Using Named Entities as a basis to share associative trails between Semantic Desktops[1]

Pat Croke, Ann Johnston, Kim Tighe

Hewlett-Packard Galway Ltd, Ballybrit Business Park, Galway, Ireland
{Pat.Croke, Ann.Johnston, Kim.Tighe}@hp.com

Abstract. In this paper we illustrate the use of Named Entities as a basis to share associative trails [1] between Semantic Desktops using a Firefox [6] extension called Trailblazer. Trailblazer can automatically detect the context of a Web page. It identifies the most relevant category for the Web page based on its context along with relevant links to additional information and notes. Named Entities belonging to that category are highlighted. Trailblazer facilitates the building and augmentation of categories with links from Named Entities to Web pages, documents, mail messages, personal notes, etc. It enables the import/export of categories and their related associative trails between users. It facilitates the distribution of discovered information through the sharing of associative trails in order to support community based knowledge. This allows faster initiation of new members to a group and information noted by one member as important can be highlighted to the rest of the group.

1 Introduction

Vannevar Bush [1] describes how the human mind operates by association. With one item in its grasp it quickly moves to the next item suggested by the association of thoughts stored as a web of trails in its brain cells. He further describes how some of these trails fade over time and are forgotten. His solution is a desktop memory extender called "Memex", which contains all of the documents and books a person had come across in their lifetime along with the various associative trails they had made through this collection. He also envisioned these trails being exchanged between people as a way of sharing knowledge and identified a new profession called trail blazers who would create associative trails to guide others through their knowledge.

In this paper we are proposing that Named Entities are a basis for storing and sharing associative trails. Named Entities as defined by [2] are unique identifiers of items such as people's names, locations, organizations. To these we would also add terminology, because when humans are reading documents they recall knowledge linked to Named Entities and terminology. Links already associated with a Named Entity can be readily associated with a Named Entity occurring in the same or a different document. Humans can distinguish between two different entities with the same name, based on the context in which it appears in a document they are reading.

We describe the Trailblazer Web browser extension which allows a user to define a Named Entity in a document within a specific context and to associate links and notes with it. It uses the Cosine Similarity Measure [7] between a document and multiple contexts to determine the appropriate context to be used for Named Entity disambiguation. Users are able to create a new category and train Trailblazer to identify pages belonging to that context by pointing it at typical documents. Trailblazer is able to import and export associative trails in Resource Description Framework (RDF) [8] format, which makes them portable amongst groups.

In this paper we first describe some scenarios to illustrate how Trailblazer can be used to link associative trails to Named Entities and then describe it's architecture. We conclude with related work and a discussion of future directions.

2 Scenario

We describe Trailblazer in terms of how a user, Ted, might experience it for the task of collecting information on horse racing. Then we extend the scenario further to include how he shares his collected information with other enthusiasts. He also uses Trailblazer to create a separate category of trails about wine to share with his friends.

Ted enjoys horse racing and to further his interest joins a horse racing club. The club consists of a group of knowledgeable racing enthusiasts who share their knowledge between themselves. They maintain this knowledge using a Firefox extension called Trailblazer which allows them to identify Horses, Trainers, Jockeys, Racecourses, etc. as Named Entities. They add associative trails to these Named Entities in the form of links to relevant Web pages and notes that they write. When Ted joins the club he is given Trailblazer and access to the club's associative trails. The Firefox browser has been altered to include the Trailblazer panel at the bottom of the screen as seen in **Figure 1**. He follows the club's instructions to import their horse racing trail and visits his favorite online racing news Web site, 'Irish Racing'. He can see from the Trailblazer panel that the page has been categorized as 'Horse Racing' as shown in **Figure 1**.

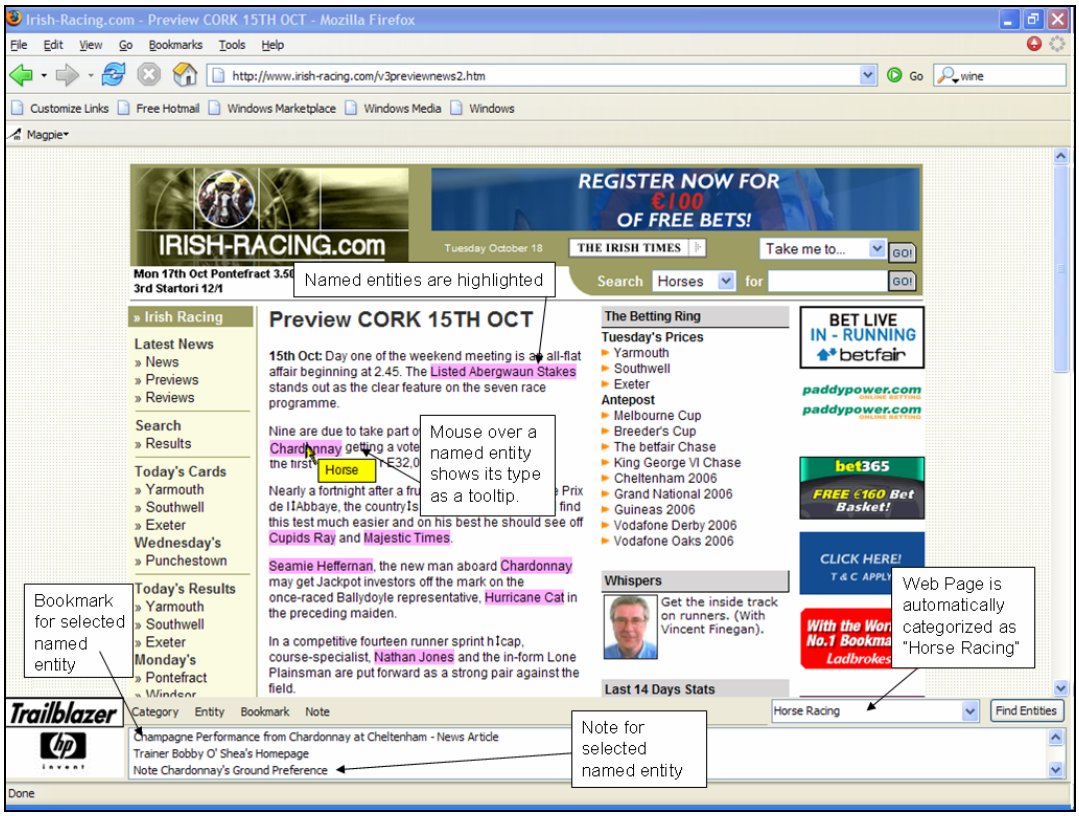

Fig. 1. Irish Racing Web site annotated and showing additional bookmarks and notes

Ted presses the 'Find Entities' button and Trailblazer highlights Named Entities of interest, for example Horses, Jockeys, Trainers, Owners, Racecourses, etc. When he does a mouse-over on a Named Entity, a tool tip appears showing its type. The mouse-over of the Named Entity 'Chardonnay' shows that it is of type 'Horse'. When he clicks on the Named Entity, bookmarks and notes associated with it are shown in the Trailblazer panel.

He follows the different links to Web pages, documents, personal notes, etc. associated with the highlighted Named Entities. Ted looks at the additional information on some of the horses he is interested in for the big race on Thursday. **Figure 2** shows an example of the extra information provided for the Named Entity, 'Chardonnay'.

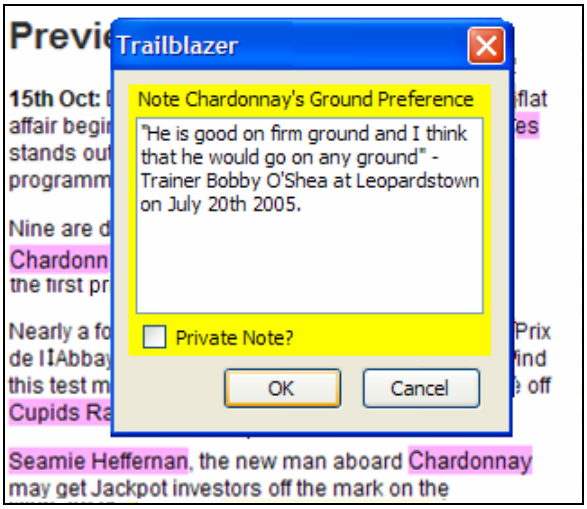

Fig. 2. Note attached to the Named Entity 'Chardonnay'

As he is studying the information Ted recalls how he overheard a Trainer at the Cork Racecourse talking about 'Chardonnay' and how the horse is recovering from a cough and will not be back to full fitness for the race on Thursday. He adds this information as a note to the Named Entity 'Chardonnay' as shown in **Figure 3**.

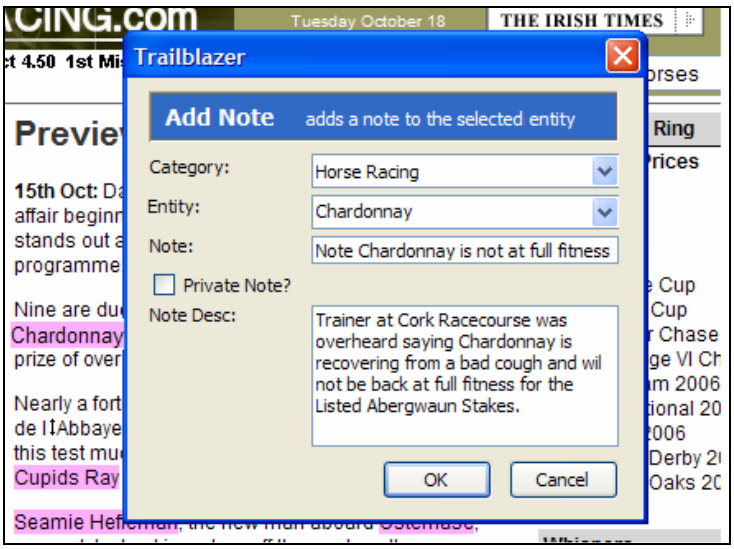

Fig. 3. Adding a note to the Named Entity 'Chardonnay'

Trailblazer enabled Ted to import the horse racing associative trail. It also allows him to export a copy of the horse racing associative trail with his added information back to the club's Web site. Ted's information on Chardonnay's cough is now available for sharing with the other club members. Other members of the club can now import this horse racing trail from the shared account and merge the additional information gathered by Ted with their horse racing trail. The local racing enthusiasts at the club can now make an informed decision with regards to the horse Chardonnay's chances of winning his race.

Ted is also a keen wine connoisseur and has many friends who share this hobby. Having seen how powerful Trailblazer is for sharing horse racing information he decides to use Trailblazer to create associative trails for wine. He creates a wine category and goes to a Web page about wine to show Trailblazer the words and phrases which typically occur on a page about wine. This is used to train Trailblazer as shown in **Figure 4**.

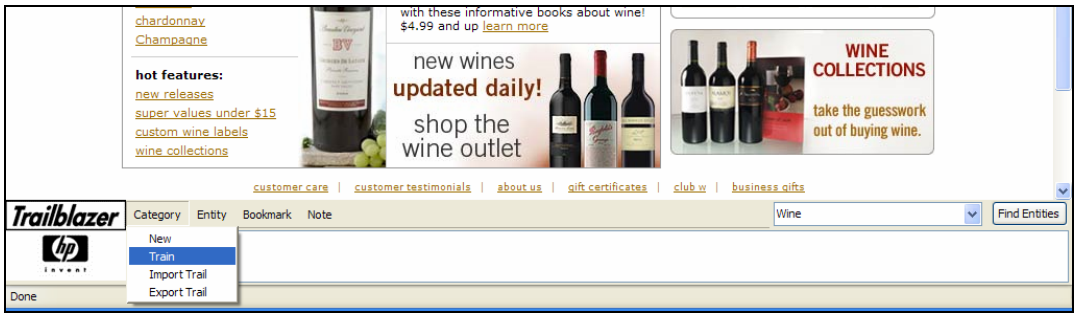

Fig. 4. Training Trailblazer about wine

Ted continues to train Trailblazer about wine using a number of typical wine pages. He then adds a number of Named Entity types to his wine category, for example 'Grape Type' and 'Region'. When he selects a Named Entity on a wine page, for example 'Chardonnay', he adds it to the category as a 'Grape Type'. He continues to add a number of Named Entities with bookmarks and notes from his own knowledge of wine as shown in **Figure 5**.

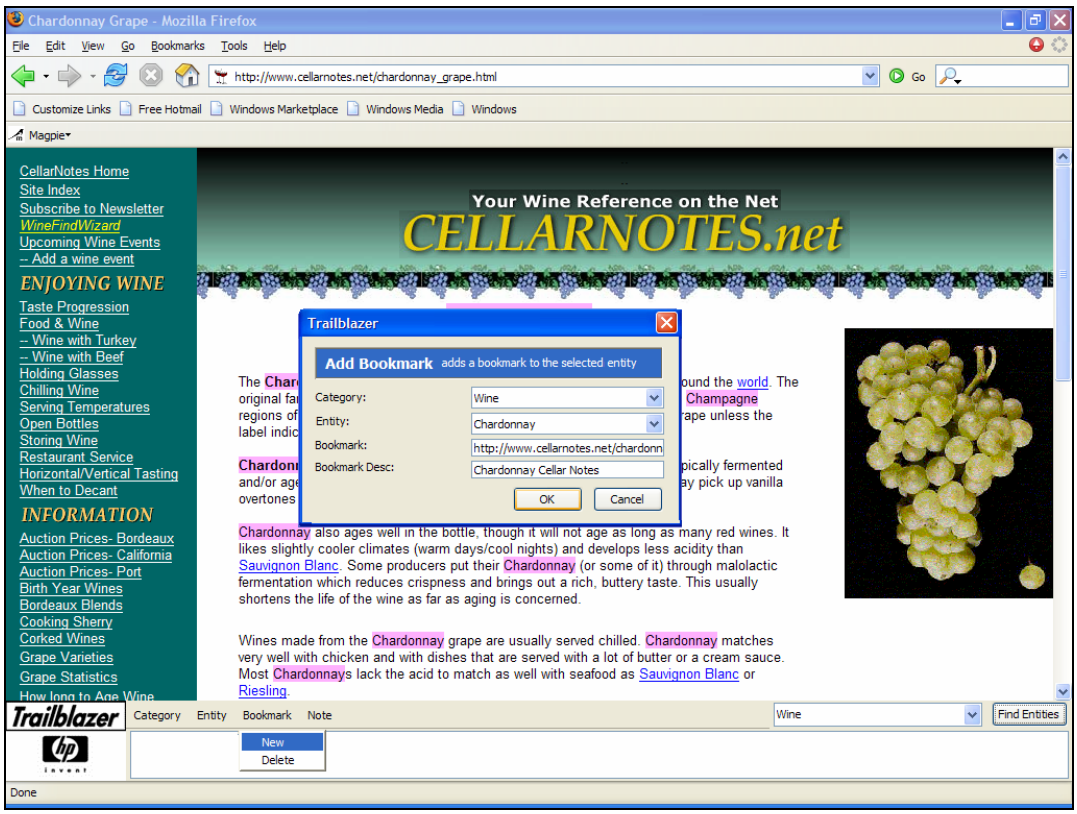

Fig. 5. Adding a New Bookmark to the 'Chardonnay' Named Entity

Using his new category, when Ted opens a Web page about wine and presses the 'Find Entities' button any of the Named Entities that are on the page will get highlighted. When he passes a mouse over the entity, a tool tip informs him of its type. When he clicks on an entity for example 'Chardonnay', the associated bookmarks and notes are displayed as shown in **Figure 6**.

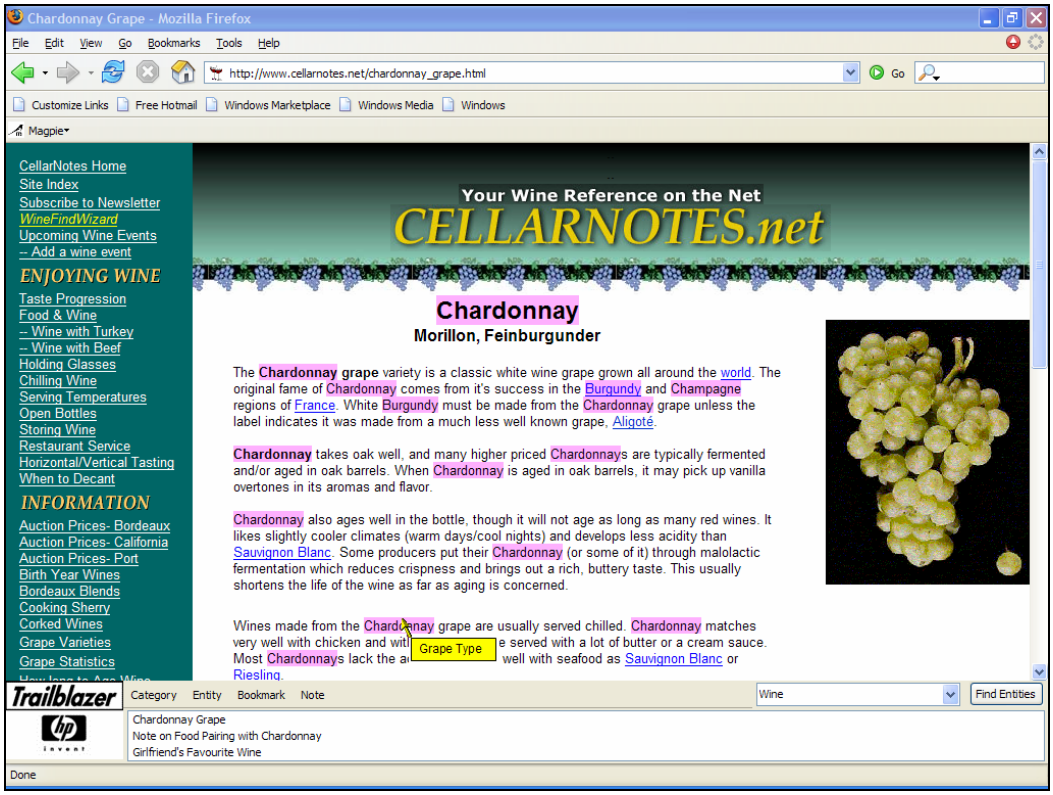

Fig. 6. New Wine Category

Ted links the selected Named Entities with bookmarks and notes information. He adds his first note on the best foods to pair with Chardonnay. His second note is a reminder that his girlfriend's favorite wine is Chablis which is made from Chardonnay grapes. However, Ted does not want to share this information with anyone so he marks it as a private note to himself and therefore only he can see the note about his girlfriend as shown in **Figure 7**.

His selected Named Entities will form the basis for his wine trail. He will then browse several other wine Web sites from which he will follow the same steps to collect more information about wine. When trained Trailblazer will then be able to automatically detect if a Web page applies to wine and will load that domain specific category with its additional information.

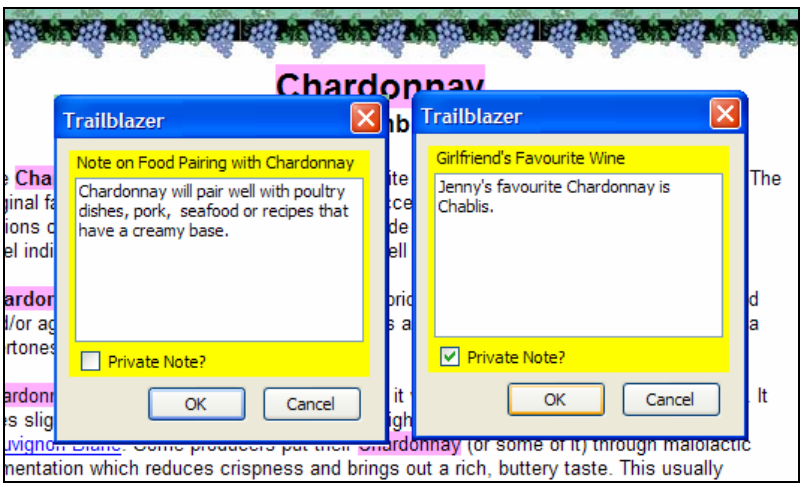

Fig. 7. Public/Private Notes

Although Ted has the Named Entity 'Chardonnay' as an instance of 'Horse' in his 'Horse Racing' category, he can also add 'Chardonnay' to his 'Wine' category as an instance of 'Grape Type'. Trailblazer differentiates between the two as shown in **Figures 8** and **9**.

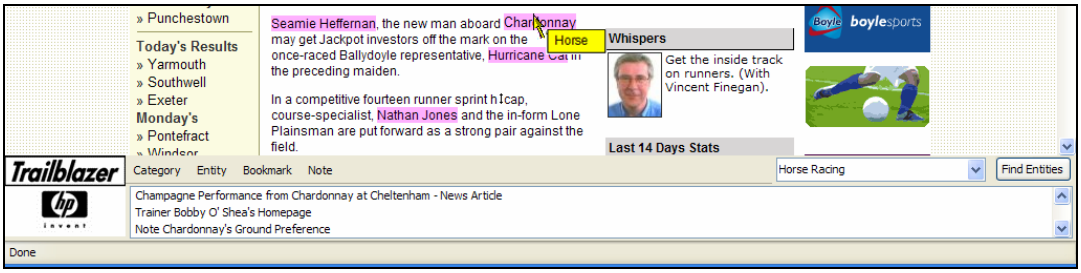

Fig. 8. 'Chardonnay' Named Entity as an instance of 'Horse'

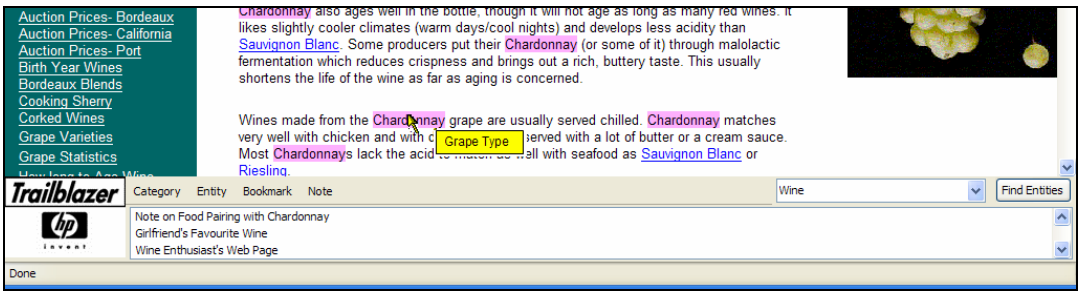

Fig. 9. 'Chardonnay' Named Entity as an instance of 'Grape Type'

Trailblazer categorizes the context of the page in **Figure 8** as a horse racing Web page so the Named Entity 'Chardonnay' is recognized as a horse. In **Figure 9,** the Web page has a context of wine so using the context of this page 'Chardonnay' is a type of grape. Ted can now share his associative trails with his wine connoisseur friends and together they can build and enhance their knowledge of wine.

3 Architecture

Trailblazer is an extension for the Mozilla Firefox [9] Web browser. Extensions are applications that can be downloaded and installed into a Firefox browser to add new functionality [6]. They can add anything from a toolbar button to a completely new feature. Extensions allow Firefox to be customized to fit the personal needs of each user if they require additional features.

Interfaces for Mozilla Firefox extensions are built using the XML User-interface Language (XUL) [10] and Cascading Style Sheets (CSS). Extensions are coded in JavaScript [12] and can link to native components using the cross-platform Component Object Model (XPCOM) [11]. JavaScript links the XUL and XPCOM components together. Firefox uses RDF as its data storage format.

JavaScript is an open, platform-independent, event-driven, interpreted programming language. RDF is an open, general-purpose data format for representing information in the Web. It is a common framework therefore application designers can leverage the availability of common RDF parsers and processing tools. This ability to exchange information between different applications means that the information may be made available to applications other than those for which it was originally created. XUL (pronounced zool, rhymes with cool) is an open, cross-platform language designed specifically for building portable user interfaces. A XUL overlay is a XUL file containing elements to be inserted into another XUL file. This insertion occurs when the other XUL file is rendered into an application interface. The browser window of Firefox is a XUL window. Extensions can be made to the browser window using an overlay file. XUL is used to overlay the browser to add the Trailblazer features. The Trailblazer overlay is rendered into the Firefox browser, generating the main Trailblazer panel as part of the browser window.

When Trailblazer is installed and the user launches Firefox, the standard browser window opens with the addition of a Trailblazer panel. In order to facilitate the addition of Named Entities and their bookmarks and notes, extra windows created using XUL dialog files are popped up by Trailblazer to request information from the user.

Firefox allows an installed extension to have additional privileges, such as reading local files and modifying user preferences. The Input/Output component of Trailblazer uses another Firefox extension called JSLib [13], which contains a library of JavaScript functions to access RDF files. This provides many types of error checking, as well as a friendly abstraction away from RDF/XML interfaces.

Page Recognition

The Trailblazer Vocabulary store contains the information to calculate the most likely match for the context of the current Web page. Each Category stored within Trailblazer contains the Named Entity types, Named Entities, and their bookmarks and notes related to that category.

When a user chooses the 'Train' option on the currently viewed page, the Trainer component extracts the text from the page to supplement the information held against the current category within the Vocabulary. Trailblazer uses a Vector Space Model (VSM) to model the collected training information.

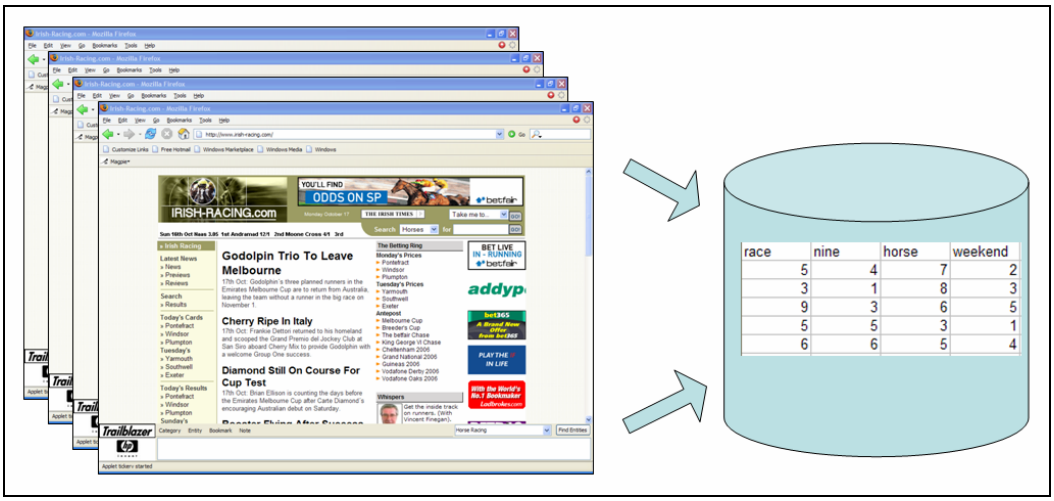

Fig. 10. Trailblazer Training Process

The VSM is an algebraic model used for information filtering and information retrieval [14]. It represents natural language documents in a formal manner by the use of vectors in a multi-dimensional space. Each dimension represents a term or concept found in the documents. The model creates a space in which documents are represented by vectors and allows decisions to be made about which documents are similar to each other. Spatial proximity is the physical distance between two items in space, for example the distance between two vectors or lines. Semantic proximity is the difference between two items in meaning, for example the context of the 'Irish Racing' Web page and the context of the 'Wine Connoisseur' Web page. The VSM uses spatial proximity to calculate semantic proximity. The Trainer uses the words and their frequency of occurrence from the current page to augment the existing training data.

When a page is loaded into the browser a comparison is made between the text on the page and the training data contained in the Trailblazer Vocabulary. A similarity measure is used to compare a vector representation of each category Trailblazer has had training on, and a vector representation of the currently viewed page. This calculates the most likely match, therefore deriving the context of the page in relation to the information obtained from training. Currently, Trailblazer is using the Cosine Similarity Measure [7] to compare the vectors of information. This is derived from the Law of Cosines, which is an extension of the Pythagorean Theorem.

The Cosine Similarity Measure is the cosine (in radians) of the angle between two vectors. For two vectors *a* and *b* the cosine similarity between *a* and *b* is given by:

$$\frac{a \cdot b}{|a| \times |b|}$$

i.e. the dot product of *a* and *b* divided by the modulus of *a* multiplied by the modulus of *b*.

Named Entity Recognition (NER)

When a user presses the 'Find Entities' button, the NER functionality processes the currently viewed Web page and highlights any discovered Named Entities from the previously derived category. NER [15] is a subtask of information extraction that seeks to locate and classify the elements in text into predefined categories such as the names of people, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. The Named Entities which have been stored for the category are grouped together on a word length basis. Each word from the Web page is processed in turn. A check is made to determine if the word could be part of a Named Entity. If it is, an attempt is made to match it by comparing it to the Named Entities beginning with the longest occurrences first. If it is possible for more than one match to be made, the longest match will always be used. For example take the two Named Entities 'Ascot' and 'Alisha's Ascot'. 'Ascot' is the name of a racetrack and 'Alisha's Ascot' is the name of a horse. When the text 'Ascot' is detected in a page, an attempt is always made to match the Named Entity 'Alisha's Ascot' first. 'Ascot' will only be annotated as a racetrack when the word 'Alisha's' does not occur before it. For efficiency the NER routine is a simple gazetteer-based process. The rare event of two different Named Entities having the same name and same type within the same context is managed by adding a comment to the Named Entity. The Trailblazer recognition process appears to be instant and does not delay the user's browsing experience. Heavier natural language processing to resolve this rare type of conflict would affect this performance. When a Named Entity is discovered, span tags are used to add a highlight to the html page. The page is reloaded with no change to its structure, only the highlight on the text is added. When the user clicks on a highlighted Named Entity, its associated bookmarks and notes are displayed in the Trailblazer panel.

Export Process

Information held in Trailblazer can be exported on a category by category basis. Export files are named using the category name and an extension of EXP. These are generated in an RDF format. Each export file stores the words and their frequency of occurrence in relation to the export category. It also stores Named Entity types and Named Entities along with their related bookmarks and notes. Private notes are not exported.

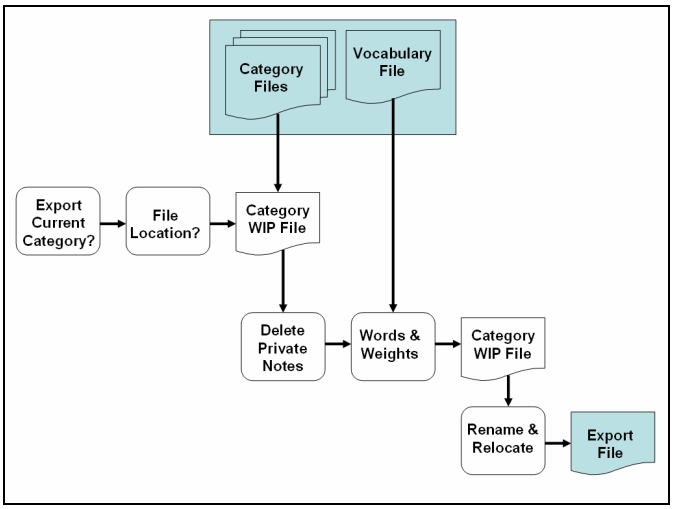

Fig. 11. Trailblazer Export Process

```
<?xml version='1.0'?>
<RDF:RDF xmlns:Trailblazer='http://hp.com/Trailblazer#'
         xmlns:NC='http://home.netscape.com/NC-rdf#'
         xmlns:RDF='http://www.w3.org/1999/02/22-rdf-syntax-ns#'>
  <RDF:Seq RDF:about='Trailblazer:Wine:InstTypes'>
    <RDF:li RDF:resource='Trailblazer:Wine:InstTypes:Region'/>
    <RDF:li RDF:resource='Trailblazer:Wine:InstTypes:Grape Type'/>
  </RDF:Seq>
<RDF:Description RDF:about='Trailblazer:Wine:ne:Sauvignon Blanc' InstType='Grape Type' />
  <RDF:Seq RDF:about='Trailblazer:Wine:ne:Sauvignon Blanc'>
    <RDF:li RDF:resource='Trailblazer:Wine:ne:Sauvignon Blanc:alias'/>
    <RDF:li RDF:resource='Trailblazer:Wine:ne:Sauvignon Blanc:bookmarks'/>
    <RDF:li RDF:resource='Trailblazer:Wine:ne:Sauvignon Blanc:notes'/>
  </RDF:Seq>
  <RDF:Seq RDF:about='Trailblazer:Wine:ne:Sauvignon Blanc:bookmarks'>
    <RDF:li RDF:resource='Trailblazer:Wine:ne:Sauvignon
Blanc:bookmarks:http://www.cellarnotes.net/sauvignon_blanc_grape.html'/>
  </RDF:Seq>
  <RDF:Description RDF:about='Trailblazer:Wine:ne:Sauvignon
Blanc:bookmarks:http://www.cellarnotes.net/sauvignon_blanc_grape.html'
                   Desc='Cellar Notes Sauvignon Blanc' />
  <RDF:Seq RDF:about='Trailblazer:Wine:ne:Sauvignon Blanc:alias'>
  </RDF:Seq>
  <RDF:Seq RDF:about='Trailblazer:Wine:ne:Sauvignon Blanc:notes'>
  </RDF:Seq>
  <RDF:Description RDF:about='Trailblazer:vocabulary:wine'
                   weight='6' />
  <RDF:Description RDF:about='Trailblazer:vocabulary:taste'
                   weight='1' />
  <RDF:Description RDF:about='Trailblazer:vocabulary:region'
                   weight='3' />
  <RDF:Description RDF:about='Trailblazer:vocabulary:grapes'
                   weight='1' />
  <RDF:Description RDF:about='Trailblazer:vocabulary:blanc'
                   weight='4' />
  <RDF:Description RDF:about='Trailblazer:vocabulary:white'
                   weight='1' />
</RDF:RDF>
```

Fig. 12. Sample Export file

Import Process

The import facility allows the user to import a file. Named Entities and their associated information are extracted from the file and appended into the users existing category file. Named Entity duplicates are not added, but their additional bookmarks and notes are merged. If a category file doesn't exist a new one is created. The words and their weights for that category are then extracted from the export file and integrated into the user's vocabulary file. This ensures that not only the annotation information is transferred between users, but also the training information to allow Trailblazer to detect the context of Web pages.

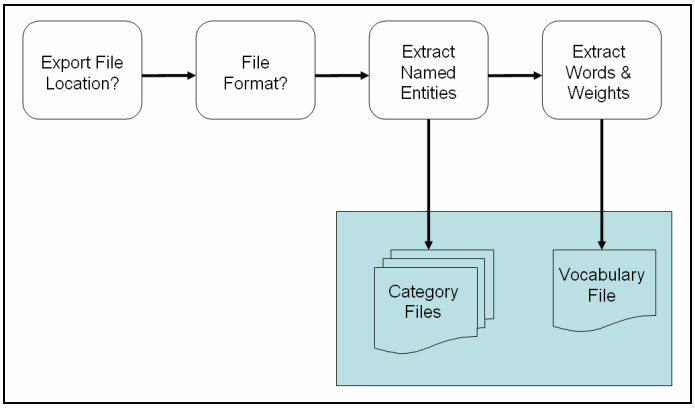

Fig. 13. Trailblazer Import Process

4 Related Work

Annotea [3] is a W3C LEAD (Live Early Adoption and Demonstration) project under Semantic Web Advanced Development (SWAD). It is currently supported by two Web browsers Amaya [4] and Annozilla [5] which, like Trailblazer, is a Firefox extension. Annotea enhances collaboration via shared metadata-based Web annotations, bookmarks [16], and their combinations. Annotations are comments, notes, explanations, or other types of external remarks that can be attached to a Web document or a selected part of the document without needing to touch the document. When the user gets the document he or she can also load the annotations attached to it from a selected annotation server or several servers and see what his peer group thinks [17]. Similarly shared bookmarks can be attached to Web documents, to help organize them under different topics, to easily find them later, to help find related material and to collaboratively filter bookmarked material. When a user accesses a Web page that has been referred to by shared bookmarks this is highlighted on the page. They are then able to see and access other pages bookmarked by other people as relevant to the topic. Trailblazer differs from Annotea in the respect that Trailblazer's bookmarks and notes are linked to a Named Entity, whereas Annotea's are linked to a document or part of a document using XPointers. This makes Trailblazer useful for pages that have not been viewed by the community before, but contain Named Entities that are known to that community. Both Trailblazer and Annotea allow annotations to be either shared or private. They also both use RDF for storing their annotations. Work [18] is in progress to extend the Annotea address scheme to beyond XPointers. This could allow Trailblazer in the future to use the Annotea schema.

Magpie [19] is a Web browser extension which uses NER based on a supplied ontology to add links to Named Entities on a Web page. Each ontology contains metadata associated with the Named Entities which, when selected by the user, are resolved by a Web server and an appropriate Web page is returned. Based on the user's interpretation of what the page is about, they choose an appropriate ontology to use to identify the Named Entities on that page and their associated links. Trailblazer differs from Magpie here in that it uses its cosine similarity logic to automatically identify the correct category to use for Named Entity identification. Magpie's links are supplied in the ontology and cannot be added to by the user. Additionally it does not have the facility to allow users to add annotations.

XP Smart Tags [20] allow Smart Tag action buttons to be attached to a document in Microsoft's Office XP. It uses a DLL (Dynamic Link Library) called a Recognizer to identify items of interest on a page and add Smart Tag action buttons. Out of the box it can recognize people's names and add a Smart Tag to allow the user to send them a mail or set up a meeting for example. Developers can create custom recognizer and action DLL's. This functionality would enable the building of read-only Trailblazer-like capability based on ontologies across the whole of the Office XP desktop environment.

5 Conclusions and Future Work

In this paper we have described the Firefox extension Trailblazer and how it can be used in a collaborative environment. We have shown how Named Entities can be used as a basis for sharing associative trails between Semantic Desktops. Future work will extend these capabilities across the whole desktop infrastructure. Leveraging the XUL overlay technology we will apply the Trailblazer extension to Thunderbird [21], which is the Mozilla mail client. This will allow the making and sharing of associative trails across Web pages, RSS [22] and Mail. By building the capability in behind Smart Tags we should get the same coverage across Windows XP. The Annotea schema appears to have good potential to be extended to meet Trailblazer's requirements. We also intend to build Annotea XPointer based bookmark and annotation capability into Trailblazer as this will allow the support of document specific associative trails. An additional area of exploration is the potential for using the Annotea server for sharing Trailblazer Named Entity based associative trails. Security needs to be addressed to ensure that links only come from trusted sources and that links are not to dangerous or undesirable Web resources.

References

[1] Vannevar Bush, *As we may think*, The Atlantic Monthly 1946.
[2] N. Chinchor, editor (1997). MUC-7 Named Entity Task Definition, Version 3.5. September 17th DARPA.
[3] Annotea http://www.annotea.org/
[4] Amaya http://www.w3.org/Amaya/
[5] Annozilla http://annozilla.mozdev.org/
[6] Firefox Extensions https://addons.mozilla.org/extensions/?application=Firefox
[7] C. J. van Rijsbergen. *Information Retrieval*, 1979.
[8] D. Brickley, R. V. Guha. 2004. *RDF Vocabulary Description Language 1.0: RDF Schema*. W3C Recommendation 10 February 2004.
[9] Firefox http://www.mozilla.org/products/Firefox/
[10] Vaughn Bullard, Kevin T. Smith, Michael C. Daconta. *Essential XUL Programming*. Wiley, 2001.
[11] XPCOM and XPConnect http://www.xulplanet.com/tutorials/xultu/xpcom.html
[12] David Flanagan. *JavaScript - The Definitive Guide.* O'Reilly, 2002.
[13] JSLib http://jslib.mozdev.org
[14] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.
[15] http://en.wikipedia.org/wiki/Named_Entity_Recognition
[16] M. Koivunen, R. Swick and E. Prud'Hommeaux. *Annotea Shared Bookmarks*. In Proc. Of KCAP 2003.
[17] J. Kahan, M. Koivunen, E. Prud'Hommeaux, and R. Swick. *Annotea: An Open RDF Infrastructure for Shared Web Annotations*. In Proc. of the WWW10 International Conference. Hong Kong, 2001
[18] http://www.w3.org/2001/Annotea/Plan/context/newcontext.html
[19] John Domingue, Martin Dzbor and Enrico Motta. *Collaborative Semantic Web Browsing with Magpie*. In Proc. of the 1st European Semantic Web Symposium (ESWS), May 2004.
[20] Smart Tags http://office.microsoft.com/en-gb/assistance/HA010347451033.aspx
[21] Thunderbird http://www.mozilla.org/products/thunderbird/
[22] Ben Hammersley. *Developing Feeds with RSS and Atom*. O'Reilly, 2005.