BotDCAT-AP: An Extension of the DCAT Application Profile for Describing Datasets for Chatbot Systems

Paolo Cappello, Marco Comerio and Irene Celino
Cefriel – Politecnico di Milano
via Fucini, 2 – Milano, Italy
E-mail: {paolo.cappello,marco.comerio,irene.celino}@cefriel.com

Abstract. Although it is still an emerging technology, the increasing usage of chatbots (also known as bots) has opened a promising touchpoint for citizen and customer engagement. A chatbot consists of a computer program aimed at simulating a conversation between humans and machines through the formulation of appropriate answers making use of external knowledge. Therefore, managing external knowledge is a crucial task for the design and development of chatbots. To facilitate the reuse of existing data sources in chatbot applications, in this paper we propose BotDCAT-AP, an extension of the Data Catalogue (DCAT) Application Profile for describing datasets for chatbots. BotDCAT-AP enables the description of *intents* (i.e., the actions users want to accomplish by interacting with a chatbot) and *entities* (i.e., individual information units associated to an intent) supported by a dataset and the *method to access* it. A practical usage of BotDCAT-AP is shown to demonstrate the value of its adoption.

1 Introduction

The W3C's Data Catalogue vocabulary (DCAT) [1] is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web. By using DCAT to describe datasets in data catalogs, publishers increase discoverability and enable applications to easily consume metadata from multiple catalogs. The DCAT Application Profile for data portals in Europe (DCAT-AP) [2] is a specification based on DCAT for describing public sector datasets in Europe. Its basic use case is to enable searches for a dataset across data portals and improve the sharing of public sector data.

Several extensions of DCAT-AP are emerging: they focus on specific types of datasets and use cases. Examples are GeoDCAT-AP [3] that makes geospatial information better searchable across borders and sectors and StatDCAT-AP [4] that provides a commonly agreed dissemination vocabulary for statistical open data.

In this paper, we propose BotDCAT-AP, an extension of DCAT-AP to describe datasets for chatbots systems. Since such systems make use of external knowledge to simulate a conversation between humans and machines, BotDCAT-AP aims at simplifying the creation of the software components managing the external knowledge.

2 Motivation

Chatbots (like AzureBot[1] or Herzi[2]) get user requests as natural language questions through different input *channels* (e.g., Instant Messaging (IM) applications, social networks). They process requests with Natural Language Understanding (NLU) engines: user questions are translated into machine understandable actions, because NLU engines are capable of interpreting users' input (*utterances*) by extracting the intent of every single request and the possible entities contained in it. The *intents* represent what the users wish to accomplish using the chatbot. The *entities* are domain specific information items extracted from the user's utterance that help in understanding the intent. The utterance in natural language is first analyzed for the intent and entities by the NLU engine and then mapped to a specific action that should be performed (e.g., access a specific *dataset* through an API) as well as the specific dialog to be returned by the chatbot. As described in [5], NLU engines are often complex, using various Natural Language Processing (NLP) models and Machine Learning techniques to provide acceptable levels of accuracy (e.g., Microsoft LUIS[3], Google API.ai[4] and Facebook Wit.ai[5]). To train NLU engines, a training set of sample utterances is used in order to support the system at run-time to correctly associate other new and unseen utterances to the correct intents and extract the relevant entities.

Let us consider a chatbot providing weather forecast. This chatbot is able to interpret utterances like "tell me the weather in Milan", "what are the weather forecasts for tomorrow?", "will it rain this weekend?". All of them are associated with the intent "get weather". Furthermore, the NLU engine extracts the entities "Milan", "tomorrow" and "weekend" that help in further understanding the intent and characterize the action to perform (querying the weather forecast data source to get information about a specific location and time frame).

Fig. *1* shows the general structure of a chatbot: even when relying on existing frameworks providing channels and NLU engines, custom development is required to create the *wrapper* that connects the chatbot components to the knowledge sources, i.e. the *datasets* (API, data dump, linked data, etc.). This wrapper is used at *design-time* to train the NLU engine to correctly identify intents and entities, and at *run-time* to retrieve the necessary information from knowledge sources to answer user questions.

To ease the development of such wrapper components, we introduce an enriched semantic description of knowledge sources with respect to our BotDCAT-AP vocabulary: this description includes the information about intents and entities supported by the available datasets and their access methods. The availability of such a description can be used to standardize the wrapper development (adding value for the chatbot developer) and to enable the reuse of datasets by multiple chatbot systems (adding value for the dataset owner). Referring to the state of the art, the proposed vocabulary does

[1] https://microsoft.github.io/AzureBot/
[2] https://devpost.com/software/herzi
[3] https://docs.microsoft.com/en-us/azure/cognitive-services/luis/home
[4] https://docs.api.ai/docs
[5] https://wit.ai/docs

not aim at overcoming open challenges for Semantic Question Answering (SQA) systems [6], which mainly deal with the internals of NLU engines, but it aims at improving the sharing of datasets useful for those systems.

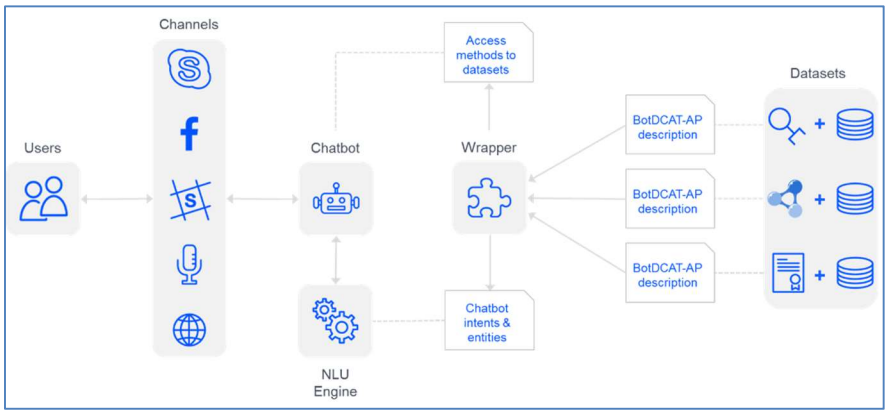

Fig. 1: How BotDCAT-AP simplifies the creation and execution of chatbot systems.

3 BotDCAT-AP

BotDCAT-AP is an RDF vocabulary, denoted by the prefix *bot* in the following and openly accessible at `http://swa.cefriel.it/ontologies/botdcat-ap`, released with a CC-BY-4.0 license. BotDCAT-AP is also listed on Linked Open Vocabularies at `http://lov.okfn.org/dataset/lov/vocabs/bot`.

The vocabulary was developed starting from the Data Catalogue vocabulary (DCAT) and its Application Profile (DCAT-AP) elaborated by a Working Group under the ISA Programme of the European Commission. BotDCAT-AP is meant to be an extension of DCAT-AP and follows all its conformance statements. The necessity of a sound and solid basis for describing the datasets is needed to deliver an easily adaptable solution with reference to a well-designed standard.

A simplified UML Class diagram of BotDCAT-AP is depicted in Fig. 2, where additions to the main classes and properties of DCAT-AP are highlighted. The *bot:Intent* class is designed to represent any possible intent supported by a dataset. The relation *bot:hasEntitiesList* connects an intent to a list of supported entities enclosed in an instance of the class *bot:EntitiesCatalog*. Entities can be represented in different ways since BotDCAT-AP allows both standard and ad-hoc entities to be specified. A first case is covered by the relation *bot:hasEntity* that is used to relate an intent to entities already specified in external ontologies. A practical use case could be a date defined in the OWL-Time ontology, or a point-of-interest (POI) in the LinkedGeoData ontology [7]. The generic *owl:Class* is used to allow the possibility to refer any concept defined in external ontologies.

The *bot:hasEntityConcept* and *bot:hasEntityDataset* relations cover the other two cases where entities are context-related and an external ontology covering such entities

is missing. The first relation targets the *skos:Concept* class and it is used when the set of possible entities is limited and there are hierarchies among them; in this case, entities can be directly added to the BotDCAT-AP description as a SKOS taxonomy. Otherwise, the Entity Catalog can be linked through *bot:hasEntityDataset* to a dataset enumerating all the possible entities. A dataset is represented as an instance of the class *dcat:Dataset*, and can optionally have multiple distributions denoted by *dcat:Distribution* accessible through a reference exposed by the relation *dcat:accessURL*.

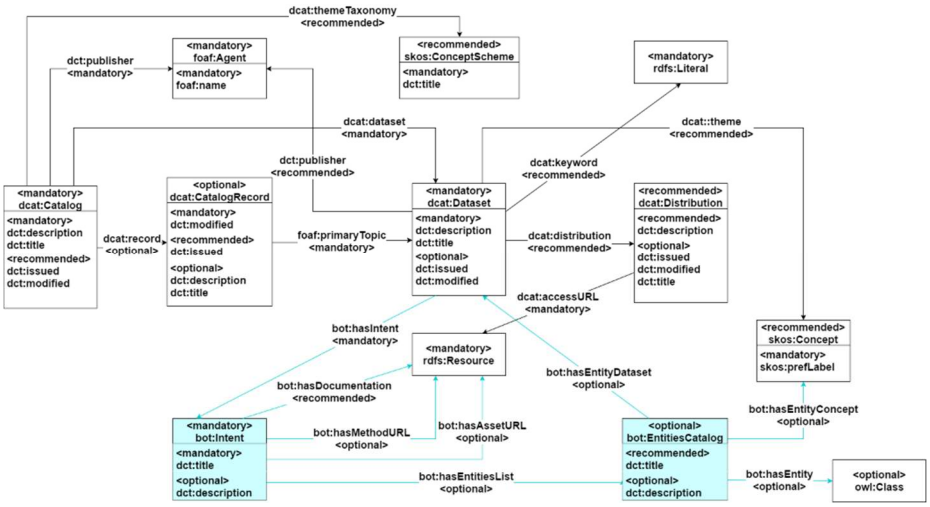

Fig. 2. BotDCAT-AP simplified UML Class diagram

As of today, DCAT-AP supports only the description of data catalogs and datasets published on the web; BotDCAT-AP overcomes this limitation giving the possibility to also define different access methods to a particular dataset [8]. This is done through the use of the relations *bot:hasMethodURL*, *bot:hasAssetURL* and *bot:hasDocumentation*, corresponding respectively to access points offered by a simple REST API, a SPARQL endpoint or any other documented method (e.g., a SOAP-based web service documented by a WSDL file). This extension supports the delivery of information that improves and speeds up the creation of the application logic needed by the chatbot system to operate at run-time.

4 Use Case

The main purpose of BotDCAT-AP is to facilitate the implementation of chatbots by providing a formal description of all the external datasets containing useful information. In the following, we explain how we adopted the proposed vocabulary to describe the data sources exploited by a bot application to provide information to final users. Additional information on BotDCAT-AP and the full versions of the dataset descriptions in RDF can be found at `http://swa.cefriel.it/bot/profiles2017_botdcat-ap.html`.

Talkin'Piazza[6] [9] is a web-based application, developed in the Piazza project[7], that aims to engage the urban community on the go to participate to the city life. Among its functionalities, Talkin'Piazza offers a bot that can be queried to get information about city events, points of interest and public transport; to reply to citizens' questions, the Talkin'Piazza bot accesses public, external and heterogeneous data sources, described with BotDCAT-AP to ease the wrapper development.

The "Milano Events" dataset contains a list of events that take place in the city of Milan. By accessing this dataset, the Talkin'Piazza chatbot is capable of replying to the user's intent proposing events filtered by category, location and price. The chatbot system can access the dataset by means of a Web API, whose reference URL is contained in the BotDCAT-AP description at `http://swa.cefriel.it/examples/botdcat-ap/Events.ttl`. Since an event is usually associated to a category stating its thematic area (e.g., sport, art, entertainment, education) and to a type of admission (e.g., free entrance, paid entrance), such units of information can be expressed by users in their utterances. The EntityCatalogs *EventsAdmissions* and *EventsCategories* contain entities associated to possible types of admission and thematic areas of the events. *EventsAdmissions* contains only the entities *FreeEntrance* and *PaidEntrance* and therefore they are simply defined as *skos:Concept*(s). The same approach would not be practical for *EventsCategories* since those entities are wide and dynamic. In this case, a reference to an external dataset *CategoriesDataset* containing the list of all the possible categories is used. In this way, the *CategoriesDataset* can be easily changed and updated without modifying the BotDCAT-AP description associated to the "Milano Events" dataset.

Talkin'Piazza is able to provide the user with information about POIs all over the city by accessing relevant data from OpenStreetMap[8]. The chatbot is trained to respond to utterances such as "where can I find an ATM?", "I'd like to know the location of the restaurants near me", "can you show me the nearest library?" and to assign them to the intent *GetPOIs*. The OpenStreetMap profile based on BotDCAT-AP is at `http://swa.cefriel.it/examples/botdcat-ap/Overpass.ttl`. Since POI categories (e.g., ATM, restaurant, kiosk, railway station, library) are well-known concepts, the entities included in the EntityCatalog *POIsCategories* are taken from the LinkedGeoData ontology [7]. In general, this solution is useful when entities express concepts already defined in external ontologies and vocabularies.

5 Conclusions

Chatbots represent one of the major rising trends, and their usage and distribution are predicted to grow over the next years. Gartner places chatbot systems in the top strategic technology trends for 2017 [10], evolving and expanding the use of Artificial intelligence and Machine learning in apps and services during the next 20 years.

---

6 The beta version still in development (Italian only, to try out the bot click on "Chiedi") of the bot application is deployed at https://ns3056488.ip-213-32-26.eu/talkinpiazza2/

7 http://www.piazza.eu

8 http://wiki.openstreetmap.org

With this growing demand and market potential for the development of chatbots, the need arises to simplify and standardize how those systems access and reuse data contained in knowledge sources. In this paper, we introduced the BotDCAT-AP vocabulary: when employed to describe datasets, it can bring benefits both to dataset owners, which enable their data to be further reused, and to chatbot developers, which are supported in the software development.

BotDCAT-AP can have a large impact by bringing value to the chatbot market, it enables and fosters reusability of datasets across chatbot systems, it is designed as an extension to DCAT-AP and it is openly available online, published and documented according to Sematic Web best practices and released with an open license. In the future, we will improve the evaluation of our proposal and we will investigate the community interest to establish an official working group and to proceed with the BotDCAT-AP standardization process.

Acknowledgement

This work is partially supported by the Piazza activity (id 16391), co-funded by EIT Digital.

References

1. Maali, F., Erickson, J., & Archer, P. (2014). Data catalog vocabulary (DCAT). W3C Recommendation. Available at: http://www.w3.org/TR/vocab-dcat/, last accessed 2017/05/10.
2. ISA working group (2015). DCAT application profile for data portals in Europe. Available at: https://joinup.ec.europa.eu/system/files/project/dcat-ap_final_v1.00_0.html, last accessed 2017/05/10.
3. ISA working group (2016). GeoDCAT-AP: A geospatial extension for the DCAT application profile for data portals in Europe. Available at: https://joinup.ec.europa.eu/node/154143/, last accessed 2017/05/10.
4. ISA working group (2016). StatDCAT-AP – DCAT Application Profile for description of statistical datasets. Available at: https://joinup.ec.europa.eu/node/157143, last accessed 2017/05/10.
5. Kar, R., & Haldar, R. (2016). Applying Chatbots to the Internet of Things: Opportunities and Architectural Elements. *Inter. Journal of Advanced Computer Science and Applications* 7(11), 147–154.
6. Höffner, K., Walter, S., Marx, E., Usbeck, R., Lehmann, J., & Ngonga Ngomo, A. C. (2016). Survey on challenges of Question Answering in the Semantic Web. *Semantic Web* (Preprint), 1-26.
7. Stadler, C., Lehmann, J., Höffner, K., & Auer, S. (2012). LinkedGeoData: A core for a web of spatial open data. *Semantic Web* 3(4), 333-354.
8. Vu, Q. H., Pham, T. V., Truong, H. L., Dustdar, S., & Asal, R. (2012). Demods: A description model for data-as-a-service. In Proc. of the IEEE 26th International Conference on Advanced Information Networking and Applications (AINA 2012), pp. 605-612.
9. Celino, I., Calegari, G. R., & Fiano, A. (2016, September). Towards Talkin'Piazza: Engaging citizens through playful interaction with urban objects. In Proc. of the IEEE International Conference on Smart Cities (ISC2 2016), pp. 1-5.
10. Gartner (2016). Top 10 Strategic Technology Trends for 2017. Gartner Report, 2016. Available at: https://www.gartner.com/doc/3471559/top--strategic-technology-trends