Linked Data for Digital Humanities Scholars and Researchers: "Rainis and Aspazija" (RunA) Collection

Uldis Bojārs[0000-0001-7444-565X], Anita Rašmane and Anita Goldberga

National Library of Latvia, Mukusalas iela 3, Riga, LV-1423, Latvia
[uldis.bojars, anita.rasmane, anita.goldberga]@lnb.lv

Abstract. This paper presents the Linked Digital Collection "Rainis and Aspazija" that demonstrates the use of Linked Data in Digital Humanities[1]. This collection offers interlinked digital objects and data from several memory institutions and private repositories related to two Latvian poets of the period of National Awakening – Rainis and Aspazija. The paper also describes the semantic annotation tool developed for cultural heritage needs that was used to create annotations in this collection. Digital object annotations are the source of links between collection's digital objects and the entities mentioned in these annotations. Information about collection's objects and entities is published as Linked Open Data facilitating reuse of this information.

Keywords: Cultural Heritage, Linked Data, Digital Collection, Text Annotation.

1 Introduction

Metadata of objects in digital collections provide means for finding the resources mainly by their external description, pre-categorization and physical parameters. Only a small part of metadata values deals with the content of resource. Scholars and researchers in Digital Humanities have to deal with large amount of texts and collections of documents, whose content has to be analyzed and processed with the intent of extracting precise meaning of mentioned entities and their interconnections. Such a depth of content description has to be and can be achieved to create a fruitful research and scholarship environment by the means of enriching written texts with annotations – marked text fragments, linked to identified entities, who in turn manifest themselves as additional metadata for a resource and can link to additional internal and external data sources.

In 2016, the National Library of Latvia (NLL) together with the National Archives of Latvia, the Institute of Literature, Folklore and Art of the University of Latvia, the Association of Memorial Museums, and the Literature and Music Museum published "Rainis un Aspazija" (RunA) – the first digital cross-sectoral cultural heritage pilot collection in Linked Data form in Latvia. RunA highlights the NLL`s efforts in developing new knowledge bases for memory institutions and researchers. During 2018-

[1] https://runa.lnb.lv/en/

2019, a special semantic annotation tool and its entity datastore was developed by the NLL to enhance RunA document annotations [2].

This paper introduces a major new release of the collection that is substantially different from the initial pilot project described in [4]. In particular, it has a new look-and-feel, additional digital objects, and annotations in the collection are created and managed using a custom-built tool for annotating cultural heritage content. An important difference from the initial release is the entity database that is part of the annotation tool and that allows users to collect and manage information about the entities mentioned in annotations.

2 Semantic Annotation Tool

Content for the "Rainis and Aspazija" collection is prepared using a custom semantic annotation tool. This tool allows users to annotate textual content, mark the mentions of important entities and provide additional information about these entities using the entity database that is a part of this tool [2]. Users of the annotation tool include domain experts and digital humanities students.

Cultural heritage content and especially historical documents may be particularly difficult for automatic entity recognition and linking because the relevant tools need to know the specific context of the documents (e.g. personal correspondence and people involved in it) and entities that are likely to be mentioned in these documents [3]. In order to create high quality annotations, the annotation tool and the RunA annotation process use manual annotation. This is due to the semantically and lexically free and unstructured use of words in personal written texts where many entities are not named directly, but are implied by pronouns ("he", "she") or generic names ("city", "sister"). In many cases, the person annotating the document is able to identify entities from context. In other cases, this may be done after examining the context of a group of documents. Another issue is the need to disambiguate between different meanings of the same text fragment. As an example, the annotated compendium to one of Aspazija's works in RunA uses the word "Aspazija" to refer to 4 different entities (person Aspazija, her work "Aspazija", etc.) which should be correctly identified when annotating this document.

The annotation tool supports three core types of annotations - simple annotations that may link to named entities, structural annotations that mark up portions of the document that have a special meaning within the context of the document (e.g. direct citation of another published material) and composite annotations for more complex use cases (e.g., for representing an event described in a document with mentions of place, time and participants, all marked and identified in their own annotations).

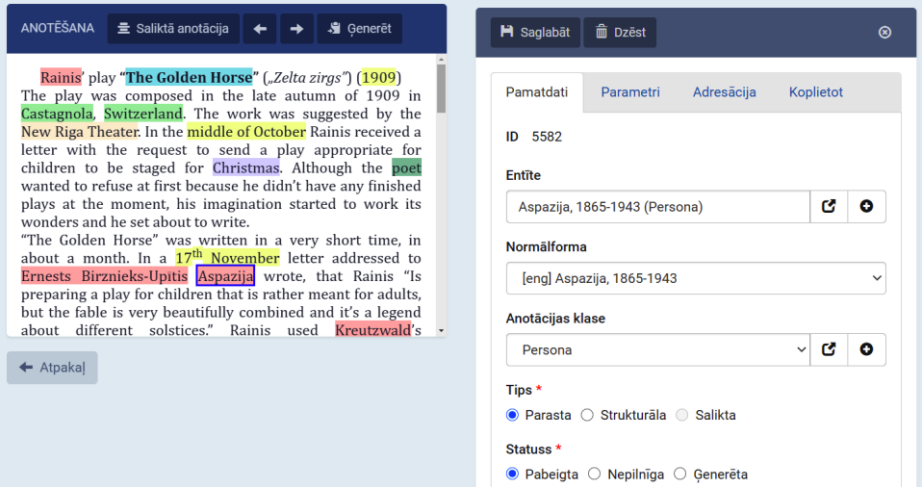

Fig. 1. Text annotation tool showing an annotation about a person (Aspazija).

Figure 1 is a screenshot of the annotation tool showing annotations in the English compendium to Rainis' play "The Golden Horse". The left side of the screen contains text with annotations while the right side shows information about the selected annotation (Aspazija) which contains a reference to the entity database record about Aspazija. Information about the annotation includes its type (simple annotation), status (completed), annotation class (person) and information about the entity associated with the annotation. Annotations may also contain user comments.

New annotations can be added by highlighting text fragments and entering information about the annotation. As a part of the workflow, users may choose an existing entity or create a new entity that the annotation will refer to.

Information about the entities referenced from annotations is maintained in a dedicated entity database that supports links between entities and can point to additional information about these entities (e.g., to Linked Data resources such as Wikidata[2]). The database provides for storing and reusing data extracted from individual annotations and those added by researchers. This allows experts to build a knowledge base about the entities referenced from annotations while annotating documents. Figure 2 shows an example of entity database information about a person – Friedrich Reinhold Kreutzwald.

The entity information can evolve as the document annotation task progresses as it is possible to enhance the entity data later on. For example, a user may create an entry for an entity that needs further research and leave comments about what is known about the entity and what is not. The entity record can later be extended with additional information (e.g. identifiers for this entity in other authoritative data sources) when it becomes available.

[2] https://www.wikidata.org/

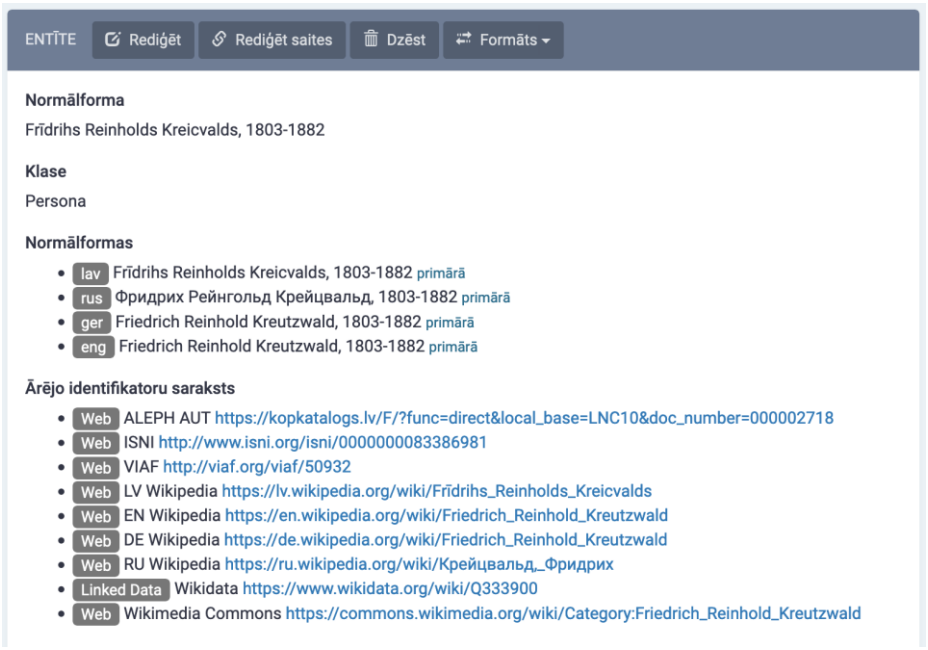

Fig. 2. Annotator's entity database information about a person.

The annotation tool can export annotated documents as rich web applications that contain the document along with annotations and information about the entities that annotations refer to. Exported annotated documents are used by the "Rainis and Aspazija" collection for embedding into collection object pages.

Further information about the annotation tool and related work can be found in [2].

3 Linked Digital Collection "Rainis and Aspazija"

"Rainis and Aspazija" is a digital collection that contains diverse, multi-format content about two Latvian poets of the period of National Awakening – Rainis (1865-1929) and Aspazija (1865-1943) – supplied by multiple cultural heritage organizations. When creating the collection, we aimed at including a wide range of content types and digital objects. The collection includes[3]:

- first editions of Rainis' and Aspazija's literary works along with their annotated compendiums (abstracts) in English and Latvian (85 works and 158 annotated compendiums);
- personal correspondence with annotations (395);
- archival documents (23);
- photos (516), audio and video recordings (37);
- posters (22), presentations (7) and cartoons (32).

[3] A summary of content types in the collection: https://runa.lnb.lv/en/par-kolekciju/

This is the second major release of the collection. Compared with the initial pilot project [4] this release:

- builds on a new annotation system that's integrated into the collection;
- contains literary work compendiums in English and Latvian;
- has additional content (extensive annotated biographies of both poets, more than 100 photos from the National Archive of Latvia, cartoons etc.).
- the user interface has English localization;
- has a new look and feel (improved search, responsive web design, etc.);
- provides interface for mobile devices.

The collection contains detailed biography pages for Rainis[4] and Aspazija[5] with summaries of collection objects (by object type) related to each person. The biography page for Aspazija is shown in Figure 3.

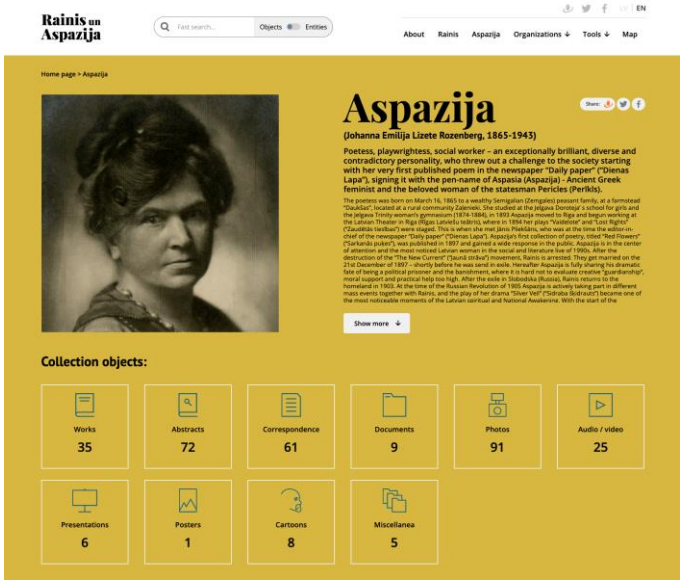

Fig. 3. Aspazija's biography page in the digital collection.

In terms of links in the collection, the most interesting types of content are (1) personal correspondence that has been meticulously transcribed and annotated, and (2) annotated compendiums (abstracts) to poets' literary works. The text of both of these content types was annotated with mentions of named entities using the annotation tool described above. Information about these entities is a part of the collection and is shown as collection's entity pages. The collection allows users to search for both digital objects and entities.

[4] Rainis: https://runa.lnb.lv/en/110023/
[5] Aspazija: https://runa.lnb.lv/en/142651/

Figure 4 shows a digital object page containing a compendium to Rainis' play "The Golden Horse"[6] (in Latvian: Zelta zirgs). The top part of the page contains object metadata including links to other objects (in this case: the first edition of this work). The bottom part contains annotated digital object text exported from the annotation tool and a list of entities that appear in these annotations. By clicking the links users can view collection's pages about these entities.

Annotations are highlighted in different colors based on the type of the entity they refer to. Most of them are simple annotations that identify the entity represented by the annotated text fragment. Collection's entity pages contain links back to the documents that mention this entity. As a result, the collection becomes a Linked Digital Collection that contains a network of objects (annotated textual content) and entities linked to one another. Users may view the network of links by clicking the "Data network" button.

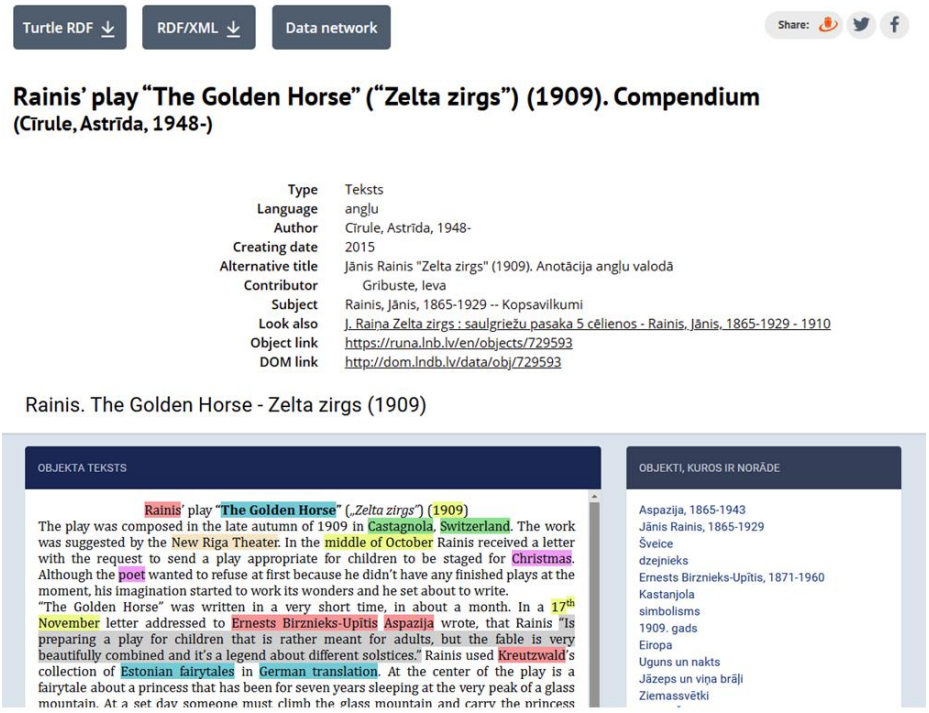

Fig. 4. Annotated document view in the digital collection.

Entity pages are part of the collection and contain information about the entity along with links to other entities and to additional resources about this entity on the web. These links may refer to web pages or to Linked Data URI identifiers. Entity pages also show information imported from Wikipedia – abstracts and images representing the entity.

[6] https://runa.lnb.lv/en/objects/729593/

Person entity pages contain links to all documents in which a person is mentioned directly (“Doriņa”) or indirectly (“beloved sister”), provided these documents contain annotations pointing to the entity. This creates a group of related texts that are immediately available to researchers who otherwise would need to examine all collection documents. Place names which depend on language and vary over time (“Pēterpils”, “Петроградъ”, “Saint Petersburg”) are another example of the difficulty that awaits researchers when examining cultural heritage documents. This difficulty is resolved by adding place name annotations.

The collection system also provides a map[7] and a timeline view[8]. The map shows the places that were important in lives of both poets. Information for the map is collected from documents annotated with the mentions of placenames. The timeline view shows events in poets’ lives in the context of important worldwide events.

An obstacle that needed to be overcome when creating this collection was that different participants (libraries, archives, museums) use different standards and data formats to represent their information. Currently, we cannot directly import metadata from museums and archives. The metadata is first converted to a spreadsheet instead, for import into NLL’s digital object management system (DOM). Perhaps an agreement between memory institutions on a single data model for named entities and a common cultural heritage entity register in which the various forms of entities are controlled could solve this problem in future.

The popularity of RunA has grown 42% from ~28 thousand page visits in 2019 to ~40 thousand visits in 2020. The information in the RunA collection is still being expanded with additional correspondence and excerpts of Rainis’ diaries. Students from the Faculty of Humanities, University of Latvia are involved in the annotation of additional poets’ correspondence. Students and researchers can use RunA as a digital resource and discover previously unknown relationships between the collection’s objects and entities. This knowledge base is also being integrated into the education process of study courses at the Faculty of Humanities, University of Latvia.

4 Linked Data

Machine-readable information about all RunA objects and entities is published according to Linked Data principles in Turtle RDF and RDF/XML format [1]: collection’s resources (objects and entities) have HTTP URI identifiers and the system responds to requests for these URIs by sending back structured RDF data which contain links to other resources. Linked Data can also be retrieved by appending a corresponding extension to the URI:

- .ttl for the Turtle RDF format (e.g. https://runa.lnb.lv/729593.ttl);
- .rdf for the RDF/XML format (e.g. https://runa.lnb.lv/729593.rdf).

[7] https://runa.lnb.lv/en/map/

[8] https://runa.lnb.lv/en/darbi-un-notikumi/

Links to RDF data are also displayed on collection's webpages. In order to facilitate the harvesting of collection's machine-readable data, the system provides an XML sitemap with URIs of its objects and entities[9].

An example of collection's RDF data can be found in Appendix 1. It describes an annotated compendium to Rainis' play "The Golden Horse" in Turtle RDF format. The vocabularies used for representing collection's resources in RDF are Dublin Core[10], FOAF[11], Bibo[12] and Schema.org. Notable properties in RDF data of the collection include:

- *bibo:annotates* property represents links between annotated documents and the original digital objects (e.g. handwritten letters);
- *dct:hasPart* property points to the files contained in the digital object (e.g. a scanned image of a letter);
- *dct:isPartOf* and *schema:isPartOf* properties point to NLL's digital object management system (DOM) collection that the resource is a part of (e.g. a digital object collection about Aspazija);
- *owl:sameAs* and *schema:sameAs* represent links to the same resource in DOM;
- *schema:mentions* property represents links between objects and entities.

Visitors of the collection can benefit from links between its objects and entities that provide a new way for exploring the collection. For example, they may want to work with a subset of documents that mention a particular entity. Advanced users can make use of collection's Linked Open Data by automatically collecting this data and using it for further analysis.

Related work includes Amsterdam Museum Linked Data project [5] which publishes objects' metadata as Linked Data but, unlike RunA, does not annotate the content of digital objects. Another example is a Singapore National Library Board project [6] in which connections between entities are automatically extracted, disambiguated and links established.

5 Conclusion

In this paper we described the Linked Digital Collection "Rainis and Aspazija" and the text annotation tool used to enrich the collection with links between its objects and entities. These links are added to the collection by annotating textual content of the collection's objects. The resulting knowledge base is available as open data and is being integrated into the education process at the University of Latvia.

The annotation tool was developed based on NLL's needs for the annotation of cultural heritage information. An integral part of the tool is its entity database where users can add and maintain information about the entities mentioned in annotations. Entity information may contain external links to webpages about the entity and Linked Data identifiers (URIs) for information about this entity in other Linked Data resources.

---

[9] https://runa.lnb.lv/sitemap.xml

[10] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

[11] http://xmlns.com/foaf/spec/

[12] http://www.bibliontology.com/

RunA allows researchers to explore the digital collection in new ways by following the links between collection's objects and entities. This collection is published as Linked Open Data, creating an opportunity for reuse of this cultural heritage knowledge base.

Acknowledgements

This research was supported by the Latvian Council of Science Project Nr. lzp-2019/1-0365 "Latvian Memory Institution Data in the Digital Space: Connecting Cultural Heritage".

References

1. Berners-Lee, T.: Linked Data - Design Issues. W3C (2006). Retrieved November 11, 2020. URL: https://www.w3.org/DesignIssues/LinkedData.html
2. Bojārs, U., Rašmane, A., Žogla, A., Bāliņa, S., Salna, E. Semantic Annotation Tool for Cultural Heritage Content. Baltic Journal of Modern Computing, Vol. 6, No. 4 (2018).
3. Bojars, U., Rasmane, A., Zogla, A. The Requirements for Semantic Annotation of Cultural Heritage Content. Proceedings of the Second Workshop on Humanities in the Semantic Web (WHiSe II), Vienna, Austria, October 22, 2017. CEUR Workshop Proceedings, vol. 2014. URL: http://ceur-ws.org/Vol-2014/
4. Bojars, U. Case Study: Towards a Linked Digital Collection of Latvian Cultural Heritage. Proceedings of the 1st Workshop on Humanities in the Semantic Web (WHiSe 2016), pp. 21-26.
5. de Boer, V. et al. (2013). Amsterdam museum linked open data. Semantic Web, 4(3), pp. 237-243.
6. Hoon E**.** (2019). Mining Text, Linking Entities - NLB's Journey. Retrieved March 3, 2021. http://library.ifla.org/2448/1/114-ee-en.pdf

Appendix 1 – Linked Data of the Compendium of “The Golden Horse” by Rainis

```
<http://runa.lnb.lv/objects/729593/>
  a bibo:Document ;
  dct:isPartOf <http://dom.lndb.lv/data/obj/59600> ;
  dct:isPartOf <http://dom.lndb.lv/data/obj/59676> ;
  dct:title “Rainis’ play ‘The Golden Horse’ (‘Zelta zirgs’) (1909).
Compendium”@en ;
  dc:type “Teksts” ;
  dct:creator <http://dom.lndb.lv/data/auth/5623> ;
  dct:language <http://id.loc.gov/vocabulary/iso639-2/eng> ;
  dct:subject <http://dom.lndb.lv/data/auth/232655> ;
  foaf:depiction <https://dom.lndb.lv/data/obj/729593.png> ;
  dct:hasPart <http://runa.lnb.lv/wp-content/uploads/ob-
jects/729593/Rainis_Zelta zirgs_1909_eng.docx> ;
  dct:date “2015” ;
  dct:source <http://dom.lndb.lv/data/obj/729593> ;
  owl:sameAs <http://dom.lndb.lv/data/obj/729593> .

<http://runa.lnb.lv/objects/729593/>
  a schema:CreativeWork ;
  schema:isPartOf <http://dom.lndb.lv/data/obj/59600> ;
  schema:isPartOf <http://dom.lndb.lv/data/obj/59676> ;
  schema:author <http://dom.lndb.lv/data/auth/5623> ;
  schema:description “Rainis’ play ‘The Golden Horse’ (‘Zelta zirgs’)
(1909). Compendium”@en ;
  schema:thumbnailUrl <https://dom.lndb.lv/data/obj/729593.png> ;
  schema:datePublished “2015” ;
  schema:sameAs <http://dom.lndb.lv/data/obj/729593> .

<http://runa.lnb.lv/objects/729593/>
  bibo:annotates <http://runa.lnb.lv/objects/61424/> .

<http://runa.lnb.lv/objects/729593/>
  schema:mentions <http://runa.lnb.lv/entities/1003/> ;
  schema:mentions <http://runa.lnb.lv/entities/1007/> ;
# ...
  schema:mentions <http://runa.lnb.lv/entities/2095/> .
```