A systematic approach towards higher quality linked open data at Nieuwe Instituut

Nora Abdelmageed[1,*], Lois Hutubessy[1]

[1]Nieuwe Instituut, Museumpark 25, 3015 CB Rotterdam, Netherlands

Abstract

Nieuwe Instituut (NI) houses the Dutch National Collection of Architecture and Urban Planning. This collection consists of about 4 million objects, including design drawings, 3D models, and photographs. As part of the program Disclosing Architecture, which is now in its sixth and final year, the Linked Open Data (LOD) project aims to share the richness of data within the collection with the public through semantic web technologies. This project will ultimately facilitate the exchange of cultural heritage data with related national and international institutions. Currently, Nieuwe Instituut (NI)'s collection management system contains inconsistent records due to changes in registration guidelines and the migration from older collection management tools. Without documentation of these guidelines, it is impossible to establish consistent rules for the entire dataset. Yet, clean data is crucial for effectively showcasing NI's collection to the public. In response, this paper introduces a framework for higher-quality LOD data, the Data Cleaning Initiative (DCI). The first implementation of the DCI is through a series of steps planned for the year 2024 with the goal of cleaning and enriching the collection data at NI.

Keywords

Linked Open Data, Cultural Heritage, Data Quality, Entity Linking, Entity Resolution

1. Introduction

The Dutch Cultural Heritage sector has undergone a significant transformation by adopting semantic web technologies and converting their data into Linked Open Data (LOD). The dataset register[1] developed by Netwerk Digitaal Erfgoed[2] lists 71 publishers that expose their datasets in one or more linked data formats[3]. Rijksmuseum is a pioneer in this area, having published its collection data as LOD as one of the first [1], thereby diversifying search results in other applications [2]. Another notable effort is by Van Wissen [3], who targeted named entity extraction from archival records with the help of LOD. However, ensuring high-standard and clean data remains an open challenge.

Nieuwe Instituut (NI)[4] houses the National Collection of Architecture and Urban Planning, which consists of approximately 4 million items ranging from design drawings to photographs

SEMANTiCS - 20th International Conference on Semantic Systems, Sep 27–19, 2024, Amsterdam, Netherlands

*Corresponding author.

n.abdelmageed@nieuweinstituut.nl (N. Abdelmageed); l.hutubessy@nieuweinstituut.nl (L. Hutubessy)

0000-0002-1405-6860 (N. Abdelmageed)

[1]https://datasetregister.netwerkdigitaalerfgoed.nl/

[2]https://netwerkdigitaalerfgoed.nl/

[3]Access June 2024 using search feature with n-triples*, rdf+xml*, turtle, and trig.

[4]https://nieuweinstituut.nl/

and 3D models. As part of the program Disclosing Architecture (AD)[5], which is now in its sixth and final year, LOD project aims to share the richness of data within the collection with the public through semantic web technologies. In parallel, NI develops a new online Collection platform that will further increase accessibility to a wider audience, facilitating discovery by design. While this platform is set to launch in November 2024, the underlying data is already exposed through a separate endpoint[6], with more than 18 million triples available in the customized Triply[7] environment.

In this paper, we propose our systematic approach Data Cleaning Initiative (DCI) at NI to enhance the data quality of our LOD. We explain its scope, tasks, and how we implement such an approach in practice. DCI is a vision that could be applied in any Cultural Heritage instituut for cleaning and enriching their LOD.

1.1. Motivation & Problem Definition

We aim to increase the exposure of Architecture and Urban Planning data. For instance, we built a central point or an encyclopedia on top of the Dutch Heritage Data collection in NI. However, due to human errors and changes in guidelines and collection registration tools, the data has become heterogeneous and inconsistent on a large scale. As such, publishing the current data version may not be the best approach. The data currently contained in the collection management system (Axiell Collections[8]), while suitable for internal purposes, would benefit from further cleaning and enrichment to ensure a higher quality of data for public use. This would facilitate collaborations with third parties and attract a wider audience.

Records in NI's collection management system are inconsistent due to changes in registration guidelines and migrations from older collection management tools. In addition, there is no documentation of these guidelines, making it difficult to establish rules for the entire dataset. Cleaning these records in their entirety is challenging for two reasons. On the one hand, understanding the meaning of sparse heterogeneous records within the same catalog is difficult. E.g., "Library catalog" registers 16 types, including books and audio materials. On the other hand, the sheer volume of heritage records adds to the complexity.

1.2. Objectives & Tasks

The Data Cleaning Initiative (DCI) has two main objectives: (i) Proposing to split catalogs containing broad relations of data into smaller, closely related datasets. This facilitates the semantic grouping of the original catalogs within the collection management system. (ii) Applying cleaning and enrichment strategies to the resulting semantic categories to improve the data quality.

We propose four categories of data cleaning and enrichment tasks in the context of DCI. We define these main categories as:

[5] https://nieuweinstituut.nl/en/projects/architectuur-dichterbij
[6] https://collectiedata.hetnieuweinstituut.nl/the-other-interface/knowledge-graph/sparql
[7] triply.cc
[8] https://www.axiell.com/solutions/product/axiell-collections/

Table 1
People & Institutions Examples

No.	use_count	name	name.type
1	10	Erp, T. van	?
2	10	Erp, Th. van	person & author
3	2	Vugt, Theo van	author
4	3	Erp, Theo van	person & author
5	3	Erp, Theodoor van	person & author

1. **Data Cleaning**: This category involves all tasks concerning primitive data cleaning. E.g., handling inconsistencies like using different formats or tackling missing values. The latter influences the guidelines for filling in this metadata. E.g., discovering a potential required field.
2. **Entity Resolution**: This category aims at similar entities discovery and grouping. Since all data is manually entered by domain experts, they might use different representations describing the same entity. E.g., *Doesburg, Theo van* is a different representation for *Doesburg, Th. van*.
3. **Entity Linking** This group maps internal records of our heritage data collections to external resources, or Knowledge Graphs (KGs), e.g., Wikidata[9]. For example, *Doesburg, Theo van* would be mapped to `wd:Q160422`[10].
4. **Entity Enrichment** This category aims to fetch external properties and pieces of information that exist in external resources and do not exist in the local collection management system. E.g., we save the image of *Doesburg* from Wikidata in our Axiell Collections.

Table 1 summarizes the scope of each task of the DCI. The first record shows an empty `name.type` where the Data Cleaning task covers this kind of issue. Records 2, 4, and 5 represent the same person in different formats; Entity Resolution groups them all as the same person. These records after grouping, Entity Linking will map these records to the Wikidata page of Theodoor van Erp -`wd:Q2759953`[11]. Finally, we can store new information in our catalog to enrich the displayed data at the application level.

2. Approach

In this section, we describe the initiative's approach. Initially, we explain the concept of semantic grouping; then, we give details of the DCI pipeline.

2.1. Semantic Grouping

Currently, domain experts enter metadata for multiple semantic categories into the same Axiell catalog. For instance, they use "People & Institutions" for registering persons, architects, publishers, universities, etc. The same case applies to the "Library catalog", where domain

[9]https://www.wikidata.org/wiki/Wikidata:Introduction
[10]http://www.wikidata.org/entity/Q160422
[11]https://www.wikidata.org/wiki/Q2759953

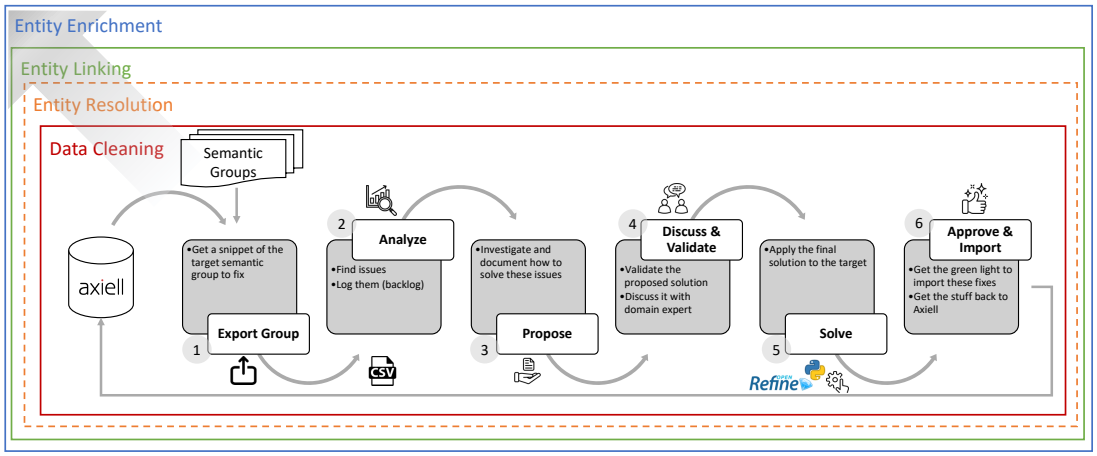

Figure 1: DCI Proposed Pipeline

experts register sixteen semantic types, including Books, serials, Audio/Visual Material, and Articles. This yields a large catalog but is sparse in terms of the metadata. Thus, we decided to split each catalog into smaller chunks that share the semantic type, i.g. extracting only Books from the Library. The main idea is the DCI relies on the semantic division of the Axiell catalogs. By this means, it facilitates human intervention and yields a better statistical view of the data. The target of this phase of the DCI is to obtain named groups to be cleaned.

Creating a semantic group requires determining representative fields for each group. To find these groups, we instigated Axiell Collections' fields for each category. In addition, we held several meetings with domain experts to determine those fields and ensure their correctness and scope. For instance, "People & Institutions" catalog contains for example: name, birthDate, birthPlace, deathDate, deathPlace, biography, ISBN_publisher_prefix. This group of fields represents two semantic groups: 1) Persons, and this group contains: name, birthDate, birthPlace, deathDate, deathPlace, biography. 2) Institutes and this group contains name and ISBN_publisher_prefix.

2.2. Pipeline & Workflow

Figure 1 depicts the proposed pipeline for DCI. The figure represents the separate steps that we follow til we reach the final goal or the end of the year for each semantic group. Our pipeline consists of six iterative steps taking into consideration four scopes of work: Data cleaning, entity resolution, entity linking, and entity enrichment. Our pipeline starts with: 1) **Export** the target semantic group given the representative group fields in a CSV file. The CSV file allows batch processing and facilitates the general overview of the exported records. 2) **Analyze** the data and log the encountered issues in our backlog system. 3) **Propose** a solution for individual issues and determine if it is possible to solve them automatically or if it needs manual intervention. 4) **Discuss** the proposed solution with domain experts and stakeholders to **validate** it. If it is not a valid solution, we go back to the proposing stage otherwise, we move to, 5) **Solve** the target issue by applying the proposed solution on the CSV file directly. Finally, 6) **Approve** and **Import**, we seek approval from our manager, and if agreed, then we import the CSV with fixes back to the Axiell Acceptance environment.

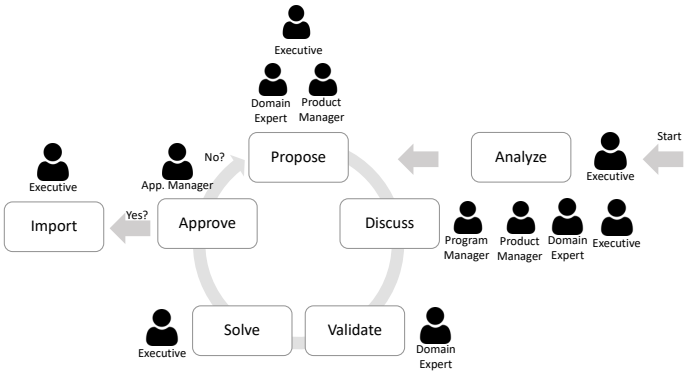

Figure 2: Simplified workflow of the DCI

Figure 2 shows a simplified data flow diagram that explains the interaction among DCI actors regarding their tasks. It starts with an executive that analyzes the target semantic group to be cleaned. Then, the executive proposes a solution that starts an iterative process (discuss, validate, solve, and approve). The domain expert is the only required actor to validate a proposed solution. If and only if the application manager approves the solution and its results, the executive can import the fixed fold back to the collections management system.

Acknowledgments

We would like to thank our domain experts, Inge van Stokkom, Christel Leenen, Ernst des Bouvrie, Evelien Dekker, Kelly James, and program manager Gijs Broos, Nieuwe Instituut (NI). Moreover, we would like to thank the Dutch Ministry of Culture, Education, and Science for funding the Disclosing Architecture program.

References

[1] C. Dijkshoorn, L. Jongma, L. Aroyo, J. van Ossenbruggen, G. Schreiber, W. ter Weele, J. Wielemaker, The rijksmuseum collection as linked data, Semantic Web 9 (2018) 221–230. URL: https://doi.org/10.3233/SW-170257. doi:10.3233/SW-170257.

[2] C. Dijkshoorn, L. Aroyo, G. Schreiber, J. Wielemaker, L. Jongma, Using linked data to diversify search results a case study in cultural heritage, in: Knowledge Engineering and Knowledge Management - 19th International Conference, EKAW 2014, Linköping, Sweden, November 24-28, 2014. Proceedings, volume 8876 of *Lecture Notes in Computer Science*, Springer, 2014, pp. 109–120. URL: https://doi.org/10.1007/978-3-319-13704-9_9. doi:10.1007/978-3-319-13704-9_9.

[3] L. van Wissen, C. Latronico, V. Zamborlini, J. Reinders, C. van den Heuvel, Unlocking the archives. a pipeline for scanning, transcribing and modelling entities of archival documents into linked open data.(short paper), in: DH Benelux 2020-Online:# Goesonline, 2020. URL: http://2020.dhbenelux.org/. doi:http://10.0.20.161/zenodo.3862817.