PaperPuppy: Sniffing the Trail of Semantic Web Publications

Jennifer Golbeck, Yarden Katz, Daniel Krech, Aaron Mannes, Taowei David Wang, James Hendler

MINDSWAP, University of Maryland, College Park, MD 20742, USA,
WWW home page: http://www.mindswap.org

Abstract. PaperPuppy is a system designed to allow the exploration and analysis of a network of ISWC conference authors and publications. Using proceeding data from the first four ISWC conferences, we show co-authorship, citation, institutional affiliation, co-depiction in photographs, and other connections supported by the underlying ontologies. We demonstrate the rich browsing experience with the PaperPuppy site, and support it with a variety of cutting-dge Semantic Web tools.

1 Introduction

It is a maxim in the intelligence community that amateurs look at content, and professionals look at traffic. Obviously, in a research community, such as the community developing the Semantic Web, content is paramount - nonetheless, examining patterns of collaboration and citation can be revealing. With these guidelines, by processing and graphing information PaperPuppy is intended to:

1. Identify influential papers, people, and institutions within the Semantic Web research community
2. Spot collaboration patterns and possible future individuals and institutions of influence
3. Find papers worth reading and individuals worth contacting to expand their knowledge of the Semantic Web research community and its work
4. To see trends in accepted papers, and help identify which may be most interesting to their own interests

The main interface is the PaperPuppy website [1], which is driven entirely by ontologies and RDF data. Within the system, a variety of tools are available for analyzing the network. Because the system is driven by Semantic Web technology, PaperPuppy is coupled with a suite of tools for ontology editing, instance creation, and rule authoring that enable users to manipulate the views and add capabilities to the system.

In this paper, we introduce PaperPuppy through the support it offers and tools it has for analysis and maintenance of the data.

[1] http://paperpuppy.mindswap.org

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2006		**00-00-2006 to 00-00-2006**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
PaperPuppy: Sniffing the Trail of Semantic Web Publications	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Maryland,MINDSWAP,College Park,MD,20742**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**8**	

2 System and Ontologies

The website is driven entirely by Semantic Web data. The navigation and organization of the site is based on multiple ontologies. Thus, it is important that users can edit and expand ontologies in the system.

The core technologies that power Paperpuppy are RDFLib [2] and Redfoot [3]. RDFLib is a Python library for working with RDF. RDFLib has a graph interface for parsing, storing, querying and serializing RDF triples. The RDF graph can be queried using triple patterns or SPARQL queries. The library contains a number of parsers and serializers for various formats including RDF/XML, N3, and RDFa. Graphs can be backed by one of several persistent stores, namely, Sleepycat, MySQL, SQLite, ZODB, sqlobject, Redland, in addition to an in-memory store. RDFLib is being used to power www.mindswap.org. Redfoot is an application layer on top of RDFLib. It bundles other third party software for the running of the site and portable server-side module management.

The PaperPuppy ontology is a small, focused ontology to structure specific information. The major classes are obvious, Person, Institution (with subclasses for Labs, Universities, and Departments), Publications (with Subclasses for Articles and Papers), outlets for Papers (Conferences and Publications), and Location. One interesting challenge was how to present researchers as connected to a specific institution. A property reflecting this affiliation was insufficient as one may be affiliated with several institutions in one's career. A Class labeled ResearchRole was created to fill this gap. ResearchRole's properties are hasResearcher (range: Person) and hasAffiliation (range: Institution). This allows papers by a researcher who has been affiliated with several institutions to be linked to their specific institution. While this class is not exposed in the portal's main navigation page, PaperPuppy would be a poorer resource without it.

The properties of the ontology have several interesting features. One challenge was Time, which was an essential property to show the changes in connections in the Semantic Web community, but can be handled in several different ways on the Semantic Web. Ultimately, using the year as the key unit was chosen as the most effective option. While a researcher may change positions in the middle of year, this would be reflected in the ResearchRole instance and their new Affiliation would still be part of the knowledge base.

Transitivity was used in several places to indicate that sub-Institutions remain under the rubric of the primary institution and in the part of the ontology dealing with locations. Inverse is a very useful property for the easing data entry. For example the property hasAuthor has the domain of publication and range of ResearchRole. It has an inverse property isAuthorOf. In this way, a site visitor entering data about a paper could enter the paper's author under the property hasAuthor. Alternately, someone visiting the site writing about a person and their ResearchRole, could enter a paper they had written under isAuthorOf.

[2] `http://redlib.net`
[3] `http://redfoot.net`

The PaperPuppy ontology imports FOAF [4]. In particular, PaperPuppy:Person is a subclass of FOAF:Person. We hope that as the site grows and more people are participating, users can submit their FOAF files in addition to their publications. With the addition of FOAF files, the possibility of comparing/contrasting a Person's social and professional relationships opens up.

3 Instance Creation

Because there is support for multiple ontologies, PaperPuppy can pull instance data from the Web in bulk and integrate it into its knowledge base. However, users of the system can add/edit data manually for finer control using a variety of tools.

Within the site itself is an easy-to-use instance editor. An important advantage of this editor is that it allows the user to enter data live on-site, without incurring additional external tools. Easy to use, the instance editor allows the user to create, delete, and merge Instances or add, edit, and delete properties. The editor provides a pair of pull-down menus, the first gives a selection of properties, the second a selection of Instances with the appropriate range.

Digital images can be annotated, submitted, and later retrieved by using PhotoStuff [1]. PhotoStuff is a digital image annotation tool that can be coupled with a Semantic Web portal site like PaperPuppy. Users can load ontologies and digital images, and annotate the latter with the former. The annotation process creates instances of the classes used, and users can additionally make assertions about these instances within PhotoStuff. These instances, along with any EXIF metadata can be submitted to a site store. Users can later load the metadata from an existing store, and reuse the pre-published instance to further describe them, or use them to describe new instances. PhotoStuff can load the PaperPuppy store to annotate photos as well as to edit existing metadata. We have created a version of PhotoStuff that have the store already preloaded when the tool starts up to facilitate the process.

4 Analyst Support

The intended functionality of this system is for analysts working with complex networks of people, places, events, objects, and transactions. Beyond simply presenting the data in the knowledge base, we provide a set of tools for discovering relationships manually and automatically.

4.1 Visualization

An ontology suitable to model and represent a particular domain may not be well suited for additional analytical tasks an analyst may want to perform. For additional synthesis of data, we provide a framework where rules can be used to

[4] `http://xmlns.com/foaf/0.1/index.rdf`

create digests for certain aspects of the data. The creation and detailed usage of rules are explained in the rules section. Here we detail how visualization driven by these rules are important to tasks that would otherwise be difficult to perform.

For each Person, we display a network depicting such Person's relationships. We have defined the "Colleague" relationship and the "co-authors with" relationship by using rules. We have also asserted that these relationships are to be displayed by graphs at the time of rule creation, so the applet would know to display them. The relationships are initially displayed to a depth of 3, but users can choose to filter the depths shown in the applet.

While the graph for each rule is useful – one can see how people are related in one view without having to link-chase for all of them – the combination of the graphs can be even more telling. For this reason, we allow users to merge graphs of any single relation in a separate workspace. Here we illustrate a use case.

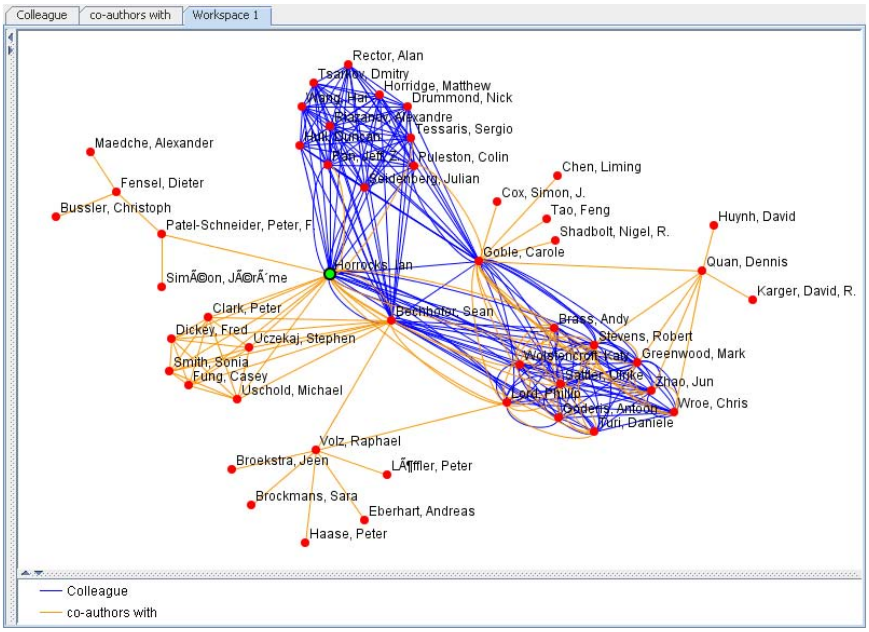

Fig. 1. Collaboration (coauthorship) among labmates is common, but here we see that some researchers (Horrocks, Goble, and Bechhofer) have a fairly large number of collaborators outside of their labmates.

4.2 Provenance Tracking

As shown in figure 2, the provenance of each statement is tracked and displayed in PaperPuppy. If a person has directly asserted a statement, the name and

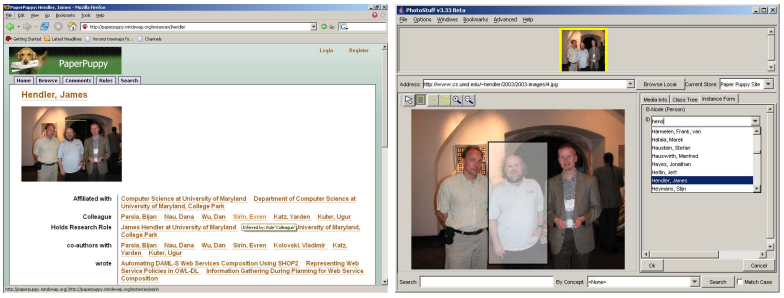

Fig. 2. Left: The author's page in PaperPuppy shows the different relationships supported by the ontology and rules. Each link is browsable, so analysis can see the people or objects to which the author is connected. Provenance is also shown on mouseover. In this example, we are shown that the colleague relationship with Evren Sirin was inferred from a specific rule. Right: Image seen on the left figure are annotated and submitted via PhotoStuff. The pull down on the right shows all the known instances of Person from the store, facilitating the annotation.

email address of the author is shown. If the statement was inferred from a rule, the name of the rule is also shown. This allows an analysts to know from where each piece of information came, and to make judgements about the fidelity of data based on that.

4.3 Comments

All pages on PaperPuppy have the option for users to make comments. Comments are stored on the system and displayed with the date and author of each comment. This allows analysts to discuss, disagree, or share information about the data.

5 Rule Systems

If the ontology is the frame that organizes the data on PaperPuppy, the rules are the engine that gives the site its power and enables the user to shape the masses of data into queries and come to a better understanding of the Semantic Web research community. Rules are employed in PaperPuppy for two primary usages - data management and presentation. Even though paper writers and editors have taken every precaution to ensure the format and content of each published article is correct, there are still many errors (e.g. spelling "university" as "univeristy"). In addition, minute differences in the names of people or institutions translate to the creation of different instances, even though we know they are the same. This problem is especially rampant as we are gathering data from multiple sources. To remedy these issues, we allow users to assert that any two instances are the same, and rules are used to merge the triples about these two instances.

The second use of rules on the site is on the presentation layer. Because ontologies are designed to model specific domains, and not for expediting the browsing, searching, and general understanding of the data itself, an effective presentation layer is important. For example, the PaperPuppy ontology has the ResearchRole class, which ties a Person and an institution together. The default display of the site is to display all triples associated with an instance. However, users will have to first click on a particular ResearchRole a person holds, and then the institution to view the institution, as opposed to one single click. Rules are employed in this case to create additional triples to facilitate common browsing tasks. Next, rules power the visualizations that help the site user come to a more intuitive understanding of the relationships within the Semantic Web community. Using rules, analysts can generate cusomized queries to explore these issues. Queries might include who, or what papers, individuals or research institutions most frequently cite and how that changes from year to year - or how often does an individual researcher cite their own work. Building on this base, specialized services have been designed. One is the "Mentor Finder", which looks for researchers who cite a colleague in multiple papers. Another advanced service based on rules is the "Allied Institutions Search", which looking at cross-institutional collaboration matches multiple instances of co-authors who are not colleagues, and whose institutions are in different locations. This rule could be expanded on, by including shared favored papers for citations, to predict future likely co-authors. As the content on PaperPuppy expands, the possibilities for new services will expand as well - presenting site visitors with new ways to explore the patterns of collaboration and research on the Semantic Web.

5.1 Rule Editing

Registered users of PaperPuppy can use the ontology editor Swoop [2] to edit, view and submit rules to the site, where these rules are applied.

The rule associated with an OWL entity defines what that OWL entity is (in addition to what the OWL ontology defines). A user defines a conjunction of antecedents (atoms) as the condition, and the consequent is the current OWL entity in view. For example, if one wants to define the hasUncle property in terms of hasSibling and hasChild using rules, the rule will look as follows. The left hand side is the consequent, and the right hand side are the antecedents.

$hasUncle(?x, ?y) : -hasChild(?z, ?x), hasSibling(?z, ?y)$

(x has uncle y if z has child x, and z has sibling y)

When a user activates the "Add" link, a popup dialog will appear for the editing of any rules associated with this OWL entity. The editing popup is shown below, using the hasUncle example.

The rule editing dialogue is shown in figure 3. Users can choose to add/delete/edit antecedents, and the bottom of the dialogue shows the full rule and its expressivity. When a rule is created, a registered user can publish it to PaperPuppy. The Pychinko [3] rule engine will process all submitted rules, and add the inferred facts to the knowledge base.

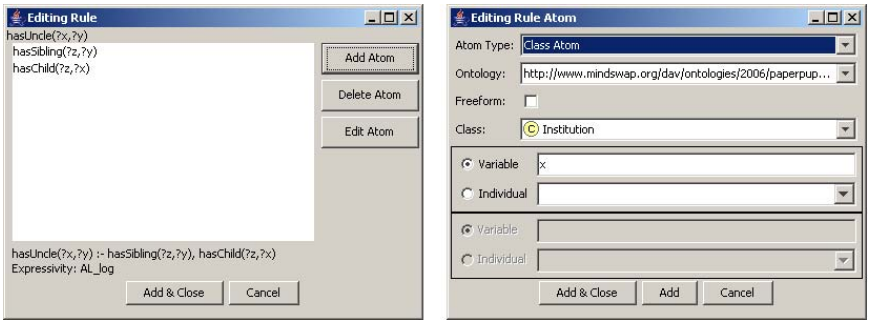

Fig. 3. Swoop's dialogues for editing rules and rule atoms.

5.2 Rule Application

Pychinko is a forward chaining rule engine, written in Python, that implements the Rete [4] algorithm. Its rules are expressed in the N3 language and its facts as RDF triples. In addition to handling plain forward-chaining rules, Pychinko also supports conjunctive query over RDF stores, and implements several useful CWM [5] math and string builtin functions.

In PaperPuppy, Pychinko is used as both a lightweight reasoner for OWL, and as a reasoner for domain-specific rules. For the first use, many of OWL's constructs – such as the behavior of inverse properties, transitive properties, or sameAs assertions – can be captured incompletely by rules. The advantage of this approach is that the overhead of heavy duty description-logic based reasoners is avoided, yet many useful inferences are still obtained. For the second use, there are many rules concerning the specific domain of paper and authorship that are used to make the website more intelligent. For example, the simple rule that if two individuals are authors of the same paper, then we can inference them to be co- authors. Some of these rules make use of useful builtins from the CWM library [5], such as case insensitive string comparisons for author names that were scraped from citations databases. Finally, PaperPuppy tracks the number of inferences yielded by each rule, and recomputes these dynamically when rules are added to or deleted from the system.

6 Requirements, Features, and Ongoing Work

PaperPuppy meets all of the minimal requirements set forth in the Semantic Web challenge. We also meet all of the additional desires, except for support for multiple languages. We are currently expanding the site's functionality to include more citation data, image codepiction and inline annotation browse support, and interfaces for FOAF/BibTex data uploads.

7 Discussion and Conclusions

The origin of PaperPuppy was with MINDSWAP's PiT (Profiles in Terror) project [6]. It is an effort to create a Semantic Web portal that provides value added for a terrorism researcher. PiT has over 5000 instances and is growing rapidly, along with a steadily growing ontology with over 150 classes and over 300 properties. Not to imply that Semantic Web researchers are of the same nature as terrorists, but the same traffic analysis can be applied across domain, where linkage of researchers and publications allow patterns of importance to emerge. We demonstrate that PaperPuppy is a Semantic Web portal integreated with a rich set of Semantic Web tools and that it offers unique analysis of ISWC paper trails.

8 Acknowledgements

Thanks to Hamid Haidarian Shahri, Keith Mantel, Marissa Moskowitz, and Sharone Horowit-Hendler for assistance in data gathering and formatting. Special kudos to Michael Grove for creating a PaperPuppy-ready version of Photo-Stuff Webstart.

This work, conducted at the Maryland Information and Network Dynamics Laboratory Semantic Web Agents Project, was funded by Fujitsu Laboratories of America – College Park, Lockheed Martin Advanced Technology Laboratory, NTT Corp., Kevric Corp., SAIC, the National Science Foundation, the National Geospatial-Intelligence Agency, DARPA, US Army Research Laboratory, NIST, and other DoD sources.

References

1. Halaschek-Wiener, C., Schain, A., Golbeck, J., Grove, M., Parsia, B., Hendler, J.: A flexible approach for managing digital images on the semantic web, 5th International Workshop on Knowledge Markup and Semantic Annotation (2005)
2. Kalyanpur, A., Parsia, B., Sirin, E., Cuenca-Grau, B., Hendler, J.: Swoop - a web ontology editing browser. Journal of Web Semantics **4**(1) (2005)
3. Katz, Y., Clark, K., Parsia, B.: Pychinko: A native python rule engine. Proceedings of PyCon International Conference (2005)
4. Forgy, C.: Rete: A fast algorithm for the many pattern/many object pattern match problem. Artif. Intell. **12** (1982) 17–37
5. Berners-Lee, T., Connolly, D., Prud'homeaux, E., Scharf, Y.: Experience with n3 rules. W3C Workshop on Rule Languages for Interoperability (2005)
6. Mannes, A., Golbeck, J., Hendler, J.: Semantic web and target-centric intelligence: Building flexible systems that foster collaboration, IUI Workshop on Intelligent User Interfaces for Intelligence Analysis (2005)