IJDC | *General Article*

Form as an Extension of Content: Strategy-Based Adaptation of the Interface of a Dataverse Installation

Benjamin Peuch
State Archives of Belgium

Laura Van den Borre
Sciensano

Elias Kruithof
Vrije Universiteit Brussel

Jens Doms
Vrije Universiteit Brussel

Jean–Paul Sanderson
Université catholique de Louvain

Youssef Ouahalou
State Archives of Belgium

Johan Van der Eycken
State Archives of Belgium

Rolande Depoortere
State Archives of Belgium

Abstract

In October 2020, a data archive for social sciences and digital humanities was launched at the Belgian national archives institution, the State Archives of Belgium. This new infrastructure, called the Social Sciences and Digital Humanities Archive (SODHA), is meant to accrue and make available datasets in any of the disciplines that constitute social sciences and digital humanities. To support this endeavor, the online platform relies on the open-source software for digital repository management Dataverse, developed by the Institute for Quantitative Social Science (IQSS) of Harvard University.

IQSS have made it possible to customise installations of their software to a large extent, especially the formal, outer aspects. This makes for interesting design choices, not just for branding an installation but also to translate any organisation's communication strategy with both broad and minute modifications. We contend that working on the design—that is, the formal aspects of an application—does not amount to "cosmetic" work; rather, it empowers the administrators of a platform to shape their front end so that it best conveys what its core functions and purposes—in other words, its content—are all about.

We present the context in which SODHA was launched, how we elaborated a particular communication strategy calibrated for researchers, and the ensuing design choices. Examples of changes and adaptations are provided and contextualised.

Submitted 26 June 2024 ~ *Accepted* 12 May 2025

Correspondence should be addressed to Johan Van der Eycken, Rue de Ruysbroeck 2, 1000 Brussels, Belgium. Email: johan.vandereycken@arch.be

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

As theorised by poets Robert Creeley and Charles Olson, “form is never more than an extension of content.”[1] Far from being an invitation to judge a book by its cover, this claim means that we should never divorce the appearance of an artwork—that is, its form; its stylistic, most external aspects—from its contents, as was customary in the early days of literary analysis. In this paper, we contend that, to some extent, the same can be said about any piece of software and its interface, be it graphical or not. Although software programmes are tools, not artworks, their outer aspects—the interfaces through which users interact with them—reflect certain choices and values, as user interfaces (UI) ultimately determine the ensuing user experience (UX) as well as, by extension, the relationship between user and provider.

To support our claim, we present as a case study the customisation of the external elements of an installation of the open-source software for data repository management Dataverse. The developers of Dataverse, from the Institute for Quantitative Social Science (IQSS)[2] of Harvard University, have made it possible to adapt several formal elements of their programme, thus enabling organisations that rely on Dataverse to shape its look-and-feel while safely retaining its core functionalities.

The installation in question is that of the Social Sciences and Digital Humanities Archive, also known as SODHA.[3] SODHA was launched in October 2020 at the national Belgian archives institution, the State Archives of Belgium,[4] and acts as national service provider for the Consortium of European Social Science Data Archives (CESSDA).[5]

In and of itself, SODHA is an “odd duck” of sorts, considering that data archives for social sciences and digital humanities are usually to be found in universities, not public records offices (Doorn & Tjalsma, 2007, p. 4). Furthermore, SODHA not only addresses two target audiences, namely social science and the digital humanities research communities; these two communities also constitute very diverse and evolving user groups. These strategic and ethical considerations were decisive factors when it came to making design choices for our platform, as we hope to show in this paper.

Sharing Data in Social Sciences

Data archives are specialised repositories that tackle the various legal, technical, and organisational challenges of archiving, documenting, and sharing research data. Most data archives have a thematic or disciplinary scope, choosing to focus their efforts on the disciplines that make up fields of study, such as molecular biology,[6] geoscience,[7] or genetics.[8] But unlike with “hard sciences”, preserving data from the social sciences and

[1] https://www.poetryfoundation.org/articles/69406/projective-verse. See also Olson (1997).

[2] https://www.iq.harvard.edu/

[3] https://www.sodha.be/

[4] https://www.arch.be/

[5] https://www.cessda.eu/

[6] For example, the Research Collaboratory for Structural Bioinformatics (RCSB)’s Protein Data Bank (PDB): https://www.rcsb.org/, or the Biobanking and Biomolecular Resources Research Infrastructure (BBMRI) European Research Infrastructure Consortium (ERIC): https://www.bbmri-eric.eu/

[7] For example, UNIDATA: https://www.unidata.ucar.edu/data/

[8] For example, the National Cancer Institute’s Genomic Data Commons (GDC): https://gdc.cancer.gov/

making them available for reuse entails specific issues. Context is always key to understanding and properly analysing scientific data of any kind, whether during or after a study; yet this requirement was shown to be even more stringent in the context of social science research, making the work of preparation and documentation a daunting task for social scientists who mean to share their data (Yoon & Kim, 2017).

An additional difficulty is the high ethical standard which social science researchers must adhere to, considering how potentially sensitive their data can be since they usually bear relation to actual (and often, living) individuals. When queried about data sharing, some of the social scientists who work on qualitative data claim that it would be just too complex to try to create shareable versions of their datasets (Jeng & Lyon, 2016; Jeng, He, & Oh, 2016). This is even truer today in the European Union (EU) in the light of the higher standards set by recent legal developments, perhaps most of all the General Data Protection Regulation (European Union, 2016). Also known as the GDPR, this text sets a very precise and demanding framework that applies to all legal entities in the world that collect data about living citizens from the European Union. Though the GDPR admits of some exceptions for historical and scientific research, this new standard has made the work of social scientists as well as data archivists and historians even more complex (Van Honacker, 2018).

SODHA

SODHA is a data archive which follows in the footsteps of specialised infrastructures that emerged as early as the 1960s (Kaase, 2013, pp. 19–20). Like other entities of that kind, SODHA seeks to acquire, document, preserve, and provide access to collections of data, or “datasets”, relevant to the variety of disciplines that constitute social sciences and digital humanities.

Despite its currently limited means, SODHA can boast several accomplishments:

- a dedicated online platform;[9]
- a robust legal framework that protects the interests of the data archive as well as of its users, both depositors and “reusers”;[10]
- a policy framework that clearly delineates the goals and the scope of the infrastructure;[11]
- documented internal procedures, which ensure that core tasks can be performed regardless of absenteeism or turnover;
- a monitoring strategy for the creation of statistical reports, as well as documented database queries for extraction of relevant information;[12]
- as of May 2024, 136 published datasets, more than 2,400 published data files, and four subrepositories.

[9] https://www.sodha.be/

[10] See the “Texts and Policies” webpage of the SODHA Guide: https://www.sodha.be/guide/Texts_and_policies.html

[11] Idem. See especially the mission statement and the dataset publishing policy.

[12] SODHA has contributed to the public list of useful database queries for Dataverse: https://docs.google.com/document/d/1-Y_iUduSxdDNeK1yiGUxe7t-Md7Fy965jp4o4m1XEoE

Target Audience(s)

Social Sciences and (Digital) Humanities

The situation of SODHA is also peculiar in the light of the fact that, because this data archive is hosted at a national archives institution, it is managed by archivists and historians, while the target audience of the infrastructure consists of sociologists, psychologists, anthropologists, demographers, and other such representatives of the social sciences. The scope of SODHA was only later expanded to the digital humanities, thus bringing in linguists, musicologists, literary scholars, archeologists, and so on, with whom historians feel a natural kinship. Social sciences, on the other hand, while not as "distant" from the humanities as, for instance, the hard sciences, are nonetheless a world of their own. There was an apprehension that social scientists, based for the most part in university research centres, would regard with skepticism the creation of a new service to archive their data at a federal institution whose reading rooms are usually filled with genealogists and history students toiling away on paper-based sources. While many representatives of the social sciences draw on historical sources for their work, a national archives center is not necessarily as typically emblematic of their fields of study as, for example, a national statistics office.

Just like historians work with various tools and methods, so do social scientists. Just like many social scientists prefer more in-depth, qualitative research methods, which involve performing close analyses (as opposed to the practice of distant reading and big data computing), so do more and more historians work with aggregated data and digital support for sweeping, corpus-based analyses of sources. The new, heavily digitised research methods ushered in by the pioneers of digital humanities have brought social scientists and digital scholars ever closer with time. As of now, "[a]lmost all existing methods of research can be found within these two domains" (Hogenaar, Tjalsma, & Priddy, 2011, p. 165).

That is why the initial fear that there might be discrepancies or misapprehensions between SODHA and its target audiences soon subsided, while even more significant in this regard was the fact that the SODHA data archive was built by an interdisciplinary team that brought together the State Archives and two university research centres, Interface Demography from *Vrije Universiteit Brussel* and the Center for Demographic Research of *Université catholique de Louvain*.[13] With this partnership, the expertise in digital archiving and platform development of the State Archives was combined with the social scientists' expert knowledge and contacts.

A broad and diverse audience

Data archives like SODHA specifically target researchers. On the one hand, high turnover means the scientific community is a very heterogeneous group: every year, large numbers of students both join and leave the research community. Therefore, information literacy training must be perpetually renewed. Even the most "anchored" users, namely tenured researchers, represent an ever-shrinking minority in the face of the growing student population (Williams, 2019, p. 210). These same users are also less likely, because of the demographic factor of age, to take to computer literacy as easily as their younger colleagues and students. For that reason, we obviously cannot assume that researchers are equally proficient with computers and data management; just like with any other user community, there is a minority of expert users, and a majority of mainstreamers, who "don't use technology for its own sake; they use it to get a job done" (Colborne, 2010/2018, p. 30).

[13] Two other universities exist in Belgium with the exact same names, only in the other language: the *Université libre de Bruxelles*, and the *Katholieke Universiteit Leuven*. To avoid ambiguity, we do not translate the names of these universities in this paper.

Communication Strategy

Data archives, when they can afford it, often take on the responsibility of putting out training materials and offering training services, especially for college students (Dale, 2013; Schneider, Katsanidou, Horton, & Wolf, 2013). SODHA unfortunately lacks the means to undertake either exhaustive or recurring training programmes, but reflection on our target audience led us to adopt a certain tone and frame our communication in a specific way. This transpired the most through the webinars with which we introduced researchers to our platform and in the changes that we made to the interface of our Dataverse installation.

Our primary concern was clarity. We felt it was paramount to present our users with tools and documentation that were as explicit, unambiguous, and user-friendly as possible. This is not saying much, considering all providers of goods and services are keen not to alienate potential users or customers by failing to convey what exactly they are offering.[14] But we were especially mindful of our communication in the light of the fact that, while we are presenting our users with a free service, we do ask them to do quite a bit of work when it comes to data deposit. The phenomenon of academic burnout has been well documented (Melendez & de Guzman, 1983) and made even worse in recent years (Gewin, 2021). Creating a dataset entails many subtasks, among which are gathering all relevant data files, going through the data in search for errors, checking notes with other researchers involved, obtaining the greenlight from various authorities (project manager, head of department, data protection officer…) to deposit the data in a repository (and which repository?), then producing adequate documentation to explain the meaning of the various codes in the data (in other words, writing the codebook) as well as give the context of the study, and, finally, going through the deposit procedure of the data archive. Preparation and polishing of data is often brought up as one of the main obstacles to data sharing (Scheuch, 2003, p. 386; Shaon, Straube, and Chowdhury, 2017, p. 151; Tenopir et al., 2011, p. 2). For that reason and many others, data sharing cannot just be imposed overnight; it must be fostered as a new academic culture (Weil & Hollander, 1990, p. 7), encouraged with rewards (Altman & Crosas, 2014, p. 66; Tenopir et al., 2015, pp. 4 and 8), and supported by training and guidance. As put by John Whalen:

> "Some of the time, what our customers want and what might benefit them are two very different things. The task before us in that case isn't easy. We have to be prepared to attract them and appeal to them using what they think they want—even if we know that something might be much more beneficial for them. Ideally, after attracting them we can educate them better in order to help the customers make an informed decision." (Whalen, 2019/2021, p. 162)

We also endeavored to make our communication as clear as possible so as to translate the concept of transparency into practice. Transparency is commonly regarded as a pillar of open science and a necessary step to actually "opening data" and making them accessible for consultation and reuse (European Union, 2019; Organisation for Economic Co-operation and Development, 2020). By making our communication as clear and user-oriented as we could, we hoped to secure our future depositors' trust. Research has shown lack of trust to be one of the main deterrents to sharing research data, especially in social sciences (Carlson & Anderson, 2007; Chauvette, Shick–Makaroff, & Molzahn, 2019; Niu & Hedstrom, 2009). That is why we sought to establish a rapport with our users by reaching out to them and give them the opportunity to ask about and even question our work via webinars during which we would present the SODHA infrastructure at length.

[14] "B2C communication failures often use over-jargony language that confuses customers […]. This failure to communicate usually stems from a business-centric perspective, resulting in overly technical language" (Whalen, 2019/2021, p. 44; see also p. 168).

We were also eager to avoid offending our users by appearing to patronise them. This was an especially delicate task considering that, inevitably, we have to explain several things to our users and, as previously argued, we are dealing with knowledge workers. While in principle researchers develop their skills and knowledge all their lives, any researcher's identity is intimately linked to their sense of self-worth and their professional respectability; that is why appearing to presume that a researcher does not know something can potentially give offence. Practically speaking, there sometimes was a temptation to resort to design techniques which have been criticised precisely as potentially patronising, as will be shown further below. That is why we tried to strike a balance between putting out a lot of information and actively reaching out to our users on the one hand, and avoiding over-explaining things and stating the obvious on the other hand. Naturally, what is obvious for us information science specialists might not be for researchers specialised in other disciplines. We had the opportunity of implementing our communication strategy thanks to the customisability of the software we rely upon for our online platform, IQSS' Dataverse.

Dataverse

Dataverse is an open-source web application for managing research data repositories. It was launched in 2006[15] and has been regularly updated ever since, with the noteworthy publication of Version 6 of the software on 8 September 2023.[16] As of February 2024, there are 112 known Dataverse installations around the world, spanning all continents.[17]

Dataverse can be used to fill out the essential tasks of a repository for research data management, namely:

- Accruing data by means of a data deposit tool and procedure;
- Documenting data, which is done in the course of the previous step by means of an extensive metadata form, the bulk of which is based on the metadata standard for social sciences, the Data Documentation Initiative (DDI);[18]
- Publishing and providing access to datasets, either in full (open data) or under certain conditions (restricted access).

Based on our experience, we contend that Dataverse is an exemplar of a *simple* tool, by which we mean:

1. On a *fundamental* level, it revolves around a limited number of key, logically bound functions and processes;
2. As we will try to show, on a *formal* level, these functions and processes are easily accessible and actionable thanks to efficient design.[19]

In other words, the *form* of Dataverse—its design—aptly reflects its *contents*—the essential business processes which this software can carry out in a computer environment. Though the SODHA team has on occasion suggested enhancements for the programme by posting on the GitHub repository of IQSS,[20] two years of activity and the positive feedback

[15] https://dataverse.org/about

[16] https://github.com/IQSS/dataverse/releases/tag/v6.0

[17] https://iqss.github.io/dataverse-installations/

[18] The Dataverse metadata form is based specifically on DDI-Codebook version 2.5. See the website of the DDI Alliance: https://ddialliance.org/Specification/DDI-Codebook/2.5/

[19] Dataverse "keeps distances short" (Tidwell, Brewer, & Valencia, 2019/2020, p. 134) between all of its core functionalities, making them easy to apprehend.

[20] https://github.com/IQSS/dataverse/issues

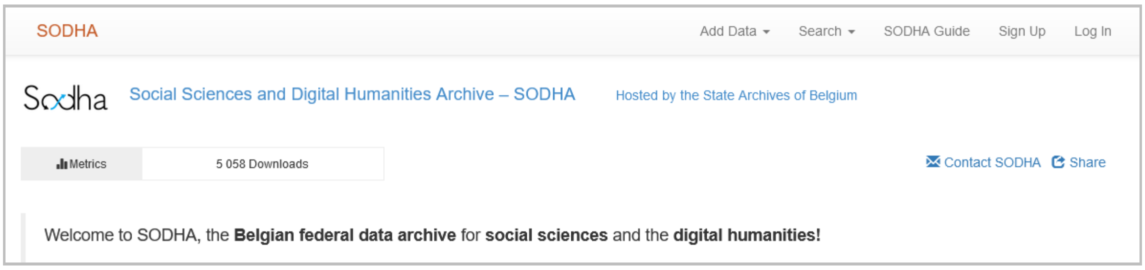
that we collected from participants to our webinars have convinced us that Dataverse currently constitutes one of the most robust and user-friendly open-source applications for research data management out there. Dataverse itself has been continually improved for several years on the basis of user feedback.[21] We find ourselves agreeing with one of the main contributors of the Dataverse project, Mercè Crosas, when she describes the interface of the software as "user-friendly [and] low-maintenance" (2011).

Customisation

Although our general appreciation of Dataverse was very positive after internal testing, we believed that certain elements of the programme could be changed in order to clarify certain technical or policy aspects as well as brand our installation. We also meant to leverage the customisation possibilities implemented by IQSS for several features of their software, having also produced extensive (and ever-growing) documentation[22] to help administrators and developers to this end. In this way, we could master the outer aspects of the key access point of our data archive and thus make its form—the online GUI—truly an extension of the content—not only the data management tools and services, but also the philosophy of our data archive—that we meant to provide.

Branding

To ensure that other organisations can truly appropriate their software, IQSS have made it possible to edit the Dataverse homepage as well as the top banner, which appears on all webpages. Organisations that rely on Dataverse can input their logo, their name, and their slogan. The way this can be done was especially helpful in the case of SODHA considering that we had to mention the host institution of the data archive but, as previously mentioned, we were afraid that overemphasising the State Archives might put off social scientists. Transparency dictated that we disclose the relation between SODHA and the State Archives, but Dataverse allowed us to do so in a very discreet manner, as shown on Figure 1:



Social Sciences and Digital Humanities Archive – SODHA

Hosted by the State Archives of Belgium

Welcome to SODHA, the **Belgian federal data archive** for **social sciences** and the **digital humanities!**

Figure 1. The upper part of the SODHA platform homepage.

As can be seen, SODHA is officially "Hosted by the State Archives of Belgium", and this text segment also contains a hyperlink that directs visitors to the website of the State Archives.[23] However, the focus is placed on SODHA, whose name appears no less than six times:

[21] IQSS largely revamped the design of their application specifically for Version 4, which came out in 2015 (Quigley, 2015).

[22] The general Dataverse Guide, organised in subsequent guides, among which are the User Guide, the Admin Guide, and the Developer Guide: http://guides.dataverse.org/en/latest/

[23] https://www.arch.be/

- in the top left corner, in the banner;
- in the logo itself, on the left;
- next to the full name;
- in the upper right part, a second time in the banner, in the label of the button directing users to the SODHA Guide;
- in the label of the button for contacting SODHA administrators;
- finally, at the outset of the introductory text ("Welcome to SODHA").

We could have removed the mention of the State Archives entirely and left it to users to find out about the host institution of SODHA by reading our texts and policies;[24] or we could have highlighted the institutional affiliation by including the name of the State Archives before SODHA's. Some would argue that the latter strategy is also viable, considering that, as a branch of the federal government, SODHA addresses all researchers regardless of their research institution. However, at the end of the brainstorming sessions that occupied the project team, we decided that it was best to call attention to SODHA in and of itself, assuming that if some of our users wanted more information on our governance or policy model, they would either soon find it in our documentation or directly contact us.

Metadata Form

Dataverse presents its users with several forms for data input, among which the one for dataset metadata is an essential waypoint of the application: it enables depositors and data curators to encode information about datasets in a structured way. Based on this input, Dataverse can produce and export standardised metadata files. The input entered into the metadata form allows Dataverse to format machine-readable and actionable XML-DDI files, among other available metadata standards. The ensuing structured documentation not only serves as a point of reference for future use; it also enables automated exchange of metadata between online platforms.[25]

Metadata encoding is made possible in Dataverse thanks to a *wizard*, that is, a "feature or component that leads the user through the interface step by step to do tasks in a prescribed order" (Tidwell et al., 2019/2020, p. 86). As Tidwell et al. put it, wizards are useful to run users through "a task that is long or complicated, and that will usually be novel for users" (2019/2020, p. 86). It would be unreasonable to ask researchers to create DDI files by themselves, as this would require them to copy paste the whole standard in a text editor, decipher the meaning of all metadata elements, identify which are relevant to their own case, and finally record the information by distributing it in the various placeholders while adhering to strict morphological and syntactic rules. That is why a GUI is absolutely necessary not only to ensure that a standard's specifications are followed, but also to make the task feasible in the first place.

Although Dataverse translates DDI into such a GUI, writing out technical documentation is always, even with the best design approach conceivable, a long and arduous process. Wizards are ideal for tasks that are "either branched or very long and tedious" (Tidwell et al., 2019/2020, p. 86):

> "By splitting up the task into a sequence of chunks, each of which can be dealt with in a discrete 'mental space' by the user, you effectively simplify the task.

[24] https://www.sodha.be/guide/Texts_and_policies.html

[25] In the case of SODHA, this is done with the CESSDA Data Catalogue, a portal for European datasets in social sciences: https://datacatalogue.cessda.eu/

> You have put together a preplanned road map through the task, thus sparing the user the effort of figuring out the task's structure—all they need to do is address each step in turn, trusting that if they follow the instructions, things will turn out OK." (Tidwell et al., 2019/2020, p. 86)[26]

The problem with wizards is that they can be perceived as "patronizing" (Tidwell et al., 2019/2020, p. 87). They are usually leveraged in situations in which "the designer of the UI [knows] more than the user does about how best to get the task done" (Tidwell et al., 2019/2020, p. 86). And indeed, before the launch of SODHA, we conducted a survey of the needs and practices of Belgian researchers in social sciences in the area of research data management, in the course of which we asked our respondents if they had ever heard about the DDI standard: the data overwhelmingly indicated that they never had (Peuch et al., 2020). Still, even if we present researchers with a sober and elegant GUI, we must streamline the whole process with great care lest it comes off as endless or fastidious.[27] That is why being able to exert control over this essential business process is a boon for organisations that rely on Dataverse.

Metadata Customisation

A whole section of the Dataverse Guide is dedicated to metadata customisation.[28] Because DDI is a standard and its consistency must be maintained, administrators may not alter it by outright deleting metadata fields, for example. However, most other actions are possible, as Dataverse administrators may:

- add fields;
- rename fields;
- hide fields;
- make fields optional or mandatory;
- choose which fields will appear upon dataset creation;[29]
- edit the description or the watermark of fields;
- add controlled vocabularies for certain fields.

We tested all these functionalities and used most of them to effect changes that are now an integral of the SODHA Dataverse installation. Of particular note are renamed fields, two of which appear at the very outset of the metadata form.

Renaming Fields

To rename metadata fields, administrators must locate the file named *Bundle.properties* on the server where their Dataverse installation is located, then find and rename certain

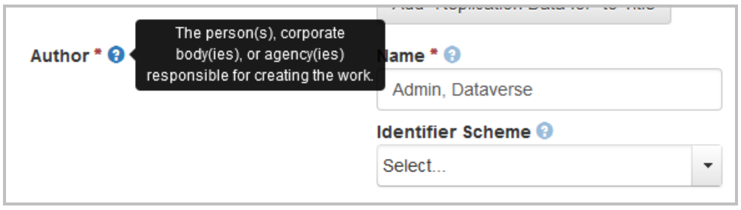
[26] Giles Colborne (2010/2018) also recommends "staged disclosure" (p. 174) as a way to pace users through procedures.

[27] Tidwell et al. make this cautionary hypothesis: "the very need for a Wizard indicates that a task might be too complicated" (2019/2020, p. 86). See also Colborne (2010/2018) on how "irksome" (p. 204) filling out forms can be.

[28] https://guides.dataverse.org/en/latest/admin/metadatacustomization.html

[29] The idea is that, at first, only a subset of metadata is presented to depositors, and they can access the rest of the fields later, once a first version of their dataset has been created.

values. In some cases, the original value can be as simple as *hostDataverse*; at other times, it can be as long as *dataset.manageTemplates.tab.action.btn.view.dialog.datasetTemplate*. Handily, the process is the same for altering the contents of infobubbles, which display the description of metadata fields:



Author *

The person(s), corporate body(ies), or agency(ies) responsible for creating the work.

ame *

Admin, Dataverse

Identifier Scheme

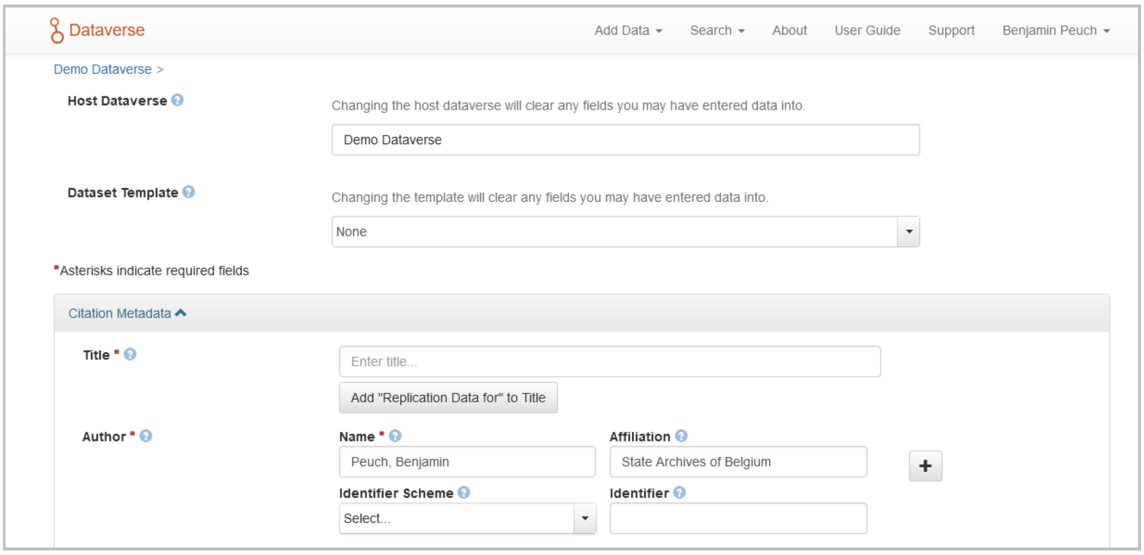
Select...

Figure 2. The contents of an infobubble displayed as the cursor hovers over it.

"Host Dataverse" → "Subrepository"

Before the launch of the platform, we asked two project partners specialised in social science research to study the metadata form closely and determine whether everything was clear to them. Afterwards, we received more user feedback by collecting the questions and reactions of the participants in the webinars we organizsd to present our platform. It appeared, based on these interactions with users, that the meaning or purpose of several fields could be ambiguous. For instance, right at the beginning of the form, the very first field, *Title*, was not unequivocal for some of our users, as they were not sure whether they should use it to encode the title of the study, or of the dataset, or of their thesis (which is based on the dataset), or likewise of their publication in a journal. That ambiguity could arise so early on as users interacted with our platform was alarming, but after we renamed the field *Title of Dataset*, the question was never asked again.[30]

Another field early in the form which we chose to rename is *Host Dataverse*, as shown in Figure 3:



Host Dataverse

Changing the host dataverse will clear any fields you may have entered data into.

Demo Dataverse

Dataset Template

Changing the template will clear any fields you may have entered data into.

None

*Asterisks indicate required fields

Citation Metadata

Title *

Enter title...

Add "Replication Data for" to Title

Author *

Name *: Peuch, Benjamin

Affiliation: State Archives of Belgium

Identifier Scheme: Select...

Identifier

Figure 3. The start of the metadata form in IQSS's demo Dataverse installation.[31]

[30] Incidentally, such doubts likely reveal the fact that, for some researchers in social sciences, datasets are quaint objects still. Surely, they are used to working with data, but the idea of constituting data collections into proper sets, with adequate documentation (metadata), an internal structure, and a self-standing title, is probably still very new, and "What title shall I give to my dataset?" is a question that they will have to ask themselves.

[31] https://demo.dataverse.org/

To understand the meaning of this field, we must turn to the Dataverse Guide, in which IQSS explains that the word *Dataverse* can designate three things:

- The computer programme;
- A certain installation, an instance of this programme;
- A data repository within this programme, which can contain other such repositories.[32]

In other words, a *Dataverse* can contain "subdataverses" of sorts, much like the Windows Explorer programme consists in a large folder that can contain many "subfolders". This is illustrated in a conceptual manner with Figure 4, from the Dataverse Guide, and Figure 5, from the SODHA Dataverse installation:

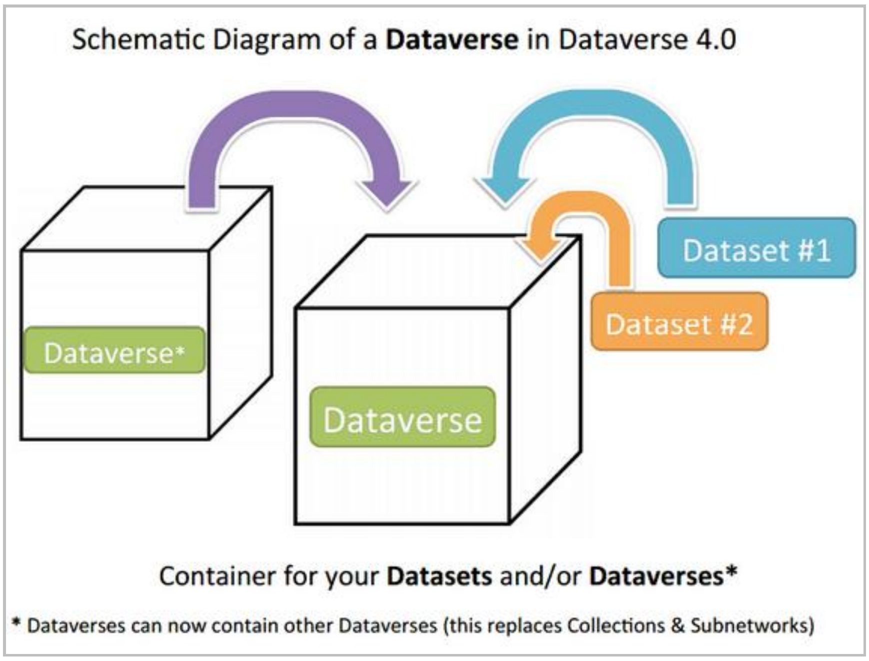
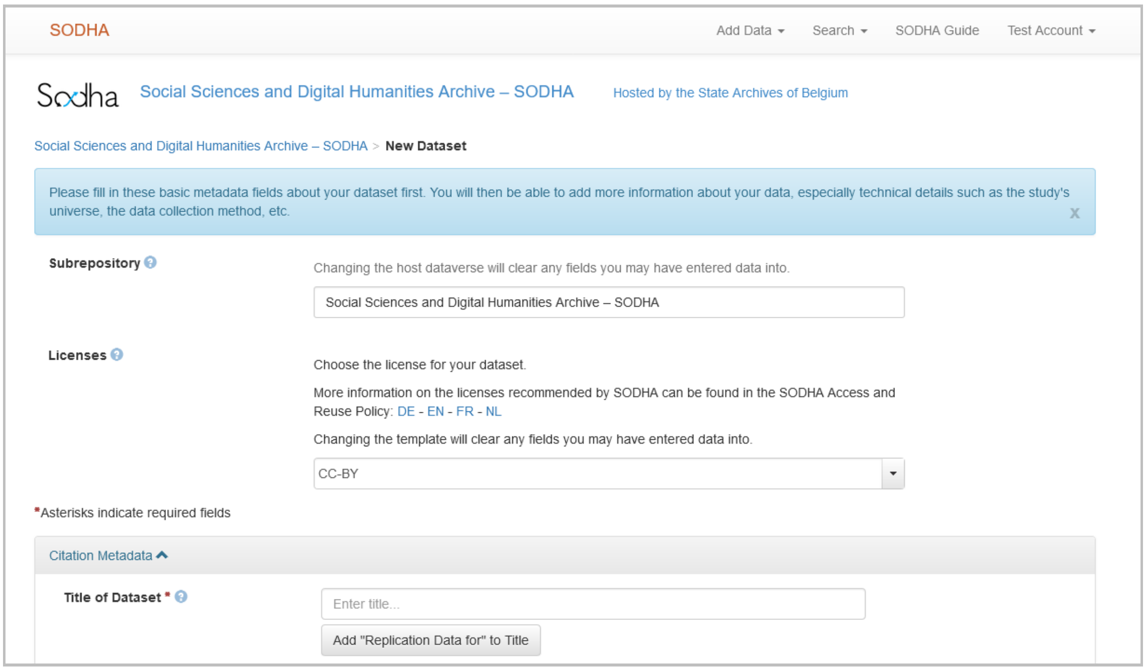

Schematic Diagram of a **Dataverse** in Dataverse 4.0
Dataverse*
Dataverse
Dataset #1
Dataset #2
Container for your **Datasets** and/or **Dataverses***
* Dataverses can now contain other Dataverses (this replaces Collections & Subnetworks)

Figure 4. An excerpt from the Dataverse User Guide which explains how Dataverse repositories can be embedded.

While this lexical choice can make sense from the perspective of a Dataverse administrator, who has developed a familiarity with the software and its jargon, we feared that this polysemy might confuse our users. Though we want to advertise Dataverse and acknowledge that we rely on it, we try not to mention the software too often to our users for fear that it might confuse them; we refer them to "the SODHA platform" instead. Those are the reasons why we renamed *Host Dataverse* as blandly as *Subrepository*, hoping this term would be more intuitive:

[32] Recently, with the publication of Version 5.9 of Dataverse, IQSS chose to use the word *collection* instead of *Dataverse* in their User Guide. Still, the name of the metadata field remains *Host Dataverse*. Compare https://guides.dataverse.org/en/5.8/user/dataverse-management.html and https://guides.dataverse.org/en/5.9/user/dataverse-management.html.



SODHA   Add Data ▾   Search ▾   SODHA Guide   Test Account ▾

Social Sciences and Digital Humanities Archive – SODHA   Hosted by the State Archives of Belgium

Social Sciences and Digital Humanities Archive – SODHA > **New Dataset**

Please fill in these basic metadata fields about your dataset first. You will then be able to add more information about your data, especially technical details such as the study's universe, the data collection method, etc.

Subrepository

Changing the host dataverse will clear any fields you may have entered data into.

Social Sciences and Digital Humanities Archive – SODHA

Licenses

Choose the license for your dataset.

More information on the licenses recommended by SODHA can be found in the SODHA Access and Reuse Policy: DE - EN - FR - NL

Changing the template will clear any fields you may have entered data into.

CC-BY

*Asterisks indicate required fields

Citation Metadata

Title of Dataset *

Enter title...

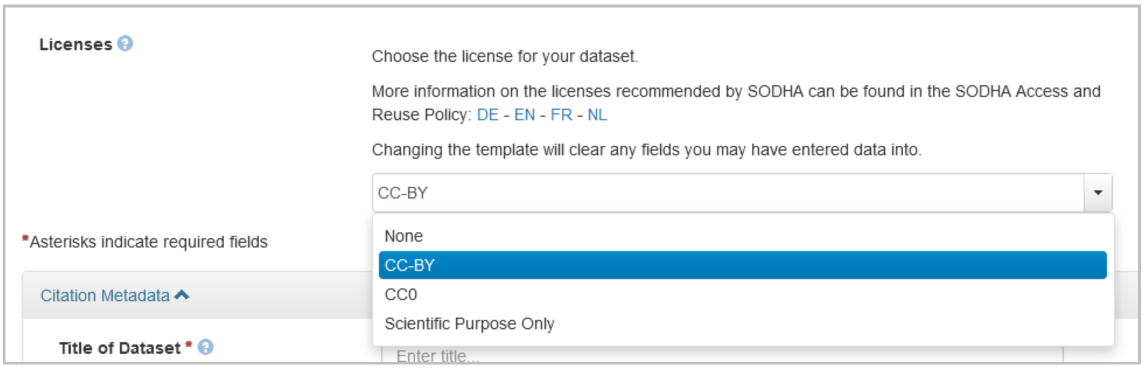
Add "Replication Data for" to Title

Figure 5. The start of the metadata form in the SODHA installation.

As shown in Figure 5, the subrepository selected by default is the *root* Dataverse, whose name here corresponds to that of the data archive and of its website, namely SODHA.

“Dataset Template” → “Licenses”

Comparison of Figures 3 and 5 reveals that another field was renamed early in the form, namely *Dataset Template*, which corresponds to *Licenses* in SODHA. That is because we have used the dataset template functionality of Dataverse to encode the text of licences which we recommend to our users. Instead of requiring that depositors seek out appropriate licences and encode them by themselves, we have selected those that seemed most relevant for sharing data and integrated them in the application:



Licenses

Choose the license for your dataset.

More information on the licenses recommended by SODHA can be found in the SODHA Access and Reuse Policy: DE - EN - FR - NL

Changing the template will clear any fields you may have entered data into.

CC-BY

*Asterisks indicate required fields

Citation Metadata

Title of Dataset *

Enter title...

Figure 6. The options of the dropdown menu of the *Licenses* field in the SODHA Dataverse installation.

This ensures not just ease of use for our users, who need only click a few buttons to find the option that will likely suit them,[33] but also data consistency, since the template functionality of Dataverse enables administrators to prepare default values from which users are free to choose. In other words, IQSS have made it possible to encode "smart defaults" (Colborne, 2010/2018, p. 112), also known as "good defaults and smart prefills" (Tidwell et al., 2019/2020, p. 519). Without this functionality, it is likely that, when researchers are not at a loss to begin with when it comes to filling out the *Terms of Use* field, we would eventually see an aggregation of differently worded text inputs that all designate the same thing, for example: "CC-BY", "Creative Commons 4.0 'CC-BY'", "Creative Commons Attribution (CC-BY)", "CC-BY 4", "CC-BY International", and so on.

Other fields

A few other fields were renamed, usually following the same logic, that is, to make sure that the objects or information that these fields point to are unambiguous for depositors:

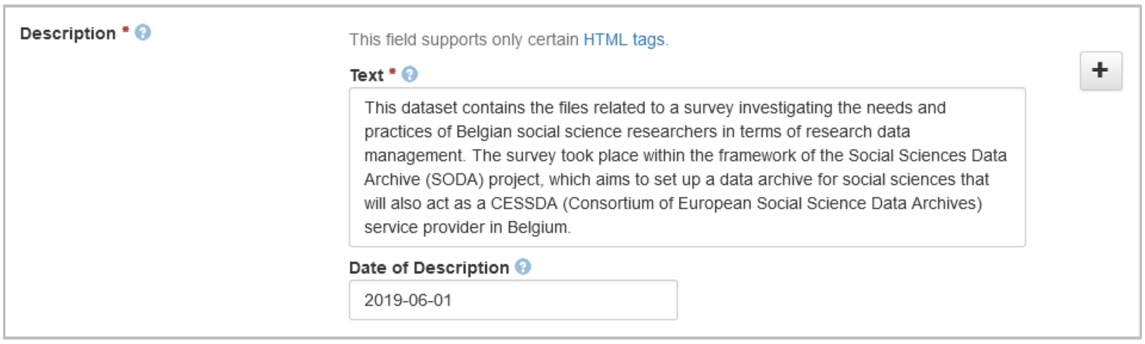
- The field *Description* has two subfields: *Description > Text*, and *Description > Date*. To specify that the latter is meant to designate the date on which the text of the description was written, we renamed it *Date of Description*;
- *Related Publication* is a very interesting field because it allows depositors and curators to reference publications linked to a dataset. Following the suggestion of a user, we renamed it *Related Publication(s)*, to highlight that depositors should feel free to mention as many relevant publications as necessary;
- *Language* was renamed *Language(s) of the Data*, to indicate likewise that more than one language can be selected and that, while the metadata are usually encoded in only one language, the data can include information in a variety of languages;[34]
- *Software* was renamed *Software Used to Create the Dataset*. We chose to sacrifice the brevity of the initial label for the sake of clarity.
- Finally, a field specific to the *Social Sciences and Humanities Metadata* called *Target Sample Size* was rephrased as *Target Sample Size / Number of Units*. While social scientists in Belgium are as conscious as researchers elsewhere of the relevance of this information, that is, the number of respondents that partook in a study, we were afraid that the highly technical English term "target sample size" might not be known by all, especially among younger researchers.

Explanatory Line

The subfield *Description > Text* has a unique property by default in Dataverse: it is preceded by an explanatory line, or "input hint" (Tidwell et al., 2019/2020, p. 489), which mentions that Hypertext Markup Language (HTML) tags can be added to the text input of the field:

[33] It is of course possible that a researcher might not be satisfied with any of the choices that we suggest, but they are free to select the option *None* and write out or copy and paste the text of a completely different licence in the appropriate metadata field (i.e., *Terms of Use*).

[34] This is even likelier to happen in the case of SODHA than in other contexts considering that Belgium has three official languages (Dutch, French, and German).



This field supports only certain HTML tags.

This dataset contains the files related to a survey investigating the needs and practices of Belgian social science researchers in terms of research data management. The survey took place within the framework of the Social Sciences Data Archive (SODA) project, which aims to set up a data archive for social sciences that will also act as a CESSDA (Consortium of European Social Science Data Archives) service provider in Belgium.

2019-06-01

Figure 7. The *Description* field as displayed in editing mode.

We thought that this additional line of text placed above the label and input box of a field was very useful and could be applied to other fields, especially those for highly technical information (such as *Target Sample Size*). But this particular feature is not comprised in the set of customisation possibilities for Dataverse: when our technical officer investigated it, he discovered that this explanatory line was hard-coded in Dataverse. This meant that applying it to other fields would require altering the fundamental code on which Dataverse runs, which would amount to forking.[35] If SODHA was endowed with a large team of developers, such a drastic course of action could have been conceivable. As it stands, with the limited means at our disposal, we soon determined that the limited added value of this isolated feature was not worth taking such risks. Perhaps it was for the best, too: although an input hint “frees users from having to guess” (Tidwell et al., 2019/2020, p. 489), making it excessively long or having too many them means that “many users’ eyes will glaze over, and they’ll ignore the text altogether” (Tidwell et al., 2019/2020, p. 489).[36]

Deposit Agreement

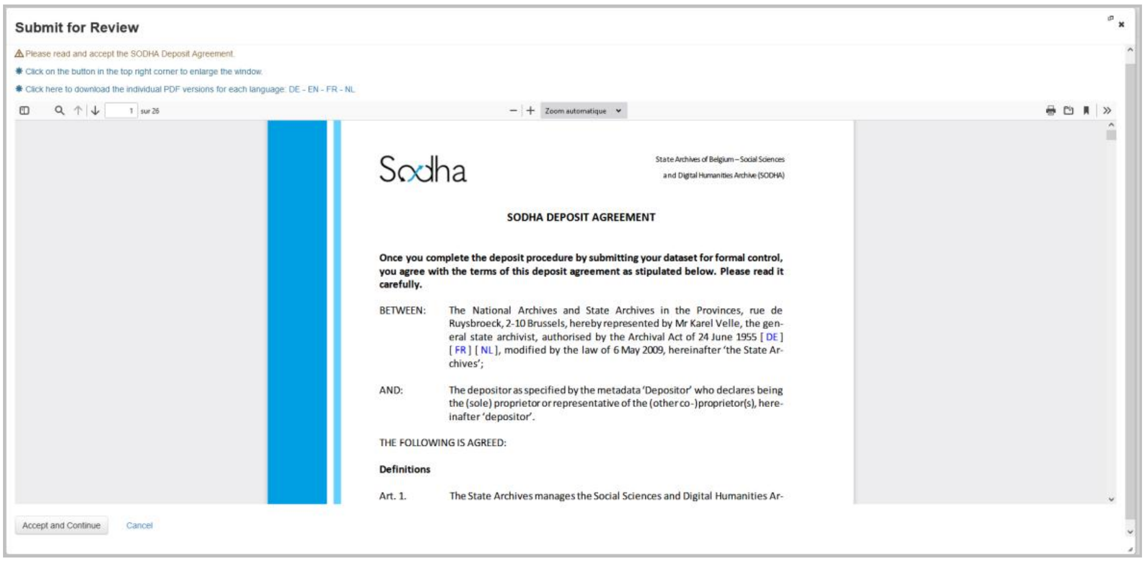
Perhaps one of our most pivotal additions came at the behest of our legal expert. By default, when depositors want to notify Dataverse administrators that they have fully gone through the data deposit procedure and that they want to submit their dataset for publication, they must click on the *Submit for Review* button. This brings up a pop-up window with which users are invited to review their choice of licence and, if applicable, custom terms of use. Our legal expert explained to us that this was the point in the procedure when we had to present our users with the generic SODHA contract for sealing a data management agreement between us and the depositor. We offered to add a line in the pop-up window to the tune of “By clicking on *Accept and Continue*, you agree with the SODHA Deposit Agreement” with a hyperlink integrated in the last three words directing users to the text in question. But out legal expert insisted that, for legal reasons, we had to *literally present* our users with the text of the contract (physically or virtually); asking that they agree to a text that was not *directly shown* to them was not enough.

This proved difficult to implement because, by default, the pop-up window contains only a few lines of text that cannot easily be formatted with headers and paragraphs. Pop-

[35] “[F]orking occurs whenever a software project splits. While the two versions remain entirely or partially compatible for some time, inevitably the unique (and now distinct) histories of each one’s development will push them apart” (St. Laurent, 2004, p. 171).

[36] “If you want to organize for simplicity, it’s important to emphasize just one or two important things. Simple organization doesn’t draw attention to itself; it lets users focus on what they’re doing” (Colborne, 2010/2018, p. 136).

up windows of the sort, also known as *modal panel*, are "disruptive" (Tidwell et al., 2019/2020, p. 159): "If the user isn't prepared to answer whatever the modal panel asks, it interrupts their workflow, possibly forcing them to make a decision about something they just don't care about" (Tidwell et al., 2019/2020, p. 159). We are used to seeing small modal panels pop-up on our screens, typically requesting that we agree to a cookie policy or that we provide our credentials to log on a website. But here, because we could not have an intermediary, stand-alone webpage presenting depositors with the data deposit agreement, we had to find a way to fit a viewer for documents in Portable Document Format (PDF) into the *Submit for Review* modal panel. This was made possible thanks to the iFrame feature of JavaScript:



Submit for Review

⚠ Please read and accept the SODHA Deposit Agreement.

- Click on the button in the top right corner to enlarge the window.
- Click here to download the individual PDF versions for each language: DE - EN - FR - NL

State Archives of Belgium – Social Sciences
and Digital Humanities Archive (SODHA)

SODHA DEPOSIT AGREEMENT

Once you complete the deposit procedure by submitting your dataset for formal control, you agree with the terms of this deposit agreement as stipulated below. Please read it carefully.

BETWEEN: The National Archives and State Archives in the Provinces, rue de Ruysbroeck, 2-10 Brussels, hereby represented by Mr Karel Velle, the general state archivist, authorised by the Archival Act of 24 June 1955 [DE] [FR] [NL], modified by the law of 6 May 2009, hereinafter 'the State Archives';

AND: The depositor as specified by the metadata 'Depositor' who declares being the (sole) proprietor or representative of the (other co-)proprietor(s), hereinafter 'depositor'.

THE FOLLOWING IS AGREED:

Definitions

Art. 1. The State Archives manages the Social Sciences and Digital Humanities Ar-

Figure 8. The modified *Submit for Review* modal panel of the SODHA Dataverse installation with an integrated PDF viewer.

Viewers are now presented with a concatenated version of our data deposit agreements written in Dutch, English, French, and German, as well as the possibility to download this version or each individual version (see the extra line "Click here to download the individual PDF versions…").

SODHA is a service within a government institution. Its purpose is to seek out, receive, and handle data produced by a variety of individuals and organisations, among which are private entities, as well as governement agencies that belong to different levels of authority. Furthermore, as a recent creation, SODHA must demonstrate its reliability and go above and beyond to prove that it closely follows the rules. Otherwise, researchers cannot be expected to entrust us with their precious data. That is why, difficult as such an adaptation proved to be, we took our legal expert's warnings to heart and looked for ways to integrate the text of the deposit agreement. There was a temptation to cut corners, but because much of the preliminary work before the launch of the data archive consisted not only in anticipating potential lawsuits but also in securing our users' trust, we decided to toe the line and not to spare any expense.

Conclusion

Fass, Revell, Stopher, and Verhoeven write: "Acting responsibly as an interface designer means considering the ethical implications of a design, including how it shapes opinions and actions" (2021, p. 11). The same authors remark that, for many people, digital interfaces have become "ubiquitous" (2021, p. 11). We took these sobering reminders to heart when we decided to work on the look-and-feel of our platform, in open disagreement with the common notion that interface design boils down to what is often designated (usually in a derisive manner) as "cosmetic" work. The phrase suggests artificiality and pointlessness; in the worst cases, it even smacks of deception. Admittedly, spending much time on deciding which shade of gray will be used for a certain button even before coding the button is counter-productive. On the other hand, design work is all too often neglected or outright dismissed with open-source technologies for lack of adequate resources.[37] Overall, we have invested little in revamping the more esthetic aspects of our online platform because we chose to focus on more decisive, operational aspects, such as data quality or legal obligations. However, this involved touching upon the very front-end aspects of the user interface, as we sought to translate our very abstract policies into concrete design choices.

Contrasting their observations on how design tends to be neglected in open-source projects (Fass et al., 2021, p. 73), Fass et al. contend: "The advantage of working in an open-source culture is that products can often be more driven to user needs than commercial goals" (2021, p. 69). Indeed, despite the lack of resources that often hamper such projects, open-source endeavors can foster relationships between users and providers that are arguably fairer, more authentic than in commercial configurations. Because open-source developers do not usually need other users to buy their product, interactions can take place in a more voluntary and transparent manner.

We see this philosophy at work both in the relationships that tie together IQSS, the original developers of Dataverse, and other organisations or individuals who install, appropriate, sometimes even further develop their software, and in the rapport that we try to establish with our target users, namely researchers in social sciences and digital humanities. Thanks to the open and flexible attitude of IQSS, who welcome user feedback, and the fact that they have made their software customisable, we have been able to make our intentions tangible in both general and minute aspects of our platform, by adapting its form to better reflect its content. Any of our users accessing our website would be wise to wonder: "What's in it for them?" While we go to great lengths to encourage our users to entrust us with their data, by advertising the many benefits of doing so, we have endeavored to convey our conviction that data sharing will promote not just our own project but scientific and social progress at large.

Acknowledgements

The SODHA infrastructure is supported by the Social Sciences Data Archive (SODA) and Belgian Infrastructure for Social Sciences and Humanities Open Science (BISHOPS) projects, which were financed by the Belgian Science Policy Office (BELSPO) under contracts FR/00/SO5 and FSIRI/32/BI1.

We would like to thank our colleagues and supervisors for their advice and guidance: Samira Hajji, Freya De Schamphelaere, Oliver Lenaerts, Thierry Eggerickx, Johan Surkyn, and Christophe Vanroelen. We would also like to kindly thank everyone at IQSS for their help and feedback, whether on GitHub, the Dataverse Google Group, or the IQSS chatroom. We are especially grateful to Philip Durbin for his invaluable assistance and for bringing up the concept of knowledge workers in the context of research data management.

[37] Fass et al. note that open-source technologies "often suffer the disadvantage of being poorly designed, offering little of the user-friendliness that is a key imperative of commercial software, and this can often be alienating for people" (2021, p. 73).

References

Altman, M., & Crosas, M. (2014). The evolution of data citation: From principles to implementation. *IASSIST Quarterly* 37(1–4), 62–70. doi:10.29173/iq504

Carlson, S., & Anderson, B. (2007). What *are* data? The many kinds of data and their implications for data re-use. *Journal of Computer-Mediated Communication* 12(2), 635–651. doi:10.1111/j.1083-6101.2007.00342.x

Chauvette, A., Schick–Makaroff, K., & Molzahn, A. E. (2019). Open data in qualitative research. *International Journal of Qualitative Methods* 18. doi:10.1177/1609406918823863

Colborne, G. (2018). *Simple and usable: Web, mobile, and interaction design* (2nd ed.). Indianapolis: New Riders. (Original work published 2010)

Crosas, M. (2011). The Dataverse Network®: An open-source application for sharing, discovering and preserving data. *D-Lib Magazine* 17(1/2). doi:10.1045/january2011-crosas

Dale, A. (2013). Advanced teaching of quantitative methods in the social sciences. In B. Kleiner, I. Renschler, B. Wernli, P. Farago, & D. Joye (eds.), *Understanding research infrastructures in the social sciences* (pp. 177–183). Zurich: Seismo Press.

Doorn, P., & Tjalsma, H. (2007). Introduction: Archiving research data. *Archiving research data*, special issue of *Archival Science* 7(1), 1–20. doi:10.1007/s10502-007-9054-6

European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union, 119*, 4 May 2016, 1–88. Retrieved from http://data.europa.eu/eli/reg/2016/679/oj

European Union (2019). Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast). *Official Journal of the European Union, 172*, 26 June 2019, 56–83. Retrieved from http://data.europa.eu/eli/dir/2019/1024/oj

Fass, J., Revell, T., Stopher, B., & Verhoeven, E. (2021). *Design & digital interfaces: Designing with aesthetic and ethical awareness*. London: Bloomsbury.

Gewin, V. (2021). Pandemic burnout is rampant in academia. *Nature* 591(7850), 489–491. doi:10.1038/d41586-021-00663-2

Hogenaar, A., Tjalsma, H., & Priddy, M. (2011). Research in the humanities and social sciences. In C. Meier zu Verl & W. Horstmann (eds.), *Studies on subject-specific requirements for open access infrastructure* (pp. 165–213). Bielefeld: Universitätsbibliothek Bielefeld. doi:10.2390/PUB-2011-7

Jeng, W., & Lyon, L. (2016). A report of data-intensive capability, institutional support, and data management practices in social sciences. *International Journal of Digital Curation,*11(1), 156–171. doi:10.2218/ijdc.v11i1.398

Jeng, W., He, D., & Oh, J. S. (2016). Toward a conceptual framework for data sharing practices in social sciences: a profile approach. *Proceedings of the Association for Information Science and Technology* 53(1), 1–10. doi:10.1002/pra2.2016.14505301037

Kaase, M. (2013). Research infrastructures in the social sciences: The long and winding road. In B. Kleiner, I. Renschler, B. Wernli, P. Farago, & D. Joye (eds.), *Understanding research infrastructures in the social sciences* (pp. 19–27). Zurich: Seismo Press.

Melendez, W. A., & de Guzman, R. M. (1983). *Burnout: The new academic disease* (ASHE–ERIC Higher Education Research Report No. 9). Retrieved from https://eric.ed.gov/?id=ED242255

Niu, J., & Hedstrom, M. (2009). Documentation evaluation model for social science data. *Proceedings of the ASIST Annual Meeting* 45(1), 1–11. doi:10.1002/meet.2008.1450450223

Olson, C. (1997). *Collected prose* (D. Allen and B. Friedlander, eds.; intro. by R. Creeley). Berkeley, CA: University of California Press.

Organisation for Economic Co-operation and Development. (2020). *OECD open, useful and re-usable data (OURdata) index: 2019*. Retrieved from OECD website: https://www.oecd.org/en/publications/open-useful-and-re-usable-data-ourdata-index-2019_45f6de2d-en.html

Peuch, B., Sanderson, J.-P., Van den Borre, L., Depoortere, R., De Schamphelaere, F., Hajji, S., . . . Naji, A. (2020). *SODA Survey: A survey of the needs and practices of Belgian researchers in social sciences in terms of research data management* [Data set]. Brussels, Belgium: Social Sciences and Digital Humanities Archive (SODHA). doi:10.34934/DVN/NYZJVM

Quigley, E. (2015). Usability testing driven redesign of Dataverse, an open source data repository [Poster]. University of Massachusetts and New England Area Librarian e-Science Symposium. doi:10.13028/3ad3-s565. Retrieved from https://repository.escholarship.umassmed.edu/entities/publication/039c57c1-cb69-4963-a1f0-45d581947e13

Scheuch, E. K. (2003). History and visions in the development of data services for the social sciences. *International Social Science Journal* 55(3), 385–399. doi:10.1111/j.1468-2451.2003.05503004.x

Schneider, S., Katsanidou, A., Horton, L., & Wolf, C. (2013). Postgraduate training for acquiring social science data skills. In B. Kleiner, I. Renschler, B. Wernli, P. Farago, & D. Joye (eds.), *Understanding research infrastructures in the social sciences* (pp. 168–176). Zurich: Seismo Press.

Shaon, A., Straube, A., & Chowdhury, K. R. (2017). Setting up a national research data curation service for Qatar: Challenges and opportunities. *International Journal of Digital Curation* 12(2), 146–156. doi:10.2218/ijdc.v12i2.515

St. Laurent, A. M. (2004). *Understanding open source and free software licensing*. Sebastopol, CA: O'Reilly Media.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., . . . Frame, M. (2011) Data sharing by scientists: Practices and perceptions. *PLoS ONE* 6(6), 1–21. doi:10.1371/journal.pone.0021101

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., . . . Dorsett, K. (2015) Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLoS ONE* 10(8), 1–24. doi:10.1371/journal.pone.0134826

Tidwell, J., Brewer, C., & Valencia, A. (2020). *Designing interfaces: Patterns for effective interaction design* (3rd ed.). Sebastopol, CA: O'Reilly Media. (Original work published 2019)

Van Honacker, K. (2018). GDPR and processing for archiving purposes in the public interest: An introduction. In Van Honacker, K. (ed.), *The right to be forgotten vs The right to remember* (pp. 11–26). Brussels, Belgium: VUBPress.

Weil, V., & Hollander, R. (1990). Sharing scientific data II: Normative issues. *IRB: Ethics and Human Research* 12(2), 7–8. doi:10.2307/3563522

Whalen, J. (2021). *Design for how people think: Using brain science to build better products* (rev. ed.). Sebastopol, CA: O'Reilly Media. (Original work published 2019)

Williams, J. D. (2019). *The decline in educational standard: From a public good to a quasi-monopoly*. Lanham, MD: Rowman & Littlefield.

Yoon, A., & Kim, Y. (2017). Social scientists' data reuse behaviors: Exploring the roles of attitudinal beliefs, attitudes, norms, and data repositories. *Library & Information Science Research* 39(3), 224–233. doi:10.1016/j.lisr.2017.07.008