



HAL
open science

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://hal-sciencespo.archives-ouvertes.fr/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](#)

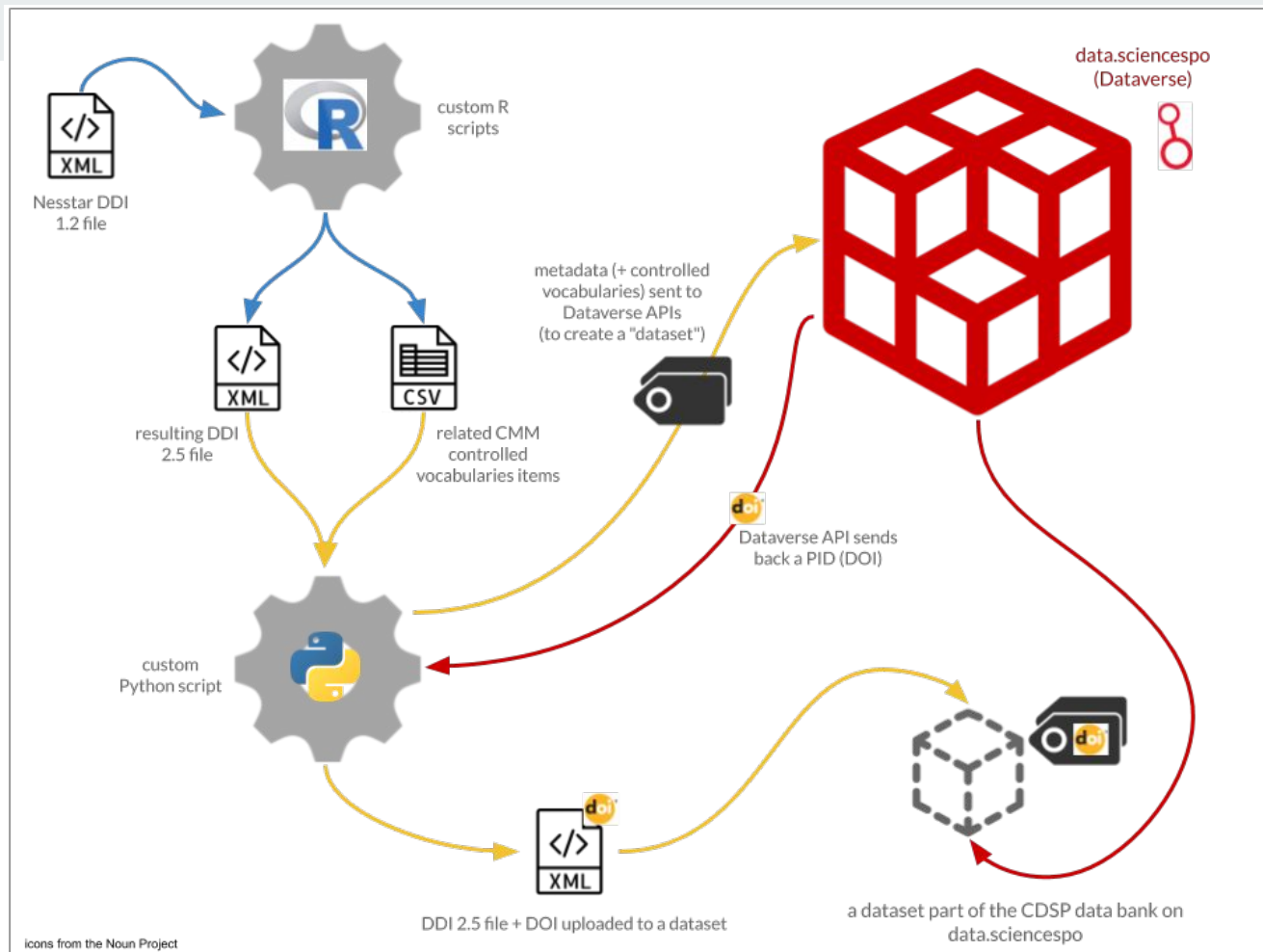
Custom Python scripts

[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

Search this dataverse...

-  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...

Chercher dans ce dataverse...

-  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

The screenshot shows the landing page of data.sciencespo, a research data repository. The browser address bar shows the URL https://data.sciencespo.fr. The page features the SciencesPo logo and the text "data.sciencespo Research data repository of Sciences Po". A prominent red button labeled "FIND AND EXPLORE DATA" is centered on the page. Below this, a paragraph explains the repository's launch in February 2020 and its purpose. Two main collection sections are visible: "Sciences Po collection (self-deposit)" and "CDSP collection". Each section includes a description, a list of key features (such as deposit type, requirements, and access delay), contact information, and a button for depositing data. The footer contains the SciencesPo address and contact details, along with a "Powered by Dataverse" logo.



User interface: consistency with our design guidelines

Metrics 674 Downloads Contact Share

L'entrepôt Institutionnel de données de la recherche de Sciences Po

Search this dataverse... Find Advanced Search Add Data

Datasets (14) 1 to 10 of 14 Results

Datasets (314)

Files (733)

Dataverse Category

Laboratory (12)

Organization or Institution (1)

Research Project (1)

Publication Year

2020 (14)

Subject

Social Sciences (5)

Résultats électoraux 8

May 5, 2020 Banque de données du CDSP

ELIPSS 8

May 5, 2020 Banque de données du CDSP

Metrics 23,135,148 Downloads Contact Share

Search this dataverse... Find Advanced Search Add Data

Datasets (1,659) Subject: Social Sciences x

Datasets (42,441) 1 to 10 of 44,100 Results

Files (0)

Dataverse Category

Researcher (634)

Research Project (530)

Organization or Institution (138)

Research Group (116)

Journal (59)

Metadata Source

Harvested (23,906)

Harvard Dataverse (20,194)

Publication Year

2020 (3,365)

2019 (2,566)

2018 (3,135)

2017 (2,299)

2016 (2,532)

Subject

Social Sciences (44,100) x

Replication Data for: From Antipetismo to Generalized Antipartisanship: The Impact of Rejection of Political Parties on the 2018 Vote for Bolsonaro, published in BPSR, Vol. 15, N. 1, 2021

Nov 26, 2020 - Brazilian Political Science Review - Dataverse

Ribeiro, Edinaldo; Borba, Julian; Fuke, Mario, 2020, "Replication Data for: From Antipetismo to Generalized Antipartisanship: The Impact of Rejection of Political Parties on the 2018 Vote for Bolsonaro, published in BPSR, Vol. 15, N. 1, 2021", <https://doi.org/10.7910/DVN/OLRFKG>, Harvard Dataverse, V1

Article's DOI: <http://doi.org/10.1590/1981-3821202100010003>

Regional Life Quality Survey - 2014

Nov 26, 2020 - Regional Life Quality Survey

Fundacion Pienza, 2020, "Regional Life Quality Survey - 2014", <https://doi.org/10.7910/DVN/BTAMXE>, Harvard Dataverse, V1, UNF:6:ijGB+NEaxt30PAhc3BO2sA== [fileUNF]

Regional Life Quality Survey -2014 version. Its universe covers 7 urban provinces from the Valparaíso region and evaluates 8 dimensions concerning public policy, Work, education, public security, health services, public transport, urban equipment, housing and life satisfaction....

Regional Life Quality Survey - 2015


Nov 26, 2020 - Regional Life Quality Survey


Fundacion Pienza, 2020, "Regional Life Quality Survey - 2015", <https://doi.org/10.7910/DVN/JDHUBA>, Harvard Dataverse, V1, UNF:6:uLjYwLjv60EUpXawWUWY4Q== [fileUNF]

Regional Life Quality Survey -2015 version. Its universe covers 7 urban provinces from the Valparaíso region and evaluates 8 dimensions concerning public policy, Work, education, public security, health services, public transport, urban equipment, housing and life satisfaction....



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

