Pundit: Creating, Exploring and Consuming Semantic Annotations

Marco Grassi[a,1], Christian Morbidoni[b,1], Michele Nucci[c,1], Simone Fonda[d,2], and Francesca Di Donato[e,3]

[1] Semedia Group, Università Politecnica delle Marche, Italy
[a]m.grassi@univpm.it, [b]christian.morbidoni@gmail.com, [c]m.nucci@univpm.it
http://www.semedia.dibet.univpm.it/

[2] NET7, Italy
[d]fonda@netseven.it - http://www.netseven.it

[3] Scuola Normale Superiore, Italy
[e]francesca.didonato@sns.it - http://www.sns.it/

Abstract. This paper presents Pundit, a novel semantic web annotation tool, and demonstrates its use in producing structured data out of users annotations. Pundit allows communities of scholars to produce machine-readable annotations that can be made public and thus consumable as web data via SPARQL and ad-hoc REST APIs. Pundit is highly configurable and can deployed in custom instances to include well-defined and agreed annotation vocabularies. Such instances can be distributed as bookmaklets to community users so they can create uniformly structured data in a certain application scenario. Basing on the provided APIs, some demonstrative applications have been developed, exploring different use scenarios, ranging from philosophy to journalism and cultural heritage. The main aim of this paper is to demonstrate how such uniformly structured annotations can be quickly re-used on the web to make information discoverable or to visualize it in interesting ways.

Keywords: Digital libraries, Semantic Web, Ontology, Data Model

1 Introduction

Annotation is a primary activity for scholars and professionals. It consists in enriching a content with some new information, which possibly helps in understanding or searching the content itself. While until few decades ago annotations were sketched by hand on the side of a book, today web technologies have the potential to make them infinitely replicable, remotely accessible and easy to share. Web annotations systems and bookmarking/clipping tools are popular nowadays both among generic users (e.g. social tagging) and among scholarly communities (e.g. Zotero[4], Mendeley[5] users). However, existing annotation systems are generally limited to textual comments, tags or predefined metadata

[4] http://www.zotero.org/
[5] http://www.mendeley.com/

templates (e.g. bibliographic records). Furthermore, annotations are often isolated into closed systems and very rarely are connected to the Web of Data. The simple idea behind our work is that of making of annotations a vehicle to create new semantic web data, actually adding links and, ultimately, knowledge to the so called Global Data Space [1]. Once annotations become available in a standard and highly expressive form, a variety of applications can be built to visualize the resulting knowledge in specific domains. Pundit is a novel annotation system that aims at implementing this vision, by enabling annotators (e.g. scholars) to use semantically specified relations and link to web of data entities, producing in fact accessible RDF graphs out of their work. Such RDF graphs are collections of annotations that we call "notebooks". Notebooks can be consumed via REST APIs or standard SPARQL endpoints. In this paper we first overview Pundit at a high level, then we focus on the issue of effectively re-using the annotations produced in Pundit to drive demonstrative use cases and address end-user needs such as sharing, exploring and visualizing annotations. Two main directions are currently being targeted. In Ask[6], we attempt at creating a portal to manage annotations, share them and explore public notebooks. We then explore, by means of some demonstrative developments, the possibility of basing on the Pundit "framework" to build vertical, specialized applications. In the latter case, the basic pattern we follow is that of configuring and deploying custom instances of Pundit, which can be distributed among users. Such instances generate annotations that conform to pre-defined data schemas and can be quickly fed into existing open-source tools to produce more interesting visualizations. Nevertheless they maintain the generality and flexibility of RDF, thus being compatible with Ask or other "general purpose" usages of data.

2 Related Works

An exhaustive state of the art in semantic annotation goes beyond the purpose of this paper and can be found in the literature [3] and this section focus only on tools related to our work. The semantic tagging paradigm, which exploits publicly available Linked Data sources to retrieve unambiguous tags, has been implemented in Faviki[7] and Europeana Connect Media Annotation Prototype (ECMAP)[4]. Other tools such as One click annotation [5], CWRC-Writer [6] and LORE (Literature Object Reuse and Exchange) [7] also allow the use of restricted vocabularies or ontologies in the annotations. Some annotations tools, as LORE and CWRC-Writer enable also the editing of more expressive annotations in the form of subject-predicate-object statements. Although not based on Semantic technologies and not supporting semantic annotations, Open Knowledge Foundation (OKFN) Annotator[8] has been conceived as a JavaScript library that can be added to any Web page, both adding it into HTML and injecting it using a bookmarklet, to make it annotatable, similarly to Pundit.

[6] http://ask.as.thepund.it
[7] http://www.faviki.com
[8] http://okfnlabs.org/annotator/

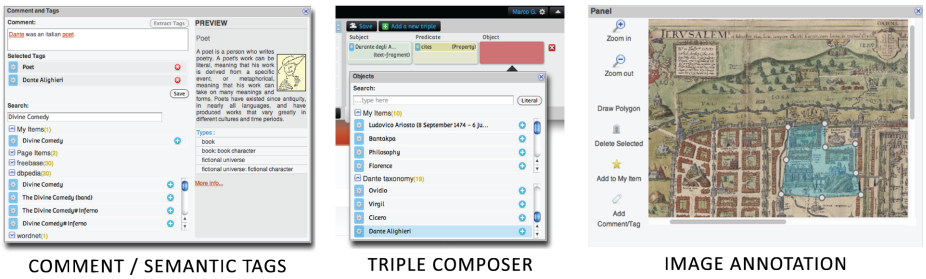

Fig. 1: Creating annotations with Pundit

3 Pundit overview

Annotations in Pundit are essentially triples that connects different kinds of items together. A triple has the form [subject - predicate - object], where the subject and object can be segments of text and images (e.g. [text - describes - image]¿) or entities from the web of data (e.g. [text - has author - Dante(from Freebase.com)] or [image - depicts - Florence(from DBPedia.org)]). The most expressive annotation interface provided by the Pundit client is the "triple composer". It allows users to drag and drop items into triples, or select them from the web page (e.g. by selecting a text or an image), as well as searching into available vocabularies and data sources. However, other annotation wizards support specific kind of annotations, as putting two segments of text in relation, or attaching tags and comments to a text segment. Image annotation of a segments of images is supported by a dedicated module as shown in 1. The Pundit client is a JavaScript application that can be deployed as a library, to then be easily included in existing web sites to make the content "annotable"[9], as well as delivered as a bookmarklet. A bookmarklet is a simple link (bookmark) that, once added to a web browser allows loading Pundit on every web page and annotating its content.

In Pundit, an annotation contains information at a twofold level. The first one is the "annotation metadata" and deals with the act of annotating, including information on the author, the time of creation and the involved web resources. Pundit bases on the Open Annotation data model (OA)[10] [8] for representing this dimension. The second, the "annotation graph", is an RDF graph resulting from metadata and relations among web resources that a user has created by annotating. In other words, it captures the semantics of the annotation representing the user's contribution in terms of "domain knowledge". For example, an annotation graph could contain Wikipedia pages corresponding to Italian writers and relevant text segments from their works on wikisource.org or other open web archives, perhaps linking each text to a number of other texts from relevant contemporary writers. We call "items" the nodes of such a graph, which represent

[9] this has been done in wittgensteinsource.org
[10] Open Annotation core specification: http://www.openannotation.org/spec/core/

the annotated web resources, being them web pages segments or other kind of entities (places, persons, etc.) While no restrictions are applied and no assumptions are made by Pundit regarding the ontologies used in annotation graphs, a certain knowledge of the structure of the single annotations into a notebook has to be owned by a developer to implement a meaningful visualization based on such "free shaped" data.

So far, one of the most successful approaches to foster the reuse of data on the web is to create a consensus around vocabularies and ontologies within a certain community. In Pundit we try to follow this pattern by make it possible to deploy customized annotation clients, in the form of JavaScript libraries or bookmarklets, which can be distributed to users by "community leaders". A custom client possibly includes a precise set of a well-defined set of "relations" to be used in annotations to create typed links among items or taxonomies where relevant web entities are collected and ready to be annotated. Both taxonomies and relations are represented in JSON and can be easily extracted from existing vocabularies (e.g. SKOS) or ontologies, as we did in the Wittgenstein's brown book pilot[11]. Aggregating items in collection, for sharing and publishing, is a common pattern in social clipping and bookmarking tools. In Pundit, annotations are collected in notebooks that users can optionally make publicly accessible. When a notebook is public, the annotations contained in it are not only shown in the Pundit client (e.g. when a user loads the Pundit bookmarklet on one of the annotated web pages) but, more interestingly, a notebook can be consumed by means of open REST APIs and accessed by a variety of web applications. Each notebook provides a SPARQL endpoint to query its content. In other words, a notebook is an independent RDF graph created by a given user in time and connecting a variety of web resources.

4 Consuming annotations

Regarding how to use annotations, and the semantic data they enclose, to drive end-user applications, there are mainly two "dimensions" that can be explored:

- Annotation centric approach. This is commonly used in clipping systems where each clip is the result of a single annotation and is shown as a "box" containing some information (e.g. pictures, links, tags) about the annotated items. We mainly based on this approach in designing Ask, a prototype web application to search over public notebooks and manage personal ones.
- Item centric approach. As annotations graphs in a notebook can be in fact consumed as a unique and bigger RDF graph, a possible way of looking at the data is that of focusing the visualization on the annotated items and their relations with other items. This approach clearly benefits from an a-priori knowledge on ontologies and custom vocabularies used in annotations, as it needs to take into account the nature of the information and deals with the "meaning" of annotations.

[11] DM2E blog, Wittgenstein Brown Book experiment, http://dm2e.eu/dm2e-to-start-work-on-wittgensteins-brown-book/

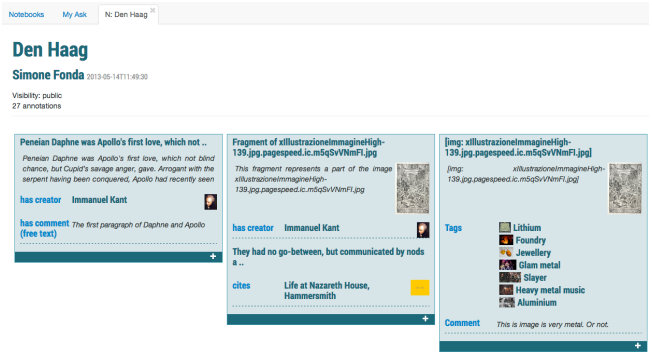

Fig. 2: Exploring notebooks with Ask.

Ask[12] is a web application where public notebooks stored in Pundit can be searched and explored. At the time of writing a new version of the tool is being released. Personal notebooks are accessible to their owner and can be made public or kept private. By default, Ask provides a general purpose visualization of notebooks where single annotations are shown as "metadata boxes". However, alternative visualizations (such as the one described in the following sections) can be easily plugged by providing a compliant REST API. Ask is currently subject to intense development, and one of the most interesting recent features is the prototypal faceted browser available in alpha version[13].

4.1 Edgemaps Visualization: A Demonstrative Use Case.

Edgemaps [10] is an open-source web tool that drives an interesting visualization demonstration in the field of philosophy[14]. The graph is generated by Freebase.com data, which includes "influences" slot in the description of authors. While for a "generic" user such a visualization is enough, we cant probably say the same for scholars that consider such relations as a matter of study and might probably ask: "Why exactly do you say that Marx influences Gramsci?", "What is the evidence of that in the actual primary sources?", "Who said that?". Structured annotations in conjunction with online open content as the one provided by Wikisource[15] make it relatively easy to bring the philosophers demo a little further: generating the graph from scholars annotations made on primary sources (thus including the evidence of the connections), rather than from centralized data. We did so proving an opportunely tuned instance of Pundit and extending, with little programming effort, the Edgempas code. The demonstration is documented on the web site[16]. The Pundit instance uses relations picked picked from

[12] http://ask.as.thepund.it
[13] http://demo.ask.thepund.it
[14] http://mariandoerk.de/edgemaps/
[15] http://wikisource.org
[16] http://www.thepund.it/visualization-demos/philosophers-demo-howto/

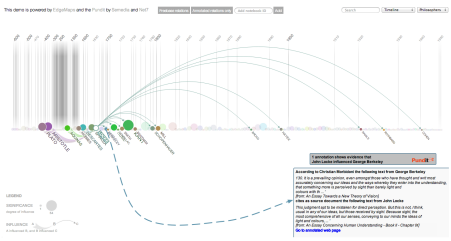

Fig. 3: Showing evidences of philosopher influence with a Edgemap Visualization

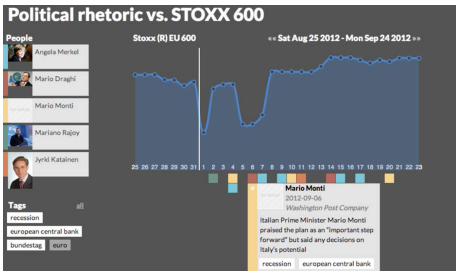

Fig. 4: Political Rhetoric vs. STOXX 600

the CiTO ontology[17] and includes predicates like "cites" and "quotes", as well as other more specific ones like "discusses", "cites as sources", "agrees with", etc. Each time two philosophers are connected by an "influenced by" relation, the corresponding annotations are shown so that the scholar can immediately get an evidence of "why the relation is there". It is also possible to load multiple notebooks from different scholars, thus in fact enabling a collaborative scenario, where annotation authorship is always tracked back and each user can decide what notebook to see or trust.

4.2 Data Journalism

The same pattern can be applied to several contexts and to address very diverse use cases, as economics or journalism. The data journalism demonstrative application shows the use of annotations, in this case quotations from politicians and public persons taken from online news papers, to produce dynamic graphics. A Pundit bookmarklet has been deployed containing a small set of relations (or properties) to tag, describe and date in time politicians' declarations. The associated visual tool has been developed in JavaScript and provided as a web API, which gets a notebook (id) as argument and builds a timeline where annotated declarations are shown along with the trend of a financial indicator. The idea is that of creating a tool for journalists to demonstrate and reveal possible existing connections among what important persons says and how the market behaves. (Fig. 4). A live demo can be found online[18].

4.3 Tracking Annotated Resources Over Time

Timeline visualization has become a common practice for showing data containing time-related information and several tools already exist that allows creating such type of visualization. Instead of developing another one, Pundit reuses TimelineJS[19]. This is an example of the advantages of decoupling annotation

[17] CiTO Ontology: http://purl.org/spar/cito/
[18] Journalism demo, http://ask.thepund.it/?#/timeline/31951d93/20120927
[19] TimelineJS: www.timeline.verite.co

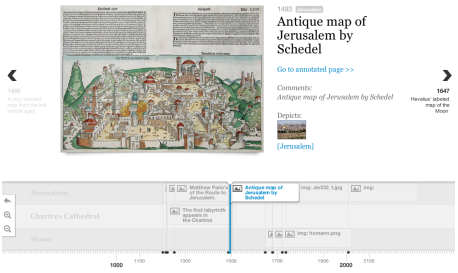

Fig. 5: Timeline visualization

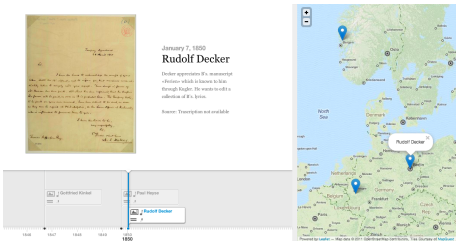

Fig. 6: Tracking annotated resources position over time and space.

creation and consumption. A pundit API has been created that allows to extract time-related information from a notebook given its id and to convert them in a TimelineJS compliant JSON to feed the timeline, as shown in Fig. 5. Annotations in the notebook simply need to have date information, i.e. they have to contain triples having as subject a text fragment or an image, as predicate "dates to" (for a date) or "start date" and "end date" (for a period in time) and as object a date.

A similar approach has been used in another experiment conducted in the context of the Burckhardtsource.org platform that aims at mapping and producing a critical edition of the extensive correspondence of 400 European intellectuals with Jacob Burckhardt over a period of more than half a century from 1842 to 1897. [20] [11]. In this case study, the resources of interest were of three types: Persons, Places and Works of art. Freebase has been used data source for such resources as it contained already several of them. In line with the principle of contributing to the Web of Data, rather than only consuming it, missing resources have been added to Freebase. As a result, at the time of writing this paper, scholars have added several hundreds of new entries to Freebase.org, providing basic metadata and descriptions. Pundit has been configured to use a simple set of properties, to cover the different relations that can occur among resources. These relations allow explicitly relating dates, places and persons with text in the letters. The Timeliner open-source tool[21] has been used to show dynamic visualizations built from the corpus of annotations that scholars created so far, mainly about places and persons, see Fig. 6. The visualization shows the letters in a timeline, based on their sending date. It also graphs all of the mentioned places and persons on a map, where person location is determined by their birthplace and their movements can be tracked over time.

5 Conclusions

In this paper we presented preliminary results in leveraging on structured semantic annotations to create interactions and visualization of collaboratively

[20] www.burckhardtsource.org
[21] http://timeliner.okfnlabs.org/

created data. In our examples we used Pundit: a customizable and flexible semantic web annotation tool. In deploying the tool for different use scenarios, we highlighted a simple pattern consisting of developing custom vocabularies, perhaps aggregating existing data, distributing a simple tool to annotate web resources of interest and, finally building on third party applications to consume the generated information and address specific data visualization needs.

6 Acknowledgments

The research activity underlaying this work is being partially funded by the European Union's Seventh Framework Programme managed by REA-Research Executive Agency[22] ([FP7/2007-2013][FP7/2007-2011]) under grant agreement n. 262301, and by the GramsciSource project of the MIUR, FIRB 2012, p. RBFR12MZ8R_003. Pundit was originally developed in the Semlib project[23].

References

1. C. Bizer, T. Heath, "Linked Data: Evolving the Web into a Global Data Space", http://linkeddatabook.com/editions/1.0/
2. C. Morbidoni, M. Grassi, M. Nucci,"Introducing SemLib Project: Semantic Web Tools for Digital Libraries". International Workshop on Semantic Digital Archives 15th International Conference on Theory and Practice of Digital Libraries (TPDL). 29.09.2011 in Berlin.
3. Andrews, P., Zaihrayeu, I., Pane, J., "A classification of semantic annotation systems. Semantic Web Journal". Online Available: http://www.semantic-Web-journal.net/content/classification-semantic-annotation-systems
4. B. Haslhofer, E. Momeni, M. Gay, and R. Simon, "Augmenting Europeana Content with Linked Data Resources', in 6th International Conference on Semantic Systems (I-Semantics), September 2010.
5. M. L. Ralf Heese, "One Click Annotation' in 6th Workshop on Scripting and Development for the Semantic Web, 2010.
6. G. Rockwell, S. Brown, J. Chartrand, S. Hesemeier, "CWRC-Writer: An In-Browser XML Editor' - Digital Humanities 2012 Conference Abstracts. University of Hamburg, Germany. July 1622, 2012
7. A. Gerber and J. Hunter, "Authoring, Editing and Visualizing Compound Objects for Literary Scholarship', Journal of Digital Information, vol. 11, 2010.
8. "Open Annotation: Alpha3 Data Model Guide' 15 October 2010 Eds. R. Sanderson and H. Van de Sompel. http://www.openannotation.org/spec/alpha3/
9. M. Grassi, C. Morbidoni, M. Nucci, S. Fonda, G. Ledda. "Pundit: Semantically Structured Annotations for Web Contents and Digital Libraries". Proceedings of the Second International Workshop on Semantic Digital Archives (SDA 2012).
10. M. Dörk, S. Carpendale, C. Williamson, "EdgeMaps: Visualizing Explicit and Implicit Relations". Proceedings of VDA 2011: Conference on Visualization and Data Analysis, IS&T/SPIE. 2011
11. F. Di Donato. "Working on scholarly contents: A semantic vision". In Proceedings of Open Platforms for Digital Humanities, 17-18 January 2013, 2013.

[22] http://ec.europa.eu/research/rea, DM2E Project: http://dm2e.eu/
[23] http://www.semlibproject.eu/, SemLib EU project