Anna: a Virtual Assistant to interact with Puglia Digital Library (Discussion Paper)

Vito Walter Anelli[1], Tommaso Di Noia[1], Eugenio Di Sciascio[1], and Azzurra Ragone[2]

[1] Polytechnic University of Bari, Bari, Italy
[2] Independent researcher, Milan, Italy
{name.surname}@poliba.it,azzurraragone@gmail.com

Abstract. In the last years, a huge amount of data has been released by private and public bodies as Linked Open Data. By their inner nature, these data contain rich semantic information that can be automatically processed by software agents and explored by humans via visual tools or structured SPARQL queries. Although they result useful for many tasks, these latter approaches miss the simplicity of the interfaces based on interactions via natural language implemented in modern virtual assistants. In this paper, we present a system able to assist the user in exploring the knowledge exposed by the *Puglia Digital Library* containing information and data associated with digital goods related to the Apulia region in Italy. We show how to interact with the Digital Library by means of a virtual assistant and how, thanks to its publication as Linked Open Data, it is possible to easily integrate it on-the-fly with external knowledge sources such as geographical ones.

Keywords: Linked Open Data, Chatbot, Vocal Assistant, Digital Libraries

1 Introduction

In 2012, Google announced its Knowledge Graph[3] as a new tool to improve the identification and retrieval of entities in return to a search query. Most of the knowledge encoded in Google Knowledge Graph originally came from Freebase which was a crowd-sourced effort to create a base of facts in all possible knowledge domains. Alongside with the development of the above-mentioned initiatives, following the original idea of a Semantic Web [4], new technologies have been developed and released with the aim of embedding structured knowledge with unambiguous semantics into Web pages in order to allow software agents to consume and elaborate information in an automated way. The original idea has been modified over the years thus making possible the creation of a full stack of semantic technologies and, more remarkably, gave birth to the

[3] https://googleblog.blogspot.it/2012/05/introducing-knowledge-graph-things-not.html

Linking Open Data initiative[4] where a community of researchers and practitioners devoted an enormous effort to build publicly available knowledge bases of machine-understandable data. By exploiting Semantic Web technologies, a knowledge base is then represented through a graph (knowledge graph) in which entities are linked to each other by binary relationships. The Linked Data initiative deserves credit for setting best practices[10] for releasing *Open Data* on the Web, and for linking different open, and structured datasets. Linked Data practices are particularly well-suited to associate high quality meta-data to resources. This is a classic issue [17] of Digital Libraries (DLs), which are complex information systems shouldering the preservation of digital documents, intellectual property right management, information filtering, information retrieval, and query answering [8]. Linked Data technologies let DLs realize a fully data-centric, and metadata-based approach in describing their resources [2]. Meta-data and thesauri are already commonly used in DLs' resources descriptions. However, linking these meta-data to Linked Open Data cloud[5] improves reusability and integration of DLs [5]. In this direction some successful examples are the German National Library[6], the British National Library[7], the National Library of Spain[8], the National Library of France[9]. Linked meta-data comes with many benefits [16] as easy cataloging, information discovery, and the possibility to build third-party applications to explore data.

Despite the efforts, technical issues prevent citizens to consume *Open Data*. In most cases *Open Data* are consumed through data-visualization tools, and tables, built upon the released data. The drawback of this approach is that, within a large catalog of items (the default case with DLs), hardly citizens will be able to discover new resources they could be interested in.

In this paper, we report the experience in making accessible the *Puglia Digital Library*[10] (PDL) through **Anna**: a virtual assistant able to explore the knowledge graph behind PDL via natural language-based interactions. In the proposed approach, speech recognition is delegated to Google Assistant, whereas Dialogflow[11] is adopted to manage the interaction with the user.

2 Related work

Chatbots are software agents specifically designed to make the interaction with the user as natural as possible, usually providing a textual and/or vocal interface to gives the user the impression of speaking to a human. This feeling was suggested by [1] who anticipated the key role of chatbots in humanizing machines. A chatbot copes with the natural way of interacting of the user, defining a new category of interfaces, Conversational User Interfaces (CUIs) [13]. The design of a chatbot involves many non-trivial operations that include NLP, pattern

[4] `http://linkeddata.org`

[5] `https://lod-cloud.net/`

[6] `http://www.dnb.de/EN/`

[7] `http://www.bl.uk/bibliograpic/datafree.html/`

[8] `http://datos.bne.es/`

[9] `http://data.bnf.fr`

[10] `http://www.pugliadigitallibrary.it/`

[11] `https://dialogflow.com/`

matching, parsing, artificial intelligence, and ontologies. In [1] a logical scheme is depicted: responder (interface), classifier (transformation), and graphmaster (knowledge extraction). The responder corresponds to the interface that is exposed to the user.

In the last years, DLs gave birth to a flourishing research field, in which a notable number of vocabularies, metadata standards, thesauri, have been designed. The number of collective and small initiatives and DLs [8] that have been put in place make almost impossible to depict the overall field. In 2000s several initiatives were put in place with the aim of managing and organizing DLs, like for instance, a global library cooperative OCLC[12], which has thousands of library members and offers services to support them. On the other side, Europeana[13] [9] is an aggregator of meta-data about more than 58 millions digital resources. Europeana represents all these digital objects through a common format and schema [5], Europeana Data Model[14] (EDM) bases on Semantic Web Languages. The Europeana data model respects these standards and ensures consistency and interoperability even though sometimes the expressiveness of original data is lost [5]. The standardization of meta-data is absolutely not a new topic in Digital Libraries field in which different organization tried to define rigorous cataloguing principles [2] like AACR, MARC[15], ISBD[16] [19], FRBR[17] [20], RDA [6]. The need for DLs to make their content freely accessible, sharable, and re-usable led generally to the adoption of `Linked Open Data`. Pioneers were the Library of the Congress, and Staford University Library, followed by Europeana and British Library [9, ?]. The different DLs take advantage of common `Linked Open Data` vocabularies that fit the need of describing the specific resources they host. In the case of digital resources like images, video, web pages, the most generally adopted vocabulary is definitely Dublin Core Metadata [12] by DCMI which is usually adopted to ensure interoperability between the different metadata vocabularies. Although there is an increasing number of chatbots that retrieve information from relational databases, the usage of *Open Data* as a knowledge source for chatbots, is an almost unexplored research direction. In [14], a question answering system is built using a question generation framework [11]. The knowledge is extracted as plain text from PDF documents using OCR techniques, whereas the matching patterns are defined using Artificial Intelligence Markup Language (AIML). OntBot [18] takes advantage of a mapping between an ontology and the knowledge stored in a relational database. Finally in [15], the authors design a chatbot to query a `Linked Open Data` dataset released by the Italian Ministry of Transport. They took advantage of IBM Bluemix and Watson Conversation to realize their query answering system.

[12] `https://www.oclc.org/`

[13] `https://www.europeana.eu/`

[14] `https://pro.europeana.eu/resources/standardization-tools/edm-documentation`

[15] `https://www.loc.gov/standards/marcxml/`

[16] `https://www.ifla.org/files/assets/cataloguing/isbd/isbd-cons_20110321.pdf`

[17] `https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf`

3 Puglia Digital Library

Puglia Digital Library is a multimedia archive of books, magazines, newspapers, photographs, sounds, videos. The aim of *Puglia Digital Library* is to preserve and to share the regional digital heritage. *Puglia Digital Library* provides an online platform to store and expose its data. The resources are accessible through a Web Portal that lets the user download the resource, consume it (preserving the integrity of fragile museum resources), and read a description. From 2016 *Puglia Digital Library* became a producer of `Linked Open Data` [7] concerning museum objects, historic, and artistical sites. In order to favor interoperability, *Puglia Digital Library* adopted several sets of meta-data[18] and controlled vocabularies, such as: DCMIType (DCMI Type Vocabulary), PICO (Thesaurus that allow interoperability with Cultura Italia), Vocabularies ICCD (Italian standard for cataloging), AAT (Art & Architecture Thesaurus), TGN (Thesaurus of Geographic Names). The ontology chosen for resources description is CIDOC-CRM[19] (ICOM[20]). However, in order to overcome some representational limitations of the above mentioned ontologies and to enrich the linking with external Linked Data datasets, other ontologies were adopted: Dublin Core[21], DBpedia[22] ,Schema.org[23] , Foaf[24] , SKOS[25], LinkedGeoData[26] and Geonames[27].

4 Anna's Architecture

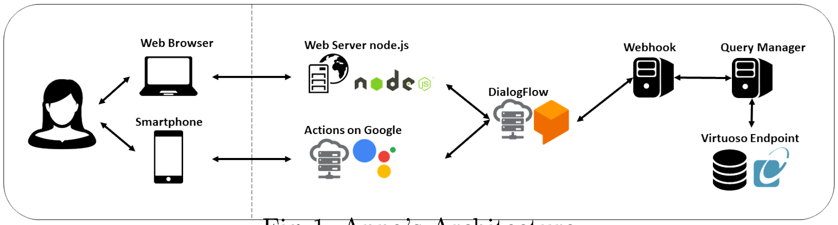

Fig. 1: Anna's Architecture

The software ecosystem behind the chatbot Anna is depicted in Figure 1. The first step to create Anna consisted of creating a SPARQL endpoint to retrieve information from the PDL knowledge base [7] in an effective way. A Virtuoso server instance was created with a SPARQL 1.1 endpoint that allows humans,

[18] DC, DDI, EAC-CPF, EAD, FGDC, ISO 19115 2003, LC-AV, LOM, MARC, MAG, METSRIGHTS, MODS, NISOIMG, PREMIS, TEIHDR, TEXTMD, VRA

[19] `http://www.cidoc-crm.org/`

[20] `http://icom.museum/`

[21] `http://dublincore.org/`

[22] `http://dbpedia.org`

[23] `http://schema.org`

[24] `http://www.foaf-project.org/`

[25] `https://www.w3.org/2004/02/skos/`

[26] `http://linkedgeodata.org/`

[27] `http://www.geonames.org/ontology`

and automated agents to query the underlying Knowledge Graph using SPARQL queries. Differently from SQL queries, SPARQL is based on the idea of graph matching. SPARQL query language lets the user express a graph pattern (in which one or more parts of the triple are substituted by variables), to retrieve the necessary knowledge.

The user can interact with the system using a Web browser or a mobile phone with the Italian Google Assistant. The Web App was implemented in `node.js` and deployed on a specific server. The Google Assistant is an *Actions on Google* project that enables users to interact with *DialogFlow*, converting the speech to text. The Web App copes with the communication protocol adopted by Google Assistant to provide a unified communication protocol. The DialogFlow module handles the conversation itself: each interaction moves the conversation from a state to another, called contexts, in which a specific choice is made by the user (intent). Contexts, and intents are at the core of the interaction mechanism and are depicted in Figure 2. The Webhook keeps the history of the iterative queries of the user, in order to refine the final query. If any further knowledge is required, the Webhook sends a message with the needed information to the Query Manager, which poses the query against the remote SPARQL endpoint. Finally, for the sake of privacy, the Webhook itself does not store any information, and once the session is terminated that specific user history is deleted. In Figure 2 the most important intents are depicted. The directed rows show the outgoing contexts. Whenever the user operates a choice, her input is analyzed and the conversation is moved to a new context. Intents basically defines the usability of

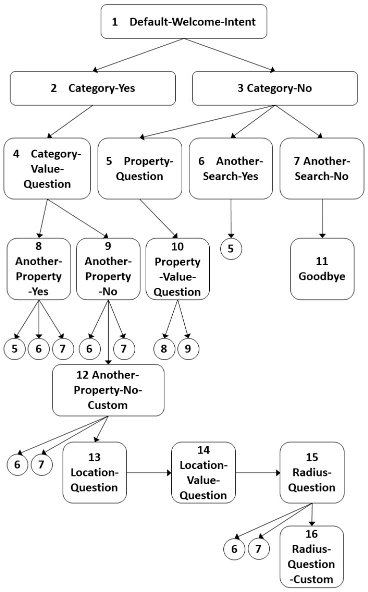

Fig. 2: Intents' Graph

the tool, hence the intents'scheme was defined under the instructions of *Puglia Digital Library* users, and managers.

4.1 A typical conversation with Anna

In order to understand the subtle interactions of the Anna ecosystem, a very effective way is to explore a real conversation. The user invokes the chatbot saying "*Talk with Digital Library*". Google Assistant opens a connection with the chatbot which greets to the user and asks if she is interested in a particular topic (Fig. 2 - Intent 1, see also `https://goo.gl/LjFZd2`). Let us suppose she answers *Yes*. This answer is sent from the chatbot to the Webhook, and then to the Query Manager. It sends a query against the Virtuoso Endpoint looking for all the possible topics for the resources stored in the knowledge base. The Query Manager extracts the results from the server's response and the Webhook prepares and sends the results to DialogFlow. At this point, Anna shows a list of available topics (Fig. 2 - Intent 2, see also `https://goo.gl/UoP1US`). User chooses "*History and Traditions*". This information is sent to the Webhook to update the constraint for this search session (Topic:*History and Traditions*). Anna asks if the user wants to choose other properties (Fig. 2 - Intent 4, see also `https://goo.gl/b3eoe9`). User says *Yes*. Anna sends this information to the Webhook, that sends all the session data to the Query Manager. This will compose a query that looks for all the properties that resources (with *"History and Traditions"* as the topic) have. Anna returns a list of options (Fig. 2 - Intent 8, see also `https://goo.gl/gzi513`) composed by "Subject", "Category", "Location". User chooses *"Subject"*. This information is sent to the Query Manager, that will look for all the possible values for the predicate *"Subject"*, considering only those resources that have *"History and Traditions"* as the topic. The results are returned to DialogFlow. Now Anna will show a long list of available topics (Fig. 2 - Intent 5, see also `https://goo.gl/fpMBFK`). User chooses "*Sanctuaries*". This information is sent to the Webhook, which stores a new association, Subject:Sanctuaries. Anna asks the user for more properties (Fig. 2 - Intent 10, see also `https://goo.gl/n1pSc5`). User says *No*. This is the signal the chatbot awaits to prepare the query for the resources. All the available information (*Topic*:*History and Traditions*; *Subject*:*Sanctuaries*) are sent to the Query Manager, which composes the query and sends it against the SPARQL endpoint. The query requires also additional resources' information: label with full name, url address of the resource, url address of the preview image, description of the resource, latitude, and longitude. This information is parsed and stored in a structured form associated to the session by the Webhook. The Webhook also prepares the carousel and sends it back to the user. Now Anna will return the carousel(Fig. 2 - Intent 9, see also `https://goo.gl/Rrjpz1`) containing: *"The Ripalta sanctuary"*, and *"Santissima Maria di Ripalta"*. User selects the first result. The Webhook receives the input of the user and prepares a card about that resource. Anna will return a Card (Fig. 2 - Intent 12, see also `https://goo.gl/vsdC9e`) containing a preview of the digital resource, the name of the selected resource and a complete description. At the end of the card (see `https://goo.gl/rfNB2x`) the user will find a link to the *Puglia Digital Library* corresponding page, with some buttons related to *"Nearest places"*, *"Another search"*, and *"Thank you"*. User selects *"Nearest places"*. Anna sends the request to the Webhook, which extracts and returns the list of considered locations' categories. Anna will return the list (Fig. 2 - Intent 13, see also `https://goo.gl/4jBeuY`): *Hotels*, *Restaurants*, *Cafes*, *Museums*. User is interested in *Restaurants*. In order to prepare a geospatial query Anna asks the user for the distance to be considered (Fig. 2 - Intent 14, see also `https://goo.gl/o4QbCa`). The user answers *"10 km"*, which is sent to the Query Manager to prepare the a special query which is sent against the LOSM [3] endpoint to retrieve this geo-referenced information in real-time. LOSM returns a list of available locations, within the considered radius, together with their coordinates. This information is used to create a new carousel. This carousel, with a list of available places (Fig. 2 - Intent 15, see also `https://goo.gl/rAJ7EV`), is returned to the user. The user chooses the *"Di Muzio"* Restaurant. Once again, the Webhook takes the input sent by the user, extracts the coordinates of the Restaurant, and prepares a card with a valid link to get directions with Google Maps. Thus Anna returns a card (Fig. 2- Intent 16, see also `https://goo.gl/hr6aBg`) with a direct link to navigate towards the Restaurant using Google Maps. At the end of this interaction the user can quit the conversation or begin a new one.

5 Conclusion

In this work, we presented Anna, a vocal assistant built upon the Linked Open Data datasets of *Puglia Digital Library*. The interface was implemented in two different ways: a web application, with a textual interaction, and through Actions on Google, which provides a mixed textual/vocal interaction. The vocal/textual interface is an innovative way of accessing *Open Data* that eases the technological barriers of modern Semantic Web technologies. A common issue for broadcasting these disruptive technologies is the lack of well-established NLP tools for languages different from English. Taking advantage of speech recognition APIs we were able to design the Assistant to be consumed by Italian citizens (the audience of *Puglia Digital Library*) in a natural way. Future work includes the creation of new vocal assistants to provide information for other digital libraries. Moreover, we want to evaluate extensively the usability of the assistant to improve the overall user experience.

Acknowledgments. The authors acknowledge partial support of the following projects: Innonetwork APOLLON, Innonetwork CONTACT, PON FLET 4.0.

References

1. Abdul-Kader, S.A., Woods, J.: Survey on chatbot design techniques in speech conversation systems. International Journal of Advanced Computer Science and Applications **6**(7) (2015).
2. Alemu, G., Stevens, B., Ross, P., Chandler, J.: Linked data for libraries: Benefits of a conceptual shift from library-specific record structures to rdf-based data models. New Library World **113** (01 2012).
3. Anelli, V.W., Calì, A., Di Noia, T., Palmonari, M., Ragone, A.: Exposing open street map in the linked data cloud. In: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. pp. 344–355. Springer (2016)
4. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific american **284**(5), 34–43 (2001)

5. de Boer, V., Wielemaker, J., van Gent, J., Hildebrand, M., Isaac, A., van Ossenbruggen, J., Schreiber, G.: Supporting linked data production for cultural heritage institutes: The amsterdam museum case study. In: Simperl, E., Cimiano, P., Polleres, A., Corcho, Ó., Presutti, V. (eds.) The Semantic Web: Research and Applications - 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27-31, 2012. Proceedings. Lecture Notes in Computer Science, vol. 7295, pp. 733–747. Springer (2012).
6. Coyle, K., Hillmann, D.: Resource description and access (RDA): cataloging rules for the 20th century. D-Lib Magazine **13**(1/2) (2007).
7. Di Noia, T., Ragone, A., Maurino, A., Mongiello, M., Marzocca, M.P., Cultrera, G., Bruno, M.P.: Linking data in digital libraries: the case of puglia digital library. In: Proceedings of the 1st Workshop on Humanities in the Semantic Web co-located with 13th ESWC Conference 2016 (ESWC 2016). pp. 27–38 (2016)
8. Fox, E.A., Marchionini, G.: Toward a worldwide digital library - introduction. Commun. ACM **41**(4), 28–32 (1998).
9. Gradmann, S.: Knowledge = information in context: on the importance of semantic contextualisation in europeana pp. 1–19 (01 2010)
10. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web, Morgan & Claypool Publishers (2011).
11. Heilman, M., Smith, N.A.: Question generation via overgenerating transformations and ranking. Tech. rep., CARNEGIE-MELLON UNIV PITTSBURGH PA LANGUAGE TECHNOLOGIES INST (2009)
12. Initiative, D.C.M., et al.: Dublin core metadata element set, version 1.1 (2012)
13. McTear, M.F.: Spoken dialogue technology - toward the conversational user interface. Springer (2004),
14. Pichponreay, L., Kim, J.H., Choi, C.H., Lee, K.H., Cho, W.S.: Smart answering chatbot based on ocr and overgenerating transformations and ranking. In: Ubiquitous and Future Networks (ICUFN), 2016 Eighth International Conference on. pp. 1002–1005. IEEE (2016)
15. Porreca, S., Leotta, F., Mecella, M., Vassos, S., Catarci, T.: Accessing government open data through chatbots. In: Garrigós, I., Wimmer, M. (eds.) Current Trends in Web Engineering - ICWE 2017 International Workshops, Liquid Multi-Device Software and EnWoT, practi-O-web, NLPIT, SoWeMine, Rome, Italy, June 5-8, 2017, Revised Selected Papers. Lecture Notes in Computer Science, vol. 10544, pp. 156–165. Springer (2017).
16. Southwick, S.B., Lampert, C.K., Southwick, R.: Preparing controlled vocabularies for linked data: Benefits and challenges. Journal of Library Metadata **15**(3-4), 177–190 (2015).
17. Tani, A., Candela, L., Castelli, D.: Dealing with metadata quality: The legacy of digital library efforts. Inf. Process. Manage. **49**(6), 1194–1205 (2013).
18. Vegesna, A., Jain, P., Porwal, D.: Ontology based chatbot (for e-commerce website). International Journal of Computer Applications **179**(14), 51–55 (Jan 2018).
19. Willer, M., Dunsire, G., Bosancic, B.: Isbd and the semantic web. JLIS. It. Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione **1**(2), 213–236 (2010).
20. Žumer, M.: Functional requirements for bibliographic records: Frbr: The end of the road or a new beginning. Bulletin of the American Society for Information Science and Technology.