Linked Art Data

Case study of the Kunsthistorisches Institut in Florenz – Max-Planck-Institut database as it appears within the PHAROS consortium.

Julius Emmel
(Writing – original draft, Data curation, Formal analysis, Investigation)[1]
Demival Vasques Filho, https://orcid.org/0000-0002-4552-0427
(Writing – review & editing, Conceptualization, Supervision)
Thorsten Wübbena, https://orcid.org/0000-0001-8172-6097
(Writing – review & editing, Conceptualization, Funding acquisition)

[1] All the role descriptions correspond to the Contributor Roles Taxonomy (CRediT), https://credit.niso.org/

Introduction – Goals

This report covers the results of the work with a dataset from the ‘Photothek’ of the ‘Kunsthistorisches Institut in Florenz’.

The given dataset consists of 480 .trig files. At the beginning there was not much more known about the dataset, then it must contain information about photos and artworks in the ‘Photothek’. The main goal of the project was to find which information the dataset contains and what could be done with it.

Therefore, the structure of the files and how the information could be extracted had to be understood. Because this is a crucial part of the whole project, this report will spend most of it’s time to concentrate on this matter. The report should explain, why certain decisions had been made and how further work with the data could go on. In a second section the report will show what information the dataset holds and recommends following steps.

The dataset was manly explored by reading the .trig files and python code. The code will also be discussed in this report.

Basics

Additionally to the dataset some more technologies and information were given.

For a more in-depth exploration of the dataset the software ‘Protégé’[2] was used. Without much knowledge about the program, single .trig files and their content could be shown in tree and graph structures. This was helpful to get an understanding of the structure and the scope of the dataset. Since there was no obvious function to import more than one .trig file into ‘Protégé’ at once, the software could not be used to access the whole dataset.

Furthermore there was the PDF-file ‘Structure of APS’ in which the structure of three APS Objects is shown. This structure could not be recognized in the dataset. Based on the APS structure some research questions were written down in an excel spreadsheet called ‘Fields and questions_Pharos_Fields’. Some of these questions referred to information which is not given by the dataset. These questions are left to be answered.

The structure of the dataset

As mentioned above, the dataset consists of 480 .trig files. The following example shows the structure of the first lines of file number 3041:

```
{
}

<https://pharos.artresearch.net/resource/khi/type/61D5FA47-8AD7-3385-9C63-
2C41047112F1/graph> {
   <https://pharos.artresearch.net/resource/khi/type/TO2BEVSN/graph>
         a      <https://pharos.artresearch.net/custom/Namedgraph> ;
         <https://pharos.artresearch.net/custom/has_provider>
               <https://pharos.artresearch.net/resource/khi/source/KHI> .

   <https://pharos.artresearch.net/resource/khi/type/61D5FA47-8AD7-3385-9C63-
```

[2] Pr4 UG - Protege Wiki (stanford.edu)

```
2C41047112F1>
        a       <http://www.cidoc-crm.org/cidoc-crm/E55_Type> ;
        <http://www.w3.org/2000/01/rdf-schema#label>
              "73D64(+5)" ;
        <http://www.cidoc-crm.org/cidoc-crm/P2_has_type>
              <https://pharos.artresearch.net/resource/khi/type/9220BFE7-066C-3521-A053-
795BAA0F74E2> ;
        <http://www.cidoc-crm.org/cidoc-crm/P67i_is_referred_to_by>
              <https://pharos.artresearch.net/resource/khi/type/TO2BEVSN/graph> ;
        <https://pharos.artresearch.net/custom/has_provider>
              <https://pharos.artresearch.net/resource/khi/source/KHI> .
}
```

The element which is the main separator between data is '}' It structures the document in different sections which will be called containers. Every file starts with an empty container. A container is also divided by a '{'. Infront of this divider stands the URI for the whole container. In this case the id is highlighted in yellow. The URI also gives a hint to a datatype. In this case it is 'type', marked in dark yellow. The URI also appears in the second section of a container. This section has as core elements pairs of data. These pairs are bundled in paragraphs and are separated by ';'. The paragraphs are separated by '.'

```
{
}

<https://pharos.artresearch.net/resource/khi/type/61D5FA47-8AD7-3385-9C63-
2C41047112F1/graph> {
   <https://pharos.artresearch.net/resource/khi/type/TO2BEVSN/graph>
        a       <https://pharos.artresearch.net/custom/Namedgraph> ;
        <https://pharos.artresearch.net/custom/has_provider>
              <https://pharos.artresearch.net/resource/khi/source/KHI> .

   <https://pharos.artresearch.net/resource/khi/type/61D5FA47-8AD7-3385-9C63-
2C41047112F1>
        a       <http://www.cidoc-crm.org/cidoc-crm/E55_Type> ;
        <http://www.w3.org/2000/01/rdf-schema#label>
              "73D64(+5)" ;
        <http://www.cidoc-crm.org/cidoc-crm/P2_has_type>
              <https://pharos.artresearch.net/resource/khi/type/9220BFE7-066C-3521-A053-
795BAA0F74E2> ;
        <http://www.cidoc-crm.org/cidoc-crm/P67i_is_referred_to_by>
              <https://pharos.artresearch.net/resource/khi/type/TO2BEVSN/graph> ;
        <https://pharos.artresearch.net/custom/has_provider>
              <https://pharos.artresearch.net/resource/khi/source/KHI> .
}
```

This basic structure is the same for every datatype. They differ in the order, content and structure of their paragraphs. For the datatype 'type' the first paragraph has two data pairs. They point to a graph with an id and the provider of the data. The attributes of the datatype are listed in the second paragraph. They are described with an CIDOC descriptor and are paired with a datatype and its id. Beside the reappearing container URI, the provider and the graph are reappearing in this paragraph. The 'P2_has_type' category points to another URI. In many cases this URI is represented with another container inside the .trig file. This is not the case here. Lastly there is also a label. For other datatypes these labels have more value.

```
{
}

<https://pharos.artresearch.net/resource/khi/type/61D5FA47-8AD7-3385-9C63-
2C41047112F1/graph> {
   <https://pharos.artresearch.net/resource/khi/type/TO2BEVSN/graph>
         a       <https://pharos.artresearch.net/custom/Namedgraph> ;
         <https://pharos.artresearch.net/custom/has_provider>
               <https://pharos.artresearch.net/resource/khi/source/KHI> .

   <https://pharos.artresearch.net/resource/khi/type/61D5FA47-8AD7-3385-9C63-
2C41047112F1>
         a       <http://www.cidoc-crm.org/cidoc-crm/E55_Type> ;
         <http://www.w3.org/2000/01/rdf-schema#label>
               "73D64(+5)" ;
         <http://www.cidoc-crm.org/cidoc-crm/P2_has_type>
               <https://pharos.artresearch.net/resource/khi/type/9220BFE7-066C-3521-A053-
795BAA0F74E2> ;
         <http://www.cidoc-crm.org/cidoc-crm/P67i_is_referred_to_by>
               <https://pharos.artresearch.net/resource/khi/type/TO2BEVSN/graph> ;
         <https://pharos.artresearch.net/custom/has_provider>
               <https://pharos.artresearch.net/resource/khi/source/KHI> .
}
```

In the following, more examples for datatypes will be shown. This takes in account the huge variety in themself.

Technique

```
<https://pharos.artresearch.net/resource/midas/vocabulary/technique/B3D29692-6F9F-30CE-9E42-558C85BE3301/graph> {
   <https://pharos.artresearch.net/resource/midas/vocabulary/technique/1HRE7ICH/graph>
         a       <https://pharos.artresearch.net/custom/Namedgraph> ;
         <https://pharos.artresearch.net/custom/has_provider>
               <https://pharos.artresearch.net/resource/khi/source/KHI> .

   <https://pharos.artresearch.net/resource/midas/vocabulary/technique/B3D29692-6F9F-30CE-9E42-558C85BE3301>
         a       <http://www.cidoc-crm.org/cidoc-crm/E29_Design_or_Procedure> ;
         <http://www.w3.org/2000/01/rdf-schema#label>
               "Pinsel" ;
         <http://www.cidoc-crm.org/cidoc-crm/P67i_is_referred_to_by>

<https://pharos.artresearch.net/resource/midas/vocabulary/technique/1HRE7ICH/graph> ;
         <https://pharos.artresearch.net/custom/has_provider>
               <https://pharos.artresearch.net/resource/khi/source/KHI> .
}
```

Actor

```
<https://pharos.artresearch.net/resource/khi/actor/07870024/graph> {
   <https://pharos.artresearch.net/resource/khi/actor/07870024/graph>
         a       <https://pharos.artresearch.net/custom/Namedgraph> ;
         <https://pharos.artresearch.net/custom/has_provider>
               <https://pharos.artresearch.net/resource/khi/source/KHI> .

   <https://pharos.artresearch.net/resource/khi/actor/07870024>
         a       <http://www.cidoc-crm.org/cidoc-crm/E21_Person> ;
         <http://www.w3.org/2000/01/rdf-schema#label>
               "Belletti, Luigi" ;
         <http://www.cidoc-crm.org/cidoc-crm/P67i_is_referred_to_by>
               <https://pharos.artresearch.net/resource/khi/actor/07870024/graph> ;
         <https://pharos.artresearch.net/custom/has_provider>
               <https://pharos.artresearch.net/resource/khi/source/KHI> .
}
```

Material

```
<https://pharos.artresearch.net/resource/midas/vocabulary/material/B84A9A38-377D-31AB-AC6D-5005CB461F0D/graph> {
   <https://pharos.artresearch.net/resource/midas/vocabulary/material/B84A9A38-377D-31AB-AC6D-5005CB461F0D>
         a       <http://www.cidoc-crm.org/cidoc-crm/E57_Material> ;
         <http://www.w3.org/2000/01/rdf-schema#label>
               "Sepia" ;
         <http://www.cidoc-crm.org/cidoc-crm/P67i_is_referred_to_by>

<https://pharos.artresearch.net/resource/midas/vocabulary/material/1HRRJQV3/graph> ;
         <https://pharos.artresearch.net/custom/has_provider>
               <https://pharos.artresearch.net/resource/khi/source/KHI> .

   <https://pharos.artresearch.net/resource/midas/vocabulary/material/1HRRJQV3/graph>
         a       <https://pharos.artresearch.net/custom/Namedgraph> ;
         <https://pharos.artresearch.net/custom/has_provider>
               <https://pharos.artresearch.net/resource/khi/source/KHI> .
}
```

Place

```
<https://pharos.artresearch.net/resource/khi/place/9E063D2C-EDB8-3772-B106-53AAE6EA4EEF/graph> {
   <https://pharos.artresearch.net/resource/khi/place/VF6EA3GH/graph>
         a      <https://pharos.artresearch.net/custom/Namedgraph> ;
         <https://pharos.artresearch.net/custom/has_provider>
               <https://pharos.artresearch.net/resource/khi/source/KHI> .

   <https://pharos.artresearch.net/resource/khi/place/9E063D2C-EDB8-3772-B106-53AAE6EA4EEF>
         a      <http://www.cidoc-crm.org/cidoc-crm/E53_Place> ;
         <http://www.w3.org/2000/01/rdf-schema#label>
               "Sarteano" ;
         <http://www.cidoc-crm.org/cidoc-crm/P2_has_type>
               <http://vocab.getty.edu/aat/300008389> ;
         <http://www.cidoc-crm.org/cidoc-crm/P67i_is_referred_to_by>
               <https://pharos.artresearch.net/resource/khi/place/VF6EA3GH/graph> ;
         <https://pharos.artresearch.net/custom/has_provider>
               <https://pharos.artresearch.net/resource/khi/source/KHI> .
}
```

Collection

```
<https://pharos.artresearch.net/resource/khi/collection/A7BBE344-528E-3511-B587-6E87B96C4A99/graph> {
   <https://pharos.artresearch.net/resource/khi/collection/A7BBE344-528E-3511-B587-6E87B96C4A99>
         a      <http://www.cidoc-crm.org/cidoc-crm/E78_Curated_Holding> ;
         <http://www.w3.org/2000/01/rdf-schema#label>
               "Kress Collection" ;
         <http://www.cidoc-crm.org/cidoc-crm/P67i_is_referred_to_by>
               <https://pharos.artresearch.net/resource/khi/collection/1H7ZRM6V/graph> ;
         <https://pharos.artresearch.net/custom/has_provider>
               <https://pharos.artresearch.net/resource/khi/source/KHI> .

   <https://pharos.artresearch.net/resource/khi/collection/1H7ZRM6V/graph>
         a      <https://pharos.artresearch.net/custom/Namedgraph> ;
         <https://pharos.artresearch.net/custom/has_provider>
               <https://pharos.artresearch.net/resource/khi/source/KHI> .
}
```

All these datatypes can be considered simple. They have a very similar structure – the biggest difference is the changing order of the paragraphs. Many of these entities are found in different .trig files. This is, since the much more important datatype ‘work’ always refers to them. When for example two artworks from the place ‘Sarteano’ are stored in two different .trig files, the container for the artwork will point to one specific object with one id. This object has to be present in both .trig files.

As mentioned above, the more interesting datatype is ‘work’. In this datatype all the simple data is connected to a larger form.

work

<https://pharos.artresearch.net/resource/khi/work/07903036/graph> {
<https://pharos.artresearch.net/resource/khi/work/07903036/dimension/-/-/5C6B8490-601D-37D8-8E15-BC933877E718>
a <http://www.cidoc-crm.org/cidoc-crm/E54_Dimension> ;
<http://www.w3.org/2000/01/rdf-schema#label>
"18,2 x 24,7 cm" .

<https://pharos.artresearch.net/resource/khi/work/07903036/visual_item>
a <http://www.cidoc-crm.org/cidoc-crm/E36_Visual_Item> ;
<http://www.cidoc-crm.org/cidoc-crm/P138_represents>
<https://pharos.artresearch.net/resource/khi/type/65C75E2E-FEE2-39BB-9A20-760B4C637C35> .

<https://pharos.artresearch.net/resource/khi/work/07903036/production>
a <http://www.cidoc-crm.org/cidoc-crm/E12_Production> ;
<http://www.cidoc-crm.org/cidoc-crm/P01i_is_domain_of>

<https://pharos.artresearch.net/resource/khi/carried_out_by/07903036/07900007> ;
<http://www.cidoc-crm.org/cidoc-crm/P2_has_type>
<https://pharos.artresearch.net/resource/midas/vocabulary/type/A9008C8A-B1E5-3B5D-A98D-C4039CA79DB6> .

<http://photothek.khi.fi.it/documents/obj/07903036>
a <http://www.cidoc-crm.org/cidoc-crm/E32_Identifier> ;
<http://www.cidoc-crm.org/cidoc-crm/P2_has_type>
<http://vocab.getty.edu/aat/300404630> .

<https://pharos.artresearch.net/resource/khi/work/07903036/graph>
a <https://pharos.artresearch.net/custom/Namedgraph> ;
<https://pharos.artresearch.net/custom/has_provider>
<https://pharos.artresearch.net/resource/khi/source/KHI> .

<https://pharos.artresearch.net/resource/khi/carried_out_by/07903036/07900007>
a <http://www.cidoc-crm.org/cidoc-crm/PC14_carried_out_by> ;
<http://www.cidoc-crm.org/cidoc-crm/P02_has_range>
<https://pharos.artresearch.net/resource/khi/actor/07900007> ;
<http://www.cidoc-crm.org/cidoc-crm/P14.1_in_the_role_of>
<https://pharos.artresearch.net/resource/midas/vocabulary/type/48378153-A40D-3533-9CD7-EF6D02595E0C> .

<https://pharos.artresearch.net/resource/khi/work/07903036/acquisition/D2BF0A6A-E33C-38CE-B18F-BA633E2EF71A>
a <http://www.cidoc-crm.org/cidoc-crm/E8_Acquisition> ;
<http://www.w3.org/2000/01/rdf-schema#label>
"Location: Biblioteca Classense Acquisition: Verwalter -" ;
<http://www.cidoc-crm.org/cidoc-crm/P2_has_type>
<https://pharos.artresearch.net/resource/midas/vocabulary/type/70F5C51D-4184-3399-847A-74125BA922CD> .

<https://pharos.artresearch.net/resource/khi/work/07903036>
a <http://www.cidoc-crm.org/cidoc-crm/E22_Man-Made_Object> ;
<http://www.w3.org/2000/01/rdf-schema#label>
"Bogenarchitektur" ;
<http://www.cidoc-crm.org/cidoc-crm/P108i_was_produced_by>
<https://pharos.artresearch.net/resource/khi/work/07903036/production> ;
<http://www.cidoc-crm.org/cidoc-crm/P138i_is_represented_by>
<https://pharos.artresearch.net/resource/khi/work/07903036/visual_item> ;

```
        <http://www.cidoc-crm.org/cidoc-crm/P1_is_identified_by>

<https://pharos.artresearch.net/resource/khi/work/07903036/appellation/817CE112-DA80-
3109-8153-79782BE9DA0D> ,
<https://pharos.artresearch.net/resource/khi/work/07903036/identifier/2AF8F7C0-C520-3BEF-
B5A8-FC84377862B0> ;
        <http://www.cidoc-crm.org/cidoc-crm/P24i_changed_ownership_through>

<https://pharos.artresearch.net/resource/khi/work/07903036/acquisition/D2BF0A6A-E33C-
38CE-B18F-BA633E2EF71A> ;
        <http://www.cidoc-crm.org/cidoc-crm/P2_has_type>
                <https://pharos.artresearch.net/resource/midas/vocabulary/type/36429239-B8D7-
3336-BC98-8983B18D007B> ,
<https://pharos.artresearch.net/resource/midas/vocabulary/type/DEAF3704-4A2A-3F4C-
9C40-35847AC7711C> , <http://vocab.getty.edu/aat/300133025> ;
        <http://www.cidoc-crm.org/cidoc-crm/P43_has_dimension>
                <https://pharos.artresearch.net/resource/khi/work/07903036/dimension/-/-
/5C6B8490-601D-37D8-8E15-BC933877E718> ;
        <http://www.cidoc-crm.org/cidoc-crm/P45_consists_of>
                <https://pharos.artresearch.net/resource/midas/vocabulary/material/C1350D83-
C54D-3880-A6AD-37A2D5B1F2F6> ;
        <http://www.cidoc-crm.org/cidoc-crm/P50_has_current_keeper>
                <https://pharos.artresearch.net/resource/khi/actor/EA8D51D0-C028-3908-9352-
1B9F9B2343E2> ;
        <http://www.cidoc-crm.org/cidoc-crm/P67i_is_referred_to_by>
                <https://pharos.artresearch.net/resource/khi/work/07903036/graph> ;
        <https://pharos.artresearch.net/custom/has_original_record>
                <http://photothek.khi.fi.it/documents/obj/07903036> ;
        <https://pharos.artresearch.net/custom/has_provider>
                <https://pharos.artresearch.net/resource/khi/source/KHI> .

    <https://pharos.artresearch.net/resource/khi/work/07903036/appellation/817CE112-DA80-
3109-8153-79782BE9DA0D>
        a       <http://www.cidoc-crm.org/cidoc-crm/E41_Appellation> ;
        <http://www.w3.org/2000/01/rdf-schema#label>
                "Bogenarchitektur" ;
        <http://www.cidoc-crm.org/cidoc-crm/P2_has_type>
                <http://vocab.getty.edu/aat/300417193> .

    <https://pharos.artresearch.net/resource/khi/work/07903036/identifier/2AF8F7C0-C520-
3BEF-B5A8-FC84377862B0>
        a       <http://www.cidoc-crm.org/cidoc-crm/E42_Identifier> ;
        <http://www.w3.org/2000/01/rdf-schema#label>
                "vol.Ms.657" ;
        <http://www.cidoc-crm.org/cidoc-crm/P2_has_type>
                <http://vocab.getty.edu/aat/300404621> .

    <http://vocab.getty.edu/aat/300133025>
        <http://www.cidoc-crm.org/cidoc-crm/P2_has_type>
                <https://pharos.artresearch.net/resource/vocab/maintype> .
}
```

The example shows a quite simple form of the work class. This work consists of eleven paragraphs, which can be described as ‘internal’ (5), ‘external’ (3), ‘hybrid’ (1), ‘combined’ (1) and ‘special’ (1). Internal paragraphs contain information which is only used in the entity itself and other entities. It is not a datatype on the same hierarchy as ‘work’ or other datatypes. This data is represented with the bluish grey. External paragraphs contain data which appears also in other entities and is represented by a datatype like ‘place’ or ‘actor’. It is represented with light yellow. Hybrid paragraphs contain internal and external data. The combined paragraph represents all internal data in the entity and is the reference for the internal paragraphs and vice versa. It also contains data which is not mentioned in the other paragraphs and complements the information we get about the artwork. The special paragraph appears also in the other datatypes and points to a ‘Namedgraph’.[3]

The example also shows a special case in the combined paragraph. Instead of simple pairs of data the descriptors ‘P1_is_identified_by’ and ‘P2_has_type’ have multiple datatypes.

What the example does not show, is the reference to a photograph. This reference is important because many questions from the excel spreadsheet refer to the data of the photographers and photographs. The following example of two paragraphs from another work entity shows how such a reference would look like.

```
<https://pharos.artresearch.net/resource/khi/work/70000491/visual_item>
        a       <http://www.cidoc-crm.org/cidoc-crm/E36_Visual_Item> ;
        <http://www.cidoc-crm.org/cidoc-crm/P128i_is_carried_by>
                <https://pharos.artresearch.net/resource/khi/negative/56520/B> ,
<https://pharos.artresearch.net/resource/khi/negative/56520/A> ,
<https://pharos.artresearch.net/resource/khi/photo/fln0610192z_p> ,
<https://pharos.artresearch.net/resource/khi/photo/fln0610191z_p> ;
        <http://www.cidoc-crm.org/cidoc-crm/P138_represents>
                <https://pharos.artresearch.net/resource/khi/type/08A15069-FD80-31BA-A05A-
0D9402F29052> .

   <https://pharos.artresearch.net/resource/khi/photo/fln0610192z_p>
        a       <http://www.cidoc-crm.org/cidoc-crm/E22_Man-Made_Object> .
```

The first paragraph contains four references to two negatives and two photographs. For every reference exists a paragraph like the second one in the example, since these references are internal. The only external reference leads to ‘type’ datatype and contains no information with greater value. There is no other information beside the ids of photos and negatives in the dataset. Questions in regard of who took the photo at which point of time, can’t be answered.

[3] This paragraph is special because it is represented in all classes and is always the same. All ‘Namedgraph’s point to a datatype ‘namedgraphs’. This datatype consists of two types of paragraphs. Type one points back to every datatype in the .trig file. Type two is a list of all datatypes in the .trig file. Fig. 1 shows a screenshot of the .trig file ‘formated-part_3041_cleaned.trig’ in which both types of paragraphs can be seen.

Code

This chapter looks at the two python scripts ‘entity_extraction.py’ and ‘dictionary_cleaning.ipynb’. These two files were used to collect and restructure all information from the .trig files. They produced three outputs with the name ‘entity_dump.json’, ‘entity_dump_cleaned.json’ and ‘entity_dump_cleaned_and_restructured.json’. Therefore the code doesn’t need to be rerun, but the understanding of it is important for the understanding of the three dump files.

‘entity_extraction.py’ was used to create the first dump ‘entity_dump.json’. Therefore all filenames had to be collected (lines 4-6). A for-loop follows which iterates over all files. This loop also defines the structure of the final dictionary, which will be dumped. This will become clearer at the end of the file.

At first the data from the file is loaded in lines 11 and 12. In line 15 a first split happens to separate the different containers. This step is combined with a transformation of certain text elements. This is important because of the labels in each datatype. In this field stands nearly every form of text. It could be a name or the title of a book or an inventory number. All separators of data are used in the labels and in different combinations. For python the loaded data is just a very long string without any structure. To separate at the right position, these separators had to be replaced or the labels to be transformed. The order of the replacements is also crucial, because some of them had to be performed before others could be.

The first replacement is needed, because the ‘.’ is used as a separator and in book titles in a group of three. The important difference between triple dots and a single dot is that the triplet has a space in front of it. Therefore, the separator ‘.’ couldn’t be differentiated from the first dot of the triplet because it has also a space in front of it. Other single dots in the label texts always have another character in front of them.

In the next step all ‘ .’ were replaced with a ‘§’. This is necessary because there are still ‘.’ In the label texts. But only the dots with a space in front of them could now be separators. Because in the next step all spaces should be eliminated, another separator had to be found or else the separator would look like a normal dot.

The following two replacements for ‘”;’ and ‘>;’ are needed to replace the separator ‘;’ since it is also used in the label texts. But it only appears with a ‘”’ and a ‘>’ in front of it when it is a separator.

In some label texts the ‘”’ were escaped via ‘\\’.[4] They had to be replaced with ‘’’ to ensure that label text only starts and ends with ‘”’.

Lastly all eventual line breaks and ‘https://” were replaced to clean the data upfront.

Further cleaning happens in lines 18 to 24. Due to the structure of the document, there will always be one empty list element at the beginning and the end. For safety all elements will be checked, if they are empty and these elements will then be popped.

In line 27 every entity will be split into the id-URI and the attributes.

[4] Example – File 3041, line 17758: "Regatta in der \"Volta de Canal\""

Afterwards in lines 31 to 47 the id-URI will be split again and the id and the datatype. At the end both elements will be brought together in a string.[5]

In lines 51 to 65 the attributes will be selected and split twice to have a list of paragraphs containing a list of data pairs.

In lines 69 to 98 these two-dimensional lists will be cleaned. The structure of the data necessitates to check for three variations. In one case an ‘a’ stands between the two parts of a data pair, in the other case nothings stands between them. The third case applies to label texts, which also need to be treated in a certain way.[6]

Lastly, in lines 103 to 115 all data will be put together in a way that it can be exported as .json.

This process will be performed for every document in the directory.

In the beginning the main idea for the structure of the .json file was, to preserve the information, which data is stored in which file. This proved to be unpracticable and unnecessary

The second script ‘dictionary_cleaning.ipynb’ applies two little changes to the ‘entity_dump.json’. They take place in the chapters ‘cleaning’ and ‘transformation’ of that notebook. The cleaning part consists of the exclusion of the dictionary key ‘namedgraphs(KHI)’. As mentioned in the chapter ‘The structure of the dataset’, ‘namedgraphs’ doesn’t have much useful data to it and takes up a lot of space since every entity in the file has to refer to it.
The transformation takes on the structure of the file and removes the division of the data into their respective files.

The other two chapters in the notebook are two early analyses of the data without the regard of the connection between the entities.

A more in depth analysis takes place in the script ‘class_statistics.ipynb’. This file should contain enough information to pass on an extended presentation here.

[5] In retrospective the lines 34 to 47 could be combined in one for loop.
[6] In the example for the ‘work’ datatype the seventh paragraph shows all three cases.

Conclusion

The data in the .trig files is highly structured. The main information containers are separated by '}'. They contain an entity of a datatype with a unique id. Each entity of information contains paragraphs, consisting of data pairs. A huge part of the data is repetitive. Many data pairs refer to other pairs in the same entity.

The datatypes have seven main categories. They are 'work', 'place', 'actor', 'type', 'technique', 'material' and 'collection'. 'Work' is the most important category because all other categories are primarily linked to it. The categories 'actor' and 'type' are very heterogenous. An 'actor' can be e. g. an artist or an institution.

To work properly with the data, it could be better to simplify it. Most parts of the different URIs are the same and can be ignored. The many layers of the data structure and the different separators make this process of simplification somewhat complex. The code which was therefore written has room for improvements. One important aspect is, that the code couldn't extract complex data pairs with a list of data as one part of the pair. In addition, the overall process that separates the huge data string in its smaller parts could be broken up into several loops. This could reduce the need for the many different separators which are created at the beginning of the script.

Figures

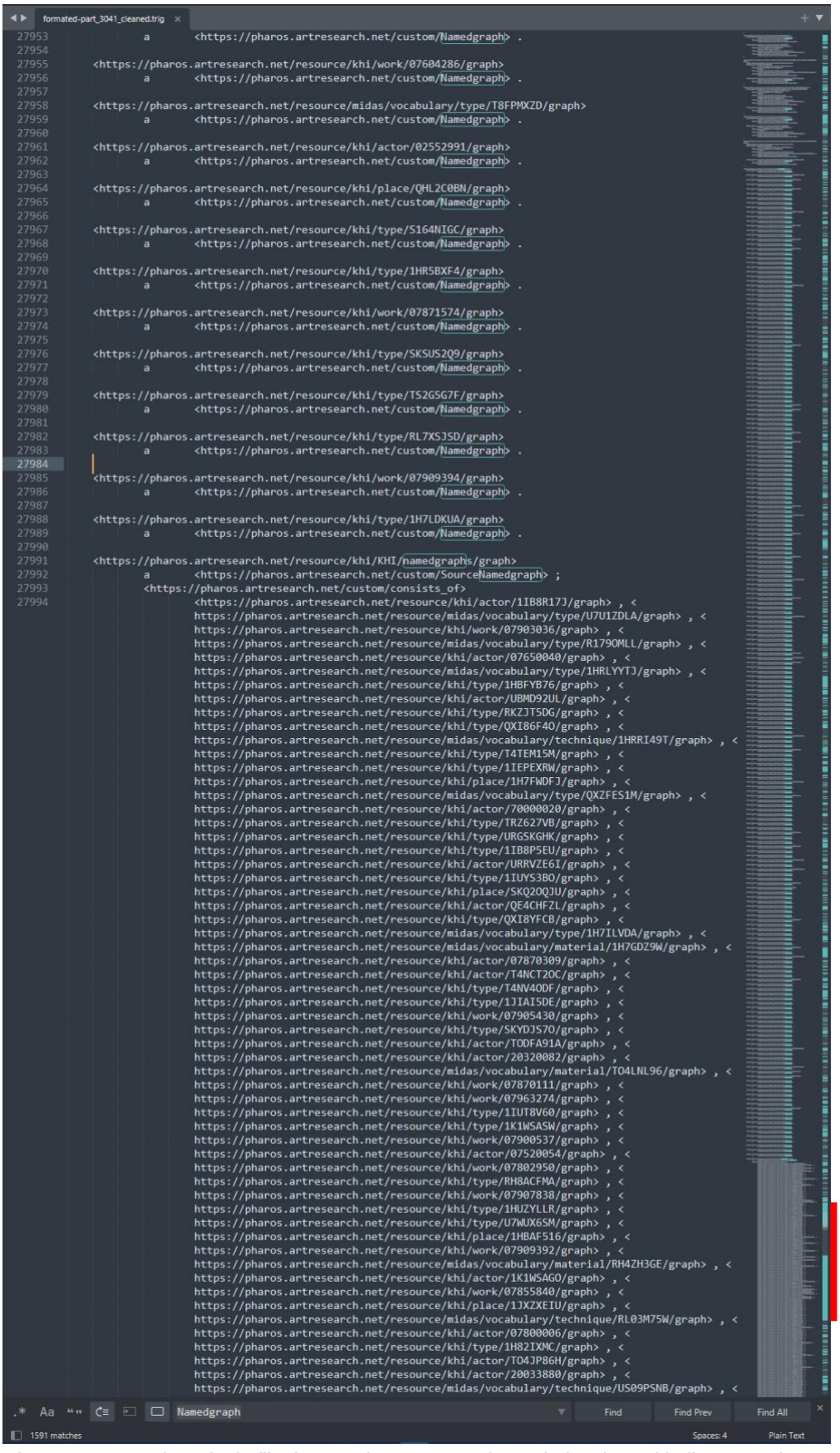

Figure 1: Screenshot of .trig file 'formated-part_3041_cleaned.trig' The red indicator on the right shows the length of the datatype ‚namedgraphs' in relation to the whole document.