



Customizing a DDI compliant repository: a detailed course with Dataverse

Geneviève Michaud, Baptiste Rouxel

► To cite this version:

Geneviève Michaud, Baptiste Rouxel. Customizing a DDI compliant repository: a detailed course with Dataverse. 12th Annual European DDI Users Conference. 2020. hal-03906364

HAL Id: hal-03906364

<https://sciencespo.hal.science/hal-03906364>

Submitted on 19 Dec 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - ShareAlike 4.0 International License

Customizing a DDI compliant repository: a detailed course with Dataverse

Baptiste Rouxel, Geneviève Michaud,
Center for Socio-Political Data

EDDI 2020, 12th Annual European DDI Users Conference

DOI: [10.5281/zenodo.4298929](https://doi.org/10.5281/zenodo.4298929)

December, 2, 2020

Customizing a DDI compliant repository: a detailed course with Dataverse

Tackling issues, sharing tools and processes

Outline

1. Metadata model customization
2. Setting up multilingualism
3. User interface, branding

Project objectives

Set up a CESSDA Metadata Model (CMM) compliant Dataverse repository

- Customize Dataverse metadata model

Migrate CDSP study level metadata from Nesstar

- Upgrade DDI files and harmonize metadata
- Keep study + variable level metadata as DDI files for each dataset in the repository

Metadata model customization

Dataverse metadata model customization: .tsv files

Customizing a Dataverse original metadata block involves modifying a [tab separated value file](#). A ".tsv" file targets a specific "metadata block". Operation include:

1. uploading the modified TSV file using the native API
2. triggering a SolR index update

Our custom TSV is available on GitHub here:



[CDSP-SCPO/dataverse-controlledvocabulary](https://github.com/CDSP-SCPO/dataverse-controlledvocabulary)

CDSP databank migration from Nesstar to Dataverse



Custom R script

[CDSP-SCPO/Nesstar2Dataverse](#)

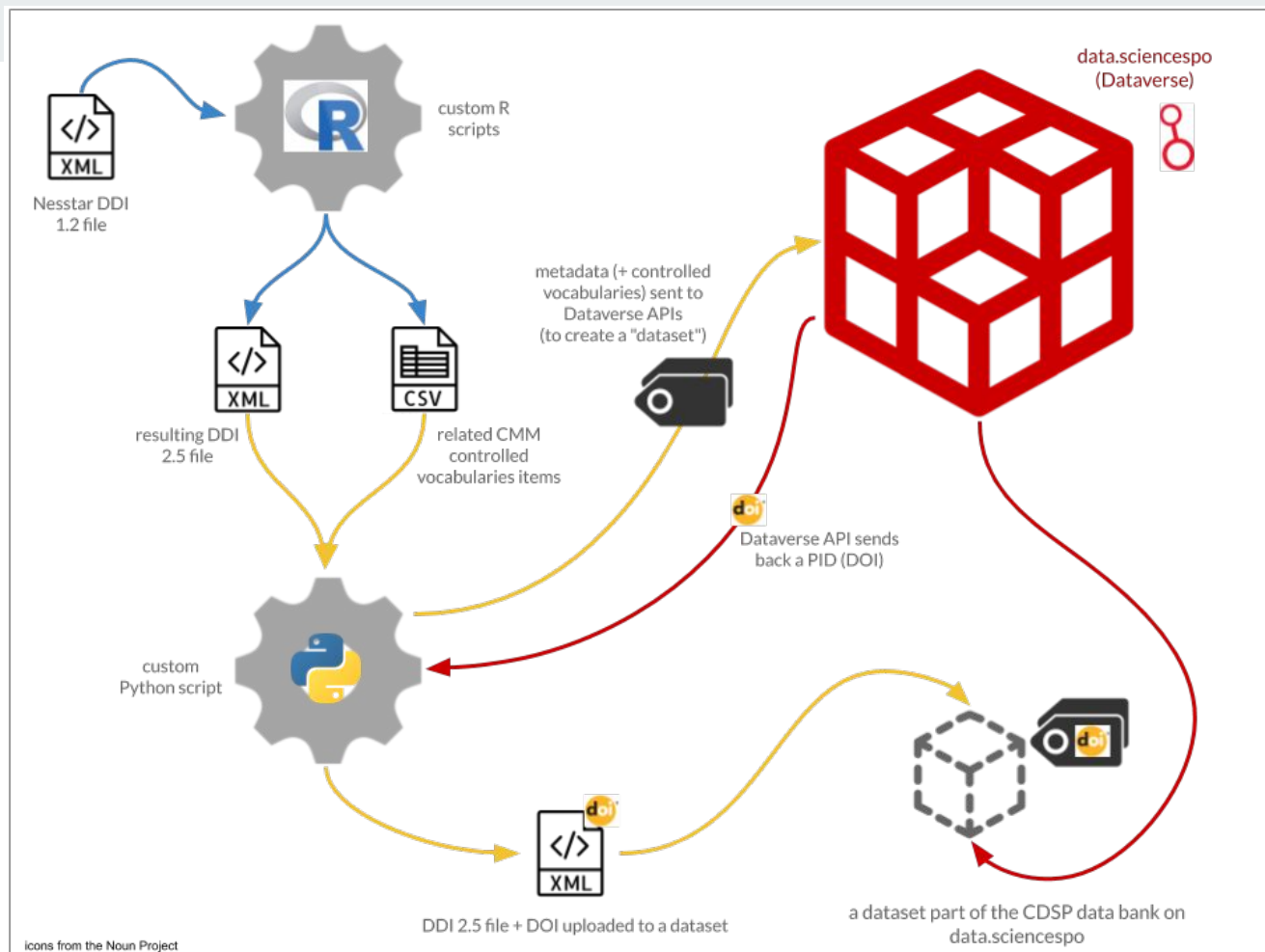
Custom Python scripts

[CDSP-SCPO/dataverse-controlledvocabulary](#)

[CDSP-SCPO/dataverse-metadataupdate](#)

[CDSP-SCPO/dataverse-jsonimport](#)

[CDSP-SCPO/dataverse-ddiimport](#)



Dataverse metadata model customization: issues

Adding metadata fields to the metadata model is quite easy (Create feature).

But this must be done **early** in the process, since revisions (Update and Delete features) require specific requests on underlying database with admin privileges. Further revisions should definitely be avoided (complexity + risk).

The DDI import API might not work properly if the metadata model is customized.

Multilingualism creates an additional layer of complexity.

Setting up multilingualism

Multilingualism: user interface

Adding languages is possible from Dataverse v4 onwards.

[Steps](#) (link to Dataverse official documentation):

- Declare new languages that will appear in the dropdown menu
- Upload a language bundle file (see the documentation to get your languages and create the bundle)



Language translations are maintained by the community and you can ask to contribute.



[GlobalDataverseCommunityConsortium/dataverse-language-packs](https://github.com/GlobalDataverseCommunityConsortium/dataverse-language-packs)

Managing multilingualism: controlled vocabularies

a language bundle = multiple language property files for each language

In the language property files, you can translate a controlled vocabulary by adding a property for each language available.

Example:

In the metadata block TSV file, if Self-administered questionnaire: Web-based (CAWI) is a controlled vocabulary item value for ModeOfCollection, add the following line in your language property file as described in the Dataverse [documentation](#):

```
controlledvocabulary.modeofcollection.self-administered_questionnaire:_web-based_(cawi)=TRANSLATED_VALUE
```

Managing multilingualism: controlled vocabularies


This process can be bit tedious and has a high risk of error if it is done manually, so we wrote a script that:


- Creates language property files for each language and different metadata blocks
- Transforms TSV controlled vocabulary values to property names
- Adds translated value to each property




[CDSP-SCPO/dataverse-controlledvocabulary](https://github.com/CDSP-SCPO/dataverse-controlledvocabulary)

Managing multilingualism: controlled vocabularies

☒  Dataverses (2)

☒  Datasets (313)

☐  Files (731)


Dataverse Category
Laboratory (1)
Research Project (1)


Publication Year
2020 (315)

Topic Classification Term
Elections (172)
Political behaviour and attitudes (84)
Energy and natural resources (33)
Government, political systems and organisations (22)
Social behaviour and attitudes (21)

[More...](#)

☒  Dataverses (2)

☒  Jeux de données (313)

☐  Fichiers (731)

Dataverse Category
Laboratoire (1)
Research Project (1)

Publication Year
2020 (315)

Classification des sujets Terme
Élections (172)
Comportements et attitudes politiques (84)
Énergie et ressources naturelles (33)
Gouvernement, systèmes et organisations politiques (22)
Comportement social et attitudes (21)

[Plus...](#)

Managing multilingualism: remaining issues

There is room for improvement in the way Dataverse follows DDI schema:

- As an example, it's not possible to save and display a **title** in more than one language.
- A specific group in Dataverse community is actively working on **flexible metadata**.

User interface, branding

Repository landing page

- Dataverse landing page, headers and footers are customizable
- A clear visual setup to help users navigate between different services
 - Search for data
 - Deposit data (self deposit)
 - Contact data and metadata curators (CDSP)
- Available in french and english

FIND AND EXPLORE DATA



data.sciencespo

SciencesPo

data.sciencespo
Research data repository of Sciences Po

FIND AND EXPLORE DATA

Launched in February 2020, data.sciencespo is a repository that offers visibility, sharing and preservation of data collected, curated and processed at Sciences Po. The repository is based on the Dataverse open-source software and organised into collections. It is indexed by re3data.

Sciences Po collection (self-deposit)

This collection, which is managed by the Direction des ressources et de l'information scientifique (DRIS), is intended to host data produced by researchers affiliated with Sciences Po, following the self-deposit process assisted by the DRIS staff. This collection is expected to be harvested by ISIDORE, BASE, OpenAIRE.

- Type of deposit: self-deposit assisted by the DRIS staff
- Deposit requirements: being affiliated with Sciences Po
- Deposit delay: immediate
- Data access delay: quick

Contact: data.bib@sciencespo.fr
Guidebooks: [deposit data](#) / [find data](#)
To find out more: [thematic guide on research data management](#)

DEPOSIT

CDSP collection

This collection managed by the Centre des données socio-politiques (CDSP) includes the catalogue of surveys, in the social science and humanities, processed and curated by CDSP engineers since 2005. This catalogue brings together surveys produced at Sciences Po and other French and international institutions. It is harvested by the French SSH portal Quetelet-PROGEDO Diffusion and by the [CESSDA catalogue](#).

- Type of deposit: at the CDSP
- Deposit requirements: deposit request and agreement
- Deposit delay: several months
- Data access delay: quick

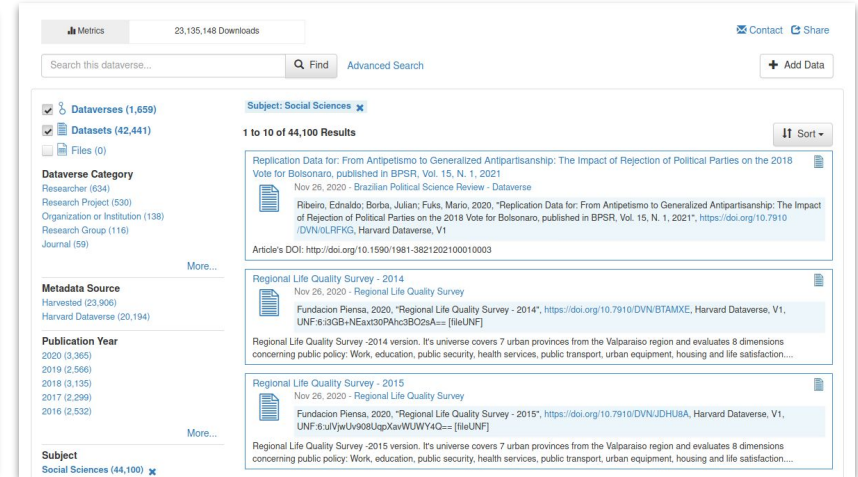
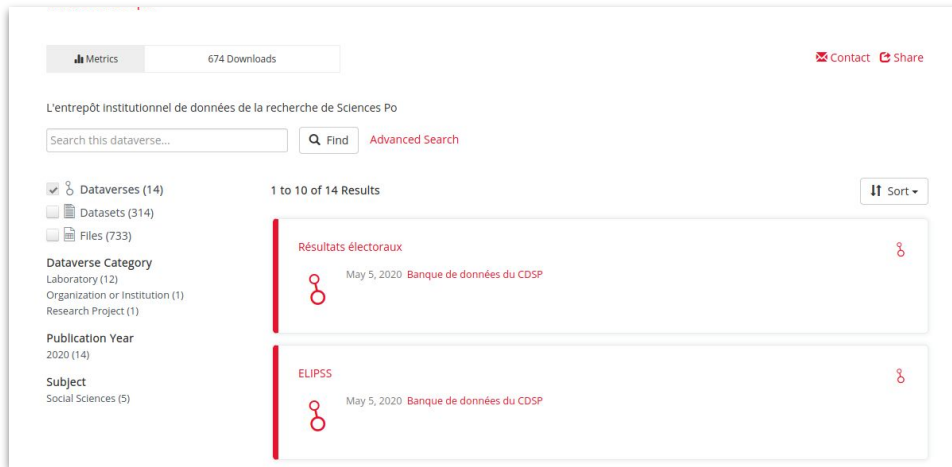
Contact: info.cdsp@sciencespo.fr
To find out more: [data deposit at the CDSP](#)

MAKE A DEPOSIT REQUEST

SciencesPo
27, rue Saint-Guillaume
75337 Paris Cedex 07
© 2020 Sciences Po
[General terms of use](#)

Powered by [Dataverse](#)

User interface: consistency with our design guidelines



User interface: consistency with our design guidelines

Livraisons des colis et Mobilités des e-consommateurs : caractérisation des pratiques et des flux (2016)

May 5, 2020 - ELIPSS

Aguilera, Anne, 2020, "Livraisons des colis et Mobilités des e-consommateurs : caractérisation des pratiques et des flux (2016)", <https://doi.org/10.21410/7E4/JOTU8V>, data.sciencespo, V2

L'enquête Livraisons des colis et Mobilités des e-consommateurs (LivMob), coordonnée par A. Aguiléra, a pour objectif d'identifier les déterminants du choix d'un mode de livraison et d'analyser tout particulièrement le rôle des territoires et des mobilités. L'enquête approfondit...

Territorial Dynamics and Wellbeing Household Survey - Colombia: Socioeconomic Data from rural-urban territories in Colombia, obtained in 2017-2018

Nov 25, 2020 - Rimisp Dataverse

Rimisp - Latin American Center for Rural Development; Universidad Iberoamericana; Universidad de los Andes, 2020, "Territorial Dynamics and Wellbeing Household Survey - Colombia: Socioeconomic Data from rural-urban territories in Colombia, obtained in 2017-2018", <https://doi.org/10.7910/DVN/YZ7Z73>, Harvard Dataverse, V1, UNF:6:7TKyDg+JcZ71XzSEaLZHbg== [fileUNF]

[English description below] Estos son los datos para Colombia de la Encuesta de Dinámicas Territoriales y Bienestar 2017-2018, parte del Programa Transformando Territorios de Rimisp, financiado por IDRC y realizado en colaboración, para su componente de investigación, con la Univ...



Next steps: improving our repository

We are currently collaborating with Dataverse on:

- Multilingualism (issue [#6607](#))
- DDI compliance (DDI import/export, OAI-PMH/OAI-DDI) (issues [#7388](#) [#7387](#) [#6751](#))

Merci !

If you have any question, contact us!

itcdsp-scpolt@sciencespo.fr

