Combining Tools with Linked Data: a Social History Example

Ivo Zandhuis$^{1,2}$

$^1$Ivo Zandhuis Research & Consultancy
$^2$International Institute of Social History

This paper presents a workflow on how to link, retrieve and create data for historical research in a Linked Data format. Through this workflow, one can create a ‘web of data’ that contains a representation of all the things that are relevant to a particular history research project. For example, the workflow can accommodate publications and archival sources (linked to online catalogues or retrieved from data set providers and registered in Zotero, Tropy and Recogito) and organizations, persons and songs (created with CoW and burgerLinker). As an example a 19th century labour history project is presented on the development of collective action among print labourers. While describing the workflow, the paper advocates the use of Linked Data and investigates how well tools are equipped to realize interoperable data with this technique, using the three Linked Data principles introduced by Tim Berners-Lee in 2009.

Keywords: Linked Open Data; methodology; reproducibility

1 The Web of Data as our research infrastructure

Collecting, organizing and analysing sources and references without computational tools is unthinkable in current historical scholarship. While finding and combining these sources and references, tools can help scholars to iteratively develop interesting research questions and arguments that support or contradict their hypotheses. Ultimately, these arguments always refer to the sources they are based upon.

Every scholar prefers their own tool set to support this process and some of the tools are unavoidable in a certain domain. Personally, I prefer to use Zotero[1] for storing and organizing references, Tropy[2] to organize photos I take in the reading room, and Recogito[3] to organize the machine readable representation of a text. Additionally, I create and use data on individuals where I wish to keep a close relation with the original source. These micro observations are published online in data sets and the

---

1 https://zotero.org
2 https://tropy.org
3 https://recogito.pelagios.org
The aforementioned combination of tools helps me to relate every digital representation and observation to the right archival material or literature reference.

Preferably, research is thought of as creating data to substantiate an argument, instead of using tools to investigate a topic. The use of tools is only a means to develop a network of data that refers to the various building blocks of the research and their provenance. This network of data, or in mathematical terms the ‘graph’, helps to create a thorough and reproducible historical explanation, independent of the tools we use. Ideally, this data should be processable in an automatic way. To be more precise: it needs to satisfy the FAIR-principles of being Findable, Accessible, Interoperable and Reusable (Wilkinson et al., 2016). Unfortunately, the tools I use each have their own environment to store and manipulate data. Without devoting special attention to creating one network of data, using the various tools would result in various unrelated data sets. That means the data would not comply to the FAIR-principles.

These principles prescribe that the data must be exported and published. By applying a standard for syntactical representation of the data and standardized protocols for publishing the data, most principles are met. Findability, Accessibility and Reusability are improved massively by doing this. Interoperability of data is achieved by adding machine readable semantics and relations: instead of a human readable code book, a data provider should add a definition of the elements in their data according to standardized syntax and protocols as well.

Because of its gradual adoption and semantic potential, Linked Data is the most obvious technique to use for creating the network of FAIR, but especially interoperable, data that we envision. The technique consists of specifications for (1) identifying concepts in the data and (2) relating them to other concepts. These specifications are independent of tools or computer platforms and enable decentralized data definition and storage. For that reason they are especially suitable for creating interoperable data.

Concretely, publishing information resources as Linked Data can be done by transforming the data stored in the used tool to the format as described by the Resource Description Framework (RDF) specification (Cyganiak et al., 2014). To explain RDF to a more broader audience than computing scientists, Tim Berners-Lee introduced three main principles for Linked Data (Berners-Lee, 2009). For a concise introduction into Linked Data I explain them here briefly and I use them to evaluate the implementation of Linked Data in the tools I discuss in this paper.

1. **All conceptual things should have a name starting with HTTP.** This means that every entity you want to publish an information resource for, must have a web address, more precisely a so-called Uniform Resource Identifier (‘URI’). The entities could, for example, be a book, the member of an association or a location.

2. **Looking up an HTTP name should return useful data about the thing in question in a standard format.** If you call the web address, the web server must serve the data about the referred conceptual thing in RDF format. In the RDF format, the data properties are expressed. These properties are for instance the title of the book, the date of birth of a person or the coordinates on a map of a location.

3. **Anything else that that same thing has a relationship with through its data, should also be given a name beginning with HTTP.** In the data, for instance, the book is related with its author using the URI of that author (e.g. http://www.wikidata.org/entity/Q80) instead of the label ("Berners-Lee, T.J."). This means that you can follow the link and use the web address of the author to collect data about
the author. In this way you can, for instance, retrieve their birth date (for Tim Berners-Lee this would be "8 June 1955"), without having to register this yourself. Linking data like this results in a "Web of Data".

In this paper, I explain how I use the tools and export the data in Linked Data format to create one graph containing a representation of all the concepts relevant to my research, like publications, archival sources, organizations and persons. I introduce an example from my 19th century Labour History project on the development of collective action among print labourers. The paper explicitly does not provide a blue print for doing (social) historical research with digital tools. The described tools are my personal (and therefore rather arbitrary) choice. The paper advocates the use of Linked Data and investigates how well these specific tools are equipped to realize interoperable data with this technique as an inspiration for tool developers to learn how this functionality can be improved in their own project. It might help researchers to understand the criteria they need while evaluating the tool set they use.

The scripts I made and the resulting graph are available on github. Remember this is a work in progress, so the repository changes over time. The remainder of this paper is organised as follows. In Section 2 I introduce the research I am currently undertaking and use as an example in the rest of the paper. After that, in Section 3 I investigate the way I am able to link and retrieve Linked Data from cultural heritage collections and data repositories and discuss my findings. Section 4 concentrates on the creation of Linked Data from the data stored in the tools I use. And again I discuss my findings. In Sections 5 and 6 I discuss the use and usefulness of Linked Data in general. Finally, Section 7 sums up the overall conclusions.

2 An example from 19th century labour history

For my project, I’m building a web of data for research into a Dutch phenomenon called ‘typografische verenigingen’ (typographical associations). During the first half of the 19th century print labourers in The Netherlands organized themselves in these local associations, comparable to the English ‘friendly societies’ and the French ‘sociétés mutuelles’ (Linden, 1996). These associations were founded to ensure health benefits and they organized a yearly feast to celebrate their identity as ‘children of Laurens Coster’, the Dutchman they believed invented printing. They were connected in a nationwide social network and organized the erection of a statue in Coster’s honour in 1856 (see Figure 1). Eventually, this led to the establishment of the first national trade union in The Netherlands in 1866 (Giele, 1972). I’m interested in how this phenomenon originated.

Modern information technology helps me to process more details to form a complete picture of the development of collective action among print labourers. I can create lists of people involved and register their attributes, like their age and marital status. If I am able to find their (family-)relations, I can use Social Network Analysis to find patterns in the diffusion of the phenomenon. Collecting the songs the members of the associations sang during their feast (which they obviously printed), helps me study potential shifts in their interests. Minutes of their meetings refer to persons and befriended associations, and provide detailed insights in their relations as well. Besides these quantitative opportunities, references to publications enable us to use historiography about the subject and refer to details in other publications.

https://github.com/ivozandhuis/typografische-verenigingen
In this paper, I present as an example, the data about J. H. Regenboog, mentioned frequently in the sources in the 1850s. We learn about his activities and find relevant family relations. This results in a part of the web of data visualized in Figure 2.

3 Linking and retrieving Linked Data

The current ‘web of documents’ is a base for the creation of the ‘web of data’ that is anticipated. Important online sources can be organized by providing simple and stable web addresses for every element in the collection of an heritage institution. Some online catalogues and data repositories already help us to link the right source.

3.1 Linking data in online catalogues

Most of the sources I need to study for my project are held by the International Institute of Social History (IISH), the Library of the University of Amsterdam (UB-UvA) and the Noord-Hollands Archief (NHA) in Haarlem. I found a lot of sources with a full-text search on Worldcat\(^5\) for publications and with a search on Archives Portal Europe\(^6\) for archival material.

I register my findings in Zotero. Zotero is a convenient tool for recording resources and for creating footnotes in the historical papers I want to write. Thanks to existing import scripts I was able to easily feed the Worldcat data into Zotero. For archival

---

\(^5\) https://www.worldcat.org/
\(^6\) https://archivesportaleurope.net/
materials I needed to insert the data in Zotero by hand. Unfortunately Zotero is unable to handle the hierarchical nature of archival descriptions, so I had to design a work-around for that. I published the overview of the sources on the Zotero website.\footnote{https://www.zotero.org/groups/2707622/typografische-verenigingen/library}

For every item I registered I returned to the original online catalogue of the institute holding the material. I checked the data and stored the web address for reference to this original catalogue. Both the IISH and the NHA provide a stable web address according to the Handle principle, while the UB-UvA has chosen to use ARK\footnote{https://lod.uba.uva.nl/UB-UVA/Catalogue/}. That way these institutes made a step towards implementing Berners-Lee’s first Linked Data principle. The same book might be held by various institutions or published on Google Books. When this is the case, I store the web addresses of the other manifestations of the same publication in Zotero as an ‘attachment’ typed ‘Link to URI’.

For example, one of the sources that are relevant for my research, is the nationwide yearbook print labourers created in 1856 \cite{Mommaas1856}. This book contains information about all the associations that existed at that moment. A short history of its origination and development is presented, as well as the founders, board members and members. I registered this yearbook in Zotero and stated that it has a digitized version on Google Books and that the original is held by the UB-UvA. Additionally, Zotero refers to a tabular data-file I created and published on GitHub, listing all members mentioned in the yearbook. One of these members is J.H. Regenboog, a board member of the typographical association in The Hague.

The UB-UvA has published information resources from their catalogues in RDF format and the URIs that they have introduced can be used to request the data on the item in the catalogue in RDF format\footnote{https://druid.datalegend.net/IISG/iisg-kg}. That way they have fully implemented Berners-Lee’s second Linked Data principle. Unfortunately, not all data is converted into RDF and the approach suggests that the RDF-version of the data is static and must be updated periodically instead of converting the data to RDF on-the-fly \cite{Koster2021}. Therefore we might be dealing with outdated metadata.

The IISH has introduced separate URIs for the RDF-version of the data in their catalogue\footnote{https://lod.uba.uva.nl/UB-UVA/Catalogue/}. This means that the persistent web addresses that were introduced according to the Handles-principle do not resolve into a representation of the data in RDF-format. As a consequence, not all requirements to implement the second Linked Data principle are met.

The NHA has no RDF representation of their data whatsoever and does not meet the requirements of the second principle.

3.2 Retrieving data from online data repositories

I want to include the people that organized themselves in the typographical associations into my social historical research: what was their origin and what were their relations? To study the individual life courses of important participants in the ‘typografische verenigingen’, I use civil registries held by various archival institutes across The Netherlands. Fortunately, these registries are accessible through online indexes created by mostly volunteers and used by genealogical enthusiasts. On a website called Open Archives, Bob Coret has aggregated all the indexes that are published as online, open data\footnote{https://www.openarch.nl/}. On this platform Coret provides an Application Programming Interface
(API) as well, which enables me to obtain the data of a particular registration.\footnote{12}

Every certificate that is included in this database has its own unique web address. This web address can be stored for a reference to the certificate on the Open Archives website. Using this web address in a web browser results in a human readable web page, but I can request the data in RDF as well. This means that both the first and second Linked Data principles are met.

Thankfully, the name of my research subject ‘Jan Hendrik Regenboog’ can also be found in the indexes on openarch.nl. Through his marriage certificate we know his age, profession and the names of his parents, wife, and parents-in-law.\footnote{13} I collect all the relevant references to important participants in the ‘typografische verenigingen’ in the civil registries in a tabular data file and retrieve the accompanying data in RDF. First I obtain a list of all civil registry certificates mentioning a person with one of the main occupations in the typographical domain: ‘letterzetter’, ‘boekbinder’, ‘boekdrukker’ and ‘drukker’. Step two is to harvest all data of the certificates by resolving the URI of the certificate.

The names of properties and classes in the data are derived from the A2A standard, originally developed for exchanging personal data in XML. There is no ontology linked to the data to be found, however. This means that only the standardization of syntax and protocol of the data is established. The RDF representation refers to various elements in the data by means of internal links. Links to other sources, like places or archival institutions are not added. Therefore the third principle is not met.

The Linked Data representation of the data set ‘History of Work’ by the IISH enables us to create an external link with a clever trick.\footnote{14} By adding a base URI to the lexical title of an occupation in the source, e.g. by combining the string ‘letterzetter’ with the prefix ‘https://iisg.amsterdam/resource/hsn/occupation/’ into the URI \url{https://iisg.amsterdam/resource/hsn/occupation/letterzetter}, a link is created to the ‘History of Work’ dataset. Thanks to this link, I relate additional data about an occupation. This contains for instance the social status of the occupation in various standards, translations of the occupational title into other languages and the HISCO grouping (Zijdeman and Lambert, 2010).

### 3.3 Discussion

In this section I investigated the four data sources that are the most relevant for my research (see Table \ref{tab:link}) and concluded that all of them made serious steps towards unique web addresses for reference to a resource. Three out of four provided data per resource in RDF.

\begin{table}[h]
\centering
\begin{tabular}{|l|c|c|c|}
\hline
Data source & Principle 1 & Principle 2 & Principle 3 \\
\hline
Int. Inst. of Social History & yes & yes, but & \\
Library Univ. of Amsterdam & yes & yes & \\
Noord-Hollands Archief & yes & no & \\
Openarch.nl & yes, but & yes & no \\
\hline
\end{tabular}
\caption{Overview of the described data sources and their compliance to the Linked Data principles}
\end{table}

\footnotetext[12]{12}{https://www.openarch.nl/api/docs/}
\footnotetext[13]{13}{https://www.openarch.nl/hga:F6619BDA-BE62-4A37-B09D-960CB8900393}
\footnotetext[14]{14}{https://druid.datalegend.net/HistoryOfWork}
There are important issues with regard to the persistency of the web addresses. And these issues result in some extra ‘but’ in the overview. Openarch.nl does not pretend to mint persistent web addresses for every resource it publishes. The web addresses are depending on the data provided by the archive holding the original data. If the identification at the archive changes, so does the web address openarch.nl provides.

The IISH provides two different web addresses for a resource: one persistent identifier according to the Handle system and one that enables providing data in RDF-format. Those two web addresses are technically unrelated.

An interesting question here that would move us beyond the scope of this paper, is whether more institutes in the Benelux and beyond use Linked Data to provide metadata about their collections. The Web of Data of heritage collections is growing fast and an overview given here would be outdated the moment it is published. At least national libraries of The Netherlands, Belgium, Great Britain and France have Linked Data facilities or plan to have them in the near future. In both The Netherlands and in Flanders, governments are stimulating the use of Linked Data in the heritage domain. This policy is executed by the Netwerk Digitaal Erfgoed in The Netherlands and meemoo in Flanders.

The observant reader notices I did not include the third Linked Data principle in my analysis. This third principle was only investigated for data I retrieved. Most of the sources I describe here, are only linked to my graph. As a reference management application, Zotero uses its own, locally stored metadata to create a reference in a paper. I chose to use the metadata of the objects conveniently stored in Zotero as the starting point for my graph rather than using the metadata that is provided in RDF via the URI of the object in the online catalogue. The alternative, i.e. using the metadata directly from the online catalogues, would have been a less practical choice. Furthermore, we have to take into account that the various catalogues each provide metadata according to different metadata standards. If I wanted to use that data, I would have needed to convert all the various types of linked data into my own model. A third argument for using the Zotero database as a starting point is that not all cultural heritage institutions — in our case for example the Noord-Hollands Archief — is providing Linked Data.

The curated data in Zotero must be transformed into Linked Data to be part of our web of data. The next section will study how this can be done, along with connecting data from other tools.

### 4 Creating Linked Data

Besides linking and retrieving the Linked Data available online, I create my own Linked Data for the graph of all relevant assets in my research. Ideally I want to create this data in convenient tools that I planned to use anyway, like the aforementioned Zotero, Tropy and Recogito. To some extent, these tools all provide the possibility to export and/or convert data into RDF. The results of these conversions will ultimately be combined into my graph.

Apart from the metadata about sources, my research needs data about entities like persons (e.g. the individual print labourers), organisations (e.g. ‘typografische verenigingen’) and songs (sung during their yearly feasts). This data is constructed in tabular form in a spreadsheet application and must be transformed into Linked Data.

---

15 [https://www.netwerkdigitaalerfgoed.nl/](https://www.netwerkdigitaalerfgoed.nl/)
16 [https://www.meemoo.be/](https://www.meemoo.be/)
Finally, I use an application to derive additional data computationally. This constructed data must be available in the graph as well.

This section studies how easy I can construct the Linked Data and the links between data created in the different applications.

4.1 Zotero

Zotero is a reference management application. This means a scholar can create and manage records to create footnotes or end notes in their publications. Fortunately, you do not need to enter all data into Zotero, but the application has functionality to obtain the data from important catalogue websites, like worldcat.org, with the click of a button. After that you can curate the metadata to your liking. That way I collected data about all relevant sources for my research and I shared the result on the Zotero-website as a ‘public library’.

By publishing my collection of references through Zotero, a web address is available for every item. The registration of the yearbook 1856, for instance, can be found using https://www.zotero.org/groups/2707622/items/P5HRJ669. In my web of data I use this web address as a basic URI to relate all information I know about this object. It relates to manifestations of this book in the Library of Leiden University, Google Books, the Library of the University of Amsterdam and refers to data available on Worldcat.

I can use a so-called Zotero Translator, built in JavaScript, to export the data from the local Zotero storage. To create Linked Data I took an available Translator and adjusted it to my needs. The script exports the data with properties standardized on Schema.org and uses the Zotero web address as a URI.

The possibility to create your own script to export the data collected in Zotero is very important. It complies with the idea of making the data more important than the tool. The available techniques meet the Linked Data principles halfway.

By creating the web address on the Zotero website for an item, I can use this as a URI for the item I want to register. On the other hand, the stability of the URI remains questionable. That way only some of the necessary aspects of the first Linked Data principle are met.

Concerning the second principle, the web address only refers to a human readable web page and is not able to deliver the underlying data in RDF. To mitigate this problem, I have created my own data set of data with the Translator. I need to publish this separately from the Zotero website.

According to the third principle I must use URIs to refer to other relevant resources. One of these is the holding institute of the item. To link items to their holding institute I use the list of International Standard Identifier for Libraries (ISIL) codes for Standardization 2019. The Dutch Royal Library and the Dutch National Archive are responsible for issuing these codes for institutes in The Netherlands. The IISH, for instance, has code ‘NL-AsdIISG’. While entering data into Zotero I use the characters ‘isil:’ in front of an ISIL code, and my script creates a URI for this data point in the RDF export. A similar trick is used to refer to typographical associations. Here I use the characters ‘typo:’ which refers to a list of typographical associations I created in a spreadsheet.

17 https://www.zotero.org/groups/2707622/typografische-verenigingen/library
18 http://schema.org/
19 https://www.bibliotheeknetwerk.nl/landelijke-digitale-infrastructuur-ldi/isil-codes
20 https://www.nationaalarchief.nl/archiveren/kennisbank/isil-codes

128
Zotero helps to create web addresses for items, but the publication of Linked Data, both according to the second and third principles, is not incorporated. Nevertheless, creating Linked Data for my graph is possible thanks to the flexible export possibility and my own built-in conversion of links to other resources.

### 4.2 Tropy

In Tropy I can organize the pictures I take in the reading rooms of the institutes I visit. To do so, I just create a ‘project’, and combine pictures into ‘items’. Of every item I can add some basic metadata, but the most important thing I register is the Zotero-URI I created. With this URI I can link the data in Tropy to the data in Zotero. One object I digitized is a booklet with songs, written on the occasion of the erection of the Coster statue in Haarlem in 1856 ([Regt and Breeman, 1856](https://www.geonames.org/)). Different members from different associations submitted their text on existing melodies. Most prominent songwriter, with six songs, was J.H. Regenboog, boardmember of the association in The Hague.

Tropy has an export function, exporting the data directly into a Linked Data format, more specifically `JSON-LD`. Unfortunately, the export function does not construct URIs for the Items and therefore the data does not comply to the first Linked Data principle. I have to create an additional Python script to correct this. In the same script I must convert the property with my reference to the Zotero URI into a property expressing the semantics that the photos are a representation of the work described in Zotero.

### 4.3 Recogito

The third tool I use is Recogito. Recogito helps users to create textual or image sources with markings for entities in the text or on the image. The entities can be typed: you can state that the marking is a person, place or event. Entities for places can be reconciled to a standardized list of geographical names, like geonames. In Recogito, I can add my own standardized list of relevant concepts, like important persons, or the typographical associations.

I created a machine-readable version of the yearbook of 1856 in Recogito and related this representation with the URI created by Zotero. In the text I marked the persons and associations that were mentioned. One of these persons is J.H. Regenboog, again as a board member of the typographical association in The Hague. He was awarded a silver medal in 1851 for his commitment to the association. From Recogito, this type of marking can be exported into a Linked Data file. Moreover, the text is hosted on the Recogito-website and the marking of J.H. Regenboog has a working web address. I can use this as an anchor for linking other information to the mentioning of “J.H. Regenboog” in the yearbook.

Of all tools discussed in this paper so far, Recogito implements the Linked Data principles best. Documents and annotations all have their own URIs, and these URIs result in a presentation on the web and can be linked to other URIs for more information on the subject it links to. Some wishes remain, though. Unfortunately the link to the marking is lost if a new version of the text is uploaded, so I am unable to correct an error in the text without breaking the existing URIs. Besides that, the annotation ontology used in the Linked Data export is not used correctly (there is no such thing as a ‘oa:Tag class’) or incomplete (‘oa:hasBody’ is missing if the marking is not reconciled) ([Sanderson et al., 2017](https://recogito.pelagios.org/annotation/51e9cac0-a07a-472d-9e9b-eef16f95995f)).

---

[^21]: [https://www.geonames.org/](https://www.geonames.org/)
[^22]: [https://recogito.pelagios.org/annotation/51e9cac0-a07a-472d-9e9b-eef16f95995f](https://recogito.pelagios.org/annotation/51e9cac0-a07a-472d-9e9b-eef16f95995f)
4.4 LDWizard and CoW

Some of the resources I need to link to in my graph are new and not available online. For example, I need resources for cultural heritage institutes, (for which I used the ISIL-codes, but which are not available as URIs online) and the typographical associations I study.

To create Linked Data for these resources, I have developed a CSV file with columns for each relevant property. For the typographical association that would for instance be the year the association was founded, or its location. To create Linked Data from this CSV file, I use the tool CoW, which stands for ‘CSV on the Web’ \(^{23}\). This tool, created in the CLARIAH program, converts the CSV into Linked Data. So the links I created in Zotero to express that an item is about a certain typographical association, refer to the URIs created in this process.

CoW needs a mapping that secures which Linked Data property must be used for each column. This mapping is stored in a JSON format, of which the creation by hand might be error prone. For that reason I used the LDWizard tool \(^{24}\). This tool provides a more user friendly environment to create such a mapping.

With my home brewed URIs, I am able to state that the yearbook of 1856 has information about the typographical association called “Door Eendracht t’ Zaam Verbonden” in The Hague. While transforming my CSV into Linked Data I complied to the first Linked Data principle because I added URIs to the things (i.e. typographical associations) I want to describe. I did not create a web server that could resolve the URIs (yet) and do not deliver the data belonging to the things I described through this URI. So Berners-Lee’s second principle is not met.

4.5 burgerLinker

Finding all family relations of Jan Hendrik Regenboog, — maybe scattered over the entire country — is very cumbersome. With help of the data retrieved from the openarch-website and a tool, developed in the CLARIAH program, called burgerLinker, I am able to find these family relations \(^{25}\).

For this I use the RDF data retrieved from openarch.nl. After harvesting the data I need to convert it into the semantics the burgerLinker tool needs. After that the transformed Linked Data is fed to the burgerLinker-tool, which finds links between mentions of the same person. burgerLinker is able to calculate family relations as well: persons with the same parents are (obviously) siblings.

In the resulting data set with family relations I’ll find a brother of J.H. Regenboog, called Christiaan Regenboog. Christiaan was born in The Hague but moved to Amsterdam and apparently took the idea of founding a typographical association and his organizational skills with him. He became the co-founder and secretary of the typographical association ‘Voorzorg en Genoegen’ in Amsterdam. His name is mentioned in several publications on the early development of the organization of labourers. (Bos 2001; Giele 1972) Although in these publications Amsterdam is considered a hotspot in labour organization, the relation with The Hague should not be neglected. And the family relations might be a good source to map relations to other towns as well.

Creating this data set with family relations requires programming skills, firstly to obtain the right data from OpenArch, and then to convert it into the Linked Data

\(^{23}\) https://github.com/CLARIAH/CoW
\(^{24}\) https://ldwizard.netwerkdigitaalerfgoed.nl/
\(^{25}\) https://github.com/CLARIAH/burgerLinker
with the properties prescribed by burgerLinker. There are initiatives that try to create more user-friendly interfaces for this process, like the aforementioned LDWizard. Furthermore, the heritage institutes in The Netherlands have the ambition to provide Linked Data directly, which simplifies and abridges this programming phase. If this development takes off, we need to agree on standardized Linked Data classes and properties, and implement an automatic transformation into the data model in burgerLinker. Otherwise the researcher still needs programming skills to convert the data from the web.

Again there is no web server that could resolve the URIs I created and deliver the data belonging to the persons I described. So I do not comply to the second Linked Data principle. I am able, though, to link mentioned persons to their occupation in the History Of Work data set. Because I use a URI to do this I do comply to Berners-Lee’s third Linked Data principle.

### 4.6 Discussion

Currently, the creation of Linked Data from the tools a researcher uses, depends on the ability to export data in a syntax complying to the RDF standard. The tool must help us create a URI for the resource that is modelled, enable the possibility to retrieve data via this URI, and encode relations by the use of URIs in other data. An overview of these three principles is presented in Table 2 for the tools where I needed to export data from. In Table 3 an overview is presented of the tools I used to created extra Linked Data.

<table>
<thead>
<tr>
<th>Tool</th>
<th>Principle 1</th>
<th>Principle 2</th>
<th>Principle 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zotero</td>
<td>yes, but</td>
<td>no</td>
<td>yes, but</td>
</tr>
<tr>
<td>Tropy</td>
<td>no</td>
<td>no</td>
<td>yes, but</td>
</tr>
<tr>
<td>Recogito</td>
<td>yes, but</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>

In the case of Zotero, coding skills are needed to construct RDF. With these skills one is able to construct a URI that results in a human readable web page, but the sustainability of the URI remains questionable. Tropy does not provide a URI at all. Therefore, I needed to write an extra script to create a URI to link to. In Recogito URIs are available and resolvable, but after updating the text in the system, the links are broken.

None of the three tools have the ability to provide data based on the URI. Finally, the links that are available in the data of resource depends heavily on the possibility to include the creation of these links in the scripting that is developed. Tools creating Linked Data from scratch are all able to comply to two out of three principles if the users walk the extra mile of using the tools with these principles in mind.
Table 3: Overview of the described tools and the ability to comply to the Linked Data principles while creating data

<table>
<thead>
<tr>
<th>Tool</th>
<th>Principle 1</th>
<th>Principle 2</th>
<th>Principle 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDWizard and CoW</td>
<td>yes, but</td>
<td>no</td>
<td>yes, but</td>
</tr>
<tr>
<td>burgerLinker</td>
<td>yes, but</td>
<td>no</td>
<td>yes, but</td>
</tr>
</tbody>
</table>

5 Using the graph

At some point in time the data and source collection phase of the research project will be finished and a more or less complete graph is available. The moment my graph is complete, I will be able to validate the coherence of the data or select data from it and do an analysis. For that, I can upload the data into a Triple Store, an application specialized in storing and retrieving Linked Data. I could use the online CLARIAH Linked Data environment ‘Druid’ \(^{26}\) which is an instantiation of TriplyDB \(^{27}\) by Triply. \(^{28}\) I could use a Triple Store on my laptop, like GraphDB \(^{29}\) with which I created Figure 2. Both systems include the option of using the query-language SPARQL to select a relevant subset for an analysis (DuCharme, 2013). The same SPARQL queries can be used in various programming environments.

![Figure 2: Part of the ‘Web of Data’ concerning Jan Hendrik and Christiaan Regenboog and their relations to the sources](https://example.com/figure2)

6 Discussion

This paper demonstrates that creating Linked Data can, with some extra consideration and coding, be achieved with standard research tools. It therefore contributes to the FAIR principles by adding interoperability to the data.

The resulting ‘web of data’ can be considered a ‘knowledge graph’ about the subject of the research at hand. That way the knowledge about the subject is encoded

\(^{26}\) https://druid.datalegend.net/
\(^{27}\) https://triplydb.com/
\(^{28}\) https://triply.cc/
\(^{29}\) https://graphdb.ontotext.com/
in a computationally processable way and can be reused and extended by future researchers.

But why is that useful and necessary?

### 6.1 Reproducibility

If the underlying sources and literature of the arguments that support or contradict the hypotheses of my research are organized in Linked Data, this means that future researchers are able to validate the research. Future researchers can easily trace conclusions back to its original sources and evaluate the quality of the path in the argumentation. It results into reproducible research: scholars who use this research can retrieve references very quickly and check them. In addition, anyone can use the graphy I created (see Figure 2) to check which sources I used to conclude that J.H. Regenboog in The Hague was related to Ch. Regenboog in Amsterdam, important for the development of the phenomenon because it says something about why it was transferred from The Hague to Amsterdam.

### 6.2 Precision

For me, the main reason for creating this graph concerning the typographical associations is to investigate whether I can store all relations I find into one system, comprising of various applications. This forces me to be extremely precise about the relations between sources and conclusions I draw from them. Being extremely precise benefits the quality of the research.

### 6.3 Short term usefulness

The short term benefits of using Linked Data has to do with the data analysis that can be applied on a data set. In our example, the use of burgerLinker and the use of the links to HISCO enable me to add extra data to the graph already available.

### 6.4 Computer Aided Historical Research

A more visionary take on the Linked Data approach is that the future might bring a kind of Computer Aided Historical Research, instead of the current googling with search terms. A system might be able to retrieve more sources relevant to your research, thanks to the semantic relations added to the web of data by your colleagues in the past. We might develop a user interface were scholars are able to ‘link up’ their own subset of the web of data, and construct and publish new links, that lead to new suggestions in other research projects. If sources can be found and combined more quickly and more precisely thanks to this automation, the historian has more time to draw conclusions. I leave the dreaming about more possibilities in the more distant future to your own imagination.

### 7 Conclusion

In this paper I investigated the available Linked Data functionality in existing tools. I did this by creating a web of data with these tools.

Every specialized tool is good in its particular function: reference management, photo management, data conversion or source enrichment. For that reason I use this
combination of tools. The data they contain is one big data system, with links between things ‘living’ in different tools. For that reason I create one graph covering this data system. Creating the combined graph needed thorough thinking and additional coding, because the links between the data exported from the tools can not be easily made into the URIs that encode the link. Future development of this setup should take the decentralised data principle of Linked Data into account and enable fields to be filled with references to URIs. This, to comply to Berners-Lee third Linked Data principle: *anything else that that same thing has a relationship with through its data, should also be given a name beginning with HTTP.*

A big issue is: how do I publish my home brewed URIs as resolvable URIs, according to the second principle: *Looking up an HTTP name should return useful data about the thing in question in a standard format.* These URIs should be persistent and dereferenceable, but at the moment they are not. I need an institution to facilitate the creation of my own dereferenceable URIs quick and easy.

After this experiment, I conclude that big steps have already been taken towards implementing the Linked Data principles. It still needs further focusing on ‘data’ instead of the tool, however. At this moment some unnecessary programming is needed, so developers should take the Linked Data functions in their tools to the next level. The benefits and future development of Linked Data are important enough to do so.

**References**


Lukas Koster. Infrastructure for heritage institutions – Open and Linked Data, June 2021. URL http://purl.org/cpl/3227


Robert Sanderson, Paolo Ciccarese, and Benjamin Young. Web Annotation Vocabulary, February 2017. URL [https://www.w3.org/TR/annotation-vocab/](https://www.w3.org/TR/annotation-vocab/).

Mark D. Wilkinson et. al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), March 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.18. URL [https://www.nature.com/articles/sdata201618](https://www.nature.com/articles/sdata201618).