The DBpedia Databus Tutorial: Increase the Visibility and Usability of Your Data

Milan Dojchinovski
DBpedia Association
Leipzig, Germany
Czech Technical University in Prague
Prague, Czech Republic
dojcinovski.milan@gmail.com

Abstract

This tutorial introduces DBpedia Databus (https://databus.dbpedia.org), a FAIR data publishing platform, to address challenges faced by data producers and consumers. It covers data organization, publishing, and consumption on the DBpedia Databus, with an exclusive focus on Linguistic Knowledge Graphs. The tutorial offers practical insights for knowledge graph stakeholders, aiding data integration and accessibility in the Linked Open Data community. Designed for a diverse audience, it fosters hands-on learning to familiarize participants with the DBpedia Databus technology.

Keywords: DBpedia, DBpedia Databus, Knowledge Graphs, LOD

1. Introduction

DBpedia (https://www.dbpedia.org) is a crowd-sourced community effort which has been initiated in 2007 with the ultimate goal to extract structured knowledge from various Wikimedia projects. This structured information resembles an open knowledge graph, the DBpedia Knowledge Graph, which is publicly available for use for everyone on the Web. Along DBpedia, large number of other knowledge graphs have been published following the Linked Data principles as part of the Linked Open Data cloud initiative. Up until now, the LOD cloud (https://lod-cloud.net/) consists of over 1,314 knowledge graphs which are publicly available under an open license. Despite of this increase, several issues have arisen. First, users find difficulties in finding relevant data due to a lack of effective search mechanisms. Second, due to the uncontrolled way of publishing the metadata, the publishers introduce various diverse metadata schema which is not aligned and very often not in line with the best practices. And third, the process for integration of new knowledge graphs and the link sets in the LOD cloud is poorly governed and outdated. All these issues have a significant negative impact on the LOD cloud ecosystem where the knowledge consumers have to invest huge amounts of effort when consuming data, while knowledge graph providers struggle with the data publishing mechanisms.

In this tutorial, we address the above-mentioned problems using the DBpedia Databus technology (https://databus.dbpedia.org), a FAIR data publishing platform. In the tutorial, first, the participants will gain basic information on the DBpedia Knowledge Graph and the DBpedia community. Then, a main focus of the tutorial will be put on the DBpedia's Databus publishing platform. In practical examples we will illustrate the potential and the benefit of using DBpedia Databus. The participants will learn:

- what is the **DBpedia Databus**,
- how the **data is organized** on the DBpedia Databus,
- how to benefit from the **Databus collections** concept,
- how to **publish data** on the DBpedia Databus,
- how to **consume data** from the DBpedia Databus,
- how to **create knowledge graphs** using the Databus and
- how to deploy a **local instance** the DBpedia Databus platform.

The tutorial will be organized as a highly interactive event. The presenters together with the participants will work together and learn how publish data on the Databus, how to organize the data on the databus and how to then consume the published data.

The domain focus of the tutorial are *Lingustic Knowledge Graphs*. The tutorial will exclusively address Linguistic Knowledge Graphs and the participants will be invited to publish linguistic datasets on the Databus.

Few weeks before the execution of the tutorial, the organizers will provide instructions to the potential participants with guides and tips which will help them benefit the most from the tutorial. In particular,

LREC-COLING 2024 Tutorials, pages 42–44

the organizers will invite participants to 'bring' their data as hosted on some public server (e.g. Zenodo) and on-site, during the event the participants will learn how to publish and register their data on the Databus.

2. Target Audience

The tutorial primarily targets knowledge graph publishers and knowledge graph consumers. We welcome stakeholders who work with open data (e.g. in the context of the LOD cloud) as well as those that maintain proprietary (commercial) knowledge graphs and would like to learn how to use the Databus technology in private settings.

In addition, the tutorial also targets existing and potential new users of DBpedia, developers that wish to learn how to replicate DBpedia Databus platform, providers interested in exploiting the DBpedia KG, data providers interested in integrating data assets with the DBpedia KG and data scientists (e.g. linguists). The tutorial is also dedicated for people from the public and private sector who are interested in implementing knowledge graph technologies, and in particular, DBpedia.

We expect about 50 on-site participants and similar number for online participants with background in linguistics, knowledge graphs, linked data and semantic web in general, knowledge engineering and knowledge extraction.

3. Outline

We will organize the DBpedia Databus tutorial with a 20% lecture-style sessions and 80% hand-on exercises. The tutorial is planned as 4h long event.

- **Intro:** Meet and greet, introduction to the tutorial (5 min)
- **Session 1:** Overview of the DBpedia technology, by Milan Dojchinovski (10 min)
- **Session 2:** The DBpedia Databus in a Nutshell, by Jan Forberg (25 min)
- **Session 3:** Your data on the Databus, by Kirill Yankov (70 min)
- **coffee break** (30 min)
- **Session 4:** Oragnizing and consuming data from the Databus, by Kirill Yankov (60 min)
- **Session 5:** Deploying own DBpedia Databus, by Jan Forberg (30 min)
- **Outro:** Q&A and wrap-up (10 min)

Note: we are flexible with the schedule and will adjust and adapt it dynamically based on the participants requirements and interests.

4. Diversity Considerations

The DBpedia Databus tutorial addresses the diversity aspects as described below.

Improved Diversity and Increased Fairness DBpedia Databus plays a crucial role in advancing diversity and fairness within the field of language technologies. By providing an extensive and openly accessible repository of knowledge, it can empower researchers to conduct more inclusive and equitable analyses. Researchers can draw from the diverse set of datasets hosted on DBpedia Databus, ensuring a more comprehensive representation of global knowledge diversity. Moreover, DBpedia Databus promotes fairness by offering a standardized and transparent approach to data handling, mitigating biases in data selection and representation. The tutorial will delve into best practices for promoting fairness in knowledge graphs research, highlighting the ethical considerations and safeguards for working with diverse datasets, ultimately contributing to the development of more equitable methodologies within the field.

Underrepresented Groups of Participants DBpedia Databus holds particular relevance for underrepresented groups of potential participants in language technologies. For instance, linguists and researchers focusing on underrepresented languages and language communities will find immense value in our tutorial. DBpedia Databus can support research on languages and dialects that have been historically marginalized, offering a platform to amplify their linguistic nuances and significance. The platform's geographic inclusivity also means that it provides an opportunity to study languages from regions that have been underrepresented in the computational linguistics community. This tutorial will provide insights and practical guidance on utilizing DBpedia Databus for these specific contexts, ensuring that the linguistic diversity and richness of underrepresented groups are recognized and studied.

Presenters from Underrepresented Groups One of the tutorial presenters have diverse and underrepresented background, further enriching the learning experience. Milan Dojchinovski brings expertise in linguistic research within underrepresented Macedonian language community.

5. Reading List

Following resources can help the participants to better understand the contents of the tutorial and its background.

- *The New DBpedia Release Cycle: Increasing Agility and Efficiency in Knowledge Extraction Workflows* (Hofer et al., 2020) Hofer et al. SEMANTICS, 2020.

- *DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia* (Lehmann et al., 2015) Lehmann et. al. 2015.
- *DBpedia - A crystallization point for the Web of Data* (Bizer et al., 2009) Bizer et al. Journal of Web Semantics, 2009.
- (Sep 13, 2023) *DBpedia tutorial co-located with the Language, Data and Knowledge conference 2023 (LDK)*[1].
- (May 2, 2022) *DBpedia tutorial co-located with the Knowledge Graph Conference (KGC) 2022*[2].
- (Apr 25, 2022) *DBpedia Tutorial at The Web Conference (WWW) 2022*[3].

6. Presenters and Organizers

Milan Dojchinovski Milan holds a Research Associate position at the Institute for Applied Informatics (InfAI) and an Assistant Professor position at the CTU in Prague. He has 10+ years experience in the computer industry in Germany, Czech Republic and Slovenia. His research interests are in Semantic Web, NLP and Knowledge Graph technologies. Since 2013 Milan is an active member of the DBpedia community project. He holds a PhD in Information Science from the Czech Technical University in Prague in the context of Linked Data, Knowledge Extraction and Web Services technologies. Milan is the main lead of the DBpedia tutorial series.
Kirill Yankov Kirill is a back-end developer at the KILT Competence Center at the Institute for Applied Informatics. Since 2021 he has been involved in DBpedia developments and contributed to the development of the DBpedia Databus. Kirill has been part of the core organization team of the series of DBpedia tutorials organized since 2021.
Jan Forberg Jan is a full stack developer at the KILT Competence Center at the Institute for Applied Informatics. Since 2016 he has been involved in DBpedia and contributed to the development of the DBpedia Databus, Dockerized DBpedia and DBpedia Lookup. Jan has been part of the core organization team of the online series of DBpedia tutorials organized since 2020.
Julia Holze Julia is head of the Organizational Development of the DBpedia Association. She holds a M.A. degree in Media & Communication Science. She will be responsible for the community outreach, support the organization of this tutorial and spread news to the DBpedia Community.
Sebastian Hellmann Sebastian is the executive director and board member of the non-profit DBpedia Association. He is a senior member of the "Agile Knowledge Engineering and Semantic Web" AKSW research center, focusing on semantic technology research. He is the head of the "KILT" Competence Center at InfAI. Sebastian is also a contributor to various open-source projects and communities such as DBpedia, NLP2RDF, DL-Learner and OWLG, and has been involved in numerous EU research projects. Sebastian will monitor and guide the tutorial preparations.

7. Ethics Statement

The tutorial is designed with a strong commitment to ethical standards to ensure that participants are equipped with knowledge and practices that will not introduce ethical issues or problems. We will emphasize ethical considerations in all aspects of the tutorial, from data selection and publishing to respectful engagement with diverse datasets. Our goal is to provide a safe and inclusive learning environment where participants can explore the intricacies of DBpedia Databus without compromising ethical standards. We are dedicated to upholding the highest ethical principles throughout the tutorial to foster a culture of responsible and ethical research.

8. Bibliographical References

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. Dbpedia-a crystallization point for the web of data. *Journal of web semantics*, 7(3):154–165.

Marvin Hofer, Sebastian Hellmann, Milan Dojchinovski, and Johannes Frey. 2020. The new dbpedia release cycle: Increasing agility and efficiency in knowledge extraction workflows. In *16th International Conference on Semantic Systems, SEMANTiCS 2020, Amsterdam, The Netherlands, September 7–10, 2020, Proceedings 16*, pages 1–18. Springer International Publishing.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195.

[1] https://www.dbpedia.org/events/dbpedia-tutorial-at-ldk-2023/
[2] https://www.dbpedia.org/events/dbpedia-tutorial-2-0-kg-conference/
[3] https://www.dbpedia.org/events/tut-at-the-web-conf/