From Paper to Digital Trail: Collections on the Semantic Web

Klijn, Edwin

Empfohlene Zitierung / Suggested Citation:
Klijn, E. (2020). From Paper to Digital Trail: Collections on the Semantic Web. *Historical Social Research*, *45*(4), 244-262. https://doi.org/10.12759/hsr.45.2020.4.244-262

From Paper to Digital Trail: Collections on the Semantic Web

Edwin Klijn *

Abstract: *»Vom Papier zur digitalen Spur: Sammlungen im Semantic Web«.* Historical research on World War II and the impact of large-scale violence largely depends on the availability of source materials: diaries, newspapers, eyewitness accounts, archival documents, photographs and videos, etc. Currently, these resources are held by a large number of memory institutions, often in analogue formats. For scholars, it can be challenging to find out which collections are relevant for their research and also what information can be found in these collections. In this article it is argued that Semantic Web technologies, together with new digital tooling to automatically open up collections and interlink their contents, have the potential to revolutionize future access and use. By making the contents of collections machine-readable and enriching them with links to reference data, a shift can be made from a "web of documents" to a "web of data." By publishing all contents as linked open data, domain experts in research infrastructures (RIs) and thematic aggregators (TAs) are enabled to add their own "thematic" layers to the data, thus empowering themselves and others to explore the data in new, more sophisticated ways. Since we are only at the start of this development, the author advocates a close cooperation between archives, libraries, and museums (ALMs) and domain experts.

Keywords: Thematic aggregator, research infrastructures, digital humanities, collections, Semantic Web, memory institutions, linked data.

1. Introduction

Employing thematic approaches to collections is not something that simply happened overnight. The use of paper catalogues, indices, collection guides, registers, and other handy analogue means via which to gain an overview of a given collection have a long tradition. What is new is that digital technologies now enable users to create copies of the analogue originals that can easily be distributed, reused, improved, or adapted. The web is a dynamic 24/7 playground in which to share and enrich data while reaching out to a global audi-

* Edwin Klijn, Netwerk Oorlogsbronnen, NIOD Institute for War, Holocaust and Genocide Studies, Herengracht 380, 1016 CJ Amsterdam, The Netherlands; edwin.klijn@oorlogsbronnen.nl.
Acknowledgements Lizzy Jongma (Netwerk Oorlogsbronnen).

ence. In a very short time, the web has become a vast network of knowledge that will outlive us all.

Pathfinding on the web has become an acquired skill. To avoid drowning in information overload, most people simply start with Google or Wikipedia. Audiences interested in World War II are often unfamiliar with professional institutions that hold collections with original objects, documents, photographs, film footage, printed documents, etc. Many archives, libraries, and museums (ALMs) only offer access to their collections via their own websites. This situation forces web users to hop from website to website, inevitably missing out on unexpected or unknown places that might also possess relevant information. Researchers looking for data on collaboration, migration, and mass violence during World War II and its aftermath have few places to search because currently, most data still remain hidden in card indexes, paper documents, and analogue publications.

In the last few years, there have been several attempts to transcend the institutional perspective and approach collections from a thematic, "bird's eye" view. Internationally, significant initiatives came from two different directions: the cultural heritage field, and the research community. In the heritage field, so-called "thematic aggregators" (TAs) emerged that harvested metadata from collections and built web portals that served as virtual guides to lead the way to the dispersed data providers. Presumably, the strongest driving force behind this development has been the EU-funded heritage platform Europeana.[1] In 2015, an aggregator was defined by Europeana as "an organization that collects, formats and manages metadata from multiple data providing partners, offering services such as supplying their own portal and acting as a data provider to Europeana" (Europeana 2015). TAs are aggregators with a scope that is confined to a particular topic, for example, fashion, the history of film, or the cultural heritage of a specific region.[2] Aggregators automatically "harvest" metadata from different collection management systems and make them searchable as a whole. Often, they add an "enrichment" layer to the original metadata to improve access to the data thematically.

In the field of science and the humanities, so-called "research infrastructures" (RIs) have emerged. According to the definition of the European Commission, RIs "are facilities, resources and services used by the science community to conduct research and foster innovation."[3] Successful RIs "enable the greatest *discoveries* in science and technology, *attract researchers* from around the *world* and build bridges between research communities" (European Com-

[1] Europeana-website <http://www.europeana.eu> (Accessed June 12, 2019).

[2] For fashion, see *Modemuze* <www.modemuze.nl> (Accessed June 12, 2019); for film, see *European Film Gateway* <http://www.europeanfilmgateway.eu> (Accessed June 12, 2019).

[3] See website European Commission <https://ec.europa.eu/info/research-and-innovation/strategy/european-research-infrastructures_en> (Accessed June 12, 2019).

mission 2019 “About Research Infrastructures”). In practice, RIs are platforms upon which data are shared and tools are developed that support scholars in their research. Within the arts and the humanities, examples of such RIs are DARIAH, CLARIN-EU, and PARTHENOS (European Commission 2017).[4]

At first sight, RIs and TAs seem worlds apart. Whereas the target audience of an RI is primarily the scholar, most TAs reach out to the general public. RIs are generally centered upon developing tools to support data-driven research, while TAs are mainly portals that refer users to the memory institutions (“data providers”) holding the collections. RIs focus on developing customized tools for research in the closed circuit of domain specialists, whereas TAs operate in the realm of cultural heritage. By thematically repackaging the collection data, they try to attract new audiences. RIs focus on the re-use of digital data, while TAs focus on optimizing searchability through the data.

Although RIs and TAs differ in many ways, upon closer inspection, similarities also become apparent. RIs and TAs share an ability to mobilize a community of people with domain-specific expertise and interests. Also, they often use the same raw materials to feed their services: data from the collections of ALMs. Currently, both RIs and TAs are busy exploring ways to guide their users through the overload of often unstructured data. Although this is not always acknowledged, the user demands of researchers are primarily not that different from those of the average end user. Both expect free, quick, online access to data, with basic features that allow them to make their own selections and download data for their own purposes.

Recently, RIs and TAs have started to grow closer to each other. Europeana’s transformation from portal to platform in the last few years illustrates this trend. “Portals are for visiting, platforms are for building on,” as Tim Sherratt from the Australian heritage search engine Trove briefly characterizes this shift in focus (Sherratt 2013). Many TAs do not restrict themselves to “harvesting” data, but often develop extra services built upon the data. At the same time, many RIs have strengthened their data collection efforts. Both have a keen interest in finding new ways to “dig” into the huge amounts of diverse data available.

Arguably the most important common denominator of RIs and TAs is that they share a strong dependency on the availability of useable data from memory institutions. The current digital data supply from ALMs is still fundamentally haphazard and inconsistent in nature. According to a survey held amongst European ALMs in 2017, approximately 40% of all collections have

[4] See *European Research Infrastructure for Language Resources and Technology (CLARIN EU)* <https://www.clarin.eu> (Accessed June 12, 2019); Digital Research Infrastructure for the Arts and Humanities (DARIAH) <http://www.dariah.eu> (Accessed June 12, 2019); PARTHENOS <http://www.parthenos-project.eu> (Accessed June 12, 2019).

not been catalogued in a collection database.[5] Although most archives are involved in digitization activities, only 10% of all their holdings have actually been digitized (Nauta, van den Heuvel, and Teunisse 2017). Then there is also the "digital drama" of the variety of different metadata standards used by memory organizations, which seriously obstructs the interoperability of collection data.[6]

Funding agencies investing in RIs or TAs are generally not inclined to spend money on mass digitization. Digitization is considered to be the responsibility of collection holding institutions. Most of them, however, lack the resources to digitize at their own expense. Consequently, both RIs and TAs are building their services on top of just a few useable corpora. For many fields of interest, including war studies, a wealth of research data remains hidden in analogue formats in archival depots. Also, there are legal and ethical barriers that complicate research into collaboration and large-scale violence during World War II. For instance, privacy legislation restricts the possibility to open up collections with indices of person names. Even if legally allowed, many ALMs are very reluctant to release ethically sensitive data in relation to collaboration, national socialism and fascism, perpetrators, etc.

While there are obvious obstacles and limitations, the direction into which the current web is developing offers memory institutions a flexible environment for interlinking, cross-collection access, and data enrichment. TAs and RIs can play an important role in speeding up the innovation process in the cultural heritage field. When, in 2001, the founding father of the web, Tim Berners-Lee, together with James Hendler and Ora Lassila, launched the concept of the so-called Semantic Web, the rationale behind the concept was both simple and visionary:

> The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation. The first steps in weaving the Semantic Web into the structure of the existing Web are already under way. In the near future, these developments will usher in significant new functionality as machines become much better able to process and "understand" the data that they merely display at present. (Berners-Lee, Hendler, and Ora Lassila 2001)

In 2003, Director Seamus Ross of Glasgow University's Humanities Advanced Technology and Information Institute (HATII) already recognized the huge potential of the Semantic Web for memory institutions: "The Semantic Web will enable the heritage sector to make its information available in meaningful

---

[5] Europeana DSI 2–Access to Digital Resources of European Heritage. D4.4.Report on ENUMERATE Core Survey 4 (2017) p. 16, see <https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/ENUMERATE/deliverables/DSI-2_Deliverable%20D4.4_Europeana_Report%20on%20ENUMERATE%20Core%20Survey%204.pdf>.

[6] "Digital drama" is a phrase coined by a Dutch national newspaper to describe the shattered landscape of digitized heritage collections (Berkhout 2011).

ways to researchers, the general public, and even its own curators" (Ross 2003, 8).

Sixteen years later, Ross's vision of "an interoperable Semantic Web for heritage resources" (Ross 2003, 8) is gradually becoming a reality. The groundbreaking Finnish project CultureSampo was one of the first initiatives in the heritage field that brought the concept of the Semantic Web into practice (Hyvönen et al. 2009). In the Netherlands, since 2015, the project Dutch War Collections (Netwerk Oorlogsbronnen, NOB) has been experimenting with creating an interlinked semantic network of World War II data, based on the philosophy and technologies that shape the Semantic Web. NOB is a cooperation between ALMs and World War II collections, with the aim of improving digital access to the joint collections as a whole.

By taking a look in the "machine room" of NOB, the high potential of Semantic Web technologies for heritage collections will be illustrated. The development of NOB over the last three years shows the transition process of a TA turning into a linked data hub developing services for both a general audience and the research community. The core of the Semantic Web concerns the transition from a web of documents to a web of data.[7] The trail from paper to linked data is still fresh and unpaved, but has a huge potential to give collection data a new life online.

2. From Paper to Digital Trail

In the Netherlands, there are over 400 ALMs with collections on World War II.[8] Together they hold a wide variety of information carriers, ranging from paper documents, photographs, film and audio recordings, diaries, newspapers, and oral history tapes. Apart from a large group of regional and local museums exclusively dedicated to World War II (approximately 83 organizations), for most ALMs only a part of their collections relates to World War II (Somers 2014). Besides the Dutch ALMs, yet to be explored collections are to be found in Germany, Australia, the United States, the United Kingdom, and many other countries.

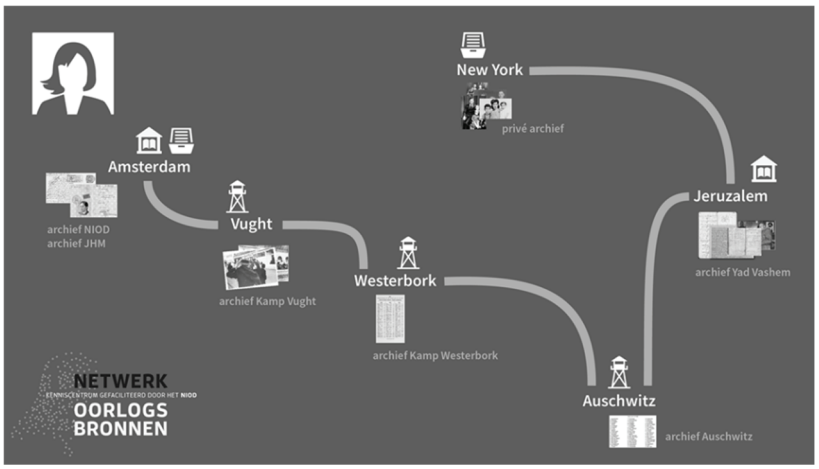
Currently, a researcher interested in, for example, Dutch volunteers in the Waffen SS, needs to investigate in advance which collections are available and the most relevant to consult (see Figure 1). A growing number of ALMs have published descriptions of their holdings on their own websites. However, they are often hard to find without prior knowledge of their existence. When search-

[7] *W3C Semantic Web Frequently Asked Questions* <www.w3.org/RDF/FAQ> (Accessed June 12, 2019).

[8] *Netwerk Oorlogsbronnen* <https://www.oorlogsbronnen.nl/organisaties> (Accessed June 12, 2019).

ing for specific information on Dutch volunteers, one runs into a number of issues. Conducting a simple Google search will return an overload of hits, most of them not referring to collections of professional organizations. Since the Waffen SS was an international organization, descriptions of archives may appear in different languages. Many search engines – even Google – are not properly equipped to cope with multilingualism. For instance, looking for an exact match for "Nederlandse Waffen ss vrijwilligers" returns 272 hits, while a similar search action for "Dutch Waffen SS volunteers" produces 2970 hits.[9] Also, many archival descriptions do not reveal much about their actual contents, for example: "This folder contains the documentation of..." or "Correspondence 1939-1944." The biggest challenge facing today's researcher is simultaneously the most hidden: nearly all ALMs have large backlogs with cataloguing, and many paper inventories have not yet been digitized. Currently, contacting other researchers or the collection specialists of an ALM by email or phone is likely to be the most viable strategy via which to start researching Dutch volunteers in the Waffen SS.

Figure 1: One Story, Many Collections



Source: Netwerk Oorlogsbronnen.

From a user's perspective, knowing which archives to consult is a fundamental first step that constitutes the starting point for further research. Currently, since most collections cannot be consulted online, the next step is to visit the reading rooms of the related institutions and manually review all of the potentially interesting documents. When browsing through the archives page by page, researchers generally do not recognize at first glance the correlations between

[9] Google search engine <http://www.google.com> (Accessed June 12, 2019).

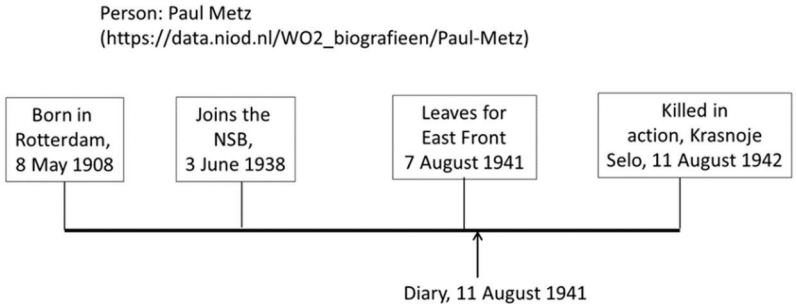
the information in the documents. Often only after analyzing all available data – also from other archives – are causal relationships discovered. The building blocks of these connections are: what, where, when, and who. Most historians make combinations of these four entry points on the basis of the available source materials and try to reconstruct past events. For instance, when studying a letter written by an individual Dutch Waffen SS volunteer, a researcher may be interested in his personal background and military career, what was happening at the same time in the village this person was staying, who his comrades were, etc.

Semantic Web technologies are extremely well suited to the contextualization of historical documents "bottom up." Linked data are the "cogs" in the Semantic Web machine. Software can be trained to identify locations, persons, dates, and domain-specific words in machine-readable texts and link them to external reference data from, for example, Wikidata, Geonames, or thesauri. These references allow computers to automatically contextualize the contents of documents. Take for example a fragment from the diary of Dutch Eastern Front volunteer Paul Metz. All references to external resources are between brackets and underlined:

> Soldaat Paul Metz [https://data.niod.nl/WO2_biografieen/paul-metz]
> Vrijwilligers Legioen Nederland
>
> [*https://data.niod.nl/WO2_Thesaurus/corporaties/4715*]
> 3e Kompagnie
>
> Maandag 11-8-1941 [date]
>
> Nadat Donderdag de trein dan eindelijk was weggereden, hadden wij gehoopt onze koffer in Rotterdam [http://sws.geonames.org/2754007/] in ontvangst te nemen, ten einde den inwendigen mensch te versterken. (…) Donderdagavond om half twaalf kwamen wij in Emmerich [http://sws.geonames.org/2930509/] aan. Wij hebben daar heerlijke soep gegeten.(..)
>
> Arys [http://sws.geonames.org//1526168/], 7-9-1941 [date]
>
> Vrijdagmiddag zijn we in Arys [http://sws.geonames.org//1526168/] aangekomen. (Metz 2011, 19, 29)

By identifying places, dates, organizations, and person names in the text and matching them to persistent identifiers of unambiguous related terms, permanent links to external sources of information are made. Additional information can be added to the excerpt from the diary. For instance, by linking to [https://data.niod.nl/WO2_biografieen/paul-metz], biographical data of Metz and references to other concepts and life events can be retrieved (Figure 2).

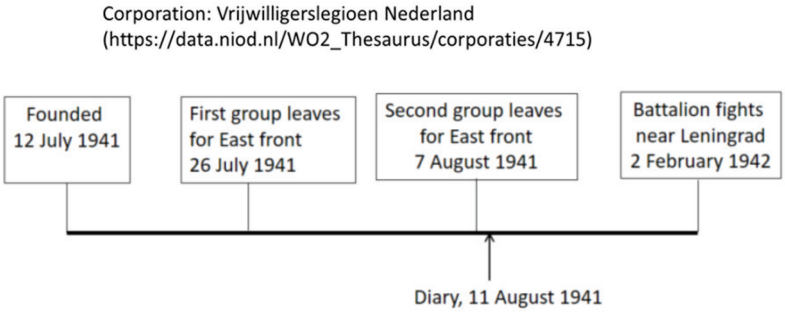
Figure 2: Timeline for Eastern Front Volunteer Paul Metz



Source: <https://data.niod.nl/WO2_biografieen/paul-metz>.

Another example of such a connection is the corporation mentioned by Metz in the fragment mentioned above, namely, *Vrijwilligerslegioen Nederland (Volunteer Legion Netherlands)*. This term also has references to events and places (see Figure 3).

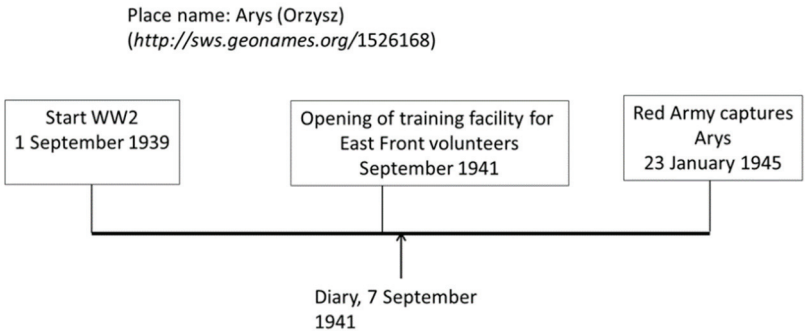
Figure 3: Timeline of Corporation Vrijwilligerslegioen Nederland



Source: <https://data.niod.nl/WO2_Thesaurus/corporaties/4715>.

In the diary excerpt, there is also a reference to Arys (today the Polish city of Orzysz). By linking it to GeoNames, Arys can be connected to World War II events in the Dutch World War II thesaurus that are connected the same GeoNames identifier (see Figure 4).

Figure 4: Timeline Geographical Location Arys

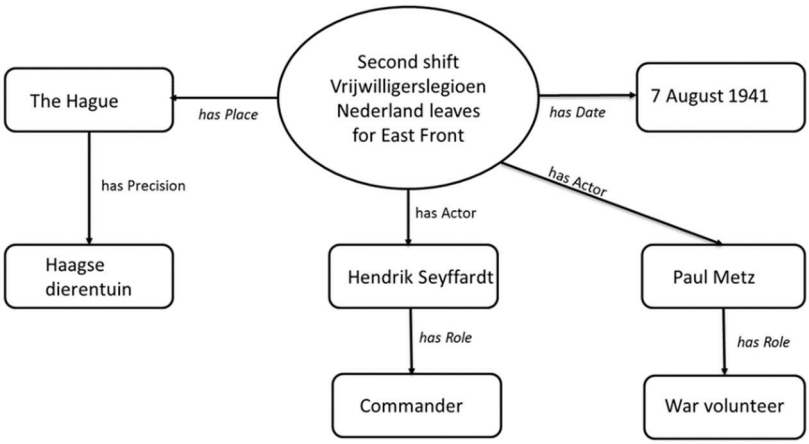


Source: <http://sws.geonames.org//1526168>.

All the nodes on the timelines can be considered as "events." They are logical units consisting of:

- related acts with a start and end date;
- related persons; and
- related geographic locations.

Figure 5: Example of Event Modelling Applied to *"Second Shift Vrijwilligerslegioen Nederland Leaves for Eastern Front"*



Source: <https://data.niod.nl/WO2_Thesaurus/events/7842>.

Thesauri are powerful knowledge systems with which to record the semantic relationships between concepts, corporations, persons, locations, and events. In 2015, a first version of the World War II thesaurus was compiled from keyword lists used at the NIOD Institute for War, Holocaust and Genocide Studies. It was published as linked open data at https://data.niod.nl/WO2_Thesaurus.

The World War II thesaurus is a work in progress and, in the last few years, it has been extended with entries on prison camps, organizations, persons, events, and concepts (for instance resistance, collaboration, etc.). NOB is cooperating with ALMs and various software companies to integrate the World War II thesaurus into frequently used collection management systems. This allows memory institutions to enrich their data themselves.

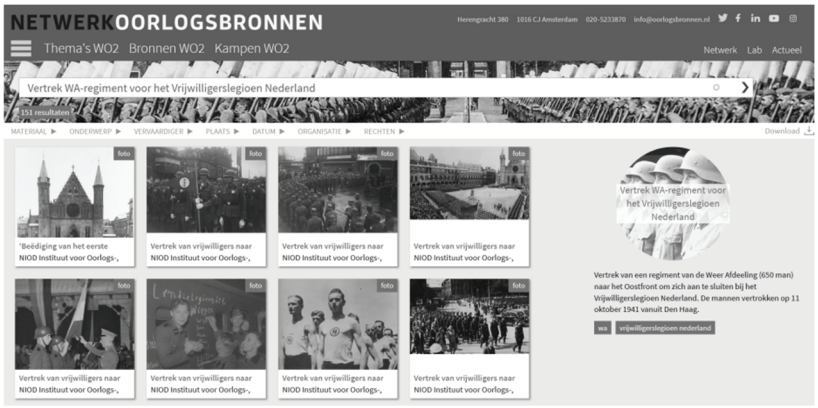
Figure 6: Tree in World War II Thesaurus with Vrijwilligerslegioen Nederland

The logical structure of a thesaurus – with narrow, related, alternate, and broader relations – is extremely well suited for improving search functions (Figure 6). Since thesauri consist of a limited number of terms, translating them into different languages is feasible. Alternate names can be very useful when dealing with spelling variants. Additionally, related terms can serve as search suggestions to end users while narrow and broader terms can be used to construct more complex search actions through so-called "query expansion." For

instance, when searching for Eastern Front, the software can be configured to automatically look for hierarchically related hits on Vrijwilligerslegioen Nederland, its German translation (*Freiwilligen Legion Niederlande*), and other battalions with Dutch volunteers. This feature can be beneficial to individual researchers and TAs when attempting to datamine large corpora and automatically filter out relevant data.

NOB's collection portal brings together 226 World War II collections. It harvests the metadata from over 100 ALMs and other organizations, and automatically matches them to the World War II thesaurus. When searching for "Vrijwilligerslegioen Nederland," 2,314 results from 26 organizations are returned.[10] These results include references to literature, newspaper articles, original billboards, archives, objects, film footage, photographs, and oral history recordings. Also, there is a scope note explaining in a few lines what the "Vrijwilligerslegioen Nederland" was, as well as further suggestions to search for "Oorlogsvrijwilligers" and "Oostfront" – both broader concepts in the thesaurus. Finally, there are links to Hendrik Seyffardt (a Dutch commander) and a number of related events (its foundation, the first battalion sent to the Eastern Front, etc.).

Figure 7: Web Interface Showing the Results of "Vertrek WA-regiment voor het Vrijwilligerslegioen Nederland"



Semantic Web technologies are potentially attractive, as they can improve access to collection data by linking. However, there are practical circumstances that obstruct its adoption in the ALMs. Most collection management systems are not properly equipped to work with external thesauri or reference data available as linked data. Also, organizations opting for linked data often have

[10] Oorlogsbronnen website <https://www.oorlogsbronnen.nl> (Accessed June 12, 2019).

to do a lot of revisionary work on their existing metadata. Currently, there are few organizations that are capable of providing linked data "from the source." TAs, with a clear interest in improving access to the collections, are the forerunners in this development.

Currently, only a few historical text-based archives are available in machine-readable formats. However, Optical Character Recognition (OCR)-technology, as well as other automated text recognition software, is progressing so rapidly that the conversion of text-based documents to machine-readable texts, in some cases, already reaches a quality that is sufficient for full-text searching. The quality of the transformation still very much depends on the condition of the original. In case of printed matter, for instance the Dutch Nazi publication *Storm SS*, the text produced after OCR'ing seems almost error free (mistakes are underlined):

> "Ik ben trotsch op mijn Nederlanders!" – Dat waren dc woorden van den Duitschen Brigadekommandcur daar boven aan den Wolchow, toen het vrijwilligerslegioen "Nederland" in een verpletterenden aanval een sowjet-divisie den terugtocht afsneed en zoodoende daadwerkelijk het zijne bijdroeg tot de vorming en liquidatie van die omsingeling, die reeds in de geschiedenis van dezen strijd in het Oosten zijn plaats gevonden heeft onder den naam: "Omsingelingsslag aan den Wolchow." (Storm SS, 13 August 1943)[11]

In 2017, in a small sample taken from digitized historical newspapers, the Koninklijke Bibliotheek (National Library of the Netherlands) measured a Word Error Rate (WER, i.e., percentage of incorrect words) of approximately 11%. With the latest version of OCR software, this was reduced to a WER of 9% (Wilms 2018). While both error margins leave room for improvement, they are low enough to create a full-text search function that is sufficient for answering most questions more accurately than was previously possible.

For archival materials, digital access is more complicated. Between 2016 and 2019, the Nationaal Archief (National Archives of the Netherlands), Huygens ING, NIOD Institute for War, Holocaust and Genocide Studies, and Netwerk Oorlogsbronnen were partners in a project called Tribunal Archives as Digital Research Facility (TRIADO). Its aim was to explore digital methods via which to transition from analogue archival documents to a digital research facility. Approximately 13.8 meters (167,197 pages) from the Central Archive of Special Jurisdiction (CABR) were scanned, after which OCR was applied. The CABR, held by the Nationaal Archief, consists of the legal case files of some 300,000 persons accused of collaborating with the German occupiers. It holds data on perpetrators, victims, the German occupation forces, and – more in general – Dutch society in extraordinary times of oppression and large-scale

[11] Storm SS. Weekblad der Germaansche SS in Nederland, vol. 3, no. 19, 13 August 1943, <https://resolver.kb.nl/resolve?urn=ddd:110529566:mpeg21:p003> (Accessed June 12, 2019).

violence. The CABR consists of approximately four kilometers of analogue documents, ranging from minutes and verdicts to membership cards, forms, and summonses (Gorter 2018).

Archival text documents, as a result of their fragile condition (tears and stains, light ink, bleed through of ink on transparent paper, folds, etc.), generally have a larger error margin than printed matter. A small test applying OCR to 150 typed or partly typed archival documents from the CABR scored a weighted average WER of 15% (TRIADO project team 2019). There are additional methods to improve OCR: pre-processing of images, machine learning with manually transcribed documents ("ground truth"), post-correction of common OCR mistakes, and applying existing lists of named entities to improve OCR. Still, even with a relatively high error rate, searching for "*oostfront*" (Eastern Front) through all documents returned 619 results, with a wide diversity of fragments:

> Voor het beeindigen der opleiding ben ik reeds naar het Oostfront vertrokken, naar ik meen was dit November…
>
> Volgens schrijven van 6-6-43 was hy met verlof in nreda. Komt'voor op ledenlijst van Oostfront-stryders
>
> …kwam het vrijwilligerslegioen geheel onder Duits ehe leiding. In Februari i944 werd ik ingezet aan het Oostfront[12]

Even with this error rate, with some fine-tuning it is possible to automatically match references to the raw data. In the TRIADO project, experiments combining the results of different OCR software were conducted (Tesseract and Abbyy Finereader). Although the number of incorrect words and characters ("noise") increased considerably, the same happened to the number of correct words. By combining the output from Abbyy Finereader and Tesseract, the "searchability" of the corpus improved, despite the higher overall number of errors.

Tests were done to automatically recognize named entities in the text to which OCR was applied. The archaic Dutch spellings, together with capitalized German nouns, caused many misinterpretations by the software. Recognizing dates was easier, due to their predicable format – for example, software can be given rules to learn that "i944" is very likely to be "1944." Matching existing lists of persons and places to the raw data proved to be a promising strategy via which to make connections to external sources of information. Although the quality of OCR was often worse than the examples of the diary of Metz or *Storm SS*, the software was still quite capable of matching place names or the names of victims mentioned in the *Nationale Database Vervolgings Slachtoffers* (National Database Persecution Victims).

[12] National Archives (NA) CABR archive 2.09.09, inventory nos. 71293, 32597 and 65853.

Currently, when doing research on the Waffen SS in the CABR, it is necessary to know in advance which persons to look for. The only name index that staff members of the Nationaal Archief have at their disposal, is a database with the names of suspects. There is no way of knowing what is mentioned in the files, except for visually inspecting the archives page by page. In a digital environment it is possible to search through all documents at page level. Names occur even if they are mentioned incidentally. In TRIADO, some experiments were done with automatically recognizing types of documents. The specific composition of the archives – legal documents accompanied by evidence such as membership cards or all kinds of forms – makes the CABR very well suited for what is called "auto-classification." For example, by feeding the software with samples of *Fragebogen* (questionnaires) filled out by Dutch volunteers signing up for the Waffen SS, identical documents can be retrieved. Although there was still an error rate of 20%, further machine learning is expected to narrow this gap (TRIADO project team 2019). Ideally, the information in these forms would be automatically converted into structured data. However, because of the complex layout, the many OCR mistakes, and the presence of handwritten text, this is still a complicated task for computers. Project READ shows promising results with interpreting tables in historical documents.[13]

Digital access to the CABR and other text-based resources would benefit from linking the raw data to external datasets with basic information about persons (perpetrators, victims, military, resistance, etc.). More generally, linking person data does not only provide more information about individuals, but also about specific groups. In our example of the Dutch volunteers of the Waffen SS, linking the conscription forms in the CABR to data from the tax office may give us valuable information about their socioeconomic background. In the Netherlands, many memory organizations have their own finding aids for persons: databases, spreadsheets, Microsoft Word documents, and in some cases also card index boxes or paper inventories. Access to these data is often restricted, due to ethical, organizational, and/or legal restrictions. Together with 13 memory institutions, NOB started a project called Oorlogslevens (War lives) with the aim to jointly build a web portal that allows access to personal data provided by the participating partners and other external sources. Data are automatically matched by applying a digital matching strategy that compares date of birth, place of birth, and the first characters of a surname to identify individuals. All connections are stored as linked data. The intranet site developed by Oorlogslevens shows the life events of these individuals – birth, arrest, imprisonment, death – with references to the data suppliers and the World War II thesaurus displayed on a timeline (see Figure 8). The approach taken by Oorlogslevens resembles that of the National Archives (UK;

[13] *Project Recognition and Enrichment of Archival Documents* (READ) <https://read.transkribus.eu/> (Accessed June 12, 2019).

Ranade 2016) and the project European Holocaust Research Infrastructure (EHRI; Nikolova 2018).

Figure 8: Computer-Generated Timeline for Cornelis Gootjes, Extracted from Eight Original Datasets

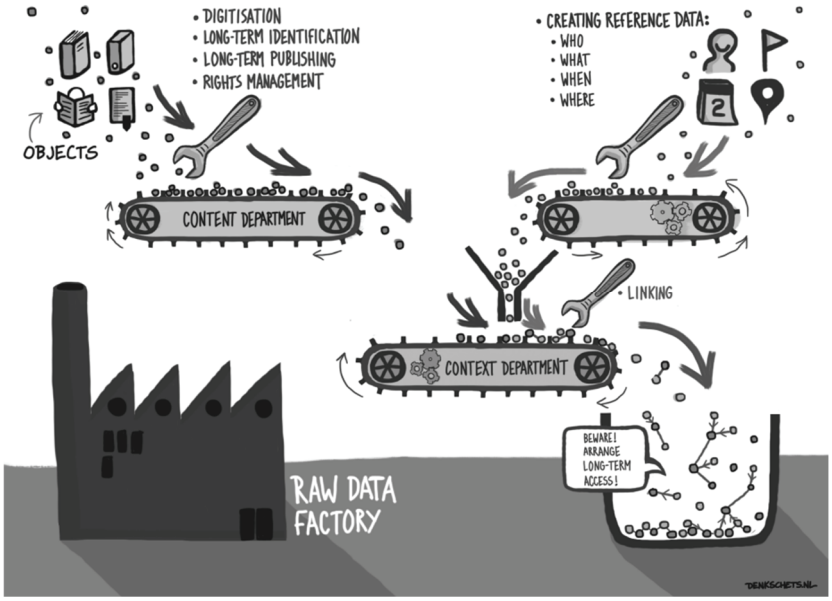
The use of Semantic Web technologies to interlink data from collections – both full content as well as metadata – have much to offer when publishing collections online. By focusing on the "who, what, where and when," logical connections can be made that contextualize individual photographs, texts, video excerpts, newspaper clippings, etc. In the digital domain, thesauri can regain their traditional function as a knowledge base. This time their main users will not be humans, but computers, which is likely to make thesauri more influential than ever before.

3. Conclusion

> It is not the amount of knowledge that makes a brain. It is not even the distribution of knowledge. It is the interconnectedness. – James Gleich, The information: A History, A Theory, A Flood, 416.

When considering collections as part of a virtual web of interconnected data, a new world of correlating perspectives emerges. Combining different chains of events – the private life of a Waffen SS volunteer, the history of the Vrijwilligerslegioen Nederland, the history of Arys during World War II – puts original resources into a variety of contexts. For ALMs, the Semantic Web offers a unique platform via which to disseminate their holdings. For TAs and RIs, the Semantic Web enables domain experts to "datify" their knowledge and put it at the disposal of a worldwide audience. Data from documents and other resources become the building blocks of all kinds of war histories: those of persons, places, organizations, events, or combinations of these elements (see Figure 9). "Liberating" the information from the documents and giving it a new life in the digital domain revitalizes the function and role of ALMs as data-driven, trusted centers of expertise.

Figure 9: The Road to Linked Data



Source: Netwerk Oorlogsbronnen.

Currently, the biggest obstacle towards "an interoperable Semantic Web for heritage resources" (Ross 2003, 8) is the limited amount of available data. So far, only a small portion of the analogue collections in Europe have been digitized. Even when collections have been digitized or catalogued, there is little consistency in the standards and formats used. ALMs and other organizations all have their own systems and conventions. Only a small number of organizations have Application Programming Interfaces (APIs) that allow computers to interact with the data, and even fewer use persistent identifiers for their digital objects.[14] Most importantly, moving from analogue to digital is a paradigm shift. The change from analogue to digital has a huge impact on the core business of a memory organization.

RIs and TAs have a shared interest in getting usable collection data. They can act as catalysts promoting the transition of ALMs into data-driven organizations. Ideally, all ALMs will provide their collection data as linked open data, although currently only a handful does. Apart from technical obstructions, for many organizations the Semantic Web remains an abstract and vague notion. RIs and TAs, in close cooperation with the ALMs, are in a unique position to showcase this technology. They can also be driving forces convincing ALMs and software suppliers to adopt linked data standards. Finally, RIs in particular include domain experts who could be involved in setting priorities for digitization, improving thesauri, etc. The tools developed in RIs might also prove extremely useful for opening up collection data in a more generic way.

Most important of all, RIs and TAs are networks of people. The trail from analogue to digital requires experts from many different fields who are willing to cross borders and understand each other's needs. We have only just begun to discover the new perspectives that the digital domain has to offer to historical research. Collections are our raw materials. Sophisticated search engine algorithms, Artificial Intelligence, and deep learning all have the potential to revolutionize digital access to collections. There is a bright future for the past on the Semantic Web.

References

Berkhout, Karel. 2011. Het digitale drama. (The digital drama) *NRC Handelsblad*, September 10 <www.nrc.nl/nieuws/2011/09/10/het-digitale-drama-12034576-a580591> (Accessed June 12, 2019).

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. The Semantic Web: A New Form of Web Content That Is Meaningful to Computers Will Unleash a Revolution of New Possibilities. *Scientific American*, May <www-sop.inria.fr/

[14] An API is a standardized entry point for automated exchange of data by computers.

acacia/cours/essi2006/Scientific%20American_%20Feature%20Article_%20The%20Semantic%20Web_%20May%202001.pdf> (Accessed June 12, 2019).
Europeana. 2014. *Portal to platform and other priorities in Europeana Business Plan 2014*. <https://pro.europeana.eu/post/portal-to-platform-and-other-priorities-in-europeana-business-pl> (Accessed June 12, 2019).
Europeana. 2015. *D.1.1: Recommendations to Improve Aggregation Infrastructure*. Version 1.0 <https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_Version3/Deliverables/EV3%20D1_1%20Aggregation%20Infrastructure.pdf> (Accessed June 12, 2019).
European Commission. 2019. *About Research Infrastructures* <https://ec.europa.eu/research/infrastructures/index.cfm> (Accessed June 12, 2019).
European Commission. 2017. *Horizon 2020 and the Research Infrastructures Landscape.* Draft version 0.1 <http://ec.europa.eu/research/infrastructures/pdf/ri_landscape_2017.pdf> (Accessed June 12, 2019).
Gleich, James. 2011. *The information: A History, A Theory, A Flood*. London: 4th Estate.
Gorter, Anne. 2018. *The Central Archive of Special Jurisdiction: A Short History* <https://www.oorlogsbronnen.nl/nieuws/central-archive-special-jurisdiction-short-history> (Accessed June 12, 2019).
van Hage, W. R., V. Malais, R. Segers, L. Hollink, and G. Schreiber. 2011. Design and Use of the Simple Event Model (SEM). *Web Semantics: Science, Services and Agents on the World Wide Web* 9: 128-36 <https://homepages.cwi.nl/~hollink/pubs/vanHage2011SEM.pdf> (Accessed June 12, 2019).
Hyvönen, Eero, Eetu Mäkelä, Tomi Kauppinen, Olli Alm, Jussi Kurki, Tuukka Ruotsalo, Katri Seppälä, Joeli Takala, Kimmo Puputti, Heini Kuittinen, Kim Viljanen, Jouni Tuominen, Tuomas Palonen, Matias Frosterus, Reetta Sinkkilä, Panu Paakkarinen, Joonas Laitio, and Katariina Nyberg. 2009. *CultureSampo-Finnish Culture on the Semantic Web 2.0: Thematic Perspectives for the End-user* <http://lib.tkk.fi/Diss/2010/isbn9789526034478/article8.pdf> (Accessed June 12, 2019).
Kenniscentrum Oorlogsbronnen. 2015. *De Nederlandse belangstelling voor de Tweede Wereldoorlog: Hedendaagse interesse en informatiebronnen*. (Dutch interest in the Second World War. Present day interests and information resources) Amsterdam: Netwerk Oorlogsbronnen.
Metz, Paul. 2011. *Mussertman Aan Het Oostfront.* Nijmegen: van Tilt.
Ministerie voor Volksgezondheid, Welzijn en Sport. 2010. *Erfgoed van de oorlog: De oogst van het programma.* (Heritage of the war: the harvest of the programme) The Hague: Koninklijke de Swart <https://www-data.oobr.nl/sites/default/files/Erfgoed%20van%20de%20Oorlog%20%28PDF%29.pdf> (Accessed June 12, 2019).
Nauta, Gertjan, Wietske van den Heuvel, and Susan Teunisse. 2017. *Europeana DSI 2- Access to Digital Resources of European Heritage* <https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/ENUMERATE/deliverables/DSI-2_Deliverable%20D4.4_Europeana_Report%20on%20ENUMERATE%20Core%20Survey%204.pdf> (Accessed June 12, 2019).
Netwerk Oorlogsbronnen. 2015. *Program Plan Network War Collections*. Amsterdam: Netwerk Oorlogsbronnen <https://www-data.oobr.nl/sites/default/

files/Programmaplan%20Netwerk%20Oorlogsbronnen_0.pdf> (Accessed June 12, 2019).
Nikolova, Ivelina. 2018. *Person Records Linking in the USHMM Survivors and Victims Database* <https://blog.ehri-project.eu/2018/05/29/person-records-linking-in-the-ushmm-survivors-and-victims-database/> (Accessed June 12, 2019).
Ranade, Sonia. 2016. *Making Connections: Tracing People Through Our Collection* <https://blog.nationalarchives.gov.uk/blog/making-connections-tracing-people-collection/> (Accessed June 12, 2019).
Ross, Seamus. 2003. Towards a Semantic Web for Heritage Resources (Position Paper). *Digicult* (3), May, pages 7-11 <www.digicult.info/downloads/ti3_high.pdf> (Accessed June 12, 2019).
Sherratt, Tim. 2013. *From portals to platforms: building new frameworks for user engagement*. Presented at the LIANZA 2013 Conference, Hamilton, New Zealand, 21 October 2013. <https://www.nla.gov.au/our-publications/staff-papers/from-portal-to-platform> (Accessed June 12, 2019).
Singhal, Amit. 2012. *Introducing the Knowledge Graph: Things, Not Strings*. Google official blog.<https://googleblog.blogspot.co. at/2012/05/introducing-knowledge-graph-things-not. html> (Accessed June 12, 2019).
Somers, Erik. 2014. *De oorlog in het museum: Herinnering en verbeelding*. (War in the museum: memory and fiction) WBOOKS: Zwolle 37.
Storm SS. *Weekblad der Germaansche SS in Nederland* 3 (19), 13 August 1943. Amsterdam: uitgeverij Storm<https://resolver.kb.nl/resolve?urn=ddd:110529566:mpeg21:a0028> (Accessed June 12, 2019).
TRIADO project team. 2019. *Final Report on TRIADO Enrichment Phase* <https://www.data.oobr.nl/sites/default/files/20190517_finalreportTRIADOenrichment.pdf> (Accessed June 12, 2019).
Wilms, Lotte. 2018. *Newspaper OCR Quality – What Do We Have and How Can We Improve It?* <http://lab.kb.nl/about-us/blog/%E2%80%8Bnewspaper-ocr-quality-%E2%80%93-what-do-we-have-and-how-can-we-improve-it> (Accessed June 12, 2019).

Historical Social Research
Historische Sozialforschung

All articles published in this Special Issue: