Towards the Russian Linked Culture Cloud: Data Enrichment and Publishing

Dmitry Mouromtsev[1], Peter Haase[2,1], Eugene Cherny[1,3], Dmitry Pavlov[4,1], Alexey Andreev[1], and Anna Spiridonova[1]

[1] ITMO University, St.Petersburg, Russia
mouromtsev@mail.ifmo.ru, eugene.cherny@niuitmo.ru, spiranna@list.ru, aandreyev13@gmail.com
[2] metaphacts GmbH, Walldorf, Germany
ph@metaphacts.com
[3] Åbo Akademi University, Turku, Finland
[4] Vismart Ltd., St.Petersburg, Russia
dmitry.pavlov@vismart.biz

Abstract. In this paper we present an architecture and approach to publishing open linked data in the cultural heritage domain. We demonstrate our approach for building a system both for data publishing and consumption and show how user benefits can be achieved with semantic technologies. For domain knowledge representation the CIDOC-CRM ontology is used. As a main source of trusted data, we use the data of the web portal of the Russian Museum. For data enrichment we selected DBpedia and the published Linked Data of the British Museum. The evaluation shows the potential of semantic applications for data publishing in contextual environment, semantic search, visualization and automated enrichment according to needs and expectations of art experts and regular museum visitors.

Keywords: semantic web, semantic data publishing, CIDOC-CRM, open data, cultural heritage

1 Introduction

The smooth and natural transfer of cultural heritage is the key factor for the preservation of national identity, which is crucial in the era of rapid globalization. At the same time the traditional mechanisms of heritage transfer from generation to generation nowadays undergo a serious change and experience a great challenge as the digital era unfolds before our own eyes. The digital era prompts developers of content and applications to use a new language of communication and a new channel to deliver the information to the consumer. Thus, cultural heritage transfer can strongly benefit from the digital movement to make it more exciting, personal and vivid. Although in order to make sure that the cultural heritage is being preserved, the digitization of content is not enough whilst adequate representation of the data starts to play a decisive role.

Progress has been made in this direction by introducing the digitization of the art works and creating large structured storage of digitized artifacts. The second step was made by creating user applications with digital data: It included establishment of large museum portals, the launch of mobile applications of various kinds and features. Some of the museums have already placed their digital collections in the open data cloud, thus opening it for querying and integration [1]. To back this trend up all the vital infrastructure was created. Of particular importance in this context is CIDOC-CRM - a Conceptual Reference Model providing definitions and a formal structure for describing the implicit and explicit concepts and relationships used in the cultural heritage domain.

In this paper, we report on the results of the first steps towards the Russian Linked Culture Cloud making the heritage data available, including the publication as Linked Data as well as through end user applications. Our long-term goal is to build the overall Russian Linked Culture Cloud by integrating data from many providers like museums and other institutions and having a powerful user interface and a set of practical tools for data acquisition, modification and publishing. The pilot project was started in cooperation with the Russian Museum in St. Petersburg, which holds the largest collection of Russian art in the world. The primary goal of our research was to demonstrate the applicability and benefits of usage of semantic data to tackle the challenges of cultural heritage transfer in the digital era. The system is meant to deliver benefits to two different target groups: the museum art experts and museum visitors. These two groups greatly differ in their needs, but the system covers the interests of both of them.

Taking into account the needs of potential users we managed to set forth the following objectives:

- *Simplified integration of external data.* While the initial effort on making the internal data open and available might look like a significant investment at the start, in the long run it holds big promises with numerous benefits achieved through integration with external data. Our challenge was to make this process easier for organizations by means of providing the mechanisms suitable for simplified acquisition of data from open sources via various APIs or by crawling and further structuring the data including smooth integration into existing data models.
- *Dealing with quality of external data.* The first challenge is directly related to the second one. Integration of external data must be accompanied with validation methodology, quality assessment, purification of acquired external data. The system must be able to perform this task easily.
- *Flexibility of data presentation.* The third challenge is to demonstrate how the employed semantic technologies can enhance the end user experience while interacting with the data. The data presentation should be adjusting in real-time to the user preferences, interests expressed either explicitly in his profile or indirectly by his actions and interaction track with system.
- *Richer representation.* Among our potential users will be the art experts that need the deeper representation of information more or less ready for

analysis. The examples of such representations might be the timelines of events and art object creation dates, the graph depicting the popularity of art movements, the maps of traces of artists and so on. Creation of such forms of visualisations involves a deep domain knowledge, clear understanding of users needs and a thorough scenario of user interaction, which altogether makes a complicated goal to achieve.

The project is still in rapid development and contributors are welcome. For collaboration we use GitHub repository: https://github.com/ailabitmo/CultureCloud-Datasets, in which one can learn the technical details of the data transformation process.

2 Overview of the System

In this section we present an overview of the created system. It has been built using the *metaphacts Knowledge Graph Workbench*[5], a platform for the development of semantic applications. The system architecture diagram is depicted in the Fig. 1.

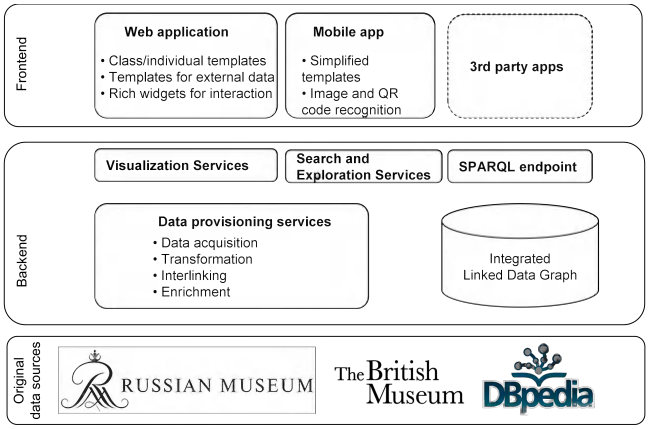

Fig. 1. Architecture of the System

Using the data provisioning services of platform, the original data sources have been transformed, interlinked, enriched and finally ingested into a triple store (a Systap Blazegraph[6] database), holding the integrated Linked Data graph. As described in detail in the subsequent section, Russian Museum relational data was transformed to RDF, represented using the CIDOC-CRM ontology. Where possible, links to DBpedia have been generated. The British

[5] http://www.metaphacts.com/
[6] http://www.blazegraph.com/

Museum thesauri were used as genre and artwork type taxonomies. The resulting data in the triple store is published via a SPARQL endpoint, accessible at http://culturecloud.ru/sparql.

Using additional backend services of the platform, e.g. visualization, search and exploration services, two applications have been built: a web application and a mobile app, as described in detail in Section 4. The applications are accessible at http://culturecloud.ru/ On the frontend side we made use of the rich templating mechanism of the platform and created templates for the relevant CIDOC-CRM classes to visualize artworks and authors. Each template also includes data from linked DBpedia entities. The main purpose of the mobile application is to provide museum visitors with additional information about art objects. It has the ability to recognize the artwork by making photo of it or by scanning a QR code. Special simplified templates were developed for this use case.

3 Publishing / Creation

3.1 Ontology model

We created and published the museum data according to the CIDOC-CRM ontology [3]. CIDOC-CRM serves as a basis for mediation of cultural heritage information and to provide the semantic 'glue' needed to transform todays disparate, localised information sources into a coherent and valuable global resource. The CIDOC-CRM ontology provides a representation aimed at harmonizing heterogeneous data, but retains the individual nature of the data - providing a semantic framework that supports the full variability and richness of the information and brings to life the concealed and implicit relationships between objects and events.

3.2 Data acquisition and transformation

In this project we agreed with Russian Museum management to work with data from one of their sites: rmgallery.ru. The original data undergoes a transformation process the main goal of which is to structure initial information into an RDF data graph conforming with the CIDOC-CRM ontology. Fig. 2 shows an example of the initial data representation in RDF and interlinking, as discussed in the next section.

CIDOC-CRM is an event-centric model. The central part of semantic representation is the event of production of some object *crm:E12_Production*. It connects all other entities that are relevant to it: A creator is connected with the *crm:P14_carried_out_by* property, an artwork is connected with the *crm:P108_has_produced* property, creation time is connected with the *crm:P4_has_time_span* property. The artwork is represented as an instance of class *crm:E22_Man-Made_Object* (Fig. 3). While not shown in the diagrams, the person's biography and artwork description are associated with *crm:E21_Person* and *crm:E22_Man-Made_Object* respectively via the *crm:P4_has_note* property.

All textual information in the dataset (names, titles, descriptions, etc.) were placed in two languages annotated with the corresponding language tag.

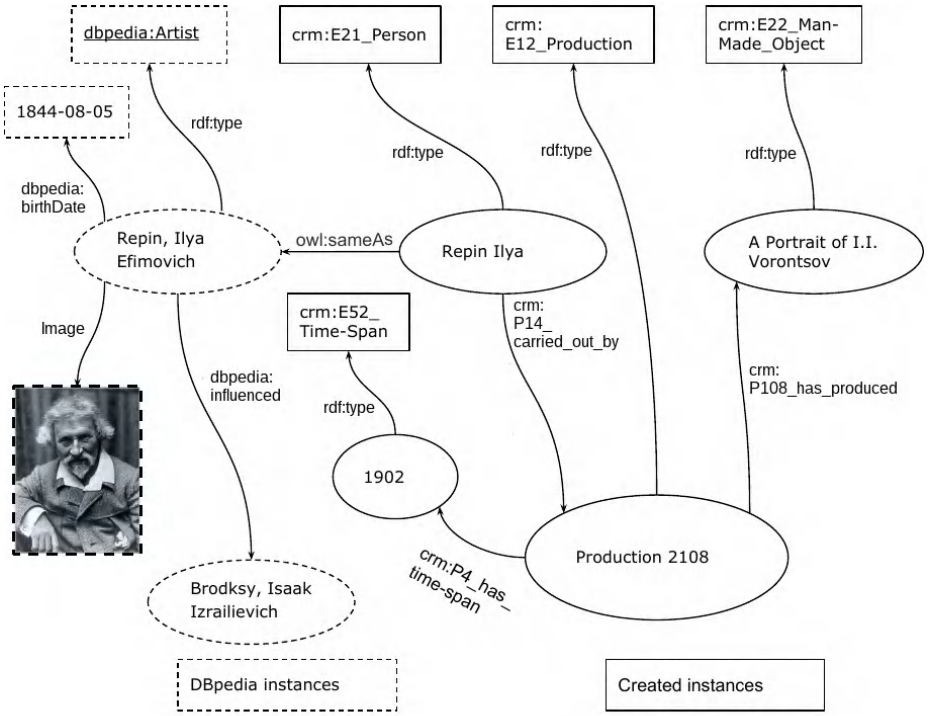

Fig. 2. Example of initial data representation and interlinking

3.3 Interlinking

To enrich the initial dataset we interlinked the authors from the Russian Museum data with persons from DBpedia. The interlinking was performed in two stages.

The first stage is implemented in semi-automatic mode: a Ruby script asks to choose from one of the options that script provides for the person's name matches. The first-step script does the following:

1. Query the Wikipedia API with the person's name
2. List query results on the screen and ask to choose the most suitable one
3. When user selects a variant, transform a chosen Wikipedia link to DBpedia one and create owl:sameAs for the *crm:E21_Person* (Fig. 2)

The second stage is carried out in automated mode and based on simple string comparison of person's initials. First of all, we extract names of the persons with type *dbpedia:Artist*. Then we transform all names to initials and performed a string comparison with names of the persons from the Russian Museum data. The second stage proves to be effective as it is shown in the Table 2. We worked with both international and Russian DBpedia datasets to interlink as many authors as possible.

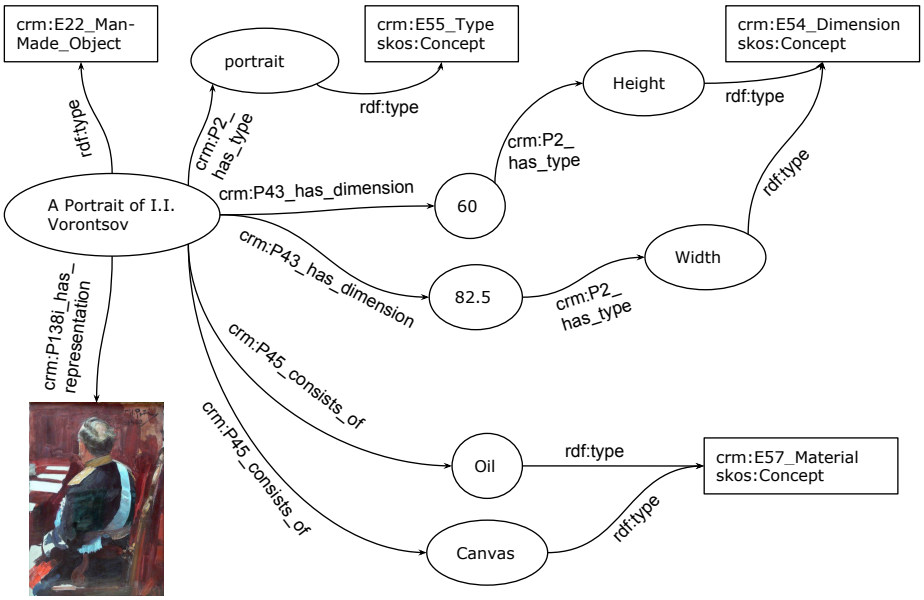

Fig. 3. Artwork representation example

3.4 Reusing thesauri of The British Museum

The British Museum has published high-quality thesauri that could be used with any museum, thus we decided to reuse them. The thesauri are based on SKOS. Every thesaurus object has the skos:Concept type and one of CIDOC-CRM more specific type. For example, the material "oil" has types of *skos:Concept* and *crm:E57_Material* and is part of "BM MATERIAL" concept scheme. The "allegory/personification" subject has types *skos:Concept* and *crm:E55_Type* and is part of "BM SUBJECT" concept scheme. We used the latter for describing the genre of the artworks. Fig. 4 shows an example of the usage.

For some genres there were no appropriate entities in British Museum dataset (illustration, caricature, theatrical scenery), for these we created additional instances following the exact same scheme.

3.5 Annotating unstructured text with DBpedia Spotlight

We have two pieces of unstructured data in the Russian Museum dataset: artwork descriptions and author biographies. We decided to contextualize this information with DBpedia Spotlight[7]. It identifies for DBpedia entities in the text and returns a text annotated with links to DBpedia resources. We replaced all initial textual information with the annotated one and added all entities in the annotation to our dataset as triples $(Entity, cc : hasAnnotation, DBpediaEntity)$,

[7] http://dbpedia-spotlight.github.io/

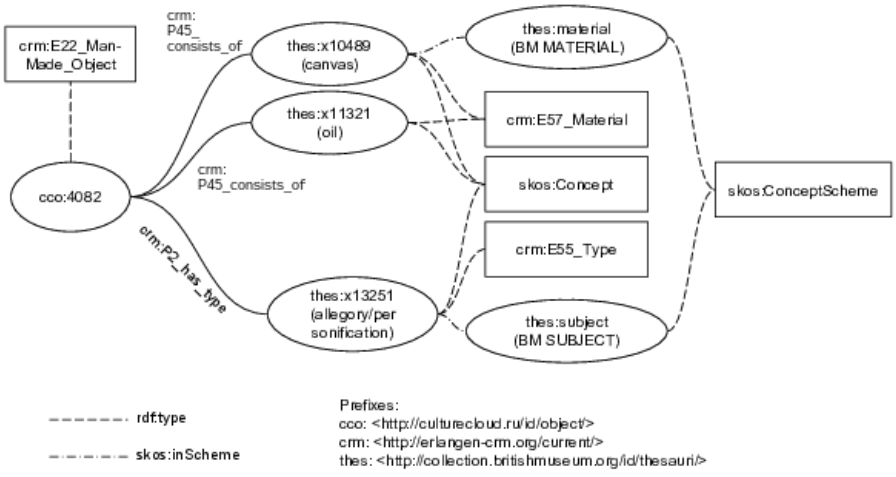

Fig. 4. Usage example of The British Museum thesauri

where *Entity* is either an artwork or an author and *DBpediaEntity* is the DBpedia resource associated with the text by Spotlight (Fig. 5).

Links in unstructured text provide additional ways for site visitor to explore existing information, besides links to the semantic entities created at transformation and enrichment stages.

4 Consumption of Data / End User Application

We created two end user applications for consuming the data: a website application and a mobile application. The website can be accessed from any mobile device or desktop computer web browser. The mobile application is created for the Android platform.

4.1 Website application

The website provides a way to navigate through the culture linked data cloud. The website is built using a wiki-based templating mechanism, where every concept of the underlying ontology is associated with a template that defines how the data is presented and which kind of interactions are possible. In the templates, rich widgets for the various data modalities are embedded, including widgets for exploring image collections, timelines for temporal data, maps for geo-spatial data, etc (Fig. 6).

The website also presents data in a number of traditional ways - text descriptions and illustrations of art works, hyperlinks connecting the web pages and so on. At the same time the system allows the integration of more effective tools for data presentation, which provide a brighter use experience and prove to be more fruitful in a process of data exploration. Some of the widgets include:

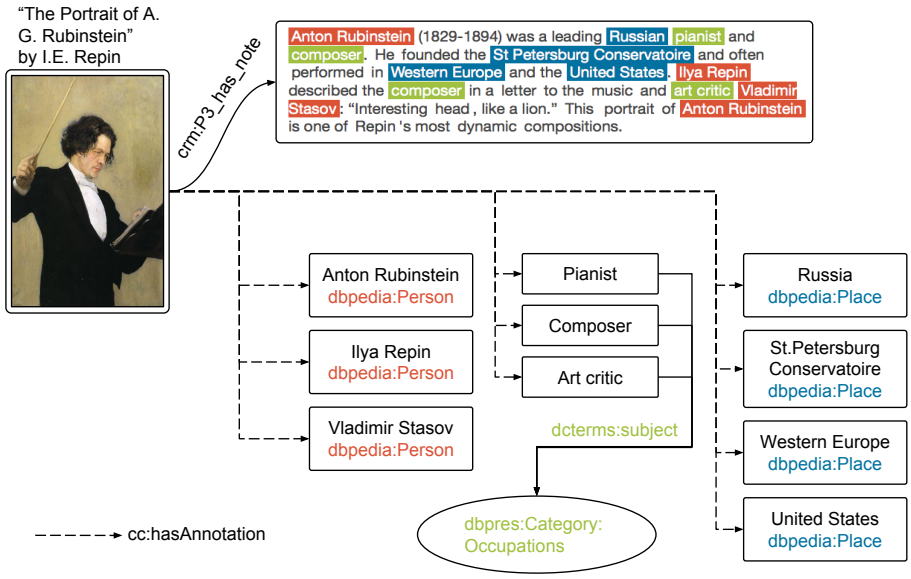

Fig. 5. Example of text annotated with DBpedia Spotlight (persons highlighted with orange, places with blue and occupations with green)

Enriched text Enriched text is a paragraph where some toponyms, people names and dates are linked against the semantic descriptions of external sources, in our case DBpedia and the British Museum. When the user clicks the link the article opens on this same site. This delivers the additional context directly to the user and keeps him on site, while in other systems he would be forced to leave the original resource.

Interactive timelines A Timeline widget enables additional visual demonstration of how long the process took place or in what sequence the things were occurring. For instance, our system employs timelines when we demonstrate the artists lifetimes in relation with the art movement to which they all belong. The other use case for timelines is to place the art objects on artists life span to display his periods of activity and inactivity. This provides a means to learn and discover the facts rather than reading the paragraphs of simple text.

Interactive influence graph The graph of influences illustrates the influences of one artist on another. From this graph many intriguing conclusions could be made: who was the most influential artist in his time, who stands aside in the cultural art process, etc. The end-user can use such graph for finding other artist that can interest him based on the artists that he already knows. The art experts can construct the more complicated graphs showing the connections between artists, art movements, countries, art school, etc.

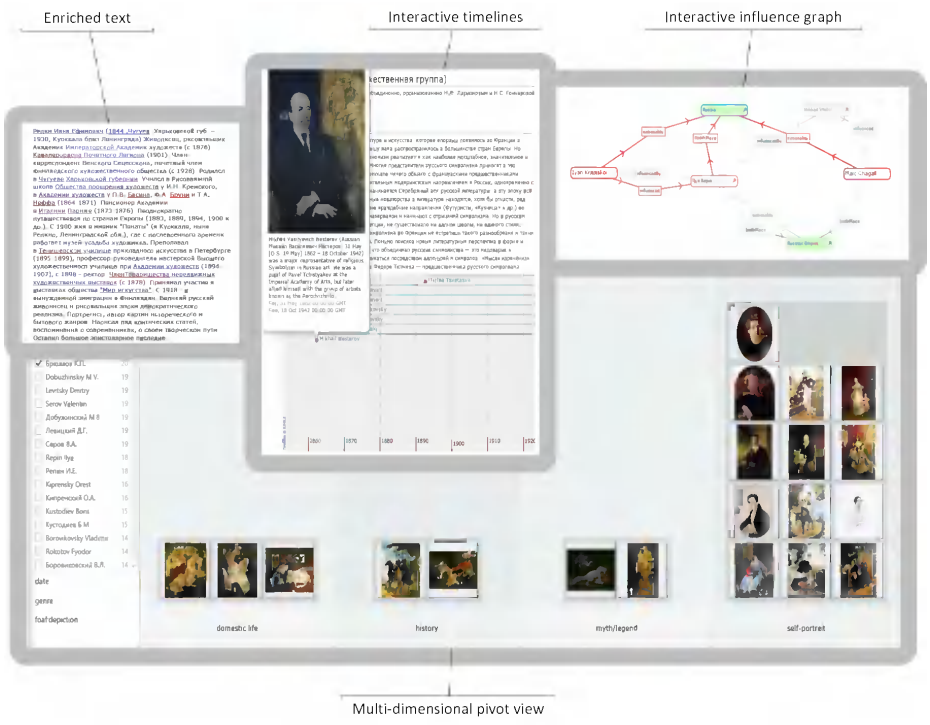

Fig. 6. Sceenshots of the web application

Multi-dimensional Pivot widget The Pivot widget allows to visually explore the data by sorting and filtering it in multiple dimensions. For instance, one can select the artists of 1890s that worked in the genres of portrait and sort their artworks by artists names. This is another exciting and interactive way to learn the content and make fascinating discoveries along the way. It is quite obvious that such tool can be of a great help and simplify the routine work when selecting the artworks for catalogues or when constructing the new exhibition.

Semantic search A search widget has been implemented that allows to visually construct structured, semantic queries against the fundamental relationships of the CIDOC-CRM ontology. The search widget provides auto-suggestions for search terms utilizing an entity index along with a suggestion of relationships that are applicable to a selected entity.

4.2 Mobile application

The mobile applications is intended to make the visit to the museum more informative. It is done by enabling automatic identification of art work by taking snapshot with a mobile device (accuracy about 80 percent). When the artwork is

identified, the mobile app brings additional annotation about it and its creator. Visitors can also rate the artwork and the rating will be shared across the social network so that the user's friends can receive recommendation based on user's ratings of what to see in the museum (We use the most popular social network in Russia - http:www.vk.com).

Hence the key features of application are:

- Artwork recognition by photo
- Provision of information about the artwork
- Ranking of artworks and sharing user rating across the social network
- Delivery of user-tailored recommendations for customized museum visits

The information is presented to the user in highly customized form, individually and deliberately sorted and filtered.

4.3 Added value for end-user

We divided the beneficiaries of our system into two major categories: Art Experts and Museum Visitors. The current state of the system already provides to the organizations e.g.museums an easy and a cheap way to maintain their website and develop information materials. As a result the number of visitors is growing. Such result is achieved due to the nature of semantic technologies, which enables simplified integration of external data sources. Secondly, the data model itself is more transparent and obvious to the experts with no technical background, which leads to easier support of such model. Then, new data acquired to expand the existing system could be integrated from the data providers and thereafter queried like internal with the reference to the internal data model. Finally, displaying external data on-site ensures the user does not leave and teaches him to have a single entry-point to all art-related content.

As for the regular visitors, the system brings them the art-related information in a more interactive, exciting and thought-provoking way. By interacting with widgets the user makes his own trajectory through the site content with regard to his own preferences and interests. The user does not consume the information in a traditional way by going from one link to another, but, more or less, makes his own unique exploration path through the materials, utilizing all the interactive tools. The other important aspect of the system for the common museum visitor is the presence of social features. Now people can follow their friends' paths through museum, share their impressions of art with others and learn more from whom they trust. This process is simplified by the automatic artwork recognition feature implemented in the mobile application allowing users instantly to learn more about the painting by taking a picture of it from their Android device.

5 Evaluation

5.1 Source data and evaluation details

In our system we used three different sources of data:

– Original data from Russian Museum[8] provides basic information about artworks and authors
– The British Museum thesauri[9] were used for describing genres of artworks
– DBpedia[10] was used for adding information about authors, such as date of birth and death, artistic movement author belongs to, persons who influenced author, etc.
– Associated Spotlight annotations were placed in the related section of the authors and artworks pages

For assessing the data quality we have created more specialized versions of several metrics listed in [5]. Our main goal was to find out how successful was the enrichment of the original dataset. A detailed description of all metrics is presented in the following section.

5.2 Metrics description

In this section we will describe metrics we used to assess quality of the resulting dataset. We split all metrics into three different categories:

1. General dataset metrics describe the volume of data we collected. These metrics consists of VoID statistics metrics (number of triples, classes, properties, entities, distinct subjects and distinct objects) and quantity of the main entities from rmgallery.ru (authors and artworks).
2. Original dataset metrics describe completeness of the data from rmgallery.ru. It shows how many artworks entities have specific information (annotation/description, size, genre, creation time) and how many of author entities have the biography information.
3. Interlinking metrics describe how much additional information we gathered by interlinking the original data with DBpedia. The first part of these metrics describes how many owl:sameAs links were created for authors and specifies how many of them were obtained using Wikipedia search or syntactical approaches. We also have counted the number of links to the international and Russian DBpedia. The second part of interlinking metrics depicts how many authors were complemented with specific information from DBpedia (birth/death date, birth/date place, art movement author considered to belong to, influenced and influenced by information).

One special note should be taken regarding a trust assessment. We did not develop objective computable metrics for that task, thus we presented subjective evaluation of provenance in the next section.

[8] http://rmgallery.ru/
[9] http://collection.britishmuseum.org/
[10] http://dbpedia.org/

5.3 Evaluation results

Available information about artworks and authors was transformed to the semantic form and we have the following statistics about its completeness: more than 90% of entities has dimension, genre, creation time and author bio information and 68% of artworks has descriptions. 85% of authors were interlinked with DBpedia which allowed us to enrich 60% of author pages with birth-death dates.

A comparison of the general and original dataset metrics (Table 1) unveils incompleteness in the Russian Museum data, as not all artworks or authors have additional information, such as genre, dimension, etc. This probably could be related to human element in curated museum data.

The results of interlinking metrics (Table 2) were in line with our expectations. Decrease of the added information number correlates to increase of the obtaining information difficulty. For example, birth and death dates are probably the most easy to find information about authors, but to find the information about art movement or persons who influenced the author one probably should look up the specialized literature.

Table 1. Evaluation of general and original dataset metrics (percentage of total number of artworks or authors)

General metrics		Original dataset metrics	
Triples	50795	Artwork descriptions	68% (628)
Classes	15	Artwork dimension	99% (911)
Properties	23	Artwork genre	94% (863)
Entities	8068	Artwork creation time	96% (887)
Distinct subjects	8081	Author's bio	98% (260)
Distinct objects	13861		
Artworks	921		
Authors	265		

The original data is poor in terms of coverage and incompleteness (rather than inconsistency) and lack of semantics. The value of the interlining is not only in the number (volume) of direct links, but in the rich additional data that provides context. The link generation algorithms have been manually tuned, as part of this the generated links have been manually validated. The improvement of the enrichment quality is based on a continuous cycle. The basic metrics allow to manage this process and evaluate results of adding new datasets (for example we are working on adding [11]).

DBpedia Spotlight annotations. The annotations of originally unstructured text from the Russian Museum data gives us a good use case for the end user, as they can observe a dataset while reading information about artworks.

[11] http://www.wikiart.org

Table 2. Evaluation of interlinking metrics (percentage of total number of *interlinked* authors)

Total number of interlinked authors	226 of 265 (85%)
incl. at the first stage	40% (90)
incl. at the second stage	60% (136)
incl. with International DBpedia	81% (183)
incl. with Russian DBpedia	80% (181)
Number of authors enriched with	
birth date	60% (136)
death date	60% (137)
birth place	22% (50)
death place	20% (46)
art movement	13% (30)
"influenced"	4% (9)
"influenced by"	4% (8)

But raw annotations are mostly unusable to perform computational reasoning over data they add, as we do not know how exactly annotations are connected to the text. For example Spotlight added annotation "Finland" to the description of Repin's painting "What an Expanse!". But in this form it is impossible to understand that Repin was inspired by some places in Finland.

In our case DBpedia Spotlight is a good solution for providing a light-weight contextualization. But to make annotations usable in meaningful dataset queries, predicates describing how exactly annotations are related to the text and entities would be needed.

6 Related work

In the cultural heritage domain, Linked Data and Semantic Web technologies have been successfully applied to publish and interlink heterogeneous, semantically rich data. Great amounts of cultural heritage data have been published in national and international portals, such as Europeana[12]. As of today, a number of different ontologies and metadata schemes are used for the representation of the data. CIDOC-CRM is the prevailing model when it comes to the representation of semantically rich cultural heritage data [2]. For example, the British Museum has published their complete data collection as Linked Open Data based on CIDOC-CRM[13]. Notable other sites that have published large collections based on CIDOC-CRM include Claros[14] and the Arches project[15]. The ResearchSpace project[16] is developing a collaborative environment for humanities and cultural heritage research using CIDOC-CRM.

[12] http://www.europeana.eu
[13] http://collection.britishmuseum.org
[14] http://www.clarosnet.org/XDB/ASP/clarosHome/
[15] http://www.getty.edu/conservation/our_projects/field_projects/arches/
[16] http://www.researchspace.org

On the data consumptions side, new applications based on the semantically rich data have been developed that enable new forms of user experience. These range from supporting semantic search in portals to mobile applications. E.g., the SMARTMUSEUM [4] system utilizes an ontology-based representation of content descriptions as a basis for context-aware, on-site access to cultural heritage in a mobile scenario. Applying context reasoning and recommendation algorithms provide users with recommendations for sites, such as museums or buildings of architectural interest, and objects on those sites, such as sculptures or other works of art, and provides explanatory descriptions and multimedia content associated with individual objects. In comparison to the related solutions our project stands out as being an external service to heritage owners, which provides interlinking and search / representation facilities to end-users and third-party applications.

7 Conclusions

In this paper we described a system for semantic publishing, enrichment, search and visualisation of cultural heritage data as a first step towards a Russian Linked Culture Cloud. The system is based on the metaphacts Knowledge Graph Workbench. As a main source of data at the initial step the virtual gallery of the Russian Museum was selected. For transformation and representation of data CIDOC-CRM Ontology was used with extended thesauri from the British Museum repository. Data enrichment is done by DBpedia. We also used the DBpedia Spotlight API to annotate and extract data from unstructured text in the initial data source (annotations, biography and so on).

The performed analyses of user benefits revealed a high demand on the flexible and extensible representation models for building applications that allow to get access to digital cultural heritage. Our system illustrates potentials of semantic technologies for creation of such solutions including semantic search and visualizations both for art experts and regular museum visitors.

One of the features we achieved is to make data deliverable to end users more informative in comparison with any data source provisioning our system. For example, the initial Russian Museum dataset does not contain much information about authors. Interlinking with external sources allowed us to show user additional information about authors, such as date of birth or person they influenced.

Our evaluations show that the enrichment of the limited original dataset was quite successful and automation of this process is efficient.

Future work Some problems raised during the project progress require additional research and further development. The most challenging problems are:

- Expand the number of data sources especially of raw data from heritage institutions. It could require extending thesauri and the CIDOC-CRM ontology in term of new kind of terms and classes.

– Support of collaborative work and contradicting facts representation in the domain ontology for art experts knowledge modelling. This will make the system more natural for the cultural heritage area. On these topics we intend to synergize with the work performed in the ResearchSpace project on argumentation and belief.
– Collecting the user statistics for tracking users trajectory through the site content and analytics of preferences and interests. Such data will allow to build an efficient recommender system.
– Developing a solution for the automated quality assessment of data sources and its trust ranking.

Acknowledgements This work was partially financially supported by Government of Russian Federation, Grant 074-U01.

References

1. Hyvnen, E.: Publishing and Using Cultural Heritage Linked Data on the Semantic Web. Synthesis Lectures on Semantic Web: Theory and Technology, Morgan & Claypool, Palo Alto, CA (2012), available as paperback and ebook (9781608459988)
2. Oldman, D., Doerr, M., de Jong, G., Norton, B., Wikman, T.: Realizing lessons of the last 20 years: A manifesto for data provisioning & aggregation services for the digital humanities (A position paper). D-Lib Magazine 20(7/8) (2014), http://dx.doi.org/10.1045/july2014-oldman
3. Oldman, D., Labs, C.: The cidoc conceptual reference model (cidoc-crm): Primer (2014)
4. Ruotsalo, T., Haav, K., Stoyanov, A., Roche, S., Fani, E., Deliai, R., Mäkelä, E., Kauppinen, T., Hyvönen, E.: SMARTMUSEUM: A mobile recommender system for the web of data. J. Web Sem. 20, 50–67 (2013), http://dx.doi.org/10.1016/j.websem.2013.03.001
5. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: A survey. Submitted to Semantic Web Journal (2014)