Close-reading of Linked Data: a practical case study on the quality of online authority files

# Introduction

More and more cultural institutions use Linked Data principles[[1]](#footnote-1) to share and connect their collection descriptions. In the archival field, initiatives emerge to exploit data contained in archival descriptions and adapt encoding standards to the semantic web[[2]](#footnote-2). In this context, online authority files can be used to enrich metadata. However, relying on a decentralized network of knowledge bases such as Wikidata, DBpedia or even Viaf has its own difficulties. This paper aims to offer a critical view of these linked authority files by adopting a *close-reading* approach. Through a practical case study, we intend to identify and illustrate the possibilities and limits of RDF triples compared to institutions’ less structured metadata.

Our paper is an invitation to travel in an unexpected way: diving through the Linked Open Data cloud[[3]](#footnote-3) by a “thought experiment”. Let’s suppose that we have a smart robot able to use SPARQL endpoints to jump from one dataset of RDF triples to another—let’s call it Alex. Alex is a kind of computerized archivist. We will ask him to use the Linked Data to get information about Henry Carton de Wiart (1869–1951), a famous Belgian personality from the early 20th century. Part of a large noble Walloon family, Carton de Wiart has been minister several times, Prime Minister, president of many councils and organizations, and a lawyer and a writer as well. The Belgian city Liège owes one of its nicknames to one of his books, *La Cité ardente*. Moreover, he has been honored by several awards and was in contact with many well-known personalities, from the poet Verlaine to the President Wilson. His life contains therefore sufficiently facets to act as a case study to compare how linked open data can reconstruct a biography compared to more traditional information sources.

While many studies about Linked Open Data adopt “big data” approaches based on methods such as data mining or network analysis, they tend to analyze the data quality of cultural heritage Linked Data only through the prism of quantitative and comparative analysis[[4]](#footnote-4). In this paper, we opt for a close-reading approach focusing on a single individuality, in order to analysis how the triple structure succeeds in capturing biographical data, compared with more traditional authority files.

# Continuum

This paper wishes to reflect on the continuum of different documentation practices and methods, from the traditional paper-based narrative to very structured data which are currently published as triples in knowledge graphs. At one end are unstructured data, like the national biography of Henry Carton de Wiart, an eight pages well-written text digitized and published online in a PDF version [cf. reference: <http://www.academieroyale.be/Academie/documents/FichierPDFBiographieNationaleTome2102.pdf>, pp. 86–94]. At the other end, stand the most structured data, which correspond to Linked Data, such as the Wikidata resource for Henry Carton de Wiart[[5]](#footnote-5), its Viaf entity[[6]](#footnote-6) or its French DBpedia page[[7]](#footnote-7). In between are established more isolated structured data like RDF triples from archives or libraries repositories [e.g. Data.bnf.fr: <http://data.bnf.fr/12062835/henry_carton_de_wiart/>], archival descriptions in XML [e.g. the EAC-cpf file from the State Archives of Belgium: the <https://search.arch.be/eac/xml/eac-BE-A0500_007556_FRE.xml>] or Wikipedia pages [e.g. the English version: <https://en.wikipedia.org/wiki/Henry_Carton_de_Wiart>].

While the most structured resources display facts (such as a birth date or death date) very clearly, in a machine-readable format, we can wonder if they provide as many details as a more classical format type, such as the national biography: do they mention his origins from Hainaut, his correspondence with an apostolic vicar living in Brazzaville, or that day in August 1914 when the Belgian king asked him to lead an extraordinary mission to the United States? This question can be mapped to the central premise of Lev Manovich’s book “The Language of New Media”, where he stressed the tension between traditional narratives and databases[[8]](#footnote-8). Through this case study, we aim to extend this thinking and confront various forms of structured sources and to observe to what extent they are consistent and complementary.

# Diving into the LOD cloud

Alex-the-robot, the main fictional character of our case study, can be seen as a cousin of the modern “robot journalists”, these systems able to write short texts—for example weather or financial reports—using structured information[[9]](#footnote-9).

Alex, for its part, would write biographical texts based on information extracted from the Linked Open Data cloud. It will proceed like a historian who collects pieces of information from various sources, to gradually be able to describe a personality. But instead of reading archives or books, Alex would perform queries on the web and then interpret the results, at least if these results are in a machine-readable format such as XML, JSON, an RDF serialization and so on. Even if it does not understand natural language, it is able to write basic sentences, for example that someone was born that year, in that location, and practiced this or that activity.

As we said, Alex’s first mission in this experiment is to collect as many triples as possible about Henry Carton de Wiart. Its starting point will be for example DBpedia, often considered as the central point of the Linked Open Data cloud, since many datasets of the LOD cloud are linked to it[[10]](#footnote-10). Thus, we will ask Alex to go to the Carton de Wiart’s English DBpedia page[[11]](#footnote-11), to collect RDF data from this page and then to follow all the links to external databases (owl:sameAs).

Figure 1 shows all the links between the knowledge bases used in this experiment. As aforementioned, the entry point is DBpedia. All the paths explored by Alex are visible. From DBpedia, Alex goes to various versions of DBpedia in other languages, then to YAGO, to Freebase (which is discontinued) and finally to Wikidata. In our figure, the size of each node is proportional to the number of outgoing links to other databases. The size of Wikidata’s bubble means it contains a lot of external links: to VIAF, BNF or the Library of Congress authority ID. Whenever our robot comes in a knowledge base, it collects all the information it finds: Henry Carton de Wiart is a politician; he was born in Brussels; he wrote the novel “La Cité ardente” and so on.

# RDF Triples Harvesting

During this process, Alex harvested more than 22,000 RDF triples about Carton de Wiart. While this amount can seem huge, it must be noted that the clear majority of these triples are meaningless. RDF can be a verbose format, requiring sometimes a lot of triples to express something quite basic[[12]](#footnote-12). Thus, once the useless triples eliminated, there are about 1,500 triples left, which use some 240 distinct properties.

However, each database using its own schemas, these properties are often redundant. To encode a person’s date of birth, for example, YAGO uses the property <http://YAGO-knowledge.org/resource/infobox/en/birthdate>, while DBpedia uses <http://dbpedia.org/ontology/birthDate>. In order to facilitate the analysis, we have roughly classified the various properties into eleven empirical classes: affiliations, appellations, category, dates, descriptions, identifiers, locations, miscellaneous, professions, relations, works. Figure 2 shows the proportion of triples in each of these classes. Triples have been placed in their respective category according to the property used in the triple. For instance, the property <[http://www.w3.org/2000/01/rdf-schema#label](http://www.w3.org/2000/01/rdf-schema" \l "label)> correspond to the “appellation” category, which encompasses all the different labels under which Henry Carton de Wiart is known. There are many, especially in VIAF, which records the different labels used by libraries: Carton de Wiart‏, Henry‏ ; Carton de Wiart, Henry,‏ Comte, 1869-1951‏ ; Carton de Wiart, Henry‏‎ (1869-1951). The “category” class is the largest because it includes Wikipedia and YAGO categories, which are numerous—for example <http://YAGO-knowledge.org/resource/wikicat\_Walloon\_people> or <http://dbpedia.org/resource/Category:Members\_of\_the\_Brussels\_Guild\_of\_Saint\_Luke>.

# Semi-Automated Biography

Inspired by the aforementioned reports produced by robot journalists, we tried to see what a biography based on RDF triples would look like. To create this biography, we have synthesized the information contained in the RDF triples collected by Alex. For the sake of clarity, we present in Figure 3 a simplified version, based only on Wikidata and three different versions of DBpedia (EN, FR, NL). This means that this experimental biography does not contain the list of books that Carton de Wiart wrote or he collaborated on.

As you can see (Figure 3), all the text fits in less than 20 lines. The result is far from Carton de Wiart’s biography displayed by the State Archives of Belgium, let alone the 8-pages biographical note in the National Biography. This confirms our first hypothesis: the overwhelming number of triples are redundant. Most of them express the same elements and often only offer basic information such as birth and death dates. We have highlighted the most problematic excerpts. The text in yellow points up contradictions. Firstly, it is not easy to know how to write Henry. The spelling changes according to the language. In Wikidata, the “given name” property is “Henry”, whereas DBpedia uses “Henri”. Secondly, the Dutch version of DBpedia[[13]](#footnote-13) wrongly asserts that Carton de Wiart was a member of the “cdH” party—a political party created in 2002, after his death. Unlike most knowledge bases which correctly indicate that Wiart belonged to the Catholic party.

Moreover, the information is not always as structured and clear as we could expect. For example, much of the information in DBpedia or YAGO is not fully explicit. Thus, the Dutch version of DBpedia mentions that Carton de Wiart belongs to the Wikipedia’s “List of Belgian ministers of Justice”. However, this piece of information would not be fully exploitable by a robot, not to mention the fact that the years of his tenure as Minister of Justice are not even indicated. Furthermore, the properties used by YAGO or DBpedia are often poorly described or not described at all. It is therefore difficult to ascertain what the “successor” property (highlighted in green in Figure 3) covers. Successor of whom? At which position?

Similarly, it is doubtful that Carton de Wiart spent his entire life at “16 rue de la Loi”, residence of the Belgian Prime Ministers presented as his home.

Finally, it appears that this text lacks essential biographical information, for example about his studies or his family. Did he have a wife, children, cousins, parents, siblings? This brings us to the next question: among this missing information, which ones could be easily added and which ones would be more difficult or impossible to translate into RDF?

# Manual Triplication

After this first exploration with the help of Alex-the-robot, we wanted to take a close look at the data contained in less structured resources. We try to do that by acting as if we were an archivist wishing to inject unstructured biographical information into the Linked Datacloud, in other words, what the options would be to “manually” triplify extra information.

Different types of resources have been used: a biographical note from the Biographie nationale de Belgique, an EAC-CPF and an EAD files from the State Archives. During this triplification process, we considered each sentence, one after the other, and tried to extract and translate every single piece of information about Carton de Wiart into RDF triples (subject, predicate, object). For example, a single sentence like “Issu d’une famille de la noblesse catholique, Henry Carton de Wiart effectua ses humanités au collège d’Alost puis au collège Saint-Michel à Bruxelles, avant d’entreprendre des études à l’Université libre de Bruxelles, où il obtint en 1890 son doctorat en droit” need at least seven different triples.

The whole triplification resulted in about 300 statements, containing far more complete and detailed information than the Linked Data RDF triples. During this process, we observed several aspects which can lead to the loss of information, which in turn renders it difficult to deliver a clear and representative description of Henry Carton de Wiart. We have identified three types of challenges, which are respectively related to the data, to the triple structure and to the vocabulary. (mieux structurer les trois points suivants)

Albeit the completeness of the data available can constitute a limit, a more challenging point is the granularity: besides practical aspects and available human resources, at what level of detail should Henry and his life be described? For example, in an archival context, should we only use the biographical section of the EAD or also take into consideration other levels of description? Thus, letters exchanged with other personalities could help to get a list of people connected to Henry. But it raises another question: what should we do with all persons related to him? Indeed, we may also include professions and personal information for friends or family. It can be difficult to determine how far it is relevant to value this contextual information in the form of triples. Will they represent a plus value for users or information overload in the LOD cloud?

The triple structure of RDF requires a different way of expressing things. For instance, the translation of one single sentence results sometimes in 5 or 6 different triples because of the reification principle.

—Additionally, if a piece of knowledge can be expressed by words in a biography, like “Henry had 4 brothers”, its full description sometimes has to be inferred from other statements in the RDF language (for example counting the number of triples using a property like “brotherOf”).

# The vocabulary

To express facts about someone in RDF triples, several RDF vocabularies can be used to specify the relationship between a subject and an object.

—Again, we have noticed some limits related to granularity.

Thus, in the family context, depending on what vocabulary we used, we were able to specify that someone was an uncle, OR merely describe the fact that he was a relative. In this case, we would lose details and have uncle and cousin described by the same vague property. In the field of social relationship, one vocabulary called “relationship” allows us to qualify “acquaintance of”, “friend of”, “a close friend of”, “has met”, etc. Although this representation is useful, sometimes it appears quite difficult to evaluate the specific nature of a relationship between two dead people…

—During the triplification process, most of the data has been translated in RDF triples without issues. In a few cases, we lacked vocabulary terms to express more atypical statements, such as details about Henry’s personality, or mentions about his activism in the context of social struggles.

Obviously, it raises the question of the long tail… to what extent should we create new properties for each specific case?

# Conclusion, Next steps

We have seen in this experiment that the linked open data cloud contains a lot of triples about Carton de Wiart, but the total amount of information is quite poor. We’ve also seen that many other biographical elements can be added as RDF triples, and that there is often a controlled vocabulary that can express them.

But creating your own RDF files requires time and skills. So, another approach might be to automatically feed Wikidata, which can be edited by users unlike DBpedia. In the coming months, we will explore these tracks. We will also try to generalize this first small experiment on a wider panel of personalities and entities, for example organizations or historical events. In other words, we will leave the close reading approach to return to a more classical distant reading.

1. Described in Berners-Lee, T. (2011). Design issues: Linked data (2006). URL: https://www.w3.org/DesignIssues/LinkedData.html (accessed 14 September 2018). Namely : 1° Use [URIs](https://en.wikipedia.org/wiki/Uniform_resource_identifier" \o "Uniform resource identifier) to name (identify) things; 2° Use [HTTP](https://en.wikipedia.org/wiki/Hypertext_Transfer_Protocol" \o "Hypertext Transfer Protocol) URIs so that these things can be looked up; 3° Provide useful information about what a name identifies when it’s looked up, using open standards such as [RDF](https://en.wikipedia.org/wiki/Resource_Description_Framework" \o "Resource Description Framework), [SPARQL](https://en.wikipedia.org/wiki/SPARQL" \o "SPARQL), etc; 4° Refer to other things using their [HTTP](https://en.wikipedia.org/wiki/HTTP" \o "HTTP) [URI](https://en.wikipedia.org/wiki/URI" \o "URI)-based names when publishing data on the Web. [↑](#footnote-ref-1)
2. Gracy, K. F. (2015). Archival description and linked data: a preliminary study of opportunities and implementation challenges. Archival Science, 15(3), 239–294. [↑](#footnote-ref-2)
3. An attempt to map the Linked Data ecosystem maintained since 2007. The August 28, 2018 version contains 1224 linked datasets: <https://lod-cloud.net/> (accessed 14 September 2018). [↑](#footnote-ref-3)
4. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2012). Quality assessment for linked data: A survey. Semantic Web, 7(1), 63–93. [↑](#footnote-ref-4)
5. [https://wikidata.org/wiki/Q14990](https://www.wikidata.org/wiki/Q14990) (accessed 14 September 2018). [↑](#footnote-ref-5)
6. <https://viaf.org/viaf/24623115> (accessed 14 September 2018). [↑](#footnote-ref-6)
7. <http://fr.dbpedia.org/page/Henry_Carton_de_Wiart> (accessed 14 September 2018). [↑](#footnote-ref-7)
8. Manovich, L., Malina, R. F., & Cubitt, S. (2001). The language of new media. MIT press. [↑](#footnote-ref-8)
9. Dierickx, L. (2017). News bot for the newsroom: how building data quality indicators can support journalistic projects relying on real-time open data. In Global Investigative Journalism Conference 2017 Academic Track. Investigative Journalism Education Consortium. [↑](#footnote-ref-9)
10. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In The semantic web (pp. 722–735). Springer. [↑](#footnote-ref-10)
11. <http://dbpedia.org/page/Henri_Carton_de_Wiart> (accessed 14 September 2018). [↑](#footnote-ref-11)
12. Especially when institutions like the Library of Congress provide metadata about triples using RDF reification. [↑](#footnote-ref-12)
13. <http://nl.dbpedia.org/page/Henri_Carton_de_Wiart> (accessed 14 September 2018). [↑](#footnote-ref-13)