Curating a Document Collection via Crowdsourcing with Pundit 2.0

Christian Morbidoni[1(✉)] and Alessio Piccioli[2]

[1] Università Politecnica delle Marche, Ancona, Italy
christian.morbidoni@gmail.com
[2] NET7 Internet Open Solutions, Pisa, Italy

Abstract. Pundit 2.0 is a semantic web annotation system that supports users in creating structured data on top of web pages. Annotations in Pundit are RDF triples that users build starting from web page elements, as text or images. Annotations can be made public and developers can access and combine them into RDF knowledge graphs, while authorship of each triple is always retrievable. In this demo we showcase Pundit 2.0 and demonstrate how it can be used to enhance a digital library, by providing a data crowdsourcing platform. Pundit enables users to annotate different kind of entities and to contribute to the collaborative creation of a knowledge graph. This, in turn, refines in real-time the exploration functionalities of the library's faceted search, providing an immediate added value out of the annotation effort. Ad-hoc configurations can be used to drive specific visualisations, like the timeline-map shown in this demo.

Keywords: Semantic annotation · Linked data · Faceted browsing · Digital humanities · Pundit

1 Introduction

Digital libraries need curated semantically structured data to provide meaningful exploration and search capabilities. However, while metadata, such as document title, authors and main topics, are usually present and well curated in digital libraries, there is a great amount of knowledge hidden in texts and that could be of great value to explore a corpus. Although automatic text annotation services are available and their performances greatly improved over the last years, there is still the need for human intervention to refine extracted data and to add information than can hardly be captured by automatic tools. Pundit[1] is a semantic annotation tool that combines powerful annotation functionalities, covering comments; tagging; semi-automatic entities markup and linking; composition of rich semantic statements by interlinking items in a web page - such as text or images - and resources from the LOD or from custom annotation vocabularies. Annotations in Pundit can be made public and then accessed - via REST

---

[1] http://thepund.it.

APIS or SPARQL queries - and combined by developers to form RDF knowledge graphs. The tool adopts a flexible data model based on RDF and an extension of the Open Annotation model[2]. Pundit is the evolution of previous systems [1,2] and is designed as a configurable annotation service. It addresses online annotation communities by allowing customisation of both user interface - by activating/deactivating annotation functionalities - and annotation vocabularies, allowing community administrators to decide what properties and resources can be used in composing annotations. Domain specific annotation environments can be deployed and made available as bookmarklets or as REST services, making it easy to connect such environments to existing web sites. Pundit 2.0 has been recently released. It features a restyled graphical user interface and additional functionalities over the previous version [3,4], such as the preconfigured annotations templates that improves productivity when annotations with the same structure have to be repeatedly created. A paper describing Pundit 2.0 in detail is currently under review and is available online[3].

In this demo we show how Pundit 2.0 can enable the collaborative creation of a knowledge graph on top of a sample digital library. The demonstrative digital library showcased in this paper, allows users to freely annotate text documents, producing RDF triples and contributing in real time to the refinement of the faceted search/browsing functionalities of the library itself. Relevant entities discovered by users and annotated with Pundit are injected in the portal as facet values, thus producing immediate added value out of the annotation effort. Pundit APIs support both open crowdsourcing scenarios, where every user can annotate, and more controlled ones, where only those annotations from authorised users - or from their specific annotation collections - are imported.

2 Description of the Online Demo

The online demo can be accessed at http://purl.org/pundit/eswc2015. Different text documents from different authors in the area of philosophy and politics (including e.g. works from Antonio Gramsci[4] and correspondence from Jacob Burckhardt[5]) have been loaded into the portal to form a small sample digital library. A simple faceted browser is provided as the main exploration mean. Some of the facets reflect structural features of the documents, such as their provider or language. Other facets show different kind of entities that the documents talk about (e.g. persons, places). When users open a document they can perform different annotation tasks. The easiest one is probably that of marking relevant entities appearing in the text. By selecting the *suggestions* mode in the annotation side bar and clicking on *scan page*, users get automatic suggestions and can review results by approving or rejecting annotations. This feature is powered by DataTXT[6], an entity linking service based on the TAGME algorithm [5].

---

[2] http://www.openannotation.org/spec/core/.
[3] http://www.semantic-web-journal.net/content/pundit-20.
[4] http://dl.gramsciproject.org.
[5] http://burckhardtsource.org.
[6] https://dandelion.eu/products/datatxt/nex/demo.
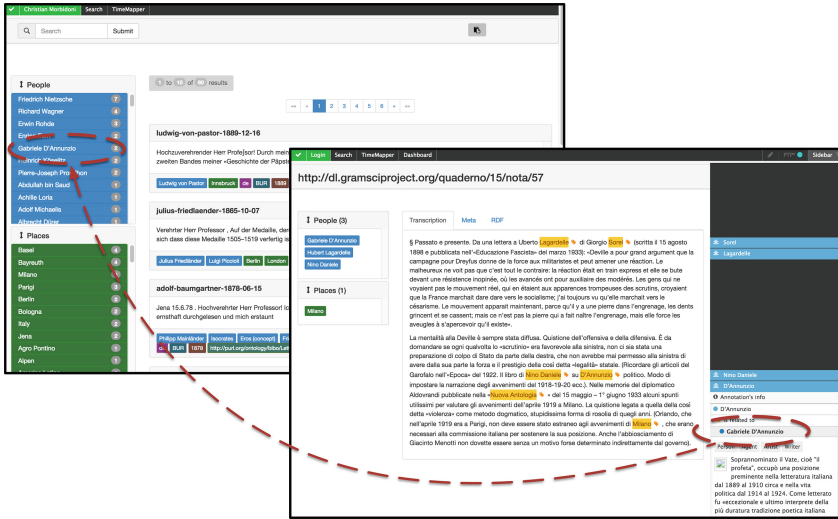
Fig. 1. A screenshot of the demo navigation portal. The facets browser (left) is populated from users annotations made on the text (right).

By annotating the documents, the user actively contribute to improve the browsing experience. For example, each time a user annotates an entity of type *place* or *person* this results in a new value appearing in the respective facet (Fig. 1). Note that for the purpose of this demo all users are entitled to do so and no quality check is performed. However, in a real world setting it would be possible to take into account annotations from trusted users only. An other way of adding annotations is by using the *annotation composer*. This allows expert users to freely compose triples. To do so, select a text in the page and choose *use as subject*': the triple composer will appear. Complete the statement by choosing an object (e.g. searching on DBpedia or in custom vocabularies) and a predicate among the proposed ones (Fig. 2).

An alternative way of exploring the document collection is the one based on TimeMapper[7], an open-source tool developed by the Open Knowledge Labs[8]. This is an example of a possible specialised view on the document collection, where texts excerpts that identify a precise event are shown in a timeline along with the main person involved in the event and the place where it occurred. To contribute to such a view, users are required to create annotations with a precise structure. While it is possible to use the annotation composer to edit such annotations, a preconfigured annotations template is provided to make the task easier and to avoid errors. To activate the *template mode*, click on the pencil icon in the Pundit top bar and then choose the *PTP* template from the dropdown menu. Once this is activated, every time the user selects a text excerpt

---

[7] http://timemapper.okfnlabs.org/.
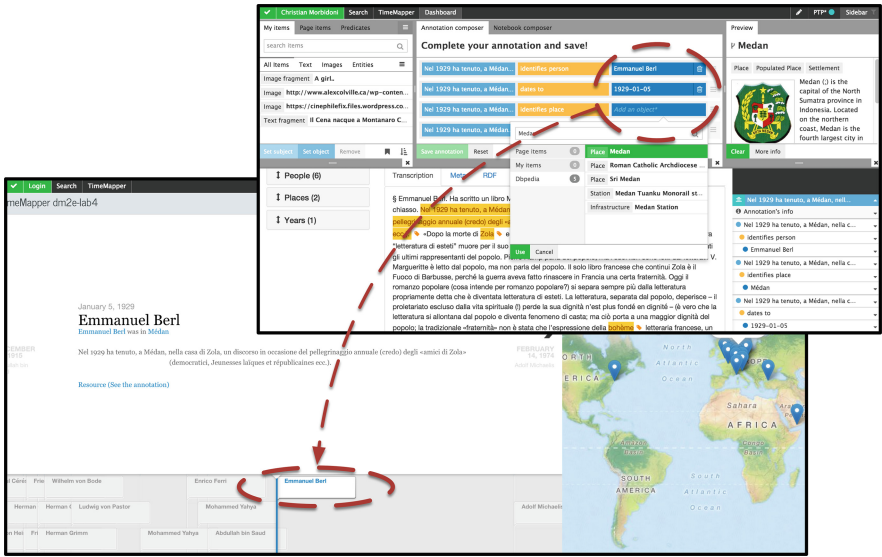[8] http://okfnlabs.org/.

Fig. 2. An annotation created with the PTP template shows up as a new slide in the timeline view.

the annotation composer will be automatically populated with three statements. Users will then have to *fill the blanks*, by searching DBpedia or Freebase for the appropriate person and place and by entering a date. On save the annotation will be shown in the sidebar and a new item will appear in the timeline, as illustrated in Fig. 2.

3   Related Works

Semantic annotation systems have been reviewed and compared in literature [6] and formal models have been developed for annotations and annotation systems [3]. In this section we shortly mention some of the annotation systems we think are more related to Pundit. Annotea [8] is one of the first systems to implement RDF based annotation, providing both client and server APIs for storing structured data. Its semantic capabilities were limited to Dublin code fields. LORE (Literature Object Reuse and Exchange) [9] is a semantic annotation system providing a Mozilla plugin to annotate content. It implements the concept of compound object which is similar to semantic annotation in Pundit. Compound objects are basically set of inter-connected resources and can be linked to web content and created with a visual graph UI. Domeo [10] implements ontology-based annotation metadata on HTML or XML document targets, using the Annotation Ontology (AO) RDF model. Semantic Turkey [11] allows to capture knowledge from web pages by associating it to reference ontologies. To our knowledge, Pundit is the first annotation tool that combines:

– Semi-automatic linking of entities in text
– Configurable semantic annotation templates, allowing to create complex annotations in few steps
– Free composition of triples to link elements in web document (e.g. words, images, images parts) to LOD entities and among each other
– Configurable annotation vocabularies of entities and relations to be used in triples
– Delivery of annotations as RDF graphs via SPARQL or via open or authenticated REST API
– Delivery of the annotation environment as-a-service, so that web applications can make their content annotatable by calling a REST API (feed.thepund.it), as well as as a bookmarklet, or simply as a javascript library.

References

1. Tummarello, G., Morbidoni, C.: Collaboratively building structured knowledge with DBin: from del.icio.us tags to an RDFS Folksonomy. In: Workshop on Social and Collaborative Construction of Structured Knowledge, CKC 2007, International World Wide Web Conference, WWW 2007, Banff, AB, Canada (2007)
2. Tummarello, G., Morbidoni, C.: The DBin platform: a complete environment for semantic web communities. J. Web Semant. **6**(4), 257–265 (2008)
3. Grassi, M., Morbidoni, C., Nucci, M., Fonda, S., Piazza, F.: Pundit: augmenting web contents with semantics. Literary Linguis. Comput. **28**, 640–659 (2013)
4. Morbidoni, C., Grassi, M., Nucci, M., Fonda, S., Ledda, G.: Introducing the semlib project: semantic web tools for digital libraries. In: 1st International Workshop on Semantic Digital Archives, SDA 2011, Berlin, Germany (2011)
5. Ferragina, P., Scaiella, U.: TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, New York (2010)
6. Andrews, P., Zaihrayeu, I., Pane, J.: A classification of semantic annotation systems. Semant. Web J. **3**(3), 223–248 (2012). http://www.semantic-Web-journal.net/content/classification-semantic-annotation-systems
7. Agosti, M., Ferro, N.: A formal model of annotations of digital content. ACM Trans. Inf. Syst. (TOIS) **26**(1), 3 (2007). TOIS Homepage archive
8. Kahan, J., Koivunen, M.R.: Annotea: an open RDF infra-structure for shared web annotations. In: Proceedings of the 10th International Conference on World Wide Web (2001)
9. Gerber, A., Hunter, J.: Authoring, editing and visualizing compound objects for literary scholarship. J. Digit. Inf. **11**(1) (2010)
10. Ciccarese, P., Ocana, M., Clark, T.: DOMEO: a web-based tool for semantic annotation of online documents. In: Bio-Ontologies 2011 (2012)
11. Pazienza, M.T., Scarpato, N., Stellato, A., Turbati, A.: Semantic turkey: a browser-integrated environment for knowledge acquisition and management. Semant. Web J. **3**(3), 279–292 (2012)