Biblioref: a Semantic Bibliographic Reference Management System

Renan Manola
Computer Science Department
Federal University of Espirito Santo
Av. Fernando Ferrari, S/N,
29060-970 – Vitória/ES
+55 27 9982 1609
rmanola@gmail.com

Renata S.S. Guizzardi
Computer Science Department
Federal University of Espirito Santo
Av. Fernando Ferrari, S/N,
29060-970 – Vitória/ES
+55 27 4009 2196
rguizzardi@inf.ufes.br

Roberta Lima Gomes
Computer Science Department
Federal University of Espirito Santo
Av. Fernando Ferrari, S/N,
29060-970 – Vitória/ES,
+55 27 4009 2130
rgomes@inf.ufes.br

ABSTRACT
This paper presents Biblioref, a system that is meant to manage content from academic professionals. The main benefits of this system can be described as: a) promoting collaboration among users through mutual access to each other's documents; b) granting user autonomy in knowledge organization, since documents are classified according to individual points of view instead of a centralized model; and c) providing mechanisms to relate the different user's classification schemes, allowing to find potential collaborators, inferred from these relations. Moreover, Biblioref is Semantic Web compliant, opening possibilities for interoperating with other systems.

Categories and Subject Descriptors
H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *data sharing, web-based services.* D.2.2 [**Software Engineering**]: Design Tools and Techniques – *modules and Interfaces.*

General Terms
Algorithms, Management, Design.

Keywords
CSCW, Content Management System, Taxonomy, RDF, Semantic Web.

1. INTRODUCTION AND MOTIVATION
In the past few years, collaborative systems presented considerable growth mainly because of its wide and almost unlimited use in the INTERNET. Besides being used for fast websites deploying, systems like Content Management Systems (CMSs), can be used as a powerful tool for promoting collaboration through the web. In particular, Knowledge Management Systems (KMS), which are aimed at supporting knowledge creation, integration and sharing, have been favored by this growth [2].

Academic articles are examples of knowledge artifacts that academic professionals tend to store. By means of these artifacts, knowledge can be shared. Often, article sharing occurs when a person recommends the reading of an article or sends it through e-mail to a colleague. Here, we propose to build a KMS (on top of a CMS) named Biblioref to support academic knowledge sharing. In fact, systems that allow the storage and retrieval of bibliographic data (such as academic articles) can be seen as simple CMSs. However, when a CMS promotes users collaboration, (for example, enabling users to see each other's stored articles and learn from it) this system becomes a KMS. This is the kind of system we intend Biblioref to be.

Currently, the organization of KMSs are often based on central repositories and classification schemes. However, this approach can lead to user's dissatisfaction because of lack of flexibility. In other words, users may not be familiar or comfortable with terms used at these static classification schemes, which can lead to system abandonment or misuse. In general, users prefer to have more control over what they are using. Here, we propose an alternative KMS based on individual taxonomies, allowing users to have more autonomy on knowledge sharing. At the same time, a central taxonomy is defined and used as a reference allowing different user's classifications to be associated.

The remaining of this article is organized as follows: section 2 presents an overview of the system's characteristics. Section 3 discusses some previous work that is somehow related to the scope of this initiative, and finally, section 4 concludes this article pointing some improvements that can be accomplished as future work.

2. SYSTEM OVERVIEW
The Biblioref[1] is a web-based system based on an open source CMS [5] that manages bibliographic content from academic professionals, which is mostly composed by scientific articles, thesis, dissertations, and other academic publications. The system allows each user to create his/her own taxonomy. A taxonomy is scheme composed of hierarchical relations between concepts. At the user's point of view, it can be seen as the directory structure used to store documents at their local computers. Dealing with the taxonomy in this way, it becomes very straight forward for the user to import his/her document organization to the system and start using it. At the same time, the system has a reference taxonomy which works as a central taxonomy used to classify the content uploaded by the users and to establish relations between terms of different users' taxonomies.

[1] Available at: http://rmanola.890m.com, for testing purposes.

2.1 Document Retrieval Based on Taxonomy Matching

At Biblioref, knowledge artifacts can be associated with different terms, which may be part of different taxonomies. In contrast, other reference management systems are usually limited to only one term per artifact. This enriched classification gives more meaning to content. When the artifact is previously bound to a taxonomy term, it is always displayed when the user views the listing page of that term. A knowledge artifact at Biblioref always has, at least, two different terms associated with it (one from the reference taxonomy and the other from user taxonomy)

There are, basically, three kinds of relation that can be established in the system: 1) Document ↔ User Taxonomy Term; 2) Reference Taxonomy Term ↔ User Taxonomy Term; 3) Document ↔ Reference Taxonomy Term.

The first is believed to be the most important by the user point of view, stating that the document being submitted is going to be part of a certain term in the user's organization. In order to establish this relation, the term must be provided by the user during document submission.

The relation between User Taxonomy Term and Reference Taxonomy Term (the second relation) constitutes one of this system's distinguishing features. It allows the mapping between user's taxonomy terms and the autonomous establishment of the third relation (Document – Reference Taxonomy Term). This association occurs when a term is created. As it is part of a taxonomy, the only inputs needed at this point is: the term *name*, the *synonyms* of this term and the *position* at the tree of user's classification (which specifies if a term is child of another term, if it is root, etc). The synonyms provided are important because the system uses them to build a rank of occurrences against the terms from the reference taxonomy (besides, each term of this taxonomy has synonyms). Based on that rank, there is an algorithm that decides to which term of the reference taxonomy the term of the user's taxonomy being submitted is going to be related. If the algorithm isn't able to automatically find out this relation, (perhaps because no synonyms or names were matched, or too much matches were found), the system shows a form to the user, requesting him to manually point out the related term from the reference taxonomy.

The third relation happens after content submission. The information provided by the user at this moment is: *title*, *abstract*, *keywords*, *authors* and *term* (of the user's taxonomy mentioned before) that best classifies the knowledge artifact's content. The algorithm is similar to the one applied at the second relation establishment, but it builds the rank from the document title and keywords against the names and synonyms of the reference taxonomy terms. If no matches or too many are found (in a way that the algorithm can't reduce them), the system does not require user's assistance. In this case, it can bind this document with the reference taxonomy term that is, on its turn, related to the user taxonomy term specified at the content creation form.

Summarizing the relations explained above, in the best case the system can infer two out of the three existing relations. At the worst case, the system infers only one (specifically, the third one).

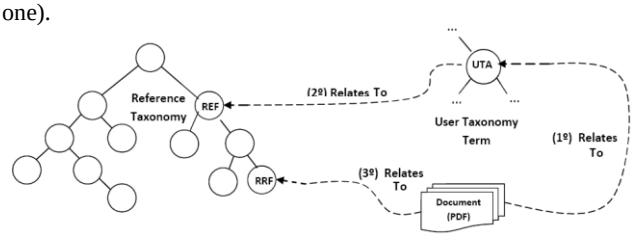

Figure 1. Illustration of the 3 kinds of existing relations.

Figure 1 shows a visual example of the best case scenario. Note that the user taxonomy term related to this document (i.e. UTA) is associated to the "REF" term in the reference taxonomy. However, the document itself is related to the "RRF" term in the reference taxonomy. This occurs because in this case, the system could automatically infer the third relation (i.e. Document – Reference Taxonomy). If that were not possible, the document would be automatically associated to "REF" (and not "RRF"). In any case, such classification comes as a suggestion and the user is able to provide the final decision regarding which term the document should be associated to.

2.2 Content View

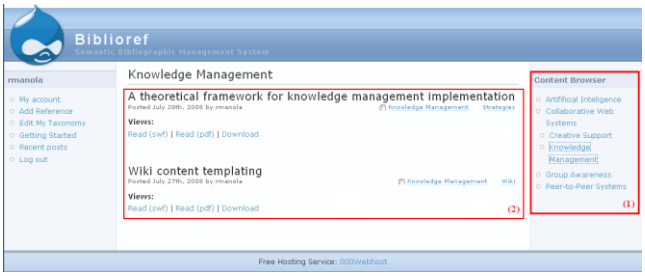

Figure 2. The (1) Taxonomy Browser Feature and (2) Documents' preview

Figure 2 shows an example of a term node that was selected at the (1) taxonomy browser (i.e. Knowledge Management) and (2) the knowledge artifacts associated with it, in this case, two different documents. This browser is focused at the usability of the user to retrieve documents, so it shows only the 2nd relation concerning Figure 2.

Figure 2 (2) also presents some *"Views"* links, i.e. the system offers three ways to access the content of the retrieved documents. The first of them consists of reading it embedded in the website using the Flash plug-in (every submitted PDF document is converted to SWF format using an external free service of Scribd [6]). The second link provides the embeddable reading using a PDF plug-in of the browser. Finally, the third option allows downloading the document itself. Every user (even if not logged in) can choose one of these access modes. As the user clicks on any of these *View* links, it is taken to the full view of the content, which displays more details about it, such as *abstract*, *keywords*, *authors* and *the classification trees* of that content. These trees are depicted at Figure 3. As a document is always related with at least two taxonomies (user and reference), we show the parental tree of the related term in each of them. Through it, the user can grasp how different people can classify the same artifact and how each term of these different classifications can be related using the reference taxonomy.

Figure 3. Classification Trees of a document

2.3 Compliance with the Semantic Web

The Semantic Web constitutes an extension of today's web architecture [2]. It has as major concern to make information available not only to human beings, but also to machines in a way that these machines are enabled to understand the context in which content is involved. This context regards extra information that can help it better process and make inferences about the content. The Biblioref system makes this information available using a microformat of the Resource Description Framework (RDF). Microformats are, in general, an approach to semantic markup that seeks the reuse of existing XHTML and HTML to convey metadata and other attributes. The applied microformat is the Embeddable RDF (eRDF), which consists in a way to encode the RDF statements embeddable with the HTML of a webpage. This way, the page that provides the information to human beings is the same that provides the context information for machines that understand this standard[2]. The biggest advantage of using eRDF is that other websites or web-tools can browse the Biblioref content and perform queries without having direct access to its database or internal organization. The information coding is in conformity with w3c (http://www.w3.org/) standards (e.g. already established standard namespaces[3] are used), the standard body which cares for the constant development of the *World Wide Web*.

3. Related Work

The main Biblioref feature is to allow a high degree of autonomy to the user in organizing, visualizing and searching for knowledge. This is accomplished by the use of taxonomies as individual and reference schemes to classify documents. There are other initiatives that make use of taxonomies in order to better manage knowledge. KARe [3] is a system that applies taxonomies to classify different knowledge artifacts. This classification is used to help the system answer to natural language questions submitted by the users. Analogously, KEEx, [4] is a KMS which allows users to share documents in a peer-to-peer fashion. Like in Biblioref, the users of KARe and KEEx organize, visualize and search for knowledge using individual user taxonomies. However, instead of applying reference taxonomy, both KARe and KEEx apply automatic algorithms to match the users' taxonomies. The main disadvantage of fully automatic approaches refer to performance, since automatic solutions tend to have high cost in terms of CPU processing, besides not always finding suitable answers. On the other hand, semi-automatic strategies, like the one we apply, although providing better performance on both terms, rely too much on the user to provide its results.

Finally, another related work is Caravela [1], a content management system with automatic information categorization used to classify submitted content. This classification is also based on taxonomies; however not individual but rather centralized ones. As already mentioned in section 1, such central schemes are considered inflexible, often leading to user dissatisfaction.

4. Conclusion and Future Work

This paper presents Biblioref, a semantic bibliographic reference management system. Our first prototype proposes a reference taxonomy regarding the CSCW domain. However, its scope can be extended to other domains and applications, fundamentally depending on the applied reference taxonomy and the type of information that better describes the documents being shared (abstract in the context of academic articles, headline in the context of news documents, etc.).

More tests should be done in order to verify the efficiency of the semi-automatic relation establishment algorithm. Adding other system functionalities is also in our research agenda. For instance, we hope to provide an in-depth search, which should provide contextual information about content classified in different levels in the user and reference taxonomy, ranking documents both in terms of these levels and in regard with the similarity to the user query.

Acknowledgements. This work has been partially funded by FAPES/MCT/CNPq/CT-INFRA #36316008/2007, and by FAPES/Universal #38874849. Renan Manola is supported by a PIBIC scholarship from CNPq. Renata S.S. Guizzardi is supported by a DCR scholarship #37274554/2007 from FAPES.

5. References

[1] Aumueller, D.; Rahm, E. "Caravela: Semantic Content Management with Automatic Information Integration and Categorization" E. Franconi, M. Kifer, and W. May (Eds.): ESWC 2007, LNCS 4519, Springer-Verlag, pp. 729–738, 2007.

[2] Davies, J.; Fensel, D.; Van Harmelen, F.: "Towards The Semantic Web: Ontology-Driven Knowledge Management", Wiley, 2003.

[3] Guizzardi, R.S.S.; Ludermir, P.G.; Sona, D. A Recommender Agent to Support Knowledge Sharing in Virtual Enterprises. In Protogeros, N. (Ed.). Agent and Web Service Technologies in Virtual Enterprises, Idea Group Publishing, 2007.

[4] Bonifacio, M., Bouquet, P., Mameli, G., and Nori, M. Peer-Mediated Distributed Knowledge Management. In van Elst, L., Dignum, V., and Abecker, A. (Eds.) Agent-Mediated Knowledge Management, LNAI 2926, Springer-Verlag, pp. 31–47, 2004.

[5] Drupal web site: http://drupal.org.

[6] Scribd web site: http://scribd.com.

[2] An Online eRDF parser that outputs XML can be found at: http://research.talis.com/2005/erdf/extract.

[3] Namespaces like RDFS and Dublin Core (DC).