

SSHOC-NL T4.2 Deliverable 2025-D10

Utility of Croissant metadata in the ODISSEI Portal: An assessment of the current implementation

Date: 2026-01-05

Version: 1.0 Published on Zenodo

DOI: <https://doi.org/10.5281/zenodo.18152866>

Authors: Ricarda Braukmann (DANS), Laura Huis in 't Veld (DANS)

Contributors: Jetze Toubert (DANS), Angelica Maineri (ODISSEI)

Goal of the Deliverable

This deliverable takes a closer look at the Croissant metadata export available in the ODISSEI Portal¹. Croissant² is a metadata format that builds on the schema.org/Dataset vocabulary and is made specifically to support Machine Learning applications.

Croissant was added to the export options of the Dataverse software³ in 2024 (version 6.4⁴) amongst others by Slava Tykhonov who worked on Croissant in the SSHOC-NL⁵ project (Tykhonov & Durbin, 2024).

Since the ODISSEI Portal makes use of Dataverse software, the Croissant metadata export is an available feature.

This deliverable aims to look closer at this export to discuss usability of the Croissant metadata in its current form. We assessed the Croissant export by comparing it to other metadata outputs currently available in the Portal like Schema.org JSON-LD and DDI Codebook. The deliverable finishes with a set of recommendations for future developments.

Introduction

Croissant format

The format specification⁶ provides detailed information about Croissant and this deliverable provides a brief description.

Croissant builds on the schema.org/Dataset vocabulary which is widely adopted. It can be seen as the successor to the format known as "Schema.org JSON-LD" available in Dataverse and used by Google Dataset Search, for instance.

Croissant's focus is to contribute to responsible AI by providing standardized ways of documenting data, in particular for datasets used in Machine Learning (ML). It proposes a machine-readable way to capture and publish metadata about such datasets and records how a dataset was created, processed and enriched and it is designed to integrate with popular ML frameworks.

With AI and ML becoming increasingly relevant across domains, making datasets "AI-ready" is an important development and Croissant has been developed to provide a standard for how ML datasets are described.

Croissant export in the ODISSEI Portal

The development of Croissant was very much welcomed by the ODISSEI Portal team as it provides an additional format with which metadata in the Portal can be exported in a standardized way that is accepted by a wide community. The Croissant export was directly enabled in the ODISSEI Portal when it became available as one of the Dataverse metadata export formats⁷.

It should be noted that Croissant's focus is to make datasets "AI-ready", for instance by providing metadata about file structures so that ML frameworks can easily integrate these datasets. The ODISSEI Portal is a metadata catalogue which does not contain any files so file information will not be present in the metadata outputs.

¹ The ODISSEI Portal is available at: <https://portal.odissei.nl>

² The Croissant documentation is available at <https://docs.mlcommons.org/croissant/docs/croissant-spec.html>

³ Information about the Dataverse software and its global consortium is available at <https://dataverse.org/>

⁴ See the release notes for version 6.4 here <https://github.com/IQSS/dataverse/releases/tag/v6.4>

⁵ The project is a collaboration between ODISSEI (Open Data Infrastructure for Social Science and Economic Innovations) and CLARIAH (Common Lab Research Infrastructure for the Arts and Humanities) that develops state of the art infrastructure for social science and humanities research and is financed by the Dutch Research Council (NWO) Large-scale Research Infrastructure Grant. <https://sshoc.nl>

⁶ The format specification is available at <https://github.com/mlcommons/croissant/blob/main/docs/croissant-spec.md>

⁷ More information about the Dataverse metadata export formats is available here:

<https://guides.dataverse.org/en/latest/developers/metadataexport.html>

In the future, one can imagine that data providers will provide such information and the Portal extends its mappings to make them available in Croissant.

Below, we describe the information that Croissant contains compared to other metadata exports for the metadata-only use case of the Portal, but we also provide an example from a dataset in a Dataverse repository to provide the full picture.

Available metadata export formats

In the Portal, there are various formats for exporting the metadata each with their own benefits and disadvantages. It should be noted that the exports provided are the formats that the Dataverse software provides by default.

The **JSON** and the **OAI_ORE** exports contain all metadata fields that a given record in the Portal has. This includes fields from Dataverse standard metadata blocks as well as fields from custom blocks. The latter are blocks which are not part of the Dataverse standard metadata blocks but have been created specifically for the ODISSEI Portal. JSON and OAI_ORE are thus exports of all the information available in the metadata blocks from the Portal.

DataCite, **OpenAIRE** and **DublinCore** provide exports following the DataCite, OpenAIRE and DublinCore standard, respectively. These exports provide a selection of fields based on a mapping of the Dataverse fields to the respective standard. The custom fields of the Portal are not included in the default mapping to these standards and are hence not visible in the exports.

The **DDI** export provides an export according to the fields identified in the Data Documentation Initiative standard. DDI is widely used in the social sciences and focuses strongly on social sciences methodology and variable information. While this export provides a selection of fields that can be mapped to the fields in DDI, it includes information specifically relevant to the social sciences such as social science-specific metadata. DDI export can also include variable information if this information is extracted from a tabular data file through the tabular data ingest functionality of Dataverse. Dataverse offers an xml version and a html version of DDI of which the latter provides a user-friendly overview of the different information available.



Figure 1. Available metadata exports in the ODISSEI Portal

Schema.org based formats

For a linked data representation of metadata, schema.org is a widely used format. In Dataverse the **Schema.Org JSON-LD** export has been available for a while. With **Croissant** now a second linked data representation based on schema.org is available. As mentioned above, Croissant provides an extension of schema.org especially geared towards datasets for AI and machine learning.

Comparing Croissant with other export formats

Croissant and Schema.org JSON-LD

As mentioned above Croissant is an extension of schema.org and it is hence logical to compare the information provided in Croissant with that provided in the Schema.org JSON-LD export.

As an example, record for this comparison we used the “*Dutch Parliamentary Election Study 2012 (DPES/NKO 2012)*” in the Portal⁸. You can download the metadata records yourself through “Export Metadata” > “Schema.org JSON-LD” and “Export Metadata” > “Croissant” but the used files will also be part of the Zenodo entry⁹ of this deliverable.

In both exports, we see information about creators, descriptions, keywords, publishers, publication dates, license and funder.

One difference is that the Croissant output starts with a more elaborate “context” section, which contains definitions that make the rest of the Croissant description less verbose. Croissant also provides information about the correct citation for the dataset which is not included in the Schema.org JSON-LD export.

Note that none of the information from the custom blocks of the Portal (e.g. enrichments, provider specific metadata) are represented in either of these exports as they follow a mapping to the standards. As mentioned above, JSON and OAI-ORE are the only exports that include information from the custom blocks.

Croissant and DDI

As the Portal focuses on Social Sciences data it is important to understand how Social Science specific information is represented in the metadata exports. Hence, we compare the Croissant export with the DDI exports.

For this we use another example that has rich information about Social Sciences methods in the Social Sciences and Humanities Metadata block, namely the CBS dataset “*Kenmerken van vorderingen van gemeenten op (ex-)ontvangers van een bijstandsuitkering*”¹⁰,

You can download the metadata records yourself through “Export Metadata” > “DDI HTML Codebook” and “Export Metadata” > “Croissant” but the used files will also be part of the Zenodo entry of this deliverable.

What we see is that in the DDI export, the information from the social sciences block is exported (for instance “Maand” as Sampling Frequency, see Figure 2), while this information is not available in the

⁸ The dataset is available on the Portal at <https://portal.odisseei.nl/dataset.xhtml?persistentId=doi:10.17026/DANS-X5H-AKDS> The DOI of the dataset is: <https://doi.org/10.17026/DANS-X5H-AKDS>

⁹ The deliverable and metadata exports are available at <https://doi.org/10.5281/zenodo.18152866>

¹⁰ The dataset is available here <https://doi.org/10.57934/0B01E4108080FA8B>

Croissant export. In the DDI export, we also see the information from the Terms fields around data access which are not included in the Croissant export.

Methodology and Processing

Frequency of Data Collection: **Maand**

Sampling Procedure: De bron van d

Figure 2. Information from the Social Sciences metadata block is presented as part of the DDI export in the ODISSEI Portal. In this example “Maand” was listed in the metadata as the Sampling Frequency.

On the other hand, Croissant includes citation information which is not included in DDI.

Croissant exports of data file information

As mentioned above, the strength of Croissant lies in the fact that it provides information about data files available in a given record in a repository. Since the Portal does not include any data files, the export also does not include information about files.

To understand how Croissant relates to other available formats, it is useful to also compare a Croissant export for a dataset from a repository where files are included.

For this, we are taking an example dataset from Borealis¹¹ which uses Dataverse version 6.8.1 and has Croissant export enabled. The example dataset we assessed is "Replication Data and Code for: Language Skills and Labor Market Outcomes of Immigrants in Europe"¹²

As above you can export the metadata files yourself on the landing page of the dataset or refer to the Zenodo package associated with this deliverable.

Croissant and Schema.org JSON-LD – file information

Both Schema.org JSON-LD and Croissant show the file objects under “distribution”. This includes information about where to find the data files. The field “contentURL” includes a URL through which the file can be directly accessed. Croissant includes a file ID which is not included in the Schema.org JSON-LD export. The IDs are used to fulfill the need for a mechanism to refer to parts of the dataset in other places.

More importantly, Croissant also includes a record set which is not available in Schema.org JSON-LD (see Figure 3A and B).

¹¹ We have chosen Borealis as they use Dataverse and include Croissant exports, the Data Station SSH currently does not have Croissant enabled and was hence less suitable for this comparison.

¹² DOI to the example dataset <https://doi.org/10.5683/SP3/HS2TEZ>

```

    @type: "Dataset"
    @id: "https://doi.org/10.5683/SP3/HS2TEZ"
    identifier: "https://doi.org/10.5683/SP3/HS2TEZ"
    name: "Replication Data and Code for: Language Skill
  ▶ creator: [ {...}, {...} ]
  ▶ author: [ {...}, {...} ]
    datePublished: "2025-11-18"
    dateModified: "2025-11-18"
    version: "1"
    description: 'The data and programs replicate tables and fi
  ▶ keywords: [ "Social Sciences" ]
    license: "http://creativecommons.org/licenses/by-nc/4.0
  ▶ includedInDataCatalog: { "@type": "DataCatalog", name: "Borealis", ur
  ▶ publisher: { "@type": "Organization", name: "Borealis" }
  ▶ provider: { "@type": "Organization", name: "Borealis" }
  ▶ distribution: (59) [ {...}, {...}, {...}, {...}, {...}, {...}, {...}, {...},

```

Figure 3A The Schema.org JSON-LD export

```

  ▶ @context: { "@language": "en", "@vocab": "https://s
    @type: "sc:Dataset"
    conformsTo: "http://mlcommons.org/croissant/1.0"
    name: "Replication Data and Code for: Language
    url: "https://doi.org/10.5683/SP3/HS2TEZ"
  ▶ creator: [ {...}, {...} ]
    description: 'The data and programs replicate tables a
  ▶ keywords: [ "Social Sciences" ]
    license: "http://creativecommons.org/licenses/by-n
    datePublished: "2025-11-18"
    dateModified: "2025-11-18"
  ▶ includedInDataCatalog: { "@type": "DataCatalog", name: "Borealis
  ▶ publisher: { "@type": "Organization", name: "Boreali
    version: "1.0"
  ▶ citeAs: "@data{SP3/HS2TEZ_2025,author = {Aydemir,
  ▶ distribution: (59) [ {...}, {...}, {...}, {...}, {...}, {...}, {...},
  ▶ recordSet: (5) [ {...}, {...}, {...}, {...}, {...} ]

```

Figure 3B The Croissant export. We can see that the Croissant export contains a record set which provides information about elements in the files of the dataset. This is not available in the Schema.org JSON-LD export (Figure 3A).

The record set in Croissant is “a set of homogeneous data records, such as a collection of images, text files, or all the rows in a table”¹³. In this example, the fields in the record set contain information about the column headers which are present in the tabular data files of the dataset. For the example dataset, the record set in Croissant starts with the column headers of the file “Cultural Distance Matrix.xlsx” (see Figure 4)

The screenshot displays the Croissant record set structure for the 'Cultural Distance Matrix.xlsx' file. The record set is a collection of fields, each representing a column in the data table. The fields are identified by their names and descriptions, and their data types are specified. The first field is named 'A' and has a description 'A'. The second field is named '40' and has a description '40'. The third field is named '56' and has a description '56'. The fourth field is named '208' and has a description '208'. The fifth field is named '246' and has a description '246'. The sixth field is named '250' and has a description '250'. The seventh field is named '276' and has a description '276'. The eighth field is named '528' and has a description '528'. The ninth field is named '578' and has a description '578'. The tenth field is named '752' and has a description '752'. The data types for all fields are 'sc:Float'.

A	40	56	208	246	250	276	528	578	752
	40	56	208	246	250	276	528	578	752

Figure 4 Screenshot (above) of the record set Croissant displays for the example dataset. Croissant provides information about elements available in the data file “Cultural Distance Matrix.xlsx”. Below is a screenshot of the headers of the corresponding data file.

¹³ See Terminology defined in the Croissant documentation <https://github.com/mlcommons/croissant/blob/main/docs/croissant-spec.md>

Dataverse extracts this information from a tabular file, like an xlsx file, through the **tabular file ingest** function¹⁴. If the file is open for download without restrictions, this information is provided in the metadata exports like Croissant and DDI (see next section).

Croissant and DDI – file information

The information about the column headers from the tabular data files of the dataset which are listed in the record set of Croissant is also available in the DDI export.

The HTML DDI export lists the variable labels and names in a dedicated block on the HTML page, the Variable Description Block (Figure 5).

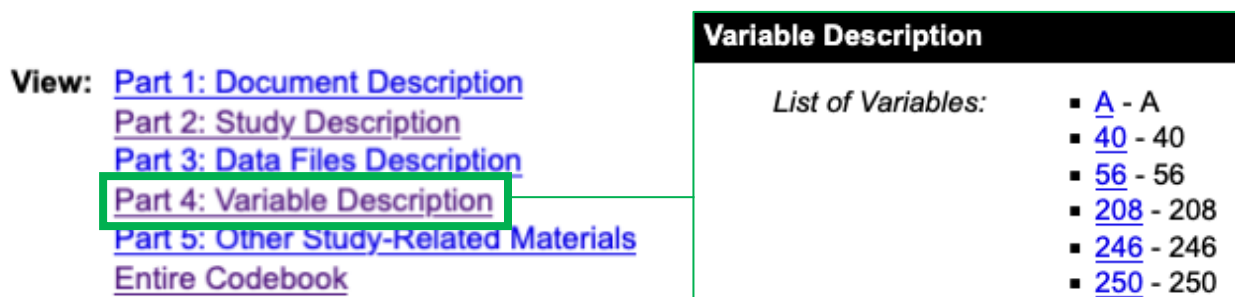


Figure 5 DDI export displaying the variable names and labels in a separate Part.

In addition, the DDI export also provides summary statistics per variable (Figure 6). As with the labels, this information is extracted through the tabular file ingest functionality of Dataverse and included in the metadata export for openly available datasets. This information is currently not included in Croissant although it has been identified as wish list issue within the Croissant community¹⁵

Summary Statistics: StDev 253.40331758782096; Valid 237.0; Min. 4.0; Mean 443.01265822784916; Max. 894.0

Variable Format: numeric

Notes: UNF:6:qUtr+v9r+VKrbGmxld+sg==

Figure 6 Screenshot of the DDI export displaying summary statistics for a given variable.

On the other hand, Croissant provides explicit information about the location of the dataset files which is not present in the DDI export. As mentioned above both Schema.org JSON-LD and Croissant provide a URL through which the file can be directly accessed. DDI does not provide this information. This makes Croissant more suitable for machines, together with the fact that the information is expressed in linked data which is not the case for the DDI export.

CONCLUSIONS

This deliverable aimed to take a closer look at the Croissant metadata export available in the ODISSEI Portal.

Croissant metadata provides an additional metadata export in linked data format based on schema.org, next to the Schema.org JSON-LD. In the Portal - which holds only metadata records - the differences

¹⁴ More information about the tabular file ingest function is available here:

<https://guides.dataverse.org/en/latest/user/tabulardataingest/index.html>

¹⁵ See this issue on the Croissant Github page <https://github.com/mlcommons/croissant/issues/640>

between the two standards are minimal with Croissant giving slightly more detailed information (e.g. more context and citation information).

In a repository holding data, however, Croissant, compared to Schema.org JSON-LD, provides more detailed information about elements present in the data files within a given dataset. As such Croissant for instance includes information about the column headers present in the tabular data files.

The goal of Croissant lies in providing more information about a dataset and its files so that the data can more easily be used within machine learning pipelines and it is hence logical that the biggest advantages for using Croissant appear when information about files is present in the system.

For both the Portal and the data repository we assessed, we do see that the DDI export provides more information than Croissant that is specifically relevant for the social sciences. In the exports of the Portal, information from the social science metadata block is included in the DDI export, while this is not available in Croissant.

With respect to file-specific information in a repository holding data, the DDI export includes summary statistics which is not yet available in Croissant. It would be valuable for Croissant to include this information, in particular the summary statistics, as they may be relevant for users. Since this issue was already raised within the Croissant community, we believe that a future version will likely address this gap.

The advantage of Croissant over DDI is that Croissant is expressed in linked data which is easier to interpret for machines. When information about data files is available, Croissant includes information about the location of each data file which is not available in DDI.

Since DDI is explicitly developed for the social sciences, whereas Croissant focuses on ML dataset, the differences we observed are not particularly surprising. It is nevertheless useful to be aware of the strengths and limitations of each of the available exports and consider how the information provided in these exports can be improved in the future. In the next section, we discuss some directions based on the findings described above.

FUTURE DIRECTIONS

In this deliverable, we took a few example (meta)datasets to illustrate how the Croissant export relates to the other metadata exports available in Dataverse-based services like the ODISSEI Portal¹⁶.

We clearly see that having **information about files** and their content enriches the information available in the metadata exports. Croissant provides more detailed information about files and their contents compared to Schema.org JSON-LD. While even more details are currently available in the DDI export, Croissant holds the benefits of expressing information in linked data and directing machines directly to the URL of a given data file.

For the ODISSEI Portal which focuses on social sciences datasets, it would be beneficial if the social science-specific information that is currently available in DDI (e.g. the fields from the social sciences metadata block and the summary statistics) would be included in Croissant as well.

¹⁶ It should be noted that we did not specifically investigate the Croissant export in the context of datasets that are explicitly focused on Machine Learning. Since these datasets are the target of Croissant, they will likely benefit most from the introduction of the standard. It would need to be further investigated what datasets in the ODISSEI Portal can be categorized as Machine Learning datasets and whether the Croissant export brings additional advantages for these datasets in the current implementation.

An important aspect for the ODISSEI Portal to investigate is whether and how the ODISSEI Portal and the exports provided could include more **information** about the **files** associated with a given dataset. In particular the information about the available variables in a data file is very relevant for the ODISSEI community but is currently either not available at all in the Portal or only included in non-standardized custom metadata block.

Addressing this issue is two-fold: On the one hand, information currently already included in the Portal is not always available in the exports. Specifically, the variable information for the CBS and LISS datasets is currently available in a custom block which is not automatically included in the mapping for the Croissant and DDI outputs. We should work towards a solution where the processing of variable information in the Portal is aligned with the tabular data ingest functionality of Dataverse. That way, this information could become part of the metadata exports through the standard Dataverse functionality.

On the other hand, variable and file information is currently not always provided by the metadata providers and thus never gets into the Portal in the first place. It will be worthwhile to assess whether more of that information can be provided from the sources. If the Portal processes are more aligned with the default Dataverse functionality, such as the tabular data ingest, likely this would also simplify the inclusion of variable information from data providers who rely on Dataverse, like DataverseNL and the DANS Data Station SSH.

REFERENCES

Aydemir, Abdurrahman; Giriskan, Ahmet, 2025, "Replication Data and Code for: Language Skills and Labor Market Outcomes of Immigrants in Europe", <https://doi.org/10.5683/SP3/HS2TEZ>, Borealis, V1, UNF:6:/7jiNracciSRzgBOcCq0TA== [fileUNF]

Centraal Bureau voor Statistiek, 2025, "Kenmerken van vorderingen van gemeenten op (ex-)ontvangers van een bijstandsuitkering", <https://doi.org/10.57934/0B01E4108080FA8B>, ODISSEI Portal, V1

H. van der Kolk; J.N. Tillie; P. van Erkel; M. van der Velden; A. Damstra, 2012, "Dutch Parliamentary Election Study 2012 (DPES/NKO 2012)", <https://doi.org/10.17026/DANS-X5H-AKDS>, DANS Data Station Social Sciences and Humanities, V2, UNF:6:WumBFQojGDGgEFQtyR+3Pg== [fileUNF]

Tykhonov, V., & Durbin, P. (2024, March 20). Croissant ML standard in the context of Dataverse, EOSC and beyond. Zenodo. <https://doi.org/10.5281/zenodo.10843668>