and problematic process of being homogenized to a canonical model, as was the case with previous versions of SARA [5].

B. *Semantic attributes*

Semantic attributes are discrete encodings of domain expertise that can be combined together and tailored to support user exploration of an information domain. They often act as abstractions and simplifications from the raw data, which are intended to make it more accessible for the ordinary, non-expert user. For instance in the music domain, semantic attributes can encompass characteristics such as *recentness*, *popularity, duration* and *similarity*. A semantic attribute may contain just a single metadata element or else combine a number into a single semantic attribute, e.g. combing the elements *bit rate, sample rate* and *file type* into a single semantic attribute *audio file quality*.

Semantic attributes can also be classified into one of three types - Expert, Template and Hybrid:

- An Expert semantic attribute is a concept that cannot be tailored by the end user. They can vary from the quite objective e.g. the semantic attribute *"Number 1 US Singles"* would only return US singles that reached number one in the US charts, to more subjective notions such as *popularity* that the expert prefers not to be amended in any way.
- In contrast a Template semantic attribute is a concept that is only personalizable and contains no expert defaults e.g. the semantic attribute *Contains Chemical Element* allows the user to search for substances that contain specific elements such as *Hydrogen* and *Oxygen*.
- A Hybrid semantic attribute contains expert default rules as well as values that can be personalized by the end user. Thus a sound engineer may use the expert defaults of *a high quality audio file* which insists on files containing uncompressed raw audio of 48,000Hz or higher. However, in the context of home listening the same semantic attribute could be tailored to include any compressed files above a minimum value bit rate.

All semantic attributes can also be sub-categorized into a number of separate ranges or parameters e.g. the semantic attribute *Price* could be divided into {*Expensive - Average – Cheap*}, and *Weight* into {*Under Weight - Normal Weight - Over Weight – Obese*}. This categorization allows non-experts to access information without detailed knowledge of the domain.

Each semantic attribute is an atomic unit of metadata and rules that can be joined together with standard logical operators by the user to form more complex queries tailored specifically to their needs. These are called semantic attribute queries, with the personalizable element being vital in enabling users to specify, if they wish to, what their interpretation is of *a high quality audio file* or *a popular song* etc. Each semantic attribute can also include default values defined by the domain expert that allow informed queries to be run quickly without personalization.

Because semantic attributes are hand-crafted by experts and not just automatically extracted from datasets it is important to allow them to be created by non-technical experts in minutes. Hence an authoring tool with a wizard GUI has been developed directly to support this. This means that SARA gets the benefit of accurate human-created semantic attributes without the cost of a lot of manual effort. Each semantic attribute created in the authoring tool is encoded in an XML model which is then imported into SARA. Another key design feature of semantic attributes is that they are completely agnostic to what technology the expert defined rules are encoded in, with the current implementation supporting SPARQL, XQuery and web service API calls.

As SARA allows new semantic attributes to be added seamlessly at any time by different users, it facilitates teams of experts in performing collaborative work, or to have different types of expertise exposed across the same domain. This diversity of expert perspectives encoded as semantic attributes empowers end users to pick and choose the semantic attributes that are best suited to their needs. In some ways this is analogous to choosing a specific critic for guidance in a domain you have interest in, but are not an expert in. Thus their subjective critiques on movies, sport, politics, finance etc. can be appropriated by the end-user to help their exploration, but more importantly they can be tailored to greater match individual preferences.

C. *Reconciliation of Results from Different Sources*

An end user connects to SARA via a client application that uses the SARA API. In this application the user may be presented with the available semantic attributes and choose the ones they are interested in by joining them together (and personalizing if necessary) with operators into a complex query e.g. *Return all Artists from my iTunes collection that have Concerts Scheduled in the USA despite their most recent top 10 Album in the USA being more than ten years ago.* This is a combination of three semantic attributes (All artists in my iTunes, All artists with concerts scheduled in the USA and All artists with top 10 albums in USA before the year 2000) each from a different source.

When SARA receives this query it breaks it down into its individual semantic attributes and sends a query to each respective source. Each source then sends back an individual result set containing ids of instances of the relevant superclass. If the data sources all use globally unique identifiers such as dereferenceable URIs (which are becoming much more apparent with the advent of the linked open data initiative), then it is these ids that are sent back. If this is not the case with one or more of the sources, then SARA keeps its own index for each instance of superclass that it receives, in order to reconcile the same instance over multiple sources. However, many domains already have schemes for unique identifiers and with the increasing need for dereferenceable URIs in the web of data, this will become less of an issue in future. Finally, using set operators, the result ids from each individual semantic attribute set get consolidated together into a final result set which gets sent to the client application.

IV. Applications of SARA

To date SARA has been applied to the music, film, photograph and academic publications domains. This section will focus on the most recent installation of SARA in the music domain, but will also highlight the others.

A. *Music Domain Case Study*

A recent installation of SARA involved five different music data sources and supported users to create complex queries reconciling information from these separate sources. The sources were a mixture of local and remote sources in three formats: An iTunes library with over 30,000 songs stored in XML in an eXist database; The entire US Singles charts from 1950-2008 stored as XML in an eXist database; The freebase.com music SPARQL endpoint [1] ; The MySpace.com SPARQL endpoint [2] ; The Last.fm web services[3].

Fig. 1 shows the implementation of SARA at design time. In step 1 you can see the five sources are registered to the source registry which in turn gets visualized in the Semantic Attribute Authoring Tool. The superclasses chosen for the domain were *Artist*, *Song*, *Album* and *Country*. In step 2 the domain expert uses the authoring tool to create semantic attributes which are stored in SARA to be used at runtime by client applications.

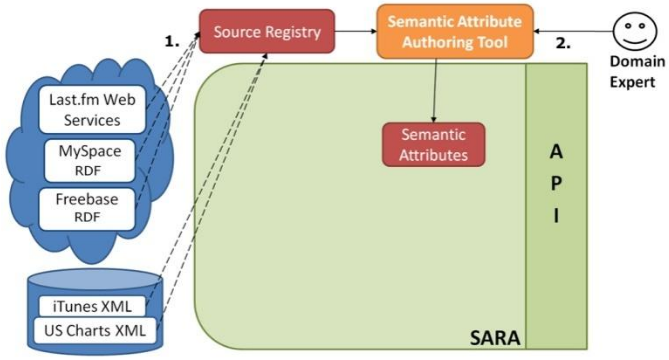

Figure 1. Design time implementation of SARA in the music domain

Fig. 2 shows the music domain implementation of SARA as it looks at runtime. Step 1 is where a user forms a query out of semantic attributes using a client application, which then gets sent to SARA via its API. For instance, a query combining three separate semantic attributes such as *Return all Artists from my iTunes collection that have Concerts Scheduled in the USA despite their most recent top 10 Album in the USA being more than ten years ago* could be sent. In step 2 SARA decomposes the query into its constituent semantic attributes and sends each constituent query to the relevant sources. In step 3 the results for each query are sent back as a set of unique identifiers to the reconciler. Because not all sources had globally unique identifiers a local index was used within SARA. In step 4 the individual sets were

[1] http://lod.openlinksw.com/sparql

[2]http://virtuoso.dbtune.org/sparql

[3] http://www.last.fm/api

reconciled into a final result set which gets sent back to the client application as XML in step 5. This XML is then rendered by the application and visualized to the end user in textual or graphical format.

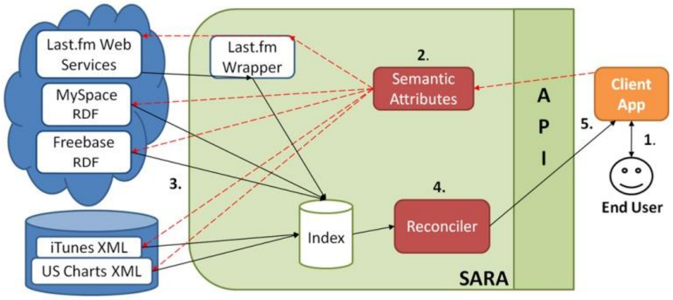

Figure 2. Runtime implementation of SARA in the music domain

B. *Other Implementations of SARA*

A number of other domains have been implemented in SARA along with accompanying client applications. X2Photo is one such application that helped users browse large image repositories with reference to the aesthetics of the photographs as well as their content. It used semantic attributes based on a photograph’s technical metadata (hue, saturation, lightness values etc.) as well as metadata and tags from Flickr. The expert rules, based on color psychology, were encoded using SARA’s Authoring Tool and allowed end users to leverage this knowledge while exploring the photographs for relevant images. The application overall received positive feedback when evaluated. Users found the more natural expressions in the system (semantic attributes) gave them much more freedom when searching for relevant images than when limited to traditional keyword searching, where they were relying on their search terms to match a relevant photo’s tags. Thus injecting expert knowledge, into a system only supporting tag-based search, allowed users to more freely express the aesthetic of the photograph they were looking for and not only the content of the image. This was largely facilitated through the use of SARA.

Two other domains that were implemented with SARA were the academic publications domain and the film domain. In the academic publications domain, a highly visual client application was developed that used semantic attributes and the SARA infrastructure to support a more explorative approach to finding relevant publications. The semantic attributes created provided a number of search axes for users to search under which were complimented by the novel visualization interface developed for the client application. Likewise, the SARA installation based in the movie domain also had an accompanying client application that was highly visual. Semantic Attributes relating to different ways of measuring the popularity of films (IMDB ratings, film profit, film grossing, film awards) were created from different sources and used to help implicitly model user preferences and to provide recommendations to similar films. In this application the semantic attributes were not displayed to users in anyway but rather used in the background as similarity metrics for the films. The user experience with the application was positive overall, and it showed how SARA can be used to help subtle domain exploration that doesn’t require the user to explicitly create complex queries.

V. Summary

This paper has described how SARA has a variety of stakeholders (the information engineer, the application developer, the domain expert and the end-user) and supports each in a different way. From the information engineer’s perspective they now have an easy way of relating a set of sources that non-technical domain experts can use as a basis to encode their expertise. Application developers can then build a variety of client tools that exploit SARA’s API to give a consolidated view over a domain. From the end users perspective, they are now abstracted away from the underlying complexity and raw metadata of the sources, and can leverage domain expertise to support complex, personalized, and semantically meaningful queries across separate data sources. This paper has highlighted four different domains that have worked successfully with SARA and future releases of SARA will further demonstrate its applicability to bespoke data source integration, as well as to ordinary users wishing to explore more easily information stored on Web 2.0 sites and on the Semantic Web.

VI. Acknowledgment

This research has been supported by The Irish Research Council for Science, Engineering and Technology: funded by the National Development Plan.

References

[1] “Linking Open Data, W3C SWEO Community Project”, [Online]. Available:http://esw.w3.org/SweoIG/TaskForces/CommunityProjects/LinkingOpenData [Accessed April 19th, 2010]

[2] C. Bizer, et al., "Linked data–the story so far," International Journal on Semantic Web and Information Systems, vol. 5, pp. 1-22, 2009.

[3] "Freebase: A social database about things you know and love",[Online]. Available: http://www.freebase.com [Accessed April 14th, 2010]

[4] C. Hampson and O. Conlan, "Supporting Personalized Information Exploration through Subjective Expert-created Semantic Attributes," 2009, Proc. IEEE International Conference on Semantic Compting pp. 384-389, 2009

[5] J. Polowinski, "Widgets for Faceted Browsing," Human Interface and the Management of Information. Desiging Information Environments. vol. 5617/2009, pp. 601-610, 2009

[6] J. Teevan, S.T. Dumais, and Z. Gutt, "Challenges for Supporting Faceted Search in Large, Heterogeneous Corpora like the Web," Proc. Human-Computer Interaction and Information Retrieval, pp. 6-8, 2008.

[7] S. J. Vaughan-Nichols, "Researchers MakeWeb Searches More Intelligent," Computer, vol. 39, pp. 16-18, 2006.

[8] J. Madhavan, et al., "Web-scale Data Integration : You can only afford to Pay As You Go," World Wide Web Internet And Web Information Systems, pp. 342-350, 2007.

[9] A. Halevy and J. Ordille, "Data Integration : The Teenage Years," Proc. 32nd International Conference on Very Large Data Bases pp. 9-16, 2006.

[10] J. C. Fagan, "Mashing up Multiple Web Feeds Using Yahoo! Pipes" Computers in Libraries, vol. 27, p. 8, 2007.

[11] “DERI Pipes: Open Source, Extendable, Embeddable Web Data Mashups”. [Online]. Available: http://pipes.deri.org/. [Accessed April 12th, 2010]