Browsing highly interconnected humanities databases through multi-result faceted browsers

Michele Pasin
Department of Digital Humanities
King's College
London, UK
michele.pasin@kcl.ac.uk

Faceted browsing is a recent paradigm in search interfaces that allows users with little familiarity of a subject domain to quickly explore the contents of databases or other structured data sources. The underlying principle of this approach can be traced back to the work of Indian librarian S.R. Ranganathan, who, in contrast with traditional top-down, taxonomical approaches to subject classification, in 1933 developed a method for organizing subjects in a bottom-up and non-hierarchical fashion. According to this model, a classification system can be created by combining together subject descriptors chosen from a number of non-exclusive and non-hierarchical *facets* – e.g., in the context of classifying books, these can be *genre*, *date* or *author*. This method supports the generation of a flexible system that better represents the multitude of perspectives we could use to represent knowledge (Broughton, 2004).

As a result of the adoption of these ideas in computer science, researchers have been creating search interfaces that allow the exploration of digital resources through the manipulation of filters describing important features of a subject domain [Broughton 2002]. A well-known pioneer in this area is Marti Hearst with her work on the Flamenco faceted browser [Hearst, 2002], followed by a number of similar approaches that, in general, aimed at creating more compelling and easy-to-use user interfaces [Hearst, 2008] [Capra, 2007] or at providing software packages that can work with different types of data sources; these may vary from manually editable JSON files [Huynh, 2007] to databases [Stuckenschmidt, 2004] and RDF triplestores

[Oren, 2006]. A number of projects have also proven the usefulness of this type of interfaces to the aim of facilitating the navigation of large repositories of humanities data, such as artwork images [Hildebrand, 2006], music resources [Bretherton, 2009] or multi-genre collections [McGann, 2007].

The success of faceted interfaces can be related to the fact that they implement a schema-less approach to classification, that is to say, they make available to the user a number of co-existing search dimensions that can be simultaneously used to browse and preview the contents of a digital repository. Many are the proven advantages of such an approach (Perugini, 2010): first, users are never asked to 'guess' the right search terms, as it happens in classic keyword search interfaces; second, inconclusive searches are prevented; third, non-experts can easily 'get a feeling' for the significance and meaning of the data available just by looking at the available facets, thus increasing their understanding of the domain. In conclusion, this type of interfaces simplify enormously the exploration of a digital repository, and, using the words of Nowivskie, they make it easy to "explore lateral relationships" to the point that they open "possibilities for algorithmic serendipity in research" (Nowivskie, 2007).

An important feature that most of the existing faceted browsers have in common is that the different facets available equally concur to the selection of a single result-type. For example, by manipulating variables such as the *color* and the *making* of cars, we can navigate a data-space of available *cars*; by choosing filters representing information about *genres*, *publishers* and *years of publication* we can easily narrow down a result list of *books*; or, in the context of a prosopographical database, by accumulating descriptors about people's *forenames*, *surnames*, or *gender* we would be able to refine our search for the *individuals* mentioned in the database.

In our work, we intended to push the boundaries of this approach by creating a faceted browsing engine that, given the same set of selectable facets, can be used to search for 'ontologically distant' entity types. For example, in the context of a prosopographical database, by manipulating the same group of filtering options, we aimed at letting users search not just for *people,* but also for *factoids* and *sources* (cf. Figure 1). In doing so, we assumed that rich and highly interconnected humanities databases call for more powerful search mechanisms; such mechanisms should be capable of revealing the intricacies of a subject domain to the

casual learner, and, at the same time, of providing a higher level of 'algorithmic serendipity' to the academic scholar. In other words, we aimed at making more visible the large number of search pathways a highly structured database can make possible - as opposed to hiding this complexity by providing a unique entry point to the wealth of data available. This means that, as shown in fig. 1, by using facets typical of the 'people' result-type (such as *surname* or *gender*) we would like to be able to search for 'sources' or 'factoids'. Or, by choosing facets typical of the 'sources' result-type (such as *document category* or *language*) we may want to filter results when searching for 'people'.

With this vision in mind we created DJFacet[1], a faceted browsing engine that lets users create powerful, multi-result search interfaces. DJFacet is written in Python and is based on Django, a popular web application framework[2] that facilitates the development of database-driven websites by providing functionalities that speed up the creation of repetitive tasks. In particular, one key component of this framework is the Object-Relational Mapper (ORM), that is, a set of functions that provide programmers with a level of abstraction between the database and the application language; as a result, it is possible to invoke complex database queries without having to write any SQL code. This makes the whole application easier to manage and more portable across different database engines.

By building on the functionalities of Django ORM, DJFacet provides a customizable and easy-to-use environment for creating database-driven faceted search applications. The underlying idea of DJFacet is that since a data-structure had already been designed and fine-tuned when the database was created, that same structure could be used to define the search dimensions of the faceted browser with little extra work required. An instance of DJFacet can run just by creating an initialization file in which we define which are the available facets and what 'behaviour' they have with respect to the database schema. The software then calculates automatically all the remaining query-paths needed to show the results in the various searches.

At the time of writing, we tested DJFacet's approach with two humanities databases.

The Paradox Of Medieval Scotland project[3] (POMS) investigates how a recognizably modern Scottish identity was formed during the period 1093-1286. Drawing on over 6000 contemporary charters, it provides

1 The software is open source and freely available online at the url http://code.google.com/p/djfacet/ .
2 http://www.djangoproject.com/
3 www.poms.ac.uk

biographical information about all known people in Scotland during that period. In this context we built a search interface that features 29 facets, organized into 5 groups (cf. figure 2). The result types are 3 (*people*, *factoids* and *sources*).

The Early Modern London Theatres[4] (EMLoT) project provides its users with a major encyclopedic resource on the early London stage, as well as a comprehensive historiographical survey of the field. EMLoT identifies, records and assesses transcriptions from primary-source materials relating to the early London stage, as found in secondary-source print and manuscript documents. In this case the faceted search interface contains 24 facets, organized into 5 groups. The result types are 7 (*transcription records*, *primary sources*, *secondary sources*, *events*, *people*, *troupes*, *venues*).

By using DJFacet it was possible to allow the formulation of queries that might not be immediately obvious to the user. For example, in POMS it became trivial to search for Charters mentioning transaction events in which 'beneficiaries' of name 'William' acquire something on the day of the 'Feast of St Patrick '. Also, the search interface provided users with more chances of coming across interesting connections in the available materials. This was made possible by the fact that the facets used in the search are ontologically distant from the respective result-types.

However, despite the fact that this approach proved to be, from the logical and computational point of view, completely feasible, it also opened up a number of research questions from the point of view of the *meaning* of these multifaceted searches across different results types. In other words, we realized that often the accumulation of filters ontologically distant from each other could be hardly translated by the end user into real-world questions; analogously, the opposite may happen, in so far as simple type of searches may be impeded by the highly structured architecture of a faceted browser.

In order to provide some answers to this issue and lay the ground for a more scientific discussion of the problem we are currently carrying out a user evaluation study with humanities scholars. The purpose of the experiment is to discover the degree to which humanities scholars can make sense of the search mechanisms provided by our faceted browser, and, indirectly, of the complex data structures often necessary for

4 www.emlot.kcl.ac.uk

representing humanities subjects. We will report on these findings at the conference, together with a deeper analysis of the implications of using multi-result faceted browsers in the context of complex humanities datasets.

References

Bretherton et al. *Integrating musicology's heterogeneous data sources for better exploration.* 10th International Society for Music Information Retrieval Conference (2009).

Broughton. *Essential classification*. Facet Publishing, London (2004).

Broughton. *Faceted classification as a basis for knowledge organization in a digital environment: the bliss bibliographic classification as a model for vocabulary management and the creation of multidimensional knowledge structures*. The New Review of Hypermedia and Multimedia (2002) vol. 7 (1) pp. 67-102.

Capra et al. *Effects of structure and interaction style on distinct search tasks*. Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries (2007) pp. 442-451.

Hearst. *UIs for Faceted Navigation: Recent Advances and Remaining Open Problems* . HCIR08 Second Workshop on Human-Computer Interaction and Information Retrieval (2008).

Hearst et al. *Finding the flow in web site search*. Communications of the ACM, Special Issue: The consumer side of search (2002) vol. 45 (9).

Hildebrand et al. */facet: A Browser for Heterogeneous Semantic Web Repositories*. ISWC'06: Proceedings of the International Semantic Web Conference (2006), pp. 272-285.

Huynh et al. *Exhibit: lightweight structured data publishing*. WWW '07: Proceedings of the international conference on World Wide Web (2007), pp. 737-746.

Oren et al. *Extending Faceted Navigation for RDF Data*. ISWC'06: Proceedings of the International Semantic Web Conference (2006), pp. 559-572.

McGann and Nowviskie. *NINES: a federated model for integrating digital scholarhip*. White paper (2005).

Nowviskie. *COLLEX: semantic collections & exhibits for the remixable web*. White paper (2007).

Perugini. *Supporting multiple paths to objects in information hierarchies: Faceted classification, faceted search, and symbolic links*. Information Processing & Management (2010) vol. 46 (1) pp. 22-43.

Stuckenschmidt et al. *A topic-based browser for large online resources*. Engineering Knowledge in the Age of the SemanticWeb, LNCS (2004) vol. 433-448 pp. 433-448.

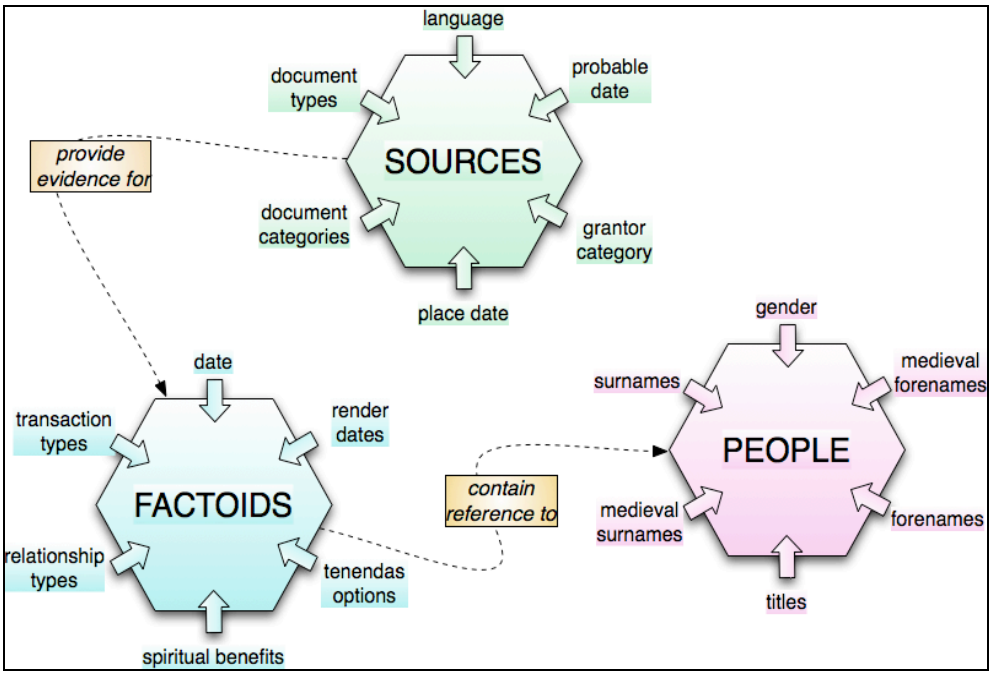

Fig. 1. The facets allowing 'entry' to a prosopographical database.

Fig. 2. Screenshot of the faceted browser for the POMS database